TOWARD FOUNDATIONS OF MULTI-TEAM LEARNING: CONVERGENCE, COORDINATION, AND EQUILIBRIUM

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

By
Ahmed Said Dönmez
December 2025

TOWARD FOUNDATIONS OF MULTI-TEAM LEARNING: CONVERGENCE, COORDINATION, AND EQUILIBRIUM
By Ahmed Said Dönmez
December 2025

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Muhammed Ömer Sayın(Advisor)

Sinan Gezici

Faruk Polat

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

TOWARD FOUNDATIONS OF MULTI-TEAM LEARNING: CONVERGENCE, COORDINATION, AND EQUILIBRIUM

Ahmed Said Dönmez
M.S. in Electrical and Electronics Engineering
Advisor: Muhammed Ömer Sayın
December 2025

Outcomes in modern socio-technical systems are increasingly shaped not by isolated individuals, but by teams of intelligent agents that coordinate internally while competing externally. This thesis focuses on the emerging grand challenge of how intelligent agents (human or artificial) learn and act as teams in complex and dynamic environments. We address two strategic environments: inter-team competition in repeated normal-form games and intra-team cooperation in dynamic, stochastic environments. We establish a unified framework based on epoch-based stochastic approximation and differential inclusions to overcome non-stationarity induced by simultaneous learning. For inter-team competition, we introduce Zero-Sum Potential Team Games (ZSPTGs), generalizing potential and zero-sum polymatrix games. We propose Team-Fictitious-Play (Team-FP), a decentralized dynamic using smoothed best responses to teammate history and opponent beliefs. We prove empirical averages under Team-FP converge almost surely to a near Team-Nash Equilibrium (TNE), quantifying error via exploration parameters. Addressing intra-team cooperation where standard dynamics often fail, we develop Logit-Q dynamics. Combining log-linear learning with Q-learning, this method allows simultaneous learning of value functions and policies without model knowledge. We prove almost sure convergence to the efficient stationary equilibrium, validated by coupling the non-stationary process with a tractable quasi-stationary reference. Finally, numerical experiments validate our theoretical findings and demonstrate effective coordination.

Keywords: Multi-agent learning, stochastic control, game theory.

ÖZET

ÇOK TAKIMLI ÖĞRENMENİN TEMELLERİNE DOĞRU: YAKINSAMA, İŞ BİRLİĞİ VE DENGE

Ahmed Said Dönmez
Elektrik Elektronik Mühendisliği, Yüksek Lisans
Tez Danışmanı: Muhammed Ömer Sayın
Aralık 2025

Modern sosyo-teknik sistemlerdeki sonuçlar, giderek artan bir şekilde izole bireyler tarafından değil, dışarıya karşı rekabet ederken kendi içlerinde koordine olan akıllı etmen takımları tarafından şekillenmektedir. Bu tez, akıllı etmenlerin (insan veya yapay) karmaşık ve dinamik ortamlarda nasıl takımlar halinde öğrendiği ve hareket ettiğine dair ortaya çıkan büyük zorluğa odaklanmaktadır. İki stratejik ortamı ele alıyoruz: tekrarlanan normal-form oyunlarda takımlar arası rekabet ve dinamik, stokastik ortamlarda takım içi işbirliği. Eşzamanlı öğrenmenin neden olduğu durağansızlığın üstesinden gelmek için dönem tabanlı stokastik yaklaşıklama ve diferansiyel içermelere dayalı birleşik bir çerçeve oluşturuyoruz. Takımlar arası rekabet için, potansiyel ve sıfır toplamlı polimatris oyunlarını genelleştiren Sıfır Toplamlı Potansiyel Takım Oyunlarını (ZSPTG'ler) tanıtıyoruz. Takım arkadaşı geçmişine ve rakip inançlarına karşı yumuşatılmış en iyi yanıtları kullanan merkeziyetsiz bir dinamik olan Takım-Hayali-Oyun'u (Team-FP) öneriyoruz. Team-FP altındaki ampirik ortalamaların, keşif parametreleri aracılığıyla hatayı nicelendirerek, bir Takım-Nash Dengesi'ne (TNE) yakın bir noktaya neredeyse kesin yakınsadığını kanıtlıyoruz. Standart dinamiklerin sıklıkla başarısız olduğu takım içi işbirliğini ele alarak, Logit-Q dinamiklerini geliştiriyoruz. Log-lineer öğrenmeyi Q-learning ile birleştiren bu yöntem, model bilgisi olmadan değer fonksiyonlarının ve politikaların eşzamanlı öğrenilmesine olanak tanır. Durağan olmayan süreci izlenebilir bir yarı-durağan referansla eşleştirerek doğruladığımız, verimli durağan dengeye neredeyse kesin yakınsamayı kanıtlıyoruz. Son olarak, sayısal deneyler teorik bulgularımızı doğrulamakta ve etkili koordinasyonu göstermektedir.

Anahtar sözcükler: Çok etmenli öğrenme, stokastik kontrol, oyun teorisi.

Acknowledgement

I would like to express my deepest gratitude to my advisor Prof. Muhammed Ömer Sayın for his continuous support, patience, and valuable guidance throughout my research. His insights and encouragement have been invaluable to my academic growth.

I would also like to thank my colleagues and friends Yüksel Arslantaş and Onur Ünlü for their contributions to this work, as well as all my lab mates and friends for creating an enjoyable research environment.

Finally, I am profoundly grateful to my family for their unwavering support and encouragement throughout my academic journey.

This thesis was supported by The Scientific and Technological Research Council of Türkiye (TUBITAK) BIDEB 2232-B International Fellowship for Early Stage Researchers under Grant Number 121C124, and TUBITAK BIDEB 2210-A National Fellowship Program for Graduate Students.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The advancement of autonomous systems in domains ranging from robotics and resource management to online gaming and financial markets has demonstrate the importance of understanding how multiple intelligent agents interact [1, 2, 3, 4, 5, 6]. A crucial feature of many such systems is the presence of teams—groups of agents that share common objectives and must coordinate their actions to succeed. Whether it is a swarm of drones competing against another swarm for aerial dominance, a group of self-driving vehicles cooperating to optimize traffic flow, or a team of financial algorithms working to maximize portfolio returns, the fundamental challenge remains the same: how can collections of individual, decentralized agents learn to act as a cohesive and effective team?

This thesis addresses this central question by investigating learning dynamics in two fundamental strategic environments that define team interactions. The first is the setting of *inter-team competition*, where multiple teams with opposing goals interact. In such adversarial scenarios, the desired outcome is a stable state of play where no team has an incentive to unilaterally deviate from its strategy. This concept is captured by the Team-Nash Equilibrium (TNE), which predicts the outcome of coordinated team interactions [7, 8, 9]. Game-theoretical equilibrium is often justified by its emergence from non-equilibrium adaptation of self-interested learners (e.g., see [10]). However, the existence of such an

equilibrium does not guarantee that it can be reached by teams of self-interested agents learning from experience. A foundational question, largely unexplored, is whether simple, decentralized learning rules can guide teams toward such a sophisticated, team-level equilibrium.

The second environment is that of *intra-team cooperation* within a single team playing a stochastic game, in which the environment is dynamics and changes based on the actions of players. Here, all agents share a common goal, but the presence of multiple equilibria may result in reaching a suboptimal solution in the absence of explicit coordination. The team may learn to coordinate on a suboptimal outcome, performing poorly despite their shared objective. The challenge, therefore, is to design learning dynamics that can provably guide agents to an efficient equilibrium that maximizes their collective utility, especially when the environment's transition dynamics are unknown.

This thesis bridges these two domains by developing and analyzing novel, decentralized learning dynamics tailored for each setting. We establish a unified framework for "Team Learning in Games", demonstrating how simple, plausible behavioral rules can empower agents to achieve complex, coordinated, and desirable team-level outcomes. By tackling both the competitive and cooperative sides of team interaction in multi-agent games, and the collaboration within a single team in a stochastic game, we aim to provide a more complete theoretical foundation for the design and analysis of multi-agent systems.

Classical game theory provides the language for analyzing strategic interactions, with the Nash Equilibrium (NE) being the most fundamental solution concept. An equilibrium is often justified by its emergence from the long-run behavior of agents adapting their strategies over time [10]. A foundational model for this process is Fictitious Play (FP), where agents play a best response to the historical frequency of their opponents' actions. While FP and its variants have shown convergence guarantees in important classes of games, their applicability is not universal [11]. For instance, FP is known to successfully reach an equilibrium in two-agent zero-sum games [12] and in potential games [13]. However, even in these cases, limitations exist. In potential games, which

are often used to model cooperative scenarios, FP does not guarantee convergence to the most efficient outcome for the team, leaving open the possibility of suboptimal coordination. An alternative, log-linear learning, has been shown to achieve efficient outcomes in such settings [14, 15], but its ability to track efficient equilibria in dynamic environments (induced by the evolving strategies of opponent teams) remains an open question. Notably, [16] addressed efficient learning under non-stationarity induced by the decaying exploration in agents' responses for the repeated play of potential games. However, this approach is orthogonal to the analysis in this thesis.

The challenge of convergence becomes more pronounced in multi-agent zero-sum games as we can transform any general-sum game to a multi-agent zero-sum game by introducing a non-effective auxiliary agent (with a single action). While several studies have explored this area [17, 18, 19], including in network games [20] and zero-sum polymatrix games [19, 21, 22], the classical FP dynamic does not readily apply to the multi-team structures central to this thesis. Specifically, in Zero-sum Potential Team Games (ZSPTGs), which extend these concepts, standard FP dynamics are not guaranteed to converge to a Team-Nash Equilibrium, highlighting the need for new learning models.

When multiple teams compete, as seen in robotics, cybersecurity, and online gaming, the strategic landscape becomes more complex. The interaction is no longer between individual agents but between coordinated groups. The appropriate solution concept is the **Team-Nash Equilibrium (TNE)**, a generalization of the team-minimax equilibrium for two-player zero-sum games [7]. In TNE, each team, acts as a single entity and jointly plays a best response to the other teams' strategies [8].

Recent studies on two-team zero-sum games and adversarial team games have primarily focused on the efficient computation of team-minimax equilibrium [23, 8, 24, 25]. A key theme in this line of work is the role of communication, with *ex ante* (pre-play) communication among team members showing particular promise [23]. Consequently, many of these studies model a team as a single, coordinated agent with imperfect recall, effectively embedding this pre-play

coordination into the problem structure [8, 24, 25]. A notable example is the Fictitious Team Play (FTP) algorithm developed for extensive-form two-team zero-sum games, where FP is applied to a simplified game representation and best responses are computed using mixed-integer linear programming [8]. Our work takes a fundamentally different approach. Instead of assuming pre-play coordination, this thesis investigates how agents can *learn* to act as a cohesive team by following simple, decentralized behavioral rules based on self-interest, much in the spirit of foundational learning dynamics like log-linear learning and FP.

In cooperative settings, often modeled as **identical-interest** or **potential games** [13], all agents' incentives are aligned with a common objective or potential function. While convergence to a Nash Equilibrium is often assured for dynamics like FP, a significant challenge arises from the existence of multiple equilibria with vastly different payoffs. Standard dynamics may converge to an inefficient equilibrium, leading to poor team performance. If we include dynamic opponents to the environment, the question of whether agents who uses behavioral learning rules can reach the TNE becomes even more challenging.

To address the issue of suboptimal equilibria in team games, researchers have developed learning dynamics that introduce exploration and inertia to escape the basins of attraction of inefficient outcomes. A prominent example is **Log-Linear Learning** (or logit dynamics) [26]. In this model, agents respond to their beliefs with some randomness (a smoothed best response) and exhibit inertia in their action updates. It has been shown that in the long run, these dynamics spend most of their time at the efficient, potential-maximizing equilibrium [14, 27, 15]. Other approaches to achieving efficiency include aspiration learning [28] and dynamics based on contentment and discontentment states [29, 30].

The aforementioned learning dynamics were primarily developed for repeated normal-form games. Extending these ideas to dynamic environments requires tools from **Multi-Agent Reinforcement Learning (MARL)** literature. The standard model for this is the **Stochastic Game**, a generalization of Markov

Decision Processes (MDPs) to multiple agents, introduced by [31]. In a stochastic game, agents' actions affect not only their immediate rewards but also the transition to subsequent states, creating a complex trade-off between present and future gains. This framework allows for modeling very complex environments in frontier artificial intelligence applications, ranging from complex board games to planning for intelligent and autonomous systems [9]. Consequently, learning in stochastic games has drawn significant interest in examining whether non-equilibrium adaptation of self-interested agents converges to equilibrium, particularly in zero-sum or identical-interest settings [32, 33, 34, 35, 36, 37]. In particular, learning without coordination or with heterogeneous learning parameters in such stochastic games is a central challenge in MARL [38, 39, 40, 41, 42]. These findings enhance the predictive power of equilibrium analysis [10]. However, for control and optimization, a drawback is that convergence guarantees apply only to certain equilibria, and agents using these dynamics may perform arbitrarily poorly even in stochastic teams with a common goal. For instance, while efficient learning dynamics have been extensively studied in repeated games [43, 28, 14, 29, 30], results for stochastic games remain limited [44].

Stochastic teams have drawn significant attention in multi-agent learning, but early studies often focused on computing Q-function estimates for efficient equilibria through coordinated maximization over joint action profiles [45, 46, 47, 48]. Such approaches stand in contrast to uncoupled learning dynamics, where agents learn to coordinate without explicit knowledge of their teammates' objectives, as seen in the repeated play of games [28, 14, 29, 30]. These decentralized dynamics are particularly valuable in non-cooperative settings, as they align with agents' self-interest while enabling rational responses to stationary opponents. A prime example is classical log-linear learning, where agents respond to their belief about others' last actions with some randomness, a behavior experimentally closer to human learning than fictitious play [49]. The common thread in these efficient learning dynamics for repeated games is the use of recency, exploration, and friction to guide agents toward optimal outcomes. While recent work has introduced fictitious play and policy

gradient methods to stochastic teams [32, 35, 50], they often do not guarantee convergence to an efficient equilibrium, leaving a critical gap.

Recent research has begun to explore dynamics that aim to bridge this gap. For example, [43] examined efficient learning but was limited to stochastic games with exogenous state transitions, thereby avoiding the core challenge of balancing immediate and continuation payoffs. Other work by [44] achieved efficient learning by restricting agents to pure stationary strategies and dividing learning into phases, which requires a degree of coordination among agents to pause their strategy updates. Similarly, another study focused on episodic logit-Q dynamics where Q-functions were held constant within episodes, making the stage games stationary but again imposing a coordination requirement [51]. Our work departs from these episodic or phased approaches by allowing for fully concurrent, uncoupled learning.

A common technical hurdle in both the competitive and cooperative settings investigated in this thesis is **non-stationarity**. In the team competition setting, an agent's learning environment is rendered non-stationary because its strategic partners and opponents are also learning and adapting their strategies simultaneously. In this setting, the beliefs an agent forms about opponent teams are constantly changing, making the environment non-stationary. In the cooperative, stochastic game setting, an agent's own estimates of the Q-functions evolve as it learns more about the environment, changing the payoffs of the effective "stage game" it plays at each state, which again leads to non-stationarity.

To handle this, some approaches have utilized *two-timescale stochastic approximation*, where Q-functions (slow timescale) and policies (fast timescale) are updated at different rates [52, 33, 35]. However, this is not directly applicable to all learning rules. Our work addresses this unifying challenge by developing a new epoch-based analysis to prove convergence even when all learning processes evolve concurrently. By dividing the horizon into epochs, we are able to deal with non-stationarity challange. On top of that, we also utilize common sophisticated tools from stochastic approximation theory such as stochastic differential inclusion approximation [53, 54].

This thesis establishes a unified framework for analyzing how decentralized agents can learn to coordinate and compete effectively in complex multi-agent environments. By addressing the distinct challenges of inter-team competition and intra-team cooperation, we provide a comprehensive theory of team learning in games. The main contributions of this thesis are as follows:

- **Novel Learning Dynamics for Strategic Team Environments:** We introduce two families of decentralized learning dynamics tailored to the fundamental settings of team interaction. For *inter-team competition*, we formulate the class of Zero-Sum Potential Team Games (ZSPTGs), which generalizes potential games and zero-sum polymatrix games. Within this framework, we propose *Team-Fictitious-Play (Team-FP)*, a dynamic where agents respond to the historical play of teammates and opponents. For *intra-team cooperation* in stochastic environments, we develop *Logit-Q dynamics*, which integrate log-linear learning with Q-learning. Together, these models demonstrate that simple, local update rules can lead to the emergence of sophisticated team behavior without the need for centralized control or explicit pre-play coordination.

- **A Unified Analytical Framework for Non-Stationary Learning:** A central theoretical contribution of this work is a robust analytical framework designed to overcome the challenge of non-stationarity, a pervasive issue in multi-agent learning. Whether arising from the adaptive strategies of opponent teams or the evolving value estimates within a stochastic team, non-stationarity typically invalidates standard convergence proofs. We address this by developing a novel epoch-based analysis combined with stochastic differential inclusion (SDI) approximation and optimal coupling techniques. This framework allows us to rigorously analyze systems where multiple learning processes—such as strategy updates and Q-value estimation—evolve concurrently. We derive explicit perturbation bounds to show that, under controlled step sizes and epoch lengths, the complex non-stationary dynamics can be approximated by tractable stationary processes, thereby ensuring asymptotic convergence.

- **Convergence to Desirable Equilibrium Concepts:** We provide rigorous convergence guarantees to solution concepts that capture both stability and efficiency. In the competitive setting, we prove that the empirical averages of play under Team-FP converge almost surely to a Team-Nash Equilibrium (TNE), demonstrating that teams can learn to coordinate their strategies effectively even without a central controller or without communication between agents against learning adversaries. In the cooperative stochastic setting, we show that Logit-Q dynamics converge not just to any equilibrium, but specifically to the *efficient* stationary equilibrium that maximizes the team's long-term collective utility. This addresses a critical limitation in existing multi-agent reinforcement learning literature, which often fails to distinguish between optimal and suboptimal equilibria. Furthermore, we quantify the approximation error of these dynamics, establishing their rationality against stationary opponents and characterizing the trade-off between exploration levels and equilibrium precision.

The rest of this thesis is organized as follows. **Chapter 2** provides the key preliminary results and background material used in this thesis for convergence proofs. **Chapter 3** focuses on inter-team competition. We introduce the Zero-Sum Potential Team Game (ZSPTG) model and detailed the Team-Fictitious-Play (Team-FP) dynamics. We present the analytical results regarding the convergence to Team-Nash Equilibrium and provide numerical simulations to validate the theoretical findings. **Chapter 4** shifts focus to intra-team cooperation within stochastic environments. We formally define the stochastic team problem and present the family of Logit-Q dynamics. We provide the convergence analysis showing how these dynamics reach efficient equilibria and discuss the implications of different implementation schemes (coordinated vs. independent). Finally, **Chapter 5** concludes the thesis, summarizing the main results and discussing potential directions for future research in team learning dynamics.

Chapter 2

Preliminary Information

This chapter provides preliminary definitions and results that will be used in the subsequent chapters. We first present the stochastic approximation with differential inclusion framework for analyzing the convergence of discrete-time stochastic processes to the solutions of differential inclusions. Then, we provide useful lemmas related to the solving the challenges in both team fictitious play and logit-Q learning dynamics, including a coupling lemma between two discrete-time stochastic processes and a proposition characterizing the stationary distributions of the action profiles in the classical and independent learning dynamics.

Definition 1. Lyapunov function We say that a continuous function $V : \mathbb{R}^m \to \mathbb{R}$ is a Lyapunov function for a subset Λ of $\mathbb{R}^m$ provided that for any trajectory of the flow φ, e.g. $\varphi_t(x)$,

- $V(y) < V(x)$ for all $x \in R^m \setminus \Lambda$, $y \in \varphi_t(x)$, $t > 0$,
- $V(y) \leq V(x)$ for all $x \in \Lambda$, $y \in \varphi_t(x)$, $t \geq 0$,

Theorem 1. *[54, Theorem 3.1] Consider a sequence $\{y_k \in Y\}_{k\geq 0}$ evolving according to*

$$y_{k+1} - y_k - \alpha_k(e_k + \omega_k) \in \alpha_k F(y_k), \tag{2.1}$$

where

- *The step sizes* $\{\alpha_k \in [0,1]\}_{k=0}^{\infty}$ *decay at a suitable rate such that* $\sum_{k=0}^{\infty} \alpha_k = \infty$ *and* $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.
- *The set valued* $F(y)$ *is a Marchaud map.*
- *The error term* $\|e_k\| \to 0$ *almost surely.*
- *The noise sequence* $\{\omega_k\}$ *is* $\mathcal{F}_{k+1}$*-measurable with respect to the filtration* $\mathcal{F}_k$ *generated by* σ*-algebra* $\sigma(y_0, \omega_0, \ldots, \omega_{k-1})$. *Furthermore,* $\{\omega_k\}$ *is a Martingale difference sequence satisfying* $\mathrm{E}[\omega_k|\mathcal{F}_k] = 0$ *and* $\mathrm{E}[\|\omega_k\|^2|\mathcal{F}_k] < W$ *for some constant* W *for all* $k \geq 0$ *almost surely.*

Then, the linear interpolation of y_k *is an asymptotic pseudo-trajectory to the differential inclusion*

$$\dot{y} \in F(y). \tag{2.2}$$

Therefore, one can characterize the limit behavior of (2.1) through (2.2) if there exists a Lyapunov function V as given in Definition 1. Note that for $\Lambda \subset Y$ and $V : Y \to \mathbb{R}$, $V(\Lambda) \subset \mathbb{R}$ has empty interior. Therefore, every internally chain transitive set of differential inclusion (2.2) and the limit set of discrete update (2.1) are contained in Λ [53, Proposition 3.27].

The following lemma provides a coupling result between two discrete-time stochastic processes that will be the key tool in this thesis to solve the non-stationarity challenge in both team fictitious play and logit-Q learning dynamics. This lemma can characterize the long-run behavior of the logit dynamics in certain non-stationary environments by coupling its evolution with the logit dynamics in stationary environments.

Lemma 1. *Consider two discrete-time processes* $\{\widehat{w}_k\}_{k\geq 0}$ *and* $\{(w_k, z_k)\}_{k\geq 0}$ *over a finite set* W *and the product space* $W \times Z$ *with some countable space* Z. *The processes* $\{\widehat{w}_k\}$ *and* $\{(w_k, z_k)\}$ *evolve, resp., according to* $\widehat{w}_k \sim \widehat{\mathbb{P}}(\cdot|\widehat{w}_{k-1})$ *and* $(w_k, z_k) \sim$

$\mathbb{P}(\cdot|(w_l, z_l)_{l<k})$. *Let* $\mathbb{P}^w(\cdot|(w_l, z_l)_{l<k})$ *denote the marginal of* $\mathbb{P}(\cdot|(w_l, z_l)_{l<k})$ *over* Z *such that* $w_k \sim \mathbb{P}^w(\cdot|(w_l, z_l)_{l<k})$. *Suppose that these processes satisfy the following conditions:*

Condition (i) *There exist* $\widehat{\varepsilon}, \varepsilon > 0$, *a destination* $w' \in W$ *and* $\kappa \in \mathbb{N}$ *such that given any* $(\widehat{w}, w, z) \in W \times W \times Z$, *where* $\widehat{w} \neq w$, *the processes* $\{\widehat{w}_k\}$ *and* $\{(w_k, z_k)\}$ *can follow, resp.,* κ*-length paths* $\widehat{w} = \widehat{h}^0, \ldots, \widehat{h}^\kappa = w'$, *and* $(w, z) = (h^0, g^0), \ldots, (h^\kappa, g^\kappa) = (w', z')$ *for some* z', *where* $\widehat{h}^l \neq h^l$ *for all* $l < \kappa$. *Furthermore, for all* k, *we have*

$$\Pr\left(\bigwedge_{l=1}^{\kappa} \widehat{w}_{k+l} = \widehat{h}^l \mid \widehat{w}_k = \widehat{w}\right) \geq \hat{\varepsilon}, \Pr\left(\bigwedge_{l=1}^{\kappa} w_{k+l} = h^l \mid (w_k, z_k) = (w, z)\right) \geq \varepsilon. \tag{2.3}$$

Condition (ii) *For any* $k = 0, \ldots, K-1$, *there exists* $\lambda \in (0, 1]$ *such that*

$$\left\| \mathbb{P}^w\left(\cdot \mid (w_l, z_l)_{l \leq k}\right) - \widehat{\mathbb{P}}(\cdot \mid w_k) \right\|_{\mathrm{TV}} \leq 1 - \lambda \quad \forall (w_l, z_l)_{l \leq k}. \tag{2.4}$$

Let $\widehat{\mu}_k \in \Delta(W)$ *and* $\mu_k \in \Delta(W \times \mathbb{Z})$ *denote, resp., the distributions of* $\widehat{w}_k$ *and* (w_k, z_k). *Furthermore,* $\widehat{\mu}^w \in \Delta(W)$ *denotes the marginal of* μ_k *over* Z. *Then, there exists a coupling over* $\{\widehat{w}_k\}$ *and* $\{(w_k, z_k)\}$ *such that*

$$\|\widehat{\mu}_k - \mu_k^w\|_1 \leq 2(1 - \varepsilon\widehat{\varepsilon})^{\frac{k}{\kappa}-1} + 2(1 - \lambda^\kappa)\frac{1 + \varepsilon\widehat{\varepsilon}}{\varepsilon\widehat{\varepsilon}} \quad \forall t = 0, \ldots, K-1. \tag{2.5}$$

We highlight that the first term on the right-hand side of (2.5) goes to zero as $k \to \infty$ whereas the second term is a constant depending on the distance between the transition probabilities. For example, if $\lambda = 1$ for $K \to \infty$, then the processes $\{\widehat{w}_k\}$ and $\{w_k\}$ are identical Markov chains and the second term is zero. Correspondingly, the proof of Lemma 1 (Section A.1) has a flavor similar to the proof of the Ergodicity Theorem for Markov chains [55, Chapter 4].

Remark 1. Existing perturbation bounds for the stationary distribution of Markov chains focus mainly on homogeneous chains using maximum entry-wise perturbations [56, 57]. However, these bounds are not tight enough for our analysis. In Lemma 1, we exploit the structure of these chains to derive sharper

bounds and address non-homogeneous cases through finite-time analysis, making Lemma 1 independently interesting.

Before establishing the next key result, we briefly distinguish between the two specific implementations of log-linear learning (logit dynamics) considered in this work: **coordinated** and **independent** learning. In both schemes, agents repeatedly play a finite potential game $\mathcal{G} = \langle (A^i, u^i)_{i\in[N]} \rangle$ with the potential function $\Phi : A \to \mathbb{R}$. The action profile at time k is denoted by $a_k = (a_k^1, \ldots, a_k^N) \in A = A^1 \times \ldots \times A^N$. The difference between the two schemes lies in how ageants that will update their strategies are selected at each time step:

1. **Coordinated (Classical) Log-Linear Learning:** In this scheme, at each discrete time step, exactly one agent is selected uniformly at random to revise their strategy, while all other agents repeat their previous actions. This constraint ensures that the system's state transitions occur only between action profiles that differ by a single agent's strategy.

2. **Independent Log-Linear Learning:** In this scheme, agents play without any central coordination. At each time step, every agent independently decides whether to revise their strategy with a fixed probability $\delta \in (0, 1)$. Consequently, multiple agents may update their strategies simultaneously, allowing the joint action profile to transition more freely in the action space.

While the coordinated dynamic is the standard model for ensuring convergence to potential maximizers in potential games, the independent dynamic is developed to make the learning process fully decentralized [15]. The following Lemma characterizes the proximity of the stationary distributions of joint actions resulting from the Markov chains induced by these two different revision processes. A preliminary version of this lemma appeared in [51].

Lemma 2. *Consider that agents follow the logit dynamics in the repeated play of a potential game* $\mathcal{G} = \langle (A^i, u^i)_{i\in[N]} \rangle$ *with the potential function* Φ*. Let* $\mu^{\mathrm{coor}}, \mu^{\mathrm{ind}} \in$

$\Delta(A)$ *denote the unique stationary distributions of the Markov chains formed by the actions profiles for the coordinated and independent revision processes of log-linear learning. Then, we have* $\mu^{\mathrm{coor}} = \underline{\mathrm{br}}(\Phi(\cdot))$ *and*

$$\|\mu^{\mathrm{coor}} - \mu^{\mathrm{ind}}\|_1 \leq \Lambda(\delta) := 2(1 - \lambda(\delta)) \cdot \frac{1 + \lambda(\delta) \cdot \bar{\epsilon}^2}{\lambda(\delta) \cdot \bar{\epsilon}^2}. \tag{2.6}$$

For independent-revision schemes, $\lambda(\delta) \in (0, 1)$ *is the probability that only one agent revises actions for* N *stages conditioned on that at least one agent revises actions and* $\bar{\epsilon} \in (0, 1)$ *is a uniform lower bound on the probability for any* N*-length trajectory where only one agent revises actions in the coordinated scheme, which depends on the inherent exploration.*

As $\delta \to 0^+$, we have $\Lambda(\delta) \to 0$ since $\lambda(\delta) \to 1$. Note the resemblance of (2.6) to the second term on the right-hand side of (2.5). Correspondingly, the proof of Lemma 2 (Section A.2) follows from Lemma 1 by quantifying the distance between the transition matrices of the coordinated and independent-revision schemes.

Chapter 3

Team-Fictitious Play for Reaching Team-Nash Equilibrium in Multi-Team Games

In this chapter[1], we investigate the dynamics of multi-team competition, a setting increasingly relevant in domains ranging from robotics to financial markets [1, 2, 3, 4, 5]. To model such interactions, we consider multiple teams with possibly different number of team members. We introduce a novel class of games called *Zero-Sum Potential Team Games* (ZSPTGs). In this game, team members have networked interconnections, as depicted in Figure 3.1. Teams have *pairwise interactions*. For any opponent team strategy, team members effectively play an underlying potential game, as in distributed optimization applications, e.g., see [59, 60, 61]. Notably, ZSPTGs reduce to zero-sum polymatrix games [19] if each team has a single agent, and to potential games [13] if there is a single team. Additionally, widely studied two-team zero-sum games, e.g., see [8, 9, 25, 62], are a special case of ZSPTGs.

Within this framework, we propose a simple behavioural learning rule called the *Team-Fictitious-Play* (Team-FP) dynamics. In Team-FP, agents adopt a

[1]This chapter is dervied from the the author's paper [58]

smoothed best response strategy based on the most recent actions of their teammates and formed beliefs regarding the strategies of neighboring opponents. We show that the Team-FP dynamics reach near TNE in ZSPTGs. This means that the empirical average of team actions converge to the near best response each team can take against the average actions of other teams. We quantify the approximation error, showing it decreases with the level of exploration in the agents' responses. Similar to the FP dynamics, Team-FP is also *rational*: teams can learn (near) optimal strategies if opponent teams play stationary strategies. These results strengthen the applicability of TNE for predicting team behavior in multi-team competition and provide a practical rule for teams of self-interested agents to learn coordination in multi-team settings.

Using this framework, we establish that under Team-Fictitious-Play dynamics, the empirical averages of the teams' action profiles converge almost surely to a near Team-Nash Equilibrium. We provide an explicit bound on the approximation error, which decays with the level of exploration in the agents' responses. Furthermore, we demonstrate the rationality of these dynamics, showing that teams learn to play near-optimally against opponent teams utilizing stationary strategies.

A key challenge in our analysis is handling the non-stationary nature of learning, as opponent teams' strategies change over time. We address this by leveraging the optimal coupling lemma (e.g. see [55, Chapter 4]) and stochastic differential inclusion approximation methods (e.g., see [53, 54]) to the repeated play of games. Motivated from the recent interest in multi-agent reinforcement learning, we can extend Team-FP dynamics to finite horizon multi-team Markov games for both model-based and model-free cases. We discuss this extension and analyze its convergence numerically in §3.6.

This chapter is organized as follows. We describe ZSPTGs and the (independent) Team-FP dynamics, resp., in §3.1 and in §3.2. We provide analytical and numerical results, resp., in §3.3 and §3.5. We conclude the chapter with some remarks in §3.7.

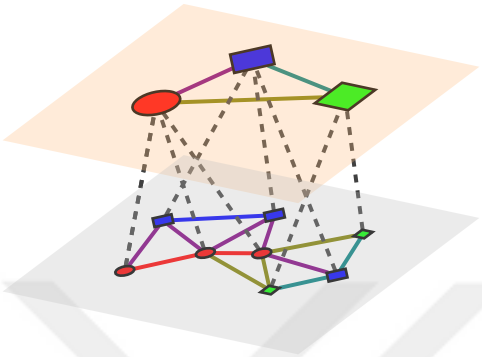


Figure 3.1: An illustration of networked interconnections agents from different teams. Nodes in bottom and top layers refer, resp., to team members and teams. Undirected edges represent the impact of actions on the payoff functions. We use different colors and shapes to represent agents from the same teams, and they are connected via dashed edges.

Notation: Given a finite set A, we let $\Delta(A)$ denote the probability simplex over A. We let $f(\mu) = \mathrm{E}_{a\sim\mu}[f(a)]$ for any probability distribution $\mu \in \Delta(A)$ and any functional $f : A \to \mathbb{R}$. Furthermore, we define the smoothed best response to any functional $q : A \to \mathbb{R}$ by

$$\mathrm{br}_\tau(q)(a) = \frac{\exp\left(q(a)/\tau\right)}{\sum_{\widetilde{a}\in A}\exp\left(q(\widetilde{a})/\tau\right)} \quad \forall a \in A \quad \Leftrightarrow \quad \mathrm{br}_\tau(q) = \arg\max_{\mu\in\Delta(A)}\{q(\mu) + \tau\mathcal{H}(\mu)\}, \tag{3.1}$$

for some temperature parameter $\tau > 0$, where $\mathcal{H}(\mu) := \mathrm{E}_{a\sim\mu}[-\log(\mu(a))]$ for all $\mu \in \Delta(A)$ is the entropy regularization.

3.1 Game Formulation

Consider a *multi-team game*, characterized by the tuple $\langle \mathcal{T}, (A^i, u^i)_{i\in\mathcal{I}}\rangle$, where $\mathcal{T}$ and $\mathcal{I}$ denote, resp., the teams' and agents' index sets, and A^i and $u^i : A \to \mathbb{R}$ (with $A := \prod_{j\in\mathcal{I}} A^j$) denote, resp., the agent i's finite action set and payoff function. Agents take their actions to maximize their payoff functions.

Definition 2 (Zero-sum Potential Team Game)**.** Let $\mathcal{I}^m$ denote the index set of agents in team m and $\underline{A}^m := \prod_{i\in\mathcal{I}^m} A^i$ denote the set of joint actions for team m. We say that a multi-team game is *zero-sum potential team game* (ZSPTG) if for

each team $m \in \mathcal{T}$, there exists a potential function $\phi^m : A \to \mathbb{R}$ such that

$$u^i(\hat{a}^i, a^{-i}, \underline{a}^{-m}) - u^i(a) = \phi^m(\hat{a}^i, a^{-i}, \underline{a}^{-m}) - \phi^m(a), \tag{3.2}$$

for all $(\hat{a}^i, a) \in A^i \times A$ and $i \in \mathcal{I}^m$, where $a^{-i} \coloneqq \{a^j\}_{j \in \mathcal{I}^m \setminus \{i\}}$ are the actions of other team members, $\underline{a}^{-m} \coloneqq \{\underline{a}^\ell\}_{\ell \neq m}$ are the action profiles of other teams, where $\underline{a}^\ell \in \underline{A}^\ell$ is team ℓ's action profile. The potential functions sum to zero, i.e., we have

$$\sum_{m \in \mathcal{T}} \phi^m(a) = 0 \quad \forall a \in A. \tag{3.3}$$

Furthermore, the actions have network separable interactions across teams such that we can separate the potential functions and correspondingly payoff functions as

$$\phi^m \equiv \sum_{\ell \neq m} \phi^{m\ell} \quad \text{and} \quad u^i \equiv \sum_{\ell \neq m} u^{i\ell} \quad \forall i \in \mathcal{I}^m, \tag{3.4}$$

for some $\phi^{m\ell} : \underline{A}^m \times \underline{A}^\ell \to \mathbb{R}$ and $u^{i\ell} : \underline{A}^m \times \underline{A}^\ell \to \mathbb{R}$.

The following example generalizes the potential game formulation for distributed optimization (e.g., see [59, 60, 61]) to two-team zero-sum potential games.

Example 1. Consider two teams of agents interacting over a network. We can represent their interactions via a graph $G = (V, E)$, where the set of vertices V refers to the agents and the set of (undirected) edges refers to their interactions. Let agent i from team m receive a local payoff $r^i : A^i \times \prod_{j:(i,j)\in E} A^j \to \mathbb{R}$ depending on the actions of the neighboring agents only. Agent i adds the neighboring team members' local payoffs whereas subtracts the other neighbors' local payoffs in his/her total payoff. In other words, his/her total payoff is given by

$$u^i \equiv \sum_{j:(i,j)\in E} \mathbb{I}_{\{j \in \mathcal{I}^m\}} r^j - \sum_{j:(i,j)\in E} \mathbb{I}_{\{j \notin \mathcal{I}^m\}} r^j. \tag{3.5}$$

This yields that the team m has the potential function

$$\phi^m \equiv \sum_{j \in \mathcal{I}^m} r^j - \sum_{j \notin \mathcal{I}^m} r^j, \tag{3.6}$$

and therefore, these potential functions sum to zero. However, the potential function is not generally the sum of the payoffs in potential games.

Remark 2 (General-sum ZSPTGs). In ZSPTGs, the underlying game can be a *general-sum* game although the team-potentials sum to zero, as described in (3.3). For example, consider two competing teams whose team members have identical payoffs corresponding to their team potentials. If the teams have different number of members, then the agents' payoffs do not sum to zero or a constant while the team potentials do so.

In the following, we introduce TNE, generalizing team-minimax equilibrium for two-team zero-sum games, e.g., see [7], to multi-team games. Particularly, at TNE, no team has an incentive to change their team strategy.

Definition 3 (Team-Nash Gap)**.** Given the strategy profile of teams $\{\pi^m \in \Delta(\underline{A}^m)\}_{m\in\mathcal{T}}$, we define the *team-Nash gap* for team m as

$$\mathrm{TNG}^m(\pi) \coloneqq \max_{\widetilde{\pi}\in\Delta(\underline{A}^m)} \left\{\phi^m(\widetilde{\pi}, \pi^{-m})\right\} - \phi^m(\pi), \tag{3.7}$$

and $\mathrm{TNG}(\pi) \coloneqq \sum_{m\in\mathcal{T}} \mathrm{TNG}^m(\pi)$, where $\pi^{-m} \coloneqq \{\pi^\ell\}_{\ell\neq m}$. Correspondingly, we say that the strategy profile of teams $\{\pi^m\}_{m\in\mathcal{T}}$ is ϵ-*TNE* if $\mathrm{TNG}(\pi) \leq \epsilon$.

3.2 Team-FP Dynamics

We first present the Team-FP dynamics combining the log-linear learning and fictitious play for learning in multi-team games played repeatedly, and then extend Team-FP to multi-team MGs in §3.6.

Let $a_k^i \in A^i$ denote the agent i's action at the kth repetition in the repeated play of the underlying ZSPTG. Correspondingly, $\underline{a}_k^m = (a_k^i)_{i\in T_m}$ denote the team m's action profile. Observing the joint actions of team m, agents $j \notin \mathcal{I}^m$ can form a belief about the team m's joint strategy. Let $\pi_k^m \in \Delta(\underline{A}^m)$ denote the belief they formed. Consider actions as pure strategy where the associated action gets played with probability 1. Then, agents $j \notin \mathcal{I}^m$ can update their beliefs about team m's strategy according to

$$\pi_{k+1}^m = \pi_k^m + \alpha_k(\underline{a}_k^m - \pi_k^m) \quad \forall k = 0, 1, \ldots \tag{3.8}$$

Algorithm 1 (Independent) Team-FP

initialize: $\{\pi_0^\ell\}_{\ell\neq m}$ and $\{a_{-1}^j\}_{j\in\mathcal{I}^m\setminus\{i\}}$ arbitrarily
for each stage $k=0,1,\ldots$ **do**
  play $a_k^i\sim \mathrm{br}_\tau(u^i(\cdot,a_{k-1}^{-i},\pi_k^{-m}))$ or $a_k^i=a_{k-1}^i$ in a coordinated way (or independently)
  update $\pi_{k+1}^\ell=\pi_k^\ell+\alpha_k(\underline{a}_k^\ell-\pi_k^\ell)$ for all $\ell\neq m$
end for

such that the belief π_{k+1}^m also corresponds to the (weighted) empirical average of the past action profiles $\{\underline{a}_0^m,\ldots,\underline{a}_k^m\}$.

Agent $i\in\mathcal{I}^m$ either takes the previous action a_{k-1}^i (i.e., $a_k^i=a_{k-1}^i$), or takes the action $a_k^i\sim\mathrm{br}_\tau(u^i(\cdot,a_{k-1}^{-i},\pi_k^{-m}))$ according to the smoothed best response (as described in (3.1)) to the previous actions of team members $a_{k-1}^{-i}:=\{a_{k-1}^j\}_{j\in\mathcal{I}^m\setminus\{i\}}$ and the beliefs $\pi_k^{-m}:=\{\pi_k^\ell\}_{\ell\neq m}$ formed about other teams. It is instructive to note that the definition of potential function $\phi^m(\cdot)$, as described in (3.2), yields that

$$\mathrm{br}_\tau\big(u^i(\cdot,a_{k-1}^{-i},\pi_k^{-m})\big)\equiv\mathrm{br}_\tau\big(\phi^m(\cdot,a_{k-1}^{-i},\pi_k^{-m})\big)\quad\forall i\in\mathcal{I}^m. \tag{3.9}$$

We introduce Team-FP and independent Team-FP dynamics depending on how agents update their actions. In the former, a single agent can get chosen randomly, as in the classical log-linear learning. In the latter, each agent can update his/her action with probability $\delta\in(0,1)$ independent of others, as in the independent log-linear learning. The latter addresses the coordination burden in the update of actions within teams. Algorithm 1 provides descriptions of these dynamics for the typical agent i from team m.

Remark 3 (Scalability). Agents can have networked interactions such that their payoff functions depend only on the actions of certain agents such as one/two-hop neighbors. For such cases, agents can form beliefs about these neighbors' strategies only, as if these neighbors play according to some stationary strategy. For example, assume that the payoff of agent $i\notin\mathcal{I}^m$ (outside team m) depends *only* on the actions of team-m members from some neighborhood $\mathcal{N}^i$, i.e., $\{j: j\in\mathcal{N}^i\cap\mathcal{I}^m\}$. Agent i can form a belief about these agents' strategies according

to

$$\pi_{k+1}^{im} = \pi_k^{im} + \alpha_k(\underline{a}_k^{im} - \pi_k^{im}) \quad \forall k = 0, 1, \dots \tag{3.10}$$

where $\underline{a}_k^{im} = \{a_k^j\}_{j \in \mathcal{N}^i \cap \mathcal{I}^m}$. Then, the linearity of the update rule yields that the local empirical average π_k^{im} corresponds to the marginalization of π_k^m such that

$$\pi_k^{im}(\{a^j\}_{j \in \mathcal{N}^i \cap \mathcal{I}^m}) = \sum_{\{a^j\}_{j \in \mathcal{I}^m \setminus \mathcal{N}^i}} \pi_k^m(\{a^j\}_{j \in \mathcal{I}^m}) \quad \forall k. \tag{3.11}$$

Therefore, local observations (within neighborhoods) would still be sufficient to follow Algorithm 1. We demonstrate the scalability of Team-FP by a large-scale experiment in §3.5, Figure 3.6.

We focus on the homogeneous cases where agents $i \notin \mathcal{I}^m$ have a common belief about team m's strategy. Homogeneous beliefs are possible if the agents have common initial beliefs and step sizes.

We moved the extension of (Independent) Team-FP to multi-team Markov games and its numerical analysis to §3.6.

3.3 Convergence Results

Team-FP and Independent Team-FP dynamics reduce, resp., to the classical log-linear learning and independent log-linear learning dynamics if there is only one team. These log-linear learning dynamics are known to reach team-optimal solution in potential games. For example, consider an n-agent potential game $\langle (A^i, u^i)_{i \in [n]} \rangle$ with the potential function $\phi(\cdot)$. In both dynamics, the action profiles played form irreducible and aperiodic Markov chains. Let $\widehat{\mu}$ and $\breve{\mu}$ denote the unique stationary distributions, resp., for the classical and independent versions. For the former, we have $\widehat{\mu} = \mathrm{br}_\tau(\phi(\cdot))$ [14, Section 3] and (3.1) yields that

$$0 \leq \max_a \{\phi(a)\} - \phi(\widehat{\mu}) \leq \tau \log |A|. \tag{3.12}$$

On the other hand, the smaller $\delta > 0$ implies closer stationary distributions in the classical and independent versions. Particularly, by utilizing Lemma 2, we

have

$$\|\widehat{\mu} - \breve{\mu}\|_1 \leq \Lambda(\delta, \epsilon_\phi), \tag{3.13}$$

for some function $\Lambda(\cdot)$ decaying to zero as $\delta \to 0^+$ for any $\epsilon_\phi > 0$, and $0 < \epsilon_\phi \leq \mathrm{br}_\tau(\phi(\cdot, a^{-i}))$ for any a^{-i} and i is a lower bound on local actions get played in the smoothed best response.

Team-FP dynamics have similar convergence guarantees for multi-team games under the following assumption on the step sizes used.

Assumption 1. The step size $\alpha_k \in [0, 1]$ satisfies the stochastic approximation conditions: $\alpha_k \to 0$ as $k \to \infty$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. Additionally, we have $\lim_{k\to\infty} \alpha_k/\alpha_{k+1} = 1$, and $\alpha_k - \alpha_{k+1} \geq \alpha_k \alpha_{k+1}$.

The last condition in Assumption 1 ensures that recent observations have comparable weight in the beliefs formed. The classical choice $\alpha_k = 1/(k+1)$, leading to the empirical averages of the actions played, satisfies Assumption 1.

The following theorem shows the convergence of Algorithm 1 to near TNE almost surely with the approximation levels (similar to (3.12) and (3.13)) inherited from the (independent) log-linear learning dynamics.

Theorem 2. *Given a ZSPTG characterized by $\langle \mathcal{T}, (A^i, u^i)_{i\in\mathcal{I}} \rangle$, let every agent follow either Team-FP or Independent Team-FP, as described in Algorithm 1. If Assumption 1 holds, then the team-Nash gap for $\pi_k := (\pi_k^m)_{m\in\mathcal{T}}$ satisfies*

$$\limsup_{k\to\infty} \mathrm{TNG}(\pi_k) \leq \begin{cases} \tau \log|A| & \textit{for Team-FP} \\ \tau \log|A| + |\mathcal{T}|^2 \overline{\phi} \cdot \Lambda(\delta, \epsilon) & \textit{for Independent Team-FP} \end{cases} \tag{3.14}$$

almost surely, where $\overline{\phi} := \max_{(m,l,a)} |\phi^{ml}(a)|$.

The key challenge in our convergence analysis is the non-stationarity arising from opponent teams adapting their strategies while the team members learn to coordinate against them. We address this by constructing a *reference scenario* where the team members' beliefs about opponent strategies are *frozen* over finite-length epochs, allowing them to hypothetically "team up" under these fixed

beliefs. By comparing the dynamics in the actual scenario and the reference scenario, and exploiting the averaging nature of belief formation, we can bound the approximation error between the two. The main proof concept, along with the Team-FP algorithm for a two-team scenario, is illustrated in Figure 3.2. This approach is similar to the one used in the next chapter to handle non-stationarity in Markov team problems. However, unlike the next chapter, we cannot directly show the error bound decay due to the lack of a contraction property in our dynamics.

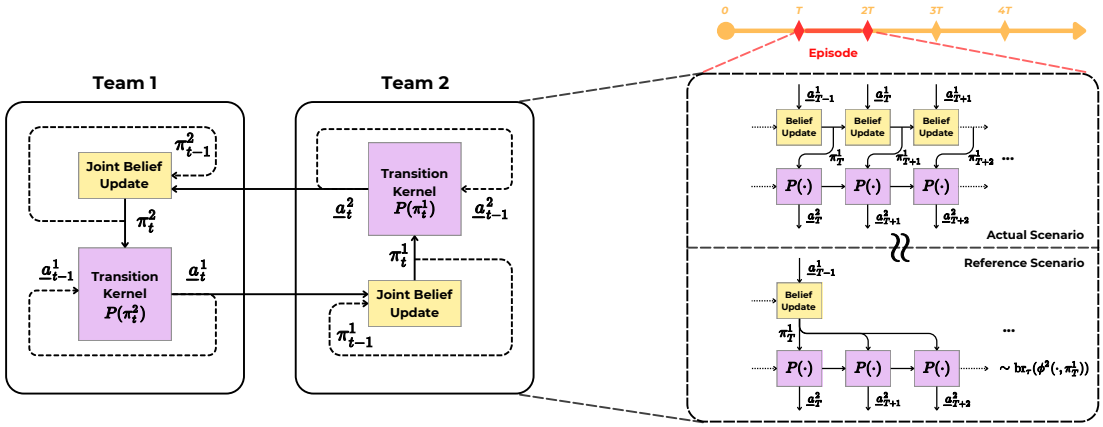


Figure 3.2: An illustration of Team-FP dynamics for two-team games on the left-hand side. Team actions change according to a transition kernel depending on the beliefs formed about the other teams. Dashed lines represent the time shift. On the right-hand side, we depict the key proof idea that we approximate the evolution of the team actions with a reference scenario where beliefs are stationary such that team actions form a homogeneous Markov chain whose unique stationary distribution corresponds to the best team response.

To overcome this limitation, we view Team-FP as smoothed fictitious play dynamics in zero-sum polymatrix games with an additive bounded error term. The additive error captures the fact that team members may not perfectly achieve team coordination. We then relax the problem by considering any approximation error within the formulated bounds, rather than the actual error. To address this relaxation, we leverage stochastic differential inclusion approximation methods [53, 54]. Finally, by constructing a suitable Lyapunov function addressing arbitrary bounded errors in continuous-time smoothed best response dynamics, we establish the convergence of the discrete-time Team-FP

updates.

3.4 Proof of Theorem 2

We can separate the proof into two main steps: (i) showing that team members can learn to team up approximately by constructing a reference scenario where beliefs got frozen, and (ii) addressing the bounded approximation error by leveraging the stochastic differential inclusion approximations.

3.4.1 Step (i) - Reference Scenario and Error Analysis

We divide the horizon into T-length epochs. Then, we can write the belief update (3.8) accumulated from $k = nT$ to $(n+1)T$ as

$$\pi^m_{(n+1)T} = \left(\sum_{k=nT}^{(n+1)T-1} (1-\alpha_k) \right) \cdot \pi^m_{nT} + \sum_{k=nT}^{(n+1)T-1} \alpha_k \left(\prod_{\ell=k+1}^{k_0+T-1} (1-\alpha_\ell) \right) \cdot \underline{a}^m_k \quad (3.15)$$

for any $n = 0, 1, \ldots$ denoting the epoch index. Let $\pi^m_{(n)} \coloneqq \pi^m_{nT}$ denote the belief about team m in epoch n. Furthermore, for notational simplicity, we define

$$\beta_k \coloneqq \alpha_k \prod_{\ell=k+1}^{(n+1)T-1} (1-\alpha_\ell) \quad \text{and} \quad \beta_{(n)} \coloneqq \sum_{k=nT}^{(n+1)T-1} \beta_k. \quad (3.16)$$

Then, we can simplify (3.15) as

$$\pi^m_{(n+1)} = (1-\beta_{(n)}) \cdot \pi^m_{(n)} + \beta_{(n)} \left(\sum_{k=nT}^{(n+1)T-1} \frac{\beta_k}{\beta_{(n)}} \cdot \underline{a}^m_k \right). \quad (3.17)$$

Lemma 3. *Due to 1, the step size $\beta_{(n)}$ decays to zero monotonically at a certain rate such that*

$$\sum_{n=0}^{\infty} \beta_{(n)} = \infty \quad \textit{and} \quad \sum_{n=0}^{\infty} \beta^2_{(n)} < \infty. \quad (3.18)$$

Let $\mathcal{F}_{(n)}$ be a filtration generated by the σ-algebra $\sigma(A_0, \ldots, A_{nT-1})$, where A_t denotes the joint actions taken by each team at stage t, i.e., $A_t = (\underline{a}_t^1, \ldots, \underline{a}_t^{|\mathcal{T}|})$. It is instructive to note that $\pi_{(n)}^m$ is $\mathcal{F}_{(n)}$-measurable. We define the joint action distributions of team m at time k in epoch n by

$$\mu_{(n),k}^m \coloneqq \mathrm{E}[\underline{a}_k^m \mid \mathcal{F}_{(n)}]. \tag{3.19}$$

Then, we can write (3.17) as in the form of stochastic approximation

$$\pi_{(n+1)}^m = (1-\beta_{(n)})\pi_{(n)}^m + \beta_{(n)} \left(\sum_{k=nT}^{(n+1)T-1} \frac{\beta_k}{\beta_{(n)}} \cdot \mu_{(n),k}^m + \omega_{(n+1)}^m \right), \tag{3.20}$$

where $\omega_{(n+1)}^m$ is a Martingale difference sequence defined by

$$\omega_{(n+1)}^m \coloneqq \sum_{k=nT}^{(n+1)T-1} \frac{\beta_k}{\beta_{(n)}} \left(\underline{a}_k^m - \mu_{(n),k}^m \right). \tag{3.21}$$

Now, let's consider a *reference scenario* for the analysis in which the beliefs (denoted by $\widehat{\pi}_t^m$) are only updated at the ends of T-length epochs. In other words, we have $\widehat{\pi}_t^m = \pi_{(n)}^m$ for all $nT \le t \le (n+1)T-1$ and $m \in \mathcal{T}$. Due to the fixed beliefs about the opponent plays, Team-FP dynamics reduce to the log-linear learning in the reference scenario. Let $\widehat{a}_{(n),k}^m$ denote the joint action of team m in the reference scenario. By the nature of the log-linear learning, $\{\widehat{a}_{(n),k}^m\}_{k=nT}^{\infty}$ form a homogeneous Markov chain (MC) even though the actual action profiles $\{a_k^m\}_{k=nT}^{\infty}$ do not necessarily do so. Denote the stationary distribution of the MC in the reference scenario by $\breve{\mu}_{(n),\star}^m$ and $\widehat{\mu}_{(n),\star}^m$ for Team-FP and Independent Team-FP, respectively. The former is given by $\breve{\mu}_{(n),\star}^m = \mathrm{br}_\tau(\phi^m(\cdot, \pi_{(n)}^{-m}))$ due to the log-linear learning update [26, 14]. Then, we can write the belief update (3.20) for team m as

$$\pi_{(n+1)}^m = (1-\beta_{(n)})\pi_{(n)}^m + \beta_{(n)} \left(\mathrm{br}_\tau(\phi^m(\cdot, \pi_{(n)}^{-m})) + \omega_{(n+1)}^m + e_{(n)}^m \right), \tag{3.22}$$

where we decompose the error $e_{(n)}^m \coloneqq \widehat{e}_{(n)}^m + \breve{e}_{(n)}^m$ as

$$\widehat{e}_{(n)}^m \coloneqq \sum_{k=nT}^{(n+1)T-1} \frac{\beta_k}{\beta_{(n)}} (\mu_{(n),k}^m - \widehat{\mu}_{(n),\star}^m) \tag{3.23}$$

$$\breve{e}_{(n)}^m \coloneqq \widehat{\mu}_{(n),\star}^m - \breve{\mu}_{(n),\star}^m. \tag{3.24}$$

The update (3.22) corresponds to the SFP dynamics of teams acting as a single decision-maker with the additive error $e^m_{(n)}$.

The following lemma, based on designing the optimal coupling between the actual and reference scenarios plays an important role in bounding $\|\widehat{e}^m_{(n)}\|$ from above. The proof of this lemma is based on Lemma 1, which is the key Lemma for the analysis in this thesis.

Lemma 4. *For some constants $c, d, \rho \geq 0$, we have*

$$\|\mu^m_{(n),k} - \widehat{\mu}^m_{(n),\star}\|_1 \leq c\,\rho^{k-nT} + d\,T\alpha_{nT} \quad \forall m \in \mathcal{T}. \tag{3.25}$$

By (3.23), Lemma 4 yields that we can bound the error $\widehat{e}^m_{(n)}$ from above by

$$\|\widehat{e}^m_{(n)}\|_1 \leq c \sum_{k=nT}^{(n+1)T-1} \frac{\beta_k}{\beta_{(n)}} \rho^{k-nT} + dT\alpha_{nT}. \tag{3.26}$$

Due to the no-recency-bias condition in 1, we have $\beta_{k+1}/\beta_k \leq 1$ by the definition (3.16). Then, we have monotonically decaying β_k, which yields

$$\frac{\beta_k}{\beta_{(n)}} \leq \frac{\beta_{nT}}{T\beta_{(n+1)T-1}} \leq \frac{\alpha_{nT}}{T\alpha_{(n+1)T}}. \tag{3.27}$$

By the assumption 1, we have

$$\lim_{n\to\infty} \frac{\alpha_{nT}}{T\alpha_{(n+1)T}} = \frac{1}{T} \lim_{n\to\infty} \prod_{k=nT}^{(n+1)T-1} \frac{\alpha_k}{\alpha_{k+1}} = \frac{1}{T}. \tag{3.28}$$

By (3.26), (3.27), (3.28), and the decaying property of α_k, we obtain

$$\limsup_{n\to\infty} \left\|\widehat{e}^m_{(n)}\right\|_1 \leq \frac{c}{T}\frac{1}{1-\rho} \quad \forall m \in \mathcal{T}. \tag{3.29}$$

On the other hand, $\breve{e}^m_{(n)}$ is non-zero only for Independent Team-FP and corresponds to the difference between the stationary distributions for the classical and independent log-linear learning. We can bound $\breve{e}^m_{(n)} \leq \Lambda(\delta, \epsilon)$ for some $\epsilon > 0$ based on (3.13). Hence, given any $\varepsilon > 0$, there exists N_ε such that

$$\left\|e^m_{(n)}\right\|_1 < C(\varepsilon, T) \quad \forall n \geq N_\varepsilon, \tag{3.30}$$

where

$$C(\varepsilon, T) \coloneqq \begin{cases} \varepsilon + \frac{c}{T}\frac{1}{1-\rho} & \text{for Team-FP} \\ \varepsilon + \frac{c}{T}\frac{1}{1-\rho} + \Lambda(\delta, \epsilon) & \text{for Independent Team-FP} \end{cases} \tag{3.31}$$

Note that $C(\varepsilon, T)$ can become arbitrarily close to $\Lambda(\delta, \epsilon)$ for sufficiently large T and small ε, which are chosen arbitrarily just for the analysis.

3.4.2 Step (ii) - Convergence Analysis with Relaxed Errors

We focus on the convergence analysis of (3.22) based on the bound (3.30). To this end, we define the set-valued mapping

$$F(\pi) \coloneqq \Big\{ \big(\mathrm{br}_\tau(\phi^m(\cdot, \pi^{-m})) - \pi^m + e^m\big)_{m\in\mathcal{T}} : \\ \|e^m\|_1 \leq C(\varepsilon, T) \text{ and } \mathrm{br}_\tau(\phi^m(\cdot, \pi^{-m})) + e^m \in \Delta(\underline{A}^m)\ \forall m \Big\} \tag{3.32}$$

for all $\pi \in \Pi \coloneqq \prod_{m\in\mathcal{T}} \Delta(\underline{A}^m)$. Then, the update (3.22) and (3.30) yield that for sufficiently large n, the empirical averages $\{\pi_{(n)}\}_{n\geq 0}$ satisfy

$$\pi_{(n+1)} - \pi_{(n)} - \beta_{(n)} \cdot \omega_{(n+1)} \in \beta_{(n)} \cdot F(\pi_{(n)}), \tag{3.33}$$

where the Martingale difference sequence $\{\omega_{(n+1)}\}$ is as described in (3.21). We have set inclusion in (3.33) different from the equality in (3.22) since we relax the update rule by considering any error satisfying the bound (3.30), rather than the actual error.

The following proposition shows that $F(\cdot)$ is a peculiar set-valued mapping. Particularly, given the sets X, Y from some underlying Euclidean space, we say that a set-valued function $\overline{F}(\cdot)$ mapping each point $x \in X$ to a set $\overline{F}(x) \subset Y$ is a *Marchaud map*, e.g., see [54, Definition 2.1] and [53, Hypothesis 1.1], if it satisfies the following conditions:

(i) $\overline{F}(\cdot)$ is upper semi-continuous, or equivalently, $\mathrm{Graph}(\overline{F}) = \{(x, y) : y \in \overline{F}(x)\}$ is a closed subset of $X \times Y$.

(ii) For all $x \in X$, the set $\overline{F}(x)$ is a non-empty, compact, and convex subset of Y.

(iii) There exists a $c > 0$ such that $\sup_{y \in \overline{F}(x)} \|y\| \leq c(1 + \|x\|)$ for all $x \in X$.

Proposition 1. *The set-valued function $F(\cdot)$ is a Marchaud map.*

Next, we approximate the relaxation (3.33) by the differential inclusion

$$\dot{\pi} \in F(\pi). \tag{3.34}$$

Particularly, due to Proposition 1 and (3.18), a linear interpolation of the iterative process $\{\pi_{(n)}\}_{n \geq 0}$ is a perturbed solution to the differential inclusion (3.34) [53, Proposition 1.4] and its limit set is internally chain transitive [53, Theorem 3.6]. Furthermore, we can characterize internally chain transitive sets (and therefore, the limit set of linear interpolations) through a Lyapunov function. Particularly, let $\Phi_t(\pi)$ be the set of solutions to (3.34) with initial point π. We say that a continuous function $\overline{V} : \Pi \to \mathbb{R}$ is a *Lyapunov function* of (3.34) for a subset $\Lambda \subset \Pi$ provided that

- $\overline{V}(\tilde{\pi}) < \overline{V}(\pi)$ for all $\pi \in \Pi \setminus \Lambda$, $\tilde{\pi} \in \Phi_t(\pi)$, $t > 0$,
- $\overline{V}(\tilde{\pi}) \leq \overline{V}(\pi)$ for all $\pi \in \Lambda$, $\tilde{\pi} \in \Phi_t(\pi)$, $t \geq 0$.

Given such a Lyapunov function for Λ, every chain internally chain transitive set L is contained in Λ if the set $\{\overline{V}(\pi) : \pi \in \Lambda\}$ has empty interior [53, Proposition 3.27]. Therfore, we propose the candidate Lyapunov function

$$V(\pi) = \max\left\{0, L(\pi) - \Xi(\lambda, \varepsilon, T)\right\}, \tag{3.35}$$

where the auxiliary terms are defined by

$$L(\pi) \coloneqq \sum_{m \in \mathcal{T}} \max_{\mu \in \Delta(\underline{A}^m)} \left\{\phi^m(\mu, \pi^{-m}) + \tau \mathcal{H}(\mu)\right\} \tag{3.36a}$$

$$\Xi(\lambda, \varepsilon, T) \coloneqq \tau \log|A| + \lambda |\mathcal{T}|^2 \overline{\phi} \cdot C(\varepsilon, T) > 0 \tag{3.36b}$$

for some arbitrary $\lambda > 1$ to characterize the long-run behavior of (3.33) based on (3.34).

Proposition 2. *The continuous function $V(\cdot)$ is a Lyapunov function of* (3.34) *for the set $\Lambda = \{\pi : V(\pi) = 0\}$.*

Proposition 2 yields that there exists some sequence $\{\zeta_k \in \mathbb{R}\}_{k\geq 0}$ such that $|\zeta_k| \to 0$ as $k \to \infty$ almost surely and

$$L(\pi_k) < \Xi(\lambda, \varepsilon, T) + \zeta_k \quad \forall k. \tag{3.37}$$

By (3.3), we can write $L(\pi_k)$ as

$$\begin{aligned} L(\pi_k) &= \sum_{m\in\mathcal{T}} \left(\max_{\mu\in\Delta(\underline{A}^m)} \left\{ \phi^m(\mu, \pi_k^{-m}) + \tau\mathcal{H}(\mu) \right\} - \phi^m(\pi_k) \right), \\ &\geq \sum_{m\in\mathcal{T}} \left(\max_{\mu\in\Delta(\underline{A}^m)} \left\{ \phi^m(\mu, \pi_k^{-m}) \right\} - \phi^m(\pi_k) \right). \end{aligned} \tag{3.38}$$

Since $\max_\mu \{\phi^m(\mu, \pi^{-m})\} - \phi^m(\pi) \geq 0$ for all π, the inequalities (3.37) and (3.38) yield that

$$0 \leq \sum_{m\in\mathcal{T}} \left(\max_{\mu\in\Delta(\underline{A}^m)} \left\{ \phi^m(\mu, \pi_k^{-m}) \right\} - \phi^m(\pi_k) \right) \leq \Xi(\lambda, \varepsilon, T) + \zeta_k \ \forall m \text{ and } n. \tag{3.39}$$

Recall that we can select $\lambda > 1$, $\varepsilon > 0$ and $T \in \mathbb{N}$ arbitrarily. The definitions (3.31) and (3.36b) yield that $\Xi(\lambda, \varepsilon, T)$ can be arbitrarily close to $\tau \log |A|$ for Team-FP and $\tau \log |A| + |\mathcal{T}|^2\overline{\phi} \cdot \Lambda(\delta, \epsilon)$ for Independent Team-FP. Furthermore, ζ_k decays to zero. Therefore, we can obtain (3.14) based on (3.39) and Definition 3. This completes the proof □.

The following corollary to the main result shows the rationality of the (independent) Team-FP dynamics.

Corollary 1. *Given a ZSPTG characterized by $\langle \mathcal{T}, (A^i, u^i)_{i\in\mathcal{I}} \rangle$, let agents from team m follow either Team-FP or Independent Team-FP while other teams play according to some stationary strategy π^{-m}. If Assumption 1 holds, then empirical average of the action profiles played by team m satisfy*

$$\limsup_{k\to\infty} \mathrm{TNG}^m(\pi_k^m, \pi^{-m}) \leq \begin{cases} \tau \log |\underline{A}^m| & \textit{for team-FP} \\ \tau \log |\underline{A}^m| + \overline{\phi} \cdot \Lambda(\delta, \epsilon) & \textit{for Independent Team-FP} \end{cases} \tag{3.40}$$

almost surely.

Theorem 2 can be generalized to the case where the rewards are random with bounded support, rather straightforwardly. Therefore, the proof of Corollary 1 follows from Theorem 2 if we view the underlying game as there is a single team receiving random rewards with bounded support.

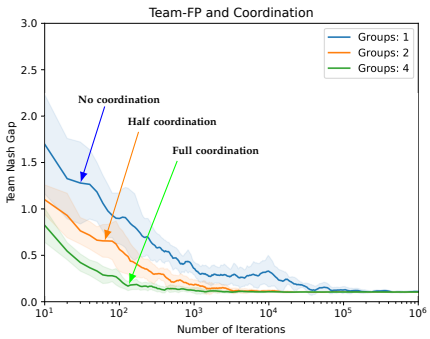
3.5 Illustrative Examples

In this section, we present various simulation results demonstrating the coordination speed of Team-FP and compare it to pure FP, no-regret algorithms, and a stationary opponent. We also observe the effect of parameters on the convergence speed. In addition, we examine the long-run behavior of team-FP in games beyond ZSPTG games, where we intuitively expect it to converge. All simulations are averaged over 10 independent trials to reduce the randomness. In all figures, the mean is plotted with a thick colorful line, while one standard deviation below and above the mean is shown with a shaded area of the same color. Also, for all simulations, temperature parameter τ is chosen to be 0.1 unless another option is mentioned. We conduct simulations for ZSPTG with two different setups: one with three teams, each consisting of three agents, and another with two teams, each consisting of four agents. Unless explicitly stated otherwise, the default setting consists of two teams, with four agents in each team. The step size is chosen to be $\alpha_k = 1/(k+1)$ for all simulations.

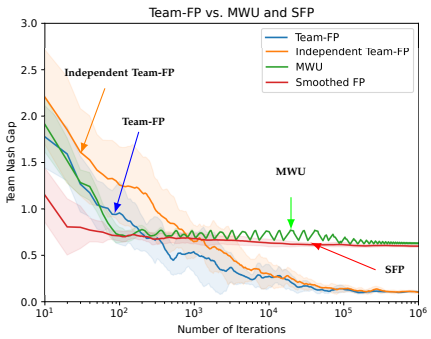
All the simulations are executed on a computer equipped with an Intel Xeon W7-3455 CPU and 128 GB RAM. Run-time for 10 independent trials over 10^6 iterations vary between 1-5 hours depending on the experiment.

3.5.0.0.1 Achieving Implicit Coordination in Team-FP In this section, we compare the performance of Team-FP to the explicit coordination of team members. We also compare Team-FP to algorithms such as Multiplicative Weights Update (MWU) and Smoothed FP (SFP), and show that these algorithms fail to achieve team coordination. We also show that Team-FP achieves near-optimum

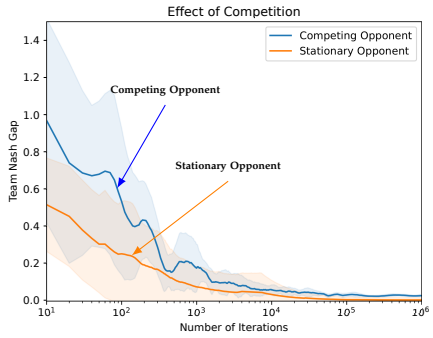
policy against a stationary opponent as stated in Corollary 1.



(a) Varying group sizes of 1, 2, and 4.



(b) Algorithm 1 (Independent) compared to SFP, and MWU



(c) Competitive vs Stationary Environment

Figure 3.3: All the above figures show the variation of TNG over time. (a) Comparison of different levels of explicit coordination for Team-FP: independent agents (group size 1), pairs of cooperating agents (group size 2), and fully coordinated teams (group size 4). (b) Performance of Team-FP and Independent Team-FP compared to MWU and SFP algorithms in a 2-team ZSPTG. (c) Convergence of Team-FP against stationary and competitive opponents in a 3-team ZSPTG.

Impact of Team Size in Team Coordination: In Team-FP self-interested team members act together without explicit communication in a coordinated way to reach TNE. We can measure how this independent cooperation compares to the explicit coordination of team members. For that, we propose an example

game with two teams and four agents in each team. For the first scenario, four agents act separately and use Team-FP. In the second scenario, the agents form groups of two and act in a coordinated way as if they are a single agent, using Team-FP. In the final scenario, all agents in a team behave as if they were a single agent, equivalent to the standard fictitious play (FP) dynamics in a two-person zero-sum game. This case mimics the ex-ante communication scheme from [8]. The scenarios are described by group sizes. Group sizes are one, two, and four respectively for these scenarios. The simulation results are presented in Figure 3.3a. As expected, the explicit coordination of all members in a team converges the fastest, followed by the explicit coordination of groups of two, and finally, Team-FP converges slowly as the agents do not coordinate explicitly.

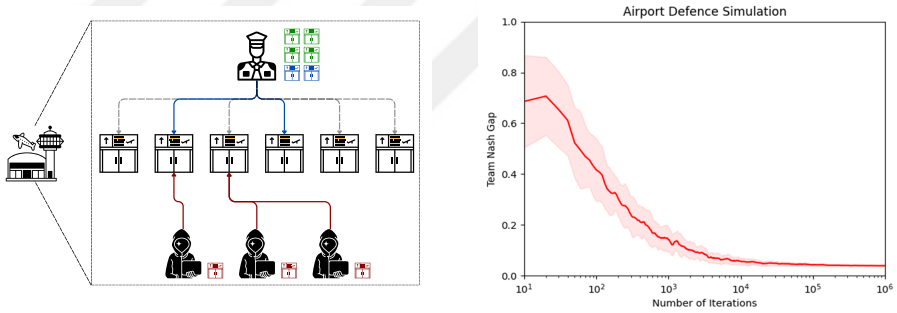
Team-FP Compared to MWU and SFP: The strong side of Team-FP is, even though the agents do not communicate, the average of joint actions of teams can reach TNE, unlike other algorithms. We compare the equilibrium behavior of team-FP with a well-known no-regret learning algorithm Multiplicative Weights Update (MWU), in which the average strategies converge to NE in zero-sum polymatrix games [19]. We also compare Team-FP with the usual SFP, where each agent holds beliefs about other agents and uses the smoothed best response against them. In Figure 3.3b, we see that both Team-FP and Independent Team-FP dynamics converge to near TNE, while the other algorithms fail to do so.

Rationality Against Stationary Opponent: In this part, we use 3-team setting where each team has 3 agents. We let a team using Team-FP compete against two other stationary teams and compared it with the performance of the same team in the same game when the opponents are also using Team-FP competitively. In Figure 3.3c, we observe that both algorithms converge, while the convergence is much faster against stationary opponents.

3.5.0.0.2 Team-FP in Application In this part, we provide an example to demonstrate that Team-FP has applications in various contexts.

Security Game Example: We model an airport security scenario as a two-team

game between defender and attacker teams. In our example, a security chief on the defender team faces three independent intruders on the attacker team, as shown in Figure 3.4a. The chief can defend a gate at a cost, while intruders decide whether to attack a gate or remain idle. Intruders gain or lose payoffs based on whether they attack undefended or defended gates, and the chief's payoffs mirror these outcomes. We conducted multiple trials and presented the evolution of the average Team-Nash Gap with standard deviations on the right side of Figure 3.4b. From a higher level, this example shows that team-minimax equilibrium can predict the outcome of games with different uncoordinated attackers. It also justifies the algorithms developed to compute team-minimax equilibrium efficiently.

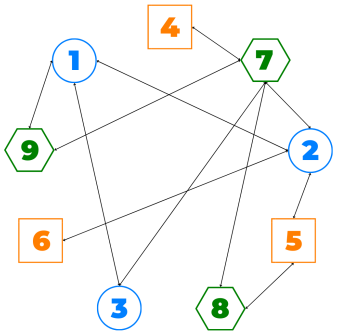


(a) (b)

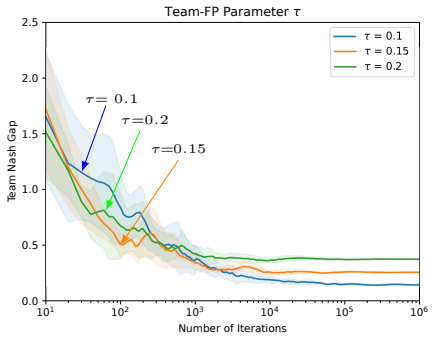
Figure 3.4: (a) The illustration of an airport security game: a security chief guarding the six gates of an airport against three different intruders making decisions autonomously. (b) The evolution of Team Nash Gap in airport security game, showing that Team-FP dynamics reach near team-minimax equilibrium.

3.5.0.0.3 Effect of parameters in Team-FP In this part, we examine how Team-FP performs for different τ and δ values. Given an example ZSPTG of three teams where each team has three agents in Figure 3.5a, we examine the evolution of TNG in the Team-FP dynamics for different values of $\tau \in \{0.1, 0.15, 0.2\}$. We also compare the evolution of NG in the Team-FP and Independent Team-FP dynamics for $\tau = 0.1$, $\delta = 0.2$, $\delta = 0.5$. All simulation results (see Figure 3.5) show convergence, and we observe lower final values

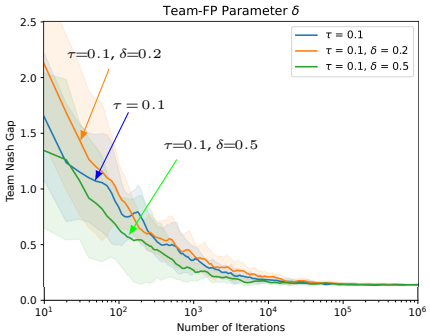
of NG(π) for smaller τ as we predicted (see. Figure 3.5b). The Independent Team-FP requires more iterations when $\delta = 0.2$ for its convergence, while it is much faster when $\delta = 0.5$ (see. Figure 3.5c). This is expected as updates are much more frequent as δ increases. However, increasing δ too much may harm the coordination behavior of the team.



(a) Simulation network for a 3 vs 3 vs 3 game.



(b) Varying τ in Team-FP



(c) Varying δ in Independent Team-FP

Figure 3.5: The 3-team experiments are tested on the randomly generated network structure (a). The other figures (b), and (c), shows the variation of TNG over iterations. (a) The simulation network for a multi-team ZSPTG, in which there are 3-teams with 3 agents in each team. (b) The impact of varying temperature parameter τ (0.1, 0.15, 0.2) in Algorithm 1 on the closeness to TNE. (c) The effect of different δ values (0.2, 0.5) in (Independent) Algorithm 1 on the convergence speed with τ fixed at 0.1

3.5.0.0.4 Large-scale Numerical Examples In this section, we give a large-scale experiment that show the scalability of Team-FP in a networked game where only 2-hop neighbors of agents affect their payoff function. We simulated a three-team game with nine agents per team, resulting in a large joint action space of size 2^{27}. After ten independent trials, we plotted the evolution of the Team-Nash Gap in Figure 3.6. Despite the problem's scale, the empirical averages of team actions converge to the Team-Nash equilibrium at a similar rate, even with sparse network interconnections, as shown in the top right of Figure 3.6.

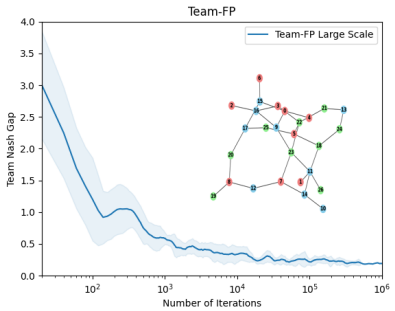
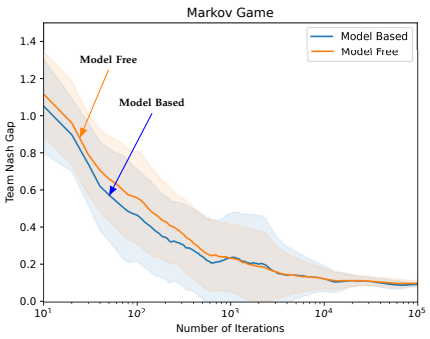


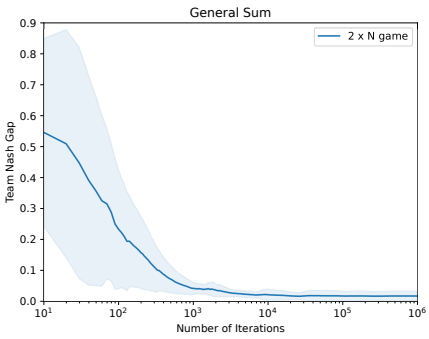
Figure 3.6: The evolution of Team Nash Gap in the large-scale example provided in the top right, showing that Team-FP dynamics reach near team-minimax equilibrium.

3.5.0.0.5 Beyond ZSPTG In this part, we try several other games for Team-FP without proof of convergence. We expect Team-FP to converge in 2xN and potential games other than zero-sum games as FP converges in these settings. For the first case, we try a game where one team has a single agent with only two actions with random rewards, while the other team has three agents with an underlying potential function. In this case, Team-FP converges (see. Figure 3.7b). In another setting, we try two teams of four agents. However, the potential functions for both teams are identical rather than summing to zero, resulting in a potential function that encompasses the individual potential functions. The team-FP algorithm converges to TNE in this case as well (see. Figure 3.7c). However, the equilibrium is not unique in this case. Finally, we create an extension of the team-FP for Markov games (see. §3.6) in an RL framework and

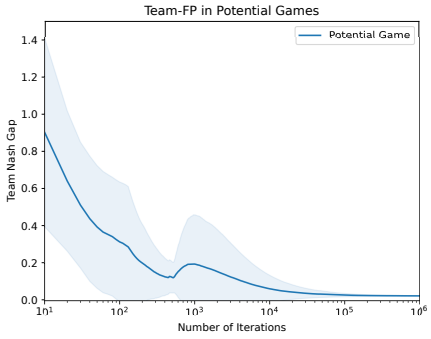
try simulations on this setting. In this case, there are two teams each having two agents, competing against each other in a finite horizon Markov game with a horizon length of ten. The number of states is two, and the state transition matrices are generated randomly. We observe that team-Nash Gap for MG's defined in (3.43), converges to near-zero.



(a) Markov Games



(b) 2xN General Sum Game



(c) Potential of Potentials Game

Figure 3.7: All the above figures describes the variation of TNG over iterations for Algorithms that are related to but outside the scope of ZSPTG. (a) The model-free and model-based Markov games of Algorithm 2, and 3, for a game of 2-team each with 2 agents, with 2 states and 10 horizon length. (b) The behavior of Team-FP dynamics in a 2xN general sum game, where a team competes against a single agent with random rewards. (c) The behavior of Team-FP dynamics in a potential game over the underlying potential functions.

3.6 Extension to Multi-team Markov Games

Markov games (MGs), introduced by [31], generalizes Markov decision processes (MDPs) to non-cooperative multi-agent environments. We can characterize a multi-team MG by the tuple $\langle H, \mathcal{T}, S, (A^i, r^i)_{i\in\mathcal{I}}, p, p_o\rangle$, where H is the horizon length, $\mathcal{T}$ and $\mathcal{I}$ again denote the index sets for the teams and agents, S denotes the *finite* set of states, and A^i and $r^i : S \times A \to \mathbb{R}$ denote, resp., the agent i's finite action set and *immediate reward* function, depending on current state and joint actions.[2] The state of the underlying game can change according to the transition kernel $p(\cdot \mid s, a)$, depending on the current state and joint actions, and the initial state is determined by the probability distribution $p_o \in \Delta(S)$.

Let each team m randomize their actions contingent on the current state $s \in S$ and stage $h \in [H] \coloneqq \{1, \ldots, H\}$ via a stationary strategy $\underline{\pi}^m : S \times [H] \to \Delta(\underline{A}^m)$. Note that team members do not necessarily randomize their actions independently. Given the strategy profile of teams $\underline{\pi} = (\underline{\pi}^m)_{m\in\mathcal{T}}$, agent i's utility function is defined by

$$U^i(\underline{\pi}) \coloneqq \mathrm{E}\left[\sum_{h=1}^{H} r^i(s_h, a_h)\right], \tag{3.41}$$

where the pair (s_h, a_h) denotes the state and action profile at stage h, the expectation is taken with respect to the randomness on these pairs (s_h, a_h) induced by the strategy profile $\underline{\pi}$ and the underlying transition kernel.

For each state s, let the reward functions $\{r^i(s, \cdot)\}_{i\in\mathcal{I}}$ induce a ZSPTG where team $m \in \mathcal{T}$ has the potential function $\phi^m(s, \cdot) : A \to \mathbb{R}$ satisfying (3.2), (3.3), and (3.4). Correspondingly, team m's utility function is given by

$$\underline{U}^m(\underline{\pi}) \coloneqq \mathrm{E}\left[\sum_{h=1}^{H} \phi^m(s_h, a_h)\right]. \tag{3.42}$$

Let $\underline{\Pi}^m \coloneqq \{\underline{\pi}^m \mid \underline{\pi}^m : S \times [H] \to \Delta(\underline{A}^m)\}$ denote the set of stationary strategy profiles for team m. Then, the following is the counterpart of Definition 3 for multi-team MGes.

[2]The results can be generalized to state-variant action sets rather straightforwardly.

Definition 4 (Team-Nash Gap for MGs)**.** Given the strategy profile of teams $\{\underline{\pi}^m \in \underline{\Pi}^m\}_{m\in\mathcal{T}}$, we define the *team-Nash gap* for team m as

$$\underline{\mathrm{TNG}}^m(\underline{\pi}) \coloneqq \max_{\widetilde{\underline{\pi}}\in\underline{\Pi}^m} \left\{\underline{U}^m(\widetilde{\underline{\pi}}, \underline{\pi}^{-m})\right\} - \underline{U}^m(\underline{\pi}), \tag{3.43}$$

and $\underline{\mathrm{TNG}}(\underline{\pi}) \coloneqq \sum_{m\in\mathcal{T}} \underline{\mathrm{TNG}}^m(\underline{\pi})$, where $\underline{\pi}^{-m} \coloneqq \{\underline{\pi}^\ell\}_{\ell\neq m}$. Correspondingly, we say that the strategy profile of teams $\{\underline{\pi}^m\}_{m\in\mathcal{T}}$ is ϵ-*TNE* if $\underline{\mathrm{TNG}}(\underline{\pi}) \leq \epsilon$.

Next, we describe the stage-game framework, going back to the introduction of Markov games [31], and also used in multi-agent reinforcement learning algorithms such as Minimax-Q [63] and Nash-Q [64]. Particularly, at each stage, the agents' joint actions determine the immediate rewards they receive and the next state, and therefore, the future rewards to be received. Given the strategy profile of teams $\underline{\pi} = (\underline{\pi}^m)_{m\in\mathcal{T}}$, let $v^i : S \times [H] \to \mathbb{R}$ denote agent i's *value function* such that $v^i(s)$ is the game value for the stage game associated with state s. Similarly, let $Q^i : S \times [H] \times A \to \mathbb{R}$ be the *Q-function* such that $Q^i(s, \cdot)$ corresponds to agent i's payoff function for the stage game associated with state s. The definition of the utility (3.41) yields that

$$Q^i(s, h, a) = r^i(s, a) + \mathbb{I}_{\{h<H\}} \sum_{s_+\in S} p(s_+ \mid s, a) \cdot v^i(s_+, h+1), \tag{3.44a}$$

$$v^i(s, h) = Q^i(s, h, \underline{\pi}(s, h, \cdot)). \tag{3.44b}$$

The Q-function and value function implicitly depend on $\underline{\pi}$.

We can extend Team-FP dynamics to MGs played repeatedly. Based on the stage game framework, agents play some stage game associated with each state $s \in S$ and stage $h \in [H]$ pair repeatedly whenever the underlying MG visits state s at stage h. Correspondingly, agents form beliefs about the opponent teams as if the opponent teams play according to some stationary strategies across these repetitions.

Let $\underline{\pi}_k^\ell : S \times [H] \to \Delta(\underline{A}^\ell)$ denote the beliefs formed by agents $i \notin \mathcal{I}^\ell$ about team ℓ, and $Q_k^i : S \times [H] \times A \to \mathbb{R}$ denote agent i's Q-function estimate at the kth repetition. If the underlying MG visits state s at stage h at the kth

Algorithm 2 Model-based (Independent) Team-FP for MGs

initialize: $\{\underline{\pi}_0^\ell(\cdot)\}_{\ell\neq m}$, $\{a_{-1}^j(\cdot)\}_{j\in\mathcal{I}^m\setminus\{i\}}$, and $Q_0^i(\cdot)$ arbitrarily for the typical agent $i\in\mathcal{I}^m$
for each repetition $k=0,1,\ldots$ **do**
  for each stage $h=1,\ldots,H$ **do**
    require: $\{\underline{\pi}_k^\ell(\cdot)\}_{\ell\neq m}$, $\{a_{k-1}^j(\cdot)\}_{j\in\mathcal{I}^m\setminus\{i\}}$, and $Q_k^i(\cdot)$
    observe current state s
    set $\bar{s}=(s,h)$
    play $a_k^i(\bar{s})\sim \mathrm{br}_\tau(Q_k^i(\bar{s},\cdot,a_{k-1}^{-i}(\bar{s}),\underline{\pi}_k^{-m}(\bar{s})))$ or $a_k^i(\bar{s})=a_{k-1}^i(\bar{s})$ in a coordinated way (or independently) simultaneously with other agents
    observe $a_{h,k}^{-i}=a_k^{-i}(\bar{s})$
    receive $r_{h,k}^i=r^i(s,a_{h,k})$
  end for
  require: trajectory $\{s_{h,k},a_{h,k}^{-i}\}_{h=1}^H$
  set

$$v_k^i(s,h)=Q_k^i(s,h,\underline{a}_k^m(s,h),\underline{\pi}_k^{-m}(s,h))$$
$$\widehat{Q}_k^i(s,h,\cdot)=r^i(s,\cdot)+\mathbb{I}_{\{h<H\}}\sum_{s_+\in S}p(s_+\mid s,\cdot)\cdot v_k^i(s_+,h+1)$$

  update the beliefs and the Q-functions

$$\underline{\pi}_{k+1}^\ell(s,h)=\underline{\pi}_k^\ell(s,h)+\mathbb{I}_{\{s=s_{h,k}\}}\alpha_{c_k(s,h)}\cdot\left(\underline{a}_k^\ell(s,h)-\underline{\pi}_k^\ell(s,h)\right)\quad\forall\ell\neq m$$
$$Q_{k+1}^i(s,h,\cdot)=Q_k^i(s,h,\cdot)+\mathbb{I}_{\{s=s_{h,k}\}}\alpha_{c_k(s,h)}\left(\widehat{Q}_k^i(s,h,\cdot)-Q_k^i(s,h,\cdot)\right)$$

  for all $(s,h)\in S\times[H]$
end for

repetition, then agent i follows Team-FP dynamics as if the payoff function is the Q-function estimate $Q_k^i(s,h,\cdot)$. Agents also recall the previous actions of their teams specific to each stage game. We denote agent i's previous action for the pair of state s and stage h until and including the kth repetition by $a_k^i(s,h)\in A^i$, with a slight abuse of notation. Then, at stage k, agent $i\in\mathcal{I}^m$ either takes the previous action for that stage game (i.e., $a_k^i(s,h)=a_{k-1}^i(s,h)$), or takes the action $a_k^i(s,h)\sim\mathrm{br}_\tau(Q_k^i(s,h,\cdot,a_{k-1}^{-i}(s,h),\underline{\pi}_k^{-m}(s,h)))$ according to the smoothed best response to the previous actions of the team members $a_{k-1}^{-i}(\cdot):=\{a_{k-1}^j(\cdot)\}_{j\in\mathcal{I}^m\setminus\{i\}}$ and the beliefs $\underline{\pi}_k^{-m}:=\{\underline{\pi}_k^\ell\}_{\ell\neq m}$ formed about other teams.

Algorithm 3 Model-free (Independent) Team-FP for MGs

initialize: $\{\underline{\pi}_0^\ell(\cdot)\}_{\ell\neq m}$, $\{a_{-1}^j(\cdot)\}_{j\in\mathcal{I}^m\setminus\{i\}}$, and $Q_0^i(\cdot)$ arbitrarily for the typical agent $i\in\mathcal{I}^m$
for each repetition $k=0,1,\ldots$ **do**
  for each stage $h=1,\ldots,H$ **do**
    require: $\{\underline{\pi}_k^\ell(\cdot)\}_{\ell\neq m}$, $\{a_{k-1}^j(\cdot)\}_{j\in\mathcal{I}^m\setminus\{i\}}$, and $Q_k^i(\cdot)$
    observe current state s
    set $\bar{s}=(s,h)$
    play $a_k^i(\bar{s})\sim \mathrm{br}_\tau(Q_k^i(\bar{s},\cdot,a_{k-1}^{-i}(\bar{s}),\underline{\pi}_k^{-m}(\bar{s})))$ or $a_k^i(\bar{s})=a_{k-1}^i(\bar{s})$ in a coordinated way (or independently) simultaneously with other agents
    observe $a_{h,k}^{-i}=a_k^{-i}(\bar{s})$
    receive $r_{h,k}^i=r^i(s,a_{h,k})$
  end for
  require: trajectory $\{s_{h,k},a_{h,k},r_{h,k}\}_{h=1}^H$
  set

$$v_k^i(s,h)=Q_k^i(s,h,\underline{a}_k^m(s,h),\underline{\pi}_k^{-m}(s,h))$$
$$\widehat{Q}_k^i(s,h,\cdot)=r_{h,k}^i+\mathbb{I}_{\{h<H\}}v_k^i(s_{h+1,k},h+1)$$

  update the beliefs and the Q-functions

$$\underline{\pi}_{k+1}^\ell(s,h)=\underline{\pi}_k^\ell(s,h)+\mathbb{I}_{\{s=s_{h,k}\}}\alpha_{c_k(s,h)}\cdot\left(\underline{a}_k^\ell(s,h)-\underline{\pi}_k^\ell(s,h)\right)\quad\forall\ell\neq m$$
$$Q_{k+1}^i(s,h,a)=Q_k^i(s,h,a)+\mathbb{I}_{\{(s,a)=(s_{h,k},a_{h,k})\}}\alpha_{c_k(s,h,a)}\left(\widehat{Q}_k^i(s,h,a)-Q_k^i(s,h,a)\right)$$

  for all $(s,h,a)\in S\times[H]\times A$
end for

Let $s_{h,k}$ and $a_{h,k}$ denote, resp., the state and action profile at stage h at the kth repetition. Then, given some reference step size $\{\alpha_c\}_{c\geq 0}$, agents $j\notin\mathcal{I}^m$ update their beliefs about team m's strategy according to

$$\underline{\pi}_{k+1}^m(s,h)=\underline{\pi}_k^m(s,h)+\lambda_k(s,h)\cdot(\underline{a}_k^m(s,h)-\underline{\pi}_k^m(s,h))\quad\forall k=0,1,\ldots, \tag{3.45a}$$
$$\lambda_k(s,h)=\mathbb{I}_{\{s=s_{h,k}\}}\alpha_{c_k(s,h)}, \tag{3.45b}$$

where $\underline{a}_k^m(\cdot):=\{a_k^j(\cdot)\}_{j\in\mathcal{I}^m}$ and $c_k(s,h)$ is the number of times state s get visited at stage h until the kth repetition. Correspondingly, by (3.44), each agent $i\in\mathcal{I}^m$ updates their Q-function estimates according to

$$Q_{k+1}^i(s,h,a)=Q_k^i(s,h,a)+\overline{\lambda}_k(s,h,a)\left(\widehat{Q}_k^i(s,h,a)-Q_k^i(s,h,a)\right), \tag{3.46}$$

where $\overline{\lambda}_k(s,h,a) \in [0,1]$ is also some step size. If the agent i knows the model of the underlying MG, then we have

$$\widehat{Q}_k^i(s,h,a) = r^i(s,a) + \mathbb{I}_{\{h<H\}} \sum_{s_+ \in S} p(s_+ \mid s,a) \cdot v_k^i(s_+, h+1), \tag{3.47a}$$

$$\overline{\lambda}_k(s,h,a) = \mathbb{I}_{\{s=s_{h,k}\}} \alpha_{c_k(s,h)}, \tag{3.47b}$$

for all $a \in A$ and

$$v_k^i(s,h) = Q_k^i(s,h,\underline{a}_k^m(s,h), \underline{\pi}_k^{-m}(s,h)) \quad \forall (s,h) \in S \times [H]. \tag{3.48}$$

If the agent does not know the model, then we have

$$\widehat{Q}_k^i(s,h,a) = r_{h,k}^i + \mathbb{I}_{\{h<H\}} v_k^i(s_{h+1,k}, h+1), \tag{3.49a}$$

$$\overline{\lambda}_k(s,h,a) = \mathbb{I}_{\{(s,a)=(s_{h,k},a_{h,k})\}} \alpha_{c_k(s,h,a)}, \tag{3.49b}$$

where $r_{h,k}^i$ denotes the reward received at stage h at the kth repetition and we approximate the expected continuation payoff by looking one stage ahead, as in the classical Q-learning algorithm, and $c_k(s,h,a)$ is the number of times the pair (s,a) gets realized at stage h until the kth repetition. Algorithms 2 and 3 provide descriptions of the extensions, resp., for the model-based and model-free cases from the perspective of the typical agent $i \in \mathcal{I}^m$ from the typical team $m \in \mathcal{T}$.

Remark 4. Algorithm 2 reduces to Algorithm 1 if $|S| = 1$ and $H = 1$.

3.7 Concluding Remarks

In this chapter, we introduced Team-FP, a novel fictitious play variant designed for multi-team games. We showed that Team-FP provably achieves near-TNE in ZSPTGs, with a quantifiable error bound. Our convergence analysis addressed the non-stationarity challenge arising from evolving opponent team strategies, by leveraging the optimal coupling lemma and stochastic differential inclusion approximation methods. We also extended Team-FP to multi-team Markov games, encompassing both model-based and model-free scenarios, with applications in multi-team reinforcement learning. Furthermore, we conducted

detailed numerical analysis of the Team-FP dynamics, focusing on the trade-off in learning to team up and competition in comparison to the classical FP and no-regret learning dynamics. We further examined the convergence of Team-FP dynamics to TNE in multi-team games beyond ZSPTGs. These results strengthen the theoretical foundations for applying TNE to predict decentralized team behavior and provide a framework for team learning in multi-team settings.

This chapter paves the way to further explore the behavior of decentralized teams in multi-team interactions. Future work could examine scenarios where team members follow different types of dynamics for teaming up within teams (other than log-linear learning) as in [65] to reach pareto efficient outcomes. Additionally, one could consider different types of dynamics for adapting to other teams' play (other than FP). Another extension of this work is on the analysis of multi-team cooperative games [66]. Numerical examples we conducted for multi-team games beyond ZSPTGs are also promising to show the provable convergence of Team-FP dynamics in multi-team (Markov) games reducing to games with the *fictitious play property* (e.g., see [67]) if the teams coordinate in acting as a single player.

Chapter 4

Logit-Q Dynamics for Efficient Learning in Stochastic Team Games

In this chapter[1], we turn our attention to the dynamics of cooperation within a single team operating in a stochastic environment. Stochastic games, introduced by [31], generalize Markov decision processes (MDPs) to multi-agent systems and serve as robust models for applications ranging from complex board games to autonomous system planning [9]. While learning in these environments has drawn significant interest, particularly regarding the convergence of non-equilibrium adaptation in zero-sum or identical-interest settings [32, 33, 34, 35], a critical gap remains: standard dynamics often fail to distinguish between optimal and suboptimal equilibria. Consequently, agents may converge to outcomes that are arbitrarily poor with respect to the team's common goal [44].

To address this challenge, we present an affirmative answer to whether efficient learning in stochastic teams can be achieved through the non-equilibrium adaptation of self-interested agents. We introduce a *new family* of algorithms called **Logit-Q dynamics**. These dynamics synthesize classical log-linear learning [14, 15] with Q-learning [69] within a *stage-game framework*. In this approach, we view the stochastic game as a sequence of repeated stage games associated

[1]This chapter is dervied from the the author's paper [68]

with the current state. The payoffs in these stage games are determined by the agents' evolving estimates of their Q-functions, allowing agents to learn Q-functions in a model-free manner akin to soft Q-learning without explicit knowledge of the underlying transition dynamics. We propose variants of this family based on whether action revisions are coordinated or independent, and whether agents estimate joint play via empirical averages or based on the most frequent joint action played.

A fundamental technical hurdle in this setting is the trade-off between immediate and continuation payoffs, which renders the learning environment **non-stationary**. Unlike in repeated games where payoffs are fixed, here the "stage game" payoffs depend on Q-function estimates that evolve as agents learn. Standard two-timescale approximation methods [33, 35, 34] are often insufficient for tracking efficient equilibria in such dynamic landscapes. To overcome this, we develop a novel **epoch-based analytical framework**. We divide the time horizon into epochs of growing length, ensuring that, under vanishing step sizes, Q-function estimates become quasi-stationary within each epoch. We then approximate the actual dynamics using a fictional scenario where Q-estimates are fixed, deriving new perturbation bounds to couple the perturbed (actual) and unperturbed (fictional) processes [56].

Using this framework, we establish that under Logit-Q dynamics, agents' Q-function estimates converge almost surely to those associated with the efficient equilibrium, and the estimated play converges to the efficient stationary policy. Furthermore, we demonstrate the rationality of these dynamics against opponents using pure stationary strategies and extend our results to single-controller stochastic games where stage payoffs align with a potential function. Such single-controller assumptions are common in multi-agent reinforcement learning applications [70, 71, 72].

The chapter is organized as follows. We describe stochastic games in Section 4.1 and the logit-Q dynamics in Section 4.2. We present the convergence result and the proofs, resp., in Section 4.3 and Section 4.4. In Section 4.5, we provide numerical examples, corroborating the convergence results in Section 4.3. We

conclude the chapter with some remarks in Section 4.6.

4.1 Stochastic Games

Formally, an N-agent stochastic game, introduced by [31], is a multi-stage game played over infinite horizon that can be characterized by the tuple $\langle S, (A^i)_{i\in[N]}, (r^i)_{i\in[N]}, p, \gamma\rangle$. At each stage $t = 0, 1, \ldots$, the game visits a state s from a *finite* set S and each agent i simultaneously takes an action a^i from a *finite* set A^i to collect *stage-payoffs*. The formulation can be extended to state-variant action sets rather straightforwardly. Agent i's stage-payoff is denoted by $r^i : S \times A \to [0, 1]$. Similar to MDPs, stage-payoffs and transition of the game across states depend only on the current state visited and current *action profile* $a = \{a^i\}_{i\in[N]}$ played. Particularly, $p(s' \mid s, a)$ for each $(s, a, s') \in S \times A \times S$ denotes the probability of transition from s to s' under action profile a, where $A := \prod_i A^i$. Lastly, $\gamma \in [0, 1)$ is the discount factor.

Assume that the agents can correlate their randomized actions and consider joint Markov stationary strategies $\pi : S \to \Delta(A)$ determining the probability of the joint actions to be played by all agents contingent on the current state.[2] Given π and the *initial* state $s \in S$, agent i's utility is given by

$$U^i(\pi; s) := \mathrm{E}_{a_t\sim\pi(s_t)}\left[\sum_{t=0}^{\infty} \gamma^t r^i(s_t, a_t) \mid s_0 = s\right], \tag{4.1}$$

where (s_t, a_t) denotes the pair of state and action profile realized at stage t. The expectation is taken with respect to the randomness on (s_t, a_t) induced by the underlying transition kernel and the joint strategy π.

Given a joint Markov stationary strategy $\pi : S \to \Delta(A)$, we denote the joint strategy of all agents except i as π^{-i}, where $\pi^{-i}(s) \in \Delta(A^{-i})$ represents the marginal distribution of $\pi(s)$ over $A^{-i} := \prod_{j\neq i} A^j$. Additionally, we define the joint strategy that randomizes actions at each state s using the product

[2] We let $\Delta(A)$ denote the probability simplex over the finite set A.

distribution $\tilde{\pi}^i(s) \times \pi^{-i}(s) \in \Delta(A)$ as $\tilde{\pi}^i \times \pi^{-i}$. Below, we introduce equilibrium concepts for such joint Markov strategies.

Definition 5. A Markov stationary strategy $\pi : S \to \Delta(A)$ is an ϵ-approximate perfect Markov coarse correlated equilibrium (abbreviated as ϵ-CCE) if it satisfies

$$U^i(\pi; s) \geq U^i(\tilde{\pi}^i \times \pi^{-i}; s) - \epsilon \tag{4.2}$$

for all $\tilde{\pi}^i : S \to \Delta(A^i)$, $i \in [N]$, and $s \in S$ [73, Definition 2.1], where $\epsilon \geq 0$.

Notice that CCEs are defined for joint Markov stationary strategies. If we restrict our attention to product strategies, we recover the definition of Nash equilibria: an ϵ-CCE becomes an ϵ-approximate perfect Markov Nash equilibrium if the joint strategy is a product strategy, i.e., $\pi = \bigtimes_i \pi^i$ for some $\pi^i : S \to \Delta(A^i)$ and $i \in [N]$ [73, Definition 2.2].

In stochastic games, a perfect Markov Nash equilibrium (also known as Markov perfect equilibrium) always exists [74], implying the existence of CCE as well. However, multiple equilibria may arise, leading to different outcomes in terms of performance metrics. In identical-interest stochastic games (also known as stochastic teams), where all agents share the same reward function, i.e., $r^i(s, a) = r(s, a)$ for all (s, a) and some common $r : S \times A \to \mathbb{R}$, a natural performance metric is the common utility, defined as

$$U(\pi; s) := \mathrm{E}_{a_t \sim \pi(s_t)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right]. \tag{4.3}$$

Correspondingly, we say that a joint Markov stationary strategy $\pi : S \to \Delta(A)$ is ϵ-efficient for some $\epsilon > 0$ if $U(\pi; s) \geq U(\tilde{\pi}; s) - \epsilon$ for all $\tilde{\pi} : S \to \Delta(A)$ and $s \in S$. By definition, an ϵ-efficient strategy is also an ϵ-CCE. Moreover, an efficient joint strategy always exists, as it corresponds to the unique solution of the Bellman optimality equations for U which are guaranteed to exist [75].

In this chapter, we present learning dynamics reaching such efficient equilibria. To this end, we focus on the *stage-game framework*, going back to the introduction of stochastic games by [31]. Particularly, given the joint Markov

stationary strategy π, let $v_\pi^i : S \to \mathbb{R}$ be the agent i's *value function* such that $v_\pi^i(s)$ is the value of state s for agent i if the game starts at state s and the agents play according to π, i.e., $v_\pi^i(s) = U^i(\pi; s)$. Correspondingly, let $Q_\pi^i : S \times A \to \mathbb{R}$ be the agent i's *Q-function* such that $Q_\pi^i(s,a)$ is the value of state-action pair (s,a) for agent i. By (4.1), the Bellman operator for a strategy yields that

$$Q_\pi^i(s,a) = r^i(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) \cdot v_\pi^i(s') \quad \forall (s,a), \tag{4.4a}$$

$$v_\pi^i(s) = \mathrm{E}_{a \sim \pi(s)} \left[Q_\pi^i(s,a) \right] \quad \forall s. \tag{4.4b}$$

Then, we can view stochastic games as the repeated play of stage games whose payoffs are the Q-functions associated with how they play in these stage games, whenever the associated state gets visited.

4.2 Logit-Q Dynamics for Stochastic Games

In this section, we introduce the *logit-Q dynamics* within the stage-game framework. In this framework, agents follow *(independent) log-linear learning* in each stage game as if the stage-game payoffs are the Q-functions estimated. They estimate the Q-functions associated with how they play in these stage games by solving the fixed-point problem (4.4) iteratively based on the game history. To this end, each agent i keeps track of

- $a_t^j(s) \in A^j$ for each $j \in [N]$ denoting agent j's latest action at state s until and including stage t, with a slight abuse of notation,
- $c_t(s,a)$ and $c_t(s) \in \{0, 1, \ldots\}$ denoting, resp., the number of times the pair (s,a) and s get realized until and including stage t,
- $Q_t^i(s,a) \in \mathbb{R}$ denoting the Q-function estimate of the pair (s,a) at stage t.

Recall the friction in the action updates of the log-linear learning, also known as the *lock-in property*. To model such friction, we consider an i.i.d. process

$\{I_t \subset [N]\}_{t\geq 0}$ determining the agents updating their actions. For example, agent i revises his/her action at stage t if and only if $i \in I_t$. For the distribution of the revision process, we consider two schemes. In the classical *coordinated* scheme, we have $I_t \sim \mathcal{P}^{\text{coor}}$ and[3] only a randomly selected agent can update his/her action. In the *independent* scheme, we have $I_t \sim \mathcal{P}^{\text{ind}}$ and agents update their actions independently with probability $\delta \in (0, 1)$. The distributions $\mathcal{P}^{\text{coor}}$ and $\mathcal{P}^{\text{ind}}$ are defined as follows:

$$\mathcal{P}^{\text{coor}}(I) := \mathbb{1}_{\{|I|=1\}} \frac{1}{N} \quad \forall I \subset [N], \quad \mathcal{P}^{\text{ind}}(I) = \delta^{|I|}(1-\delta)^{N-|I|} \quad \forall I \subset [N], \tag{4.5}$$

Let $s_t \in S$ be the state at stage t. Each agent $i \in I_t$ revises his/her action according to the *smoothed* best response *as if* the game payoffs are $Q_t^i(s_t, \cdot)$ and the other agents are playing the latest actions $a_{t-1}^{-i}(s_t) = \{a_{t-1}^j(s_t)\}_{j\neq i}$. Particularly, given $q \in \mathbb{R}^B$ for some finite set B, we define the smoothed best response by

$$\underline{\text{br}}(q)(b) := \frac{\exp(q(b)/\tau)}{\sum_{b'\in B} \exp(q(b')/\tau)} \quad \forall b \in B. \tag{4.6}$$

Then, we have $a_t^i \sim \underline{\text{br}}(Q_t^i(s_t, \cdot, a_{t-1}^{-i}(s_t)))$ and the parameter $\tau > 0$ balances exploration vs exploitation. For example, agent i plays best response(s) with higher probabilities (whose sum goes to 1 as $\tau \to 0$) and can explore any action with some positive probabilities (going to $1/|A^i|$ as $\tau \to \infty$). On the other hand, each agent $j \notin I_t$, not revising their actions, plays their latest action at that state, i.e., $a_t^j = a_{t-1}^j(s_t)$.

Agents estimate the joint Markov stationary strategy $\pi : S \to \Delta(A)$ from the game history in one of two ways. In the *averaging* approach, the estimated play is the average of the joint actions, and in the *exploration-free* approach, which mitigates exploration effects (due to $\tau > 0$ in (4.6)), the estimated play is the most frequently played actions. For both cases, we define π_t^{ave}, and π_t^{free}

[3]For a finite set I, we denote its number of elements by $|I|$.

respectively as follows:

$$\pi_t^{\text{ave}}(s)(a) = \frac{c_t(s,a)}{c_t(s)} \quad , \quad \pi_t^{\text{free}}(s)(a) = \begin{cases} 1/|A_t(s)| & \text{if } a \in A_t(s), \\ 0 & \text{otherwise}, \end{cases} \quad \forall (s,a), \tag{4.7}$$

with $A_t(s) = \arg\max_a \{c_t(s,a)\} \subset A$.

Let $\pi_t^{\text{ave}}(s)$ and $\pi_t^{\text{free}}(s)$ be uniform for $c_t(s) = 0$, without loss of generality.

To compute the Q-function associated with the estimated play, the agents can solve (4.4) iteratively and use *one-stage look ahead* to address the unknown transition kernel, as in the classical Q-learning. After observing the next state s_{t+1} at stage $t+1$, each agent i updates his/her Q-function estimate for the pair (s_t, a_t) according to

$$Q_{t+1}^i(s_t,a_t) = Q_t^i(s_t,a_t) + \beta_{c_t(s_t,a_t)} \left(r_t^i + \gamma v_t^i(s_{t+1}) - Q_t^i(s_t,a_t) \right), \tag{4.8}$$

where $r_t^i := r^i(s_t,a_t)$ is the stage-payoff received and v_t^i denotes the value function estimate. Here, the agents use the value $v_t^i(s_{t+1})$ of the next state as an approximation of the expected continuation payoff $\sum p(s' \mid s,a) v_t^i(s')$ in (4.4a). Given the estimated play π_t, the value v_t^i is given by

$$v_t^i(s_{t+1}) = \mathrm{E}_{a \sim \pi_t(s_{t+1})}[Q_t^i(s_{t+1},a)] \tag{4.9}$$

due to (4.4b). The *reference* step size $\beta_c \in (0,1)$ for all $c = 0,1,\ldots$ implies that the agents update their Q-function estimates to convex combinations of the previous estimates Q_t^i and the stage-games' (estimated) payoffs $r_t^i + \gamma \mathrm{E}_{a \sim \pi_t(s_{t+1})}[Q_t^i(s_{t+1},a)]$. Since agents do not observe the counterfactuals about the joint actions that are not played and the states that are not visited in the current stage, they do not update the other Q-function estimates, i.e., $Q_{t+1}^i(s,a) = Q_t^i(s,a)$ for all $(s,a) \neq (s_t,a_t)$.

Algorithm 4 describes a family of logit-Q dynamics that differ in their revision processes $\mathcal{P}$ and estimated plays π_t, as tabulated in Table 4.1. Recall that agents update their Q-functions after observing the next state and they initialize

Algorithm 4 Logit-Q Dynamics

Require: each agent can observe (s_t, a_t) and keep track of $a_t(\cdot)$ and $c_t(\cdot)$

1: **Input:** initializations $Q_0^i(s,a) = 0$ for all $(i,s,a) \in [N] \times S \times A$.
2: **for** each stage $t = 0, 1, \ldots$ **do**
3: s_t gets realized
4: **if** $t = 0$ **then** jump to Step 6
5: each agent i **updates** the Q-function estimates for stage $t-1$ according to

$$Q_t^i(s,a) = Q_{t-1}^i(s,a) + \mathbb{1}_{\{(s,a)=(s_{t-1},a_{t-1})\}} \cdot \beta_{c_{t-1}(s,a)} \cdot (\widehat{Q}_t^i - Q_{t-1}^i(s,a)),$$

where $\widehat{Q}_t^i := r_{t-1}^i + \gamma \mathrm{E}_{a \sim \pi_{t-1}(s_t)}[Q_{t-1}^i(s_t, a)]$ and $\pi_{t-1}(\cdot)$ depends on $c_{t-1}(\cdot)$

6: $I_t \sim \mathcal{P}$ gets realized
7: each agent $i \in I_t$ **plays** $a_t^i \sim \underline{\mathrm{br}}(Q_t^i(s_t, \cdot, a_{t-1}^{-i}(s_t)))$
8: each agent $i \notin I_t$ **plays** $a_t^i = a_{t-1}^i(s_t)$
9: **end for**

Table 4.1: A family of the logit-Q dynamics with the parameters $\mathcal{P}$ and π_t, as described in (4.5), and (4.7)

Logit-Q Dynamics	Revision Process $\mathcal{P}$	Estimated Play π_t
Averaged Logit-Q (AL-Q)	$\mathcal{P}^{\mathrm{coor}}$	π_t^{ave}
Exploration-free Logit-Q (EL-Q)	$\mathcal{P}^{\mathrm{coor}}$	π_t^{free}
Averaged Independent Logit-Q (AIL-Q)	$\mathcal{P}^{\mathrm{ind}}$	π_t^{ave}
Exploration-free Independent Logit-Q (EIL-Q)	$\mathcal{P}^{\mathrm{ind}}$	π_t^{free}

Q-function estimates as $Q_0^i(s,a) = 0$ for all (i,s,a) rather than arbitrarily. While such coordination can be challenging for ad hoc team formations, the step sizes $\beta_c \in (0,1)$ suppress the impact of the initialization on the estimates at the future stages and we examine the impact of arbitrary initializations numerically later in Section 4.5. Moreover, this initialization and $\beta_c \in (0,1)$ ensure that $0 \leq Q_t^i(s,a) \leq 1/(1-\gamma)$ for all (i,s,a) (since $r^i : S \times A \to [0,1]$). Hence, agents revising their actions always select any action with a probability uniformly bounded away from zero by (4.6).

4.3 Convergence Results

In this section, we characterize the convergence properties of Algorithm 4 based on the following assumption on the step size β_c.

Assumption 2. The reference step size $\beta_c \in (0, 1)$ satisfies

(i) Monotonic decay: $\beta_c \to 0$ as $c \to \infty$ and $\beta_c > \beta_{c'}$ for all $c < c'$,

(ii) Sufficiently slow decay: $\sum_c \beta_c = \infty$,

(iii) Sufficiently fast decay: $c\beta_{\lfloor mc \rfloor} \to 0$ as $c \to \infty$ for any $m \in (0, 1]$.[4]

Assumption 2-(i) and Assumption 2-(ii) are standard in stochastic approximation, and Assumption 2-(iii) is standard for asynchronous two-timescale stochastic approximation if we view $1/(c + 1)$ as the step size for the dynamics evolving at the fast timescale [76, Section 4.2]. For example, $\beta_c = 1/((c + 1) \log(c + 1))$ satisfies Assumption 2. We also note that Assumption 2-(iii) implies $\sum_c \beta_c^2 < \infty$, playing an important role in addressing the stochastic approximation noise.

The following theorem characterizes the convergence properties of the logit-Q dynamics for stochastic teams.

Table 4.2: Error bounds in Theorem 3, where $\Lambda(\delta)$ and $\Xi(\delta)$ as described later in (4.53) and (4.60), respectively. Both $\Lambda(\delta)$ and $\Xi(\delta)$ decay to zero as $\delta \to 0^+$.

	AL-Q	AIL-Q	EL-Q	EIL-Q
e_Q (4.10a)	$\frac{\gamma}{1-\gamma} \cdot (\tau \log \lvert A \rvert)$	$\frac{\gamma}{1-\gamma} \cdot (\tau \log \lvert A \rvert + \frac{1}{1-\gamma} \cdot \Lambda(\delta))$	0	$\frac{\gamma}{1-\gamma} \cdot \Xi(\delta)$

Theorem 3. *Given a stochastic team $\langle S, (A^i)_{i\in[N]}, r, p, \gamma\rangle$ with the common objective $U(\cdot)$, consider that every agent follow Algorithm 4 with the same revision-process and estimated-play schemes, as in Table 4.1. Suppose that Assumption 2 holds and the underlying stochastic game is*

[4] Let $\lfloor r \rfloor := \max\{z \in \mathbb{Z} : z \leq r\}$ for any $r \in \mathbb{R}$ denote the floor function.

irreducible in the sense that there exist at least one Markov stationary strategy, for which the states visited form an irreducible Markov chain.[5] *Then, for each* $i \in [N]$*, we have*

$$\limsup_{t\to\infty} \left| Q_t^i(s,a) - Q_*(s,a) \right| \leq e_Q \quad \forall (s,a) \tag{4.10a}$$

$$\limsup_{t\to\infty} \left| \max_{\pi} \{U(\pi; s)\} - U(\pi_t; s) \right| \leq \frac{1+\gamma}{\gamma(1-\gamma)} \cdot e_Q \quad \forall s \tag{4.10b}$$

almost surely, where the team-optimal Q-function $Q_* : S \times A \to \mathbb{R}$ *is the unique fixed point of the following Q-function variant of the Bellman optimality operator:*

$$Q_*(s,a) = r(s,a) + \gamma \sum_{s'} p(s' \mid s,a) \max_{a' \in A} \{Q_*(s',a')\} \quad \forall (s,a) \tag{4.11}$$

and the error bound $e_Q \geq 0$ *is as tabulated in Table 4.2 for different revision-process and estimated-play schemes.*

Theorem 3 shows that in Algorithm 4, the Q-function estimates and the estimated plays almost surely converge, resp., to the team-optimal Q-functions and efficient stationary equilibrium *approximately*, with quantifiable error bounds (see Table 4.2). The proof is moved to Section 4.4.

The following corollary to Theorem 3 shows the rationality of Algorithm 4 against others playing according to pure stationary strategies.

Corollary 2. *Given a general-sum stochastic game* $\langle S, (A^i, r^i)_{i\in[N]}, p, \gamma \rangle$*, consider that agent* i *follows AIL-Q or EIL-Q while the other agents* $j \neq i$ *play according to pure Markov stationary strategy, e.g.,* $\pi^{-i} : S \to A^{-i}$*. Suppose that Assumption 2 holds and the states visited under the strategy* $\tilde{\pi}^i \times \pi^{-i}$ *form an irreducible Markov chain for any pure Markov stationary strategy* $\tilde{\pi}^i : S \to A^i$*. Then, we have*

$$\limsup_{t\to\infty} |Q_t^i(s, a^i, \pi^{-i}(s)) - q_*^i(s,a^i)| \leq e_Q \quad \forall (s,a^i) \tag{4.12a}$$

$$\limsup_{t\to\infty} |\max_{\tilde{\pi}^i} \{U^i(\tilde{\pi}^i \times \pi^{-i}; s)\} - U^i(\pi_t; s)| \leq \frac{1+\gamma}{\gamma(1-\gamma)} \cdot e_Q \quad \forall s \tag{4.12b}$$

[5] *Such irreducibility assumptions are widely used for stochastic games, e.g., see [77, 34].*

almost surely, where $q^i_* : S \times A^i \to \mathbb{R}$ is the local Q-function associated with the best response to π^{-i} and is the unique fixed point of the Bellman operator variant

$$q^i_*(s, a^i) = r^i(s, a^i, \pi^{-i}(s)) + \gamma \sum_{s'} p(s'|s, a^i, \pi^{-i}(s)) \max_{\tilde{a}^i \in A^i} \{q^i_*(s', \tilde{a}^i)\} \quad \forall (s, a^i). \tag{4.13}$$

The error bound $e_Q \geq 0$ is as in Table 4.2 for $A = A^i$.

Proof. Pure Markov stationary strategy π^{-i} yields that the underlying stochastic game reduces to a *single-agent* stochastic team characterized by $\langle S, A^i, \bar{r}^i, \bar{p}, \gamma \rangle$ where $\bar{r}^i(s, a^i) = r^i(s, a^i, \pi^{-i}(s))$ and $\bar{p}(s'|s, a^i) = p(s'|s, a^i, \pi^{-i}(s))$. Then, the convergence of logit-Q dynamics in the single-agent stochastic team $\langle S, A^i, \bar{r}^i, \bar{p}, \gamma \rangle$ follows from Theorem 3. □

4.3.1 Beyond Stochastic Teams

Next, consider an N-agent game characterized by $\mathcal{G} = \langle A^i, u^i \rangle_{i \in [N]}$, where A^i is the finite action set of agent i, $A := \prod_{j \in [N]} A^j$ is the set of action profiles, and $u^i : A \to \mathbb{R}$ is the payoff function of agent i. We say that $\mathcal{G}$ is a potential game if there exists a potential function $\Phi : A \to \mathbb{R}$ satisfying

$$\Phi(a^i, a^{-i}) - \Phi(\tilde{a}^i, a^{-i}) = u^i(a^i, a^{-i}) - u^i(\tilde{a}^i, a^{-i}) \tag{4.14}$$

for all $(a^i, a^{-i}, \tilde{a}^i)$ and $i \in [N]$ [67]. Note that identical-interest games are potential games with the common payoff as the potential function.

For potential games, if agents follow (independent) log-linear learning dynamics, then by (4.6) and (4.14), the smoothed best response can be written as

$$\frac{\exp(u^i(a^i, a^{-i})/\tau)}{\sum_{\tilde{a}^i \in A^i} \exp(u^i(\tilde{a}^i, a^{-i})/\tau)} = \frac{\exp(\Phi(a^i, a^{-i})/\tau)}{\sum_{\tilde{a}^i \in A^i} \exp(\Phi(\tilde{a}^i, a^{-i})/\tau)}. \tag{4.15}$$

This implies that we can view the (independent) log-linear learning dynamics as if the agents have the potential function as their objective without loss of generality. Therefore, the convergence of the (independent) log-linear learning

dynamics to near efficient equilibrium with respect to the potential function follows from the near efficient equilibrium convergence of the dynamics in identical-interest games.

However, the convergence of Algorithm 4 to near efficient stationary equilibrium in stochastic teams, as shown in Theorem 3, does not imply the convergence in cases where the stage-payoffs induce a potential game in general. We can list the challenges as follows:

Challenge (i). The stage games may not be potential games even though the stage payoffs induce potential games. By (4.14), for each state s, stage payoffs $r^i(s,\cdot)$'s induce a potential game if there exists a potential function $\Phi(s,\cdot)$ such that

$$r^i(s,a^i,a^{-i}) - r^i(s,\tilde{a}^i,a^{-i}) = \Phi(s,a^i,a^{-i}) - \Phi(s,\tilde{a}^i,a^{-i}), \tag{4.16}$$

for all $(\tilde{a}^i,a^i,a^{-i})$ in joint action spaces, and for all $i \in [N]$. However, this does not imply that the auxiliary stage game with the payoffs $(Q^i_\pi(s,\cdot))_{i\in[N]}$ is a potential game since there may not exist a potential function $\Psi_\pi(s,a)$ such that for all $(\tilde{a}^i,a^i,a^{-i})$ and $i\in[N]$, we have

$$Q^i_\pi(s,a^i,a^{-i}) - Q^i_\pi(s,\tilde{a}^i,a^{-i}) = \Psi_\pi(s,a^i,a^{-i}) - \Psi_\pi(s,\tilde{a}^i,a^{-i}). \tag{4.17}$$

More explicitly, by (4.4a) and (4.16), we have

$$\begin{aligned}
Q^i_\pi(s,a^i,a^{-i}) - Q^i_\pi(s,\tilde{a}^i,a^{-i}) &= r^i(s,a^i,a^{-i}) - r^i(s,\tilde{a}^i,a^{-i}) \\
&\quad + \gamma \sum_{s_+} p(s_+ \mid s,a^i,a^{-i}) v^i_\pi(s_+) \\
&\quad - \gamma \sum_{s_+} p(s_+ \mid s,\tilde{a}^i,a^{-i}) v^i_\pi(s_+) \\
&= \Phi(s,a^i,a^{-i}) - \Phi(s,\tilde{a}^i,a^{-i}) \\
&\quad + \gamma \sum_{s_+} p(s_+ \mid s,a^i,a^{-i}) v^i_\pi(s_+) \\
&\quad - \gamma \sum_{s_+} p(s_+ \mid s,\tilde{a}^i,a^{-i}) v^i_\pi(s_+) \\
&= \Psi_\pi(s,a^i,a^{-i}) - \Psi_\pi(s,\tilde{a}^i,a^{-i}),
\end{aligned} \tag{4.18}$$

for all $(\tilde{a}^i,a^i,a^{-i})$ and $i\in[N]$. Due to the continuation payoffs v^i_π's, the existence of a potential function Φ in the game induced by stage payoffs r^i's is not a

sufficient condition for the existence of a potential function Ψ_π for the auxiliary stage game with payoffs Q^i_π's.

Challenge (ii). Even when stage games are potential games, reaching the efficient equilibrium in the stage games with respect to the potential function does not imply the maximization of the local stage-game payoffs. Observe that maximization of the potential function $\Psi_\pi(s,\cdot)$ does not necessarily imply the maximization of the stage game payoffs $Q^j_\pi(s,\cdot)$'s since there is no general guarantee that

$$\arg\max_{a^j} \Psi_\pi(s, a^j, a^{-j}) = \arg\max_{a^j} Q^j_\pi(s, a^j, a^{-j}). \tag{4.19}$$

The following corollary to Theorem 3 (similar to [35, Corollary 4]) shows that we can address these challenges for a certain class of stochastic games.

Corollary 3. *Consider a general-sum stochastic game $\langle S, (A^j, r^j)_{j\in[N]}, p, \gamma\rangle$ where for some specific agent i, the stage-payoffs of agents $j \neq i$ satisfy*

$$r^i(s, a^j, a^{-j}) - r^i(s, \tilde{a}^j, a^{-j}) = r^j(s, a^j, a^{-j}) - r^j(s, \tilde{a}^j, a^{-j}) \quad \forall (s, a^j, a^{-j}, \tilde{a}^j) \tag{4.20}$$

similar to (4.14)*, and the underlying transition kernel satisfies*

$$p(s' \mid s, a^i, a^{-i}) = p(s' \mid s, a^i, \tilde{a}^{-i}) \quad \forall (s, a^i, a^{-i}, \tilde{a}^{-i}, s') \tag{4.21}$$

such that only agent i controls the state transitions.

Suppose that every agent follow Algorithm 4, Assumption 2 holds and the underlying stochastic game is irreducible, as described in Theorem 3. Then, for the single-controller agent i, we have

$$\limsup_{t\to\infty} \left| \, Q^i_t(s,a) - Q^i_*(s,a) \, \right| \leq e_Q \quad \forall (s,a) \tag{4.22a}$$

$$\limsup_{t\to\infty} \left| \, \max_\pi \{U^i(\pi; s)\} - U^i(\pi_t; s) \, \right| \leq \frac{1+\gamma}{\gamma(1-\gamma)} \cdot e_Q \quad \forall s \tag{4.22b}$$

almost surely, where the Q-function $Q^i_* : S \times A \to \mathbb{R}$ is the unique fixed point of the following Bellman operator variant:

$$Q^i_*(s,a) = r^i(s,a) + \gamma \sum_{s'} p(s' \mid s, a) \max_{a' \in A} \{Q^i_*(s', a')\} \quad \forall (s,a). \tag{4.23}$$

The error bounds are as described in Table 4.2.

Proof. By (4.21), we define $\bar{p}(s_+ \mid s, a^i) := p(s_+ \mid s, a^i, a^{-i})$ and $\Psi_\pi(s, a^i, a^{-i}) := Q^i_\pi(s, a^i, a^{-i})$ for all (s, a^i, a^{-i}, s_+). Note that i refers to the single controller rather than any agent. Then, (4.4a) yields that

$$\Psi_\pi(s, a^i, a^{-i}) := r^i(s, a^i, a^{-i}) + \gamma \sum_{s' \in S} \bar{p}(s' \mid s, a^i) \cdot v^i_\pi(s'). \tag{4.24}$$

By the definition $\Psi_\pi \equiv Q^i_\pi$, Ψ_π is aligned with Q^i_π and (4.18) holds for the single controller. For agents $j \neq i$, we have

$$\begin{aligned} Q^j_\pi(s, a^j, a^{-j}) - Q^j_\pi(s, \tilde{a}^j, a^{-j}) &= r^j(s, a^j, a^{-j}) - r^j(s, \tilde{a}^j, a^{-j}) \\ &\quad + \gamma \sum_{s'} \bar{p}(s' \mid s, a^i) v^j_\pi(s') \\ &\quad - \gamma \sum_{s'} \bar{p}(s' \mid s, a^i) v^j_\pi(s') && (4.25) \\ &\overset{(a)}{=} r^i(s, a^j, a^{-j}) - r^i(s, \tilde{a}^j, a^{-j}) && (4.26) \\ &\overset{(b)}{=} r^i(s, a^j, a^{-j}) - r^i(s, \tilde{a}^j, a^{-j}) \\ &\quad + \gamma \sum_{s'} \bar{p}(s' \mid s, a^i) v^i_\pi(s') \\ &\quad - \gamma \sum_{s'} \bar{p}(s' \mid s, a^i) v^i_\pi(s') && (4.27) \\ &\overset{(c)}{=} \Psi_\pi(s, a^i, a^{-i}) - \Psi_\pi(s, \tilde{a}^i, a^{-i}), && (4.28) \end{aligned}$$

for all $(\tilde{a}^i, a^i, a^{-i})$ and $i \in [N]$, where (a) follows from (4.20). Since the last two terms on the right-hand side of (4.25) cancel each other, (b) follows from adding and subtracting $\gamma \sum_{s'} \bar{p}(s' \mid s, a^i) v^i_\pi(s')$, and (c) is due to (4.24). Therefore, (4.18) holds also for agents $j \neq i$. This implies that the auxiliary stage games with the payoffs $(Q^j_\pi)_{j \in [N]}$ are potential games with the potential functions Ψ_π. Furthermore, the maximization of the potential function Ψ_π in the stage games would imply the maximization of the stage game payoffs Q^i_π of the single controller i since $\Psi_\pi \equiv Q^i_\pi$. Due to (4.15), the logit-Q dynamics for the stochastic game $\langle S, (A^j, r^j)_{j \in [N]}, p, \gamma \rangle$ would lead to the same logit-Q dynamics in the stochastic team $\langle S, (A^j)_{j \in [N]}, r^i, p, \gamma \rangle$. Then, the proof follows from Theorem 3. □

4.4 Proof of Theorem 3

The proof follows by showing that the estimated plays π_t track the (soft) maximizers of the Q-function estimates Q_t^i's. Then, the Q-update (4.8) approximates the (soft) Q-learning algorithm in which the entire team coordinates on the (soft) maximizer, and under standard conditions on step sizes and transition probabilities, Q-learning converges to the Q-function associated with the optimal team strategy.

In the following, we show this *tracking result* by focusing on the evolution of the estimated plays across the repeated play of stage games. Recall that in the logit-Q dynamics, agents follow the *logit dynamics* in the stage games specific to each state whenever the associated state gets visited as if the stage-game payoffs are their Q-function estimates. Hence, we first characterize the convergence properties of the logit dynamics for non-stationary environments, which might be of independent interest.

4.4.1 Convergence Properties of the Logit Dynamics

In the logit dynamics, agents revise their actions according to a revision process, e.g., either in a coordinated or independent way. The agents revising their actions play according to the smoothed best response to the others' previous actions.

Notably, the action profiles played in the logit dynamics form an irreducible and aperiodic Markov chain [14]. The following proposition characterizes the dependence of the (long-run) behavior of this Markov chain on the game payoffs.

Proposition 3. *Consider that agents follow the logit dynamics in the repeated play of a potential game $\mathcal{G} = \langle (A^i, u^i)_{i\in[N]} \rangle$ with the potential function Φ. Then, the transition matrix and the unique stationary distribution for the Markov chain formed by the action profiles are locally Lipschitz continuous functions of the potential function.*

Proof. In the logit dynamics, the transition probabilities of the Markov chain formed by the action profiles can be written as

$$\Pr(a_+ \mid a) = \sum_{I \subset [N]} \mathcal{P}(I) \cdot \prod_{i \in I} \underline{\mathrm{br}}(\Phi(\cdot, a^{-i}))(a^i_+) \cdot \prod_{j \notin I} \mathbb{1}_{\{a^j_+ = a^j\}} \quad \forall (a, a_+) \in A \times A, \tag{4.29}$$

by noting that $\underline{\mathrm{br}}(\Phi(\cdot, a^{-i})) = \underline{\mathrm{br}}(u^i(\cdot, a^{-i}))$ for all $i \in [N]$, where $\mathcal{P}$ is the distribution of the revision process, e.g., as described in (4.5). Since $\underline{\mathrm{br}}(\cdot)$ is a continuously differentiable function, (4.29) yields that the transition matrix is a locally Lipschitz continuous function of the potential function. Furthermore, in irreducible and aperiodic Markov chains, the *unique* stationary distributions are Lipschitz continuous functions of their transition matrices [56, Section 3]. □

This proposition plays an important role in the convergence analysis since the stage-game payoffs, i.e., Q-function estimates, can evolve in time due to (4.8). Small changes in the Q-function estimates can result in (relatively) small changes in the transition probabilities and the corresponding stationary distributions due to the Lipschitz continuity.

In the following, we quantify the impact of these changes on the long-run behavior of the dynamics in comparison to the cases where the underlying stage game is stationary. However, the coupling of the dynamics across states poses a challenge. To address this challenge, we focus on the dynamics specific to a *fixed* state s and model the dynamics specific to other states as some external (possibly coupled) processes. Let w_k and z_k denote, resp., the action profile played at the kth visit to s and the trajectory of the state-action pairs realized in between the kth and $(k+1)$th visit to s, as illustrated in Figure 4.1. Then, we have $w_k \in A$ and $z_k \in Z$, where Z represents the *countable* set of all possible trajectories in between consecutive visits to s.

Let t_k be the stage for the kth visit to s. Then, we have $\sigma\left((s_\tau, a_\tau)_{\tau < t_k}, s_{t_k}\right) = \sigma\left((w_l, z_l)_{l<k}\right)$. The Q-function estimate Q_{t_k} is $\sigma\left((w_l, z_l)_{l<k}\right)$-measurable since Q_{t_k} is $\sigma\left((s_\tau, a_\tau)_{\tau < t_k}, s_{t_k}\right)$-measurable by (4.8). Therefore, the randomness on (w_k, z_k) is contingent only on $(w_l, z_l)_{l<k}$ and there exists a transition kernel $\mathbb{P}$ such that $(w_k, z_k) \sim \mathbb{P}\left(\cdot \mid (w_l, z_l)_{l<k}\right)$.

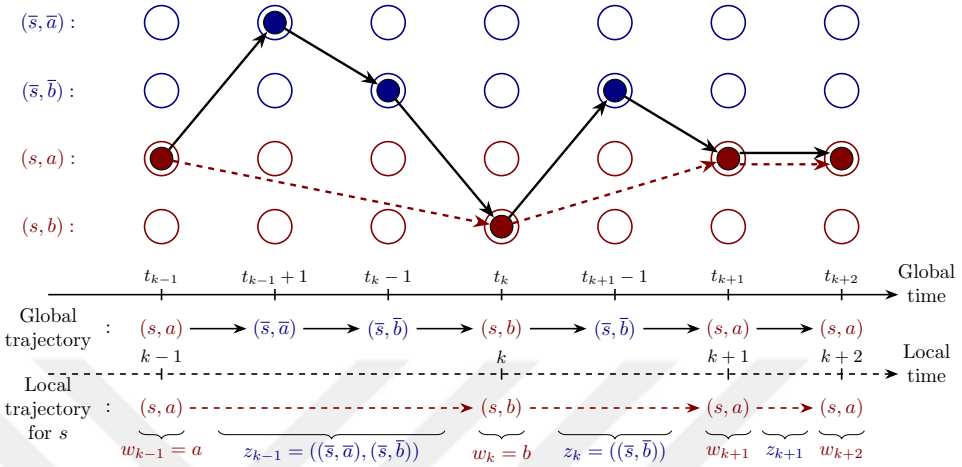


Figure 4.1: Consider two states s and $\overline{s}$ with action profiles $\{a,b\}$ and $\{\overline{a},\overline{b}\}$, respectively. This is a figurative illustration for the *global* trajectory of the pairs $(s_t,a_t)_{t\geq 0}$ evolving over the global timescale $t=0,1,\ldots$, and the *local* trajectory of the pairs $(w_k,z_k)_{k\geq 0}$ specific to s and evolving over the local timescale $k=1,2,\ldots$ whenever s gets visited. Let t_k be the time of the kth visit to s. In the local trajectory, w_k and z_k denote, resp., the action profile played at the kth visit to s and the trajectory of the state-action pairs realized in between the kth and $(k+1)$th visit to s.

In the following, we leverage Proposition 3, Lemma 2 and Lemma 1 to prove that the estimated plays can track the (soft) maximizers of the Q-function estimates.

4.4.2 Tracking Result for the Estimated Plays

For the estimated plays, we can focus on π_t^{ave} since we can compute π_t^{free} based on π_t^{ave} according to

$$\pi_t^{\text{free}}(s)(a)>0 \quad \Leftrightarrow \quad a\in A_t(s)=\arg\max_{\tilde{a}\in A}\{\pi_t^{\text{ave}}(s)(a)\} \quad \forall(s,a). \tag{4.30}$$

Let $\mu(u)\in\Delta(A)$ denote the unique stationary distribution of the action profiles for the logit dynamics followed in the identical-interest game $\mathcal{G}=\langle(A^i)_{i\in[N]},u\rangle$. Then, the following proposition shows that the empirical averages π_t^{ave} can track the stationary distributions associated with the stage-game payoffs Q_t's.

Proposition 4 (Tracking Result). *We have* $\|\pi_t^{\text{ave}}(s) - \mu(Q_t(s,\cdot))\|_1 \to 0$ *as* $t \to \infty$ *almost surely for each* s.

Proof. If the stage-game payoffs, i.e., the Q-function estimates, remained fixed over time, the empirical distribution of action profiles, π_t^{ave}, would converge to the stationary distribution of the homogeneous Markov chain governing these profiles, thanks to the ergodicity of the chain.

In our setting, however, the Q-function estimates evolve as the game is played. On the other hand, the use of a vanishing step size (as specified in Assumption 2-(i)) and the boundedness of the iterates by the nature of the update (4.8) yield that *one-step* update to the Q-function becomes progressively smaller over time, decaying to zero.

Because the logit dynamics we employ are designed such that small changes in the payoffs (Q-functions) induce also (relatively) small changes in the transition probabilities (as established in Proposition 3), the stage games behave almost as if they are stationary over short time intervals. In these "quasi-stationary" periods, the changes accumulated in the transition probabilities are negligible enough that the system's short-run behavior mimics that of a stationary process.

Over longer time periods, even these small changes can add up, leading to significant shifts in the dynamics. To address this, we partition the overall time horizon into epochs whose lengths grow at a sufficiently slow rate. More explicitly, fix some state s and divide the horizon into epochs such that the state s gets visited only for a certain number of times specific to the epoch and the first state of every epoch is s. For clear exposition, we use k and (n), resp., as subscripts for the parameters associated with the kth visit to s and the nth epoch for s. For example, t_k, $t_{(n)}$, and $k_{(n)}$ denote the stage that s gets visited for the kth time, the stage that the nth epoch starts, and the number of times s gets visited until and including stage $t_{(n)}$, respectively.

For each epoch n, we define the filtration $\mathcal{F}_{(n)} = \sigma(\{s_t, a_t\}_{t<t_{(n)}}, s_{t_{(n)}})$ and

$$\mu_k := \mathbb{E}[a_{t_k} \mid \mathcal{F}_{(n)}] \in \Delta(A) \tag{4.31}$$

denotes the distribution over the action profiles played at the kth visit to s conditioned on $\mathcal{F}_{(n)}$. We can invoke Lemma 1 to characterize the distribution μ_k for $k = k_{(n)}, \dots, k_{(n+1)} - 1$ within epoch n in comparison to the stationary distribution of the logit dynamics in a *fictional* (stationary) scenario. To this end, we focus on the evolution of the action profiles across consecutive visits to state s while modeling the state-action trajectories in between the consecutive visits to s as some external (possibly coupled) processes, as illustrated in Figure 4.1.

Let $Q_{(n)}(\cdot) := Q_{t_{(n)}}(\cdot)$ and $c_{(n)}(\cdot) := c_{t_{(n)}}(\cdot)$.[6] For the fictional scenario, we consider the logit dynamics followed in the repeated play of $\mathcal{G}_{(n)} = \langle (A^i)_{i\in[N]}, Q_{(n)}(s,\cdot)\rangle$ at every visit to state s in the main scenario. Correspondingly, $\mu(Q_{(n)}(s,\cdot)) \in \Delta(A)$ is the unique stationary distribution of the action profiles for the logit dynamics followed in $\mathcal{G}_{(n)}$. On the other hand, by (4.8) and Assumption 2-(i), we can bound the changes on the Q-function estimates within epoch n in the main scenario by

$$|Q_t(s,a) - Q_{(n)}(s,a)| \le HK_{(n)}\beta_{(n)} \quad \forall a \in A,\ t \in [t_{(n)}, t_{(n+1)}), \tag{4.32}$$

where we define $\beta_{(n)} := \max_a\{\beta_{c_{(n)}(s,a)}\}$, $K_{(n)} := k_{(n+1)} - k_{(n)}$, and $H := 2/(1-\gamma)$ is a uniform upper bound on $|r_t - \gamma v_t(s_{t+1}) - Q_t(s,a)|$ for all (s,a) and t.

The following lemma characterizes μ_k in terms of $\mu(Q_{(n)}(s,\cdot))$ by coupling the evolution of the dynamics in the main and fictional scenarios based on (4.32).

Lemma 5. *For sufficiently large n, we have*

$$\|\mu_k - \mu(Q_{(n)}(s,\cdot))\|_1 \le C\rho^{k-k_{(n)}} + DK_{(n)}\beta_{(n)} \quad \forall k \in [k_{(n)}, k_{(n+1)}) \tag{4.33}$$

for some constants $C, D > 0$ and $\rho \in (0,1)$.

[6] We drop the agent index i in Q_t^i since the agents have common estimates.

The proof (Section C.1) follows from Proposition 3 and Lemma 1. We highlight the resemblance of (4.33) to (2.5).

By (4.7), the empirical average $\pi_t^{\text{ave}}(s)$ evolves according to

$$\pi_{t+1}^{\text{ave}}(s) = \pi_t^{\text{ave}}(s) + \mathbb{1}_{\{s_t=s\}} \cdot \frac{1}{c_t(s)} \cdot (a_t - \pi_t^{\text{ave}}(s)). \tag{4.34}$$

Henceforth, we view actions $a \subset \Delta(A)$ as pure strategies, or probability vectors, in which the associated action gets played with probability 1. By (4.34), $\pi_t^{\text{ave}}(s)$ gets updated only when s gets visited, i.e., $\pi_t^{\text{ave}}(s) = \pi_{t_k}^{\text{ave}}(s)$ for all $t \in [t_k, t_{k+1})$. Then, we can write (4.34) as

$$\pi_{t_{k+1}}^{\text{ave}}(s) = \pi_{t_k}^{\text{ave}}(s) + \alpha_k \cdot (a_{t_k} - \pi_{t_k}^{\text{ave}}(s)), \tag{4.35}$$

where $\alpha_k := 1/k$.

The following lemma shows that accumulated changes can still form a stochastic approximation recursion for certain (growing) epoch lengths.

Lemma 6. *Consider the iterative update rule* (4.35) *with some step sizes* $\{\alpha_k \in [0,1]\}$ *satisfying the standard conditions that* $\alpha_k \to 0$ *monotonically,* $\sum_k \alpha_k = \infty$ *and* $\sum_k \alpha_k^2 < \infty$*. Then,* (4.35) *yields that we can write the accumulated changes in* $\pi_{(n)}^{\text{ave}}(s) := \pi_{t_{(n)}}^{\text{ave}}(s)$ *across epochs* $n = 1, 2, \ldots$ *as*

$$\pi_{(n+1)}^{\text{ave}}(s) = \pi_{(n)}^{\text{ave}}(s) + \overline{\alpha}_{(n)} \cdot \left(\sum_{k=k_{(n)}}^{k_{(n+1)}-1} \frac{\overline{\alpha}_k}{\overline{\alpha}_{(n)}} \cdot a_{t_k} - \pi_{(n)}^{\text{ave}}(s) \right), \tag{4.36}$$

where the auxiliary step sizes are defined by

$$\overline{\alpha}_{(n)} := \sum_{k=k_{(n)}}^{k_{(n+1)}-1} \overline{\alpha}_k \quad \textit{and} \quad \overline{\alpha}_k := \alpha_k \prod_{l=k+1}^{k_{(n+1)}-1} (1-\alpha_l) \quad \forall k \in [k_{(n)}, k_{(n+1)}). \tag{4.37}$$

Furthermore, there exist epoch lengths $K_{(n)} := k_{(n+1)} - k_{(n)}$ *growing at a sufficiently slow rate that* $K_{(n)} \to \infty$ *and* $K_{(n)}/k_{(n)} \to 0$ *as* $n \to \infty$ *while the step sizes* $\{\overline{\alpha}_{(n)} \in [0,1]\}$ *satisfy the standard conditions that* $\overline{\alpha}_{(n)} \to 0$, $\sum_n \overline{\alpha}_{(n)} = \infty$, *and* $\sum_n \overline{\alpha}_{(n)}^2 < \infty$.

As an illustrative example for Lemma 6, consider $L_n \coloneqq \lfloor \log_2(n) \rfloor$ and a step size sequence given by $\alpha_k = 1/k$. Define the epoch lengths as $K_{(n)} = 1 + L_n$. Then, we have $k_{(n+1)} = nL_n + L_n + n - 2^{L_n+1} + 2 \geq n(L_n - 1)$, which leads to the accumulated step size estimate $\overline{\alpha}_{(n)} = K_{(n)}/(k_{(n+1)} - 1)$. From this, we derive the bounds $L_n/(n(L_n + 1)) \leq \overline{\alpha}_{(n)} \leq L_n/(n(L_n - 1))$. Taking the limit as $n \to \infty$, we obtain $\lim_{n\to\infty} K_{(n)}/k_{(n)} = 0$, and $\lim_{n\to\infty} \overline{\alpha}_{(n)} = 0$. Moreover, the accumulated step size $\overline{\alpha}_{(n)}$ behaves asymptotically like $1/n$, ensuring that $\sum_n \overline{\alpha}_{(n)} = \infty$, and $\sum_n \overline{\alpha}^2_{(n)} < \infty$. Lemma 6 is a general result asserting that for any step size sequence α_k satisfying the lemma's conditions, we can always determine an appropriate epoch length $K_{(n)}$. The proof of Lemma 6 is provided in Section C.2.

For $\alpha_k = 1/k$, the accumulated update (4.36) can be written as

$$\pi^{\text{ave}}_{(n+1)}(s) = \pi^{\text{ave}}_{(n)}(s) + \overline{\alpha}_{(n)} \cdot \left(\mu(Q_{(n)}(s,\cdot)) + e_{(n+1)} + \omega_{(n+1)} - \pi^{\text{ave}}_{(n)}(s)\right), \quad (4.38)$$

where $\overline{\alpha}_{(n)} = K_{(n)}/(k_{(n+1)} - 1)$ by (4.37) and the auxiliary error terms are defined by

$$e_{(n+1)} \coloneqq \frac{1}{K_{(n)}} \sum_{k=k_{(n)}}^{k_{(n+1)}-1} (\mu_k - \mu(Q_{(n)}(s,\cdot))), \quad \omega_{(n+1)} \coloneqq \frac{1}{K_{(n)}} \sum_{k=k_{(n)}}^{k_{(n+1)}-1} (a_{t_k} - \mu_k). \quad (4.39)$$

Both $e_{(n+1)}$ and $\omega_{(n+1)}$ depend on s implicitly for notational simplicity. The bound on $\|\mu_k - \mu(Q_{(n)}(s,\cdot))\|$ in Lemma 5 yields that the error $e_{(n+1)}$ is bounded by

$$\left\|e_{(n+1)}\right\|_1 \leq \frac{C}{1-\rho} \cdot \frac{1}{K_{(n)}} + D \cdot K_{(n)}\beta_{(n)} \quad (4.40)$$

for sufficiently large n, due to the triangle inequality. On the other hand, $\omega_{(n+1)}$ forms a square-integrable Martingale difference sequence by the definition (4.31).

At the same timescale with (4.38), we can write the accumulated changes for the Q-update (4.8) as

$$Q_{(n+1)}(s,\cdot) = Q_{(n)}(s,\cdot) + \overline{\alpha}_{(n)} \cdot \epsilon_{(n+1)} \quad (4.41)$$

for some error term $\epsilon_{(n+1)} \in \mathbb{R}^{|A|}$, bounded by

$$\|\epsilon_{(n+1)}\|_\infty \leq H\beta_{(n)}k_{(n+1)} \quad (4.42)$$

due to (4.32).

The following lemma shows that the error terms are asymptotically negligible.

Lemma 7. *We have* $\beta_{(n)}k_{(n+1)} \to 0$*, and therefore,* $\|e_{(n)}\| \to 0$ *and* $\|\epsilon_{(n)}\| \to 0$ *as* $n \to \infty$ *almost surely.*

Proof. Recall that each agent revising their action chooses any available action with a positive probability that is uniformly bounded away from zero, and every agent revises their action with some fixed positive probability. Since there are finitely many states, at least one state is visited infinitely often; call this state s. Because logit dynamics allow one action change per agent at a time (as is also possible in the independent case), any action profile can be realized with positive probability after s is visited N times. Therefore, for any $a \in A$, the state-action pair (s, a) is realized infinitely often, and the frequencies of played actions are comparable. Moreover, by the irreducibility assumption stated in Theorem 3, there is at least one action profile a that leads from s to some reachable state s'. Since (s, a) is visited infinitely often, it follows that s' is also visited infinitely often. By induction, and given that agents can play any action profile at comparable frequencies, the irreducibility assumption in the statement of Theorem 3 implies that every state s is visited infinitely often i.e., $c_t(s) \to \infty$ as $t \to \infty$ almost surely. Correspondingly, any pair (s, a) gets realized infinitely often, i.e., $c_t(s, a) \to \infty$ as $t \to \infty$ almost surely. Therefore, if we follow similar lines with the proof of [35, Lemma 6.1], Assumption 2-(iii) yields that

$$\lim_{t\to\infty} c_t(s)\beta_{c_t(s,a)} = 0 \quad \forall(s, a), \tag{4.43}$$

with probability 1.

Recall also that $k_{(n)} = c_{(n)}(s)$ and $c_{(n)}(s, a)$ denote, resp., the number of times s and (s, a) get realized in the first stage of epoch n. By (4.43), $k_{(n)}\beta_{c_{(n)}(s,a)}$ decays to zero for all (s, a), and therefore, $k_{(n)}\beta_{(n)}$ decays to zero since there are finitely many pairs and $\beta_{(n)} = \max_a\{\beta_{c_{(n)}(s,a)}\}$.

Let $K_{(n)}$ be chosen according to Lemma 6. Then, such a choice of $K_{(n)}$ satisfies $k_{(n)} \geq n \geq K_{(n)}$ as $\{K_{(n)}\}$ grows at a sub-linear rate and $K_{(n)} \geq 1$ for all n. Therefore, $K_{(n)}\beta_{(n)}$ and $k_{(n+1)}\beta_{(n)} = K_{(n)}\beta_{(n)} + k_{(n)}\beta_{(n)}$ decay to zero almost surely. By (4.40) and (4.42), the error terms also decay to zero almost surely. $\square$

Lemma 7 yields that we can rewrite the accumulated changes (4.38) and (4.41), resp., on $\pi^{\text{ave}}_{(n)}(s)$ and $Q_{(n)}(s,\cdot)$ across epochs as

$$\pi^{\text{ave}}_{(n+1)}(s) = \pi^{\text{ave}}_{(n)}(s) + \overline{\alpha}_{(n)} \cdot (\mu(Q_{(n)}(s,\cdot)) - \pi^{\text{ave}}_{(n)}(s) + \omega_{(n+1)} + \text{a.n.e.}) \tag{4.44a}$$

$$Q_{(n+1)}(s,\cdot) = Q_{(n)}(s,\cdot) + \overline{\alpha}_{(n)} \cdot \text{a.n.e.}, \tag{4.44b}$$

where a.n.e. refers to asymptotically negligible error terms. Since step sizes $\overline{\alpha}_{(n)}$ satisfies the stochastic approximation conditions by Lemma 6, the function $\mu(\cdot)$ is locally Lipschitz continuous by Proposition 3, $\omega_{(n+1)}$ forms a square-integrable Martingale difference sequence, and the iterates $(\pi^{\text{ave}}_{(n)}(s), Q_{(n)}(s,\cdot))$ are bounded, we can approximate the discrete-time accumulated updates (4.44) with the following continuous-time flow

$$\dot{\pi} = \mu(Q) - \pi \quad \text{and} \quad \dot{Q} = 0 \tag{4.45}$$

such that (4.44) converges to an internally chain-recurrent set of (4.45) [78]. Furthermore, we can characterize its internally chain-recurrent set by formulating a Lyapunov function [78]. For example, the Lyapunov function $V(\pi, Q) = \|\mu(Q) - \pi\|_2^2$ yields that

$$\lim_{n \to 0} \|\pi^{\text{ave}}_{(n)}(s) - \mu(Q_{(n)}(s,\cdot))\|_1 = 0 \tag{4.46}$$

almost surely as norms are comparable in finite-dimensional spaces.

We have the tracking result (4.46) for the first stages $t_{(n)}$ of each epoch n. We can also show the tracking result for each t since

$$\begin{aligned}\|\pi^{\text{ave}}_t(s) - \mu(Q_t(s,\cdot))\|_1 \leq \|\pi^{\text{ave}}_t(s) - \pi^{\text{ave}}_{(n)}(s)\|_1 + \|\mu(Q_{(n)}(s,\cdot)) - \mu(Q_t(s,\cdot))\|_1 \\ + \|\pi^{\text{ave}}_{(n)}(s) - \mu(Q_{(n)}(s,\cdot))\|_1\end{aligned}$$

for all $t \in [t_{(n)}, t_{(n+1)})$, by the triangle inequality. The first term on the right-hand side is bounded by

$$\|\pi_t(s) - \pi_{(n)}(s)\|_1 \leq 2K_{(n)}/k_{(n)} \quad \forall t \in [t_{(n)}, t_{(n+1)}) \tag{4.47}$$

by (4.34) and decays to zero by Lemma 6. The second one is bounded by

$$\|\mu(Q_t(s,\cdot)) - \mu(Q_{(n)}(s,\cdot))\|_1 \leq LHK_n\beta_{(n)} \quad \forall t \in [t_{(n)}, t_{(n+1)}) \tag{4.48}$$

for some constant L, by Proposition 3 and (4.32), and decays to zero by Lemma 7. The last one decays to zero by (4.46). This completes the proof of Proposition 4 since the fixed state s is arbitrary. □

4.4.3 Convergence of the Q-function Estimates

The convergence of the Q-function estimates to Q_*, as described in (4.11), follows from showing that the value function estimates track the joint maximizer of the Q-function estimates approximately, as shown in the following corollary to Proposition 4.

Corollary 4. *We have* $\limsup_{t\to\infty} |v_t(s) - \max_a\{Q_t(s,a)\}| \leq e_Q \cdot (1-\gamma)/\gamma$ *almost surely, where* $e_Q \geq 0$ *is as described in Table 4.2.*

Proof. For AL-Q, we have $\pi_t \equiv \pi_t^{\text{ave}}$ and $\|\pi_t(s) - \mu^{\text{coor}}(Q_t(s,\cdot))\|_1 = \|\pi_t(s) - \underline{\text{br}}(Q_t(s,\cdot))\|_1 \to 0$ as $t \to \infty$ by Lemma 2 and Proposition 4. This implies that

$$\limsup_{t\to\infty} |v_t(s) - \max_a\{Q_t(s,a)\}| \leq \tau \log|A| \tag{4.49}$$

since $|\max_a\{Q_t(s,a)\} - \mathbb{E}_{a\sim\underline{\text{br}}(Q_t(s,\cdot))}[Q_t(s,a)]| \leq \tau\log|A|$.

For AIL-Q, we have $\pi_t \equiv \pi_t^{\text{ave}}$ and $\|\pi_t(s) - \mu^{\text{ind}}(Q_t(s,\cdot))\|_1 \to 0$ as $t \to \infty$ by Proposition 4. By the triangle inequality, we have

$$|v_t(s) - \max_a\{Q_t(s,a)\}| \leq |\mathbb{E}_{a\sim\pi_t(s)}[Q_t(s,a)] - \mathbb{E}_{\mu^{\text{ind}}(s)}[Q_t(s,a)]| \tag{4.50}$$

$$|\mathbb{E}_{\mu^{\text{ind}}(s)}[Q_t(s,a)] - \mathbb{E}_{\mu^{\text{coor}}(s)}[Q_t(s,a)]|$$

$$|\mathbb{E}_{\mu^{\text{coor}}(s)}[Q_t(s,a)] - \max_a\{Q_t(s,a)\}|. \tag{4.51}$$

The first term on the right-hand side decays to zero while the last is bounded by $\tau\log|A|$. Since the Q-function estimates are bounded by $1/(1-\gamma)$, if we invoke

Lemma 2 for the second term, we obtain

$$\limsup_{t\to\infty} |v_t(s) - \max_a\{Q_t(s,a)\}| \le \tau \log|A| + \frac{\Lambda(\delta)}{1-\gamma}, \tag{4.52}$$

where $\Lambda(\delta)$ is defined as follows:

$$\begin{aligned}\Lambda(\delta) &:= 2(1-\lambda(\delta)) \cdot \frac{1+\lambda(\delta)\cdot\bar{\epsilon}^2}{\lambda(\delta)\cdot\bar{\epsilon}^2}, \\ \lambda(\delta) &:= \frac{N\delta(1-\delta)^{N-1}}{1-(1-\delta)^N}, \\ \bar{\epsilon} &:= \left(\frac{\epsilon}{N}\right)^N.\end{aligned} \tag{4.53}$$

For EIL-Q, we have $\pi_t \equiv \pi_t^{\text{free}}$ and

$$\begin{aligned}\|\pi_t^{\text{ave}}(s) - \underline{\text{br}}(Q_t(s,\cdot))\| \le \|\pi_t^{\text{ave}}(s) - \mu^{\text{ind}}(Q_t(s,\cdot))\| \\ + \|\mu^{\text{ind}}(Q_t(s,\cdot)) - \mu^{\text{coor}}(Q_t(s,\cdot))\|\end{aligned} \tag{4.54}$$

for any norm by the triangle inequality. Then, Proposition 4 and Lemma 2 yield

$$\limsup_{t\to\infty} \|\pi_t^{\text{ave}}(s) - \underline{\text{br}}(Q_t(s,\cdot))\|_\infty \le \Lambda(\delta). \tag{4.55}$$

as $\|\cdot\|_\infty \le \|\cdot\|_1$. Let $a^1 \in \arg\max_a\{\underline{\text{br}}(Q_t(s,\cdot))(a)\}$ and $a^2 \in \arg\max_a\{\pi_t^{\text{ave}}(s)(a)\}$. Then, we can bound the difference of the probabilities of these action profiles in $\underline{\text{br}}(Q_t(s,\cdot))$ by

$$\begin{aligned}0 &\le \underline{\text{br}}(Q_t(s,\cdot))(a^1) - \underline{\text{br}}(Q_t(s,\cdot))(a^2) \\ &\le \underline{\text{br}}(Q_t(s,\cdot))(a^1) - \pi_t^{\text{ave}}(s)(a^1) + \pi_t^{\text{ave}}(s)(a^2) - \underline{\text{br}}(Q_t(s,\cdot))(a^2) \\ &\le 2\Lambda(\delta) + \text{a.n.e.},\end{aligned} \tag{4.56}$$

where the first and second inequalities follow since a^1 and a^2 are, resp., the maximizers of the distributions $\underline{\text{br}}(Q_t(s,\cdot))$ and $\pi_t^{\text{ave}}(s)$, and the last inequality follows from (4.55).

The boundedness of the Q-iterates by $1/(1-\gamma)$ and the smoothed best response (4.6) guarantee that

$$\underline{\text{br}}(Q_t(s,\cdot))(a) \ge \epsilon := \frac{1}{|A|}\exp\left(-\frac{1}{\tau(1-\gamma)}\right) \quad \forall(s,a). \tag{4.57}$$

Therefore, we can bound the difference of these action profiles in $Q_t(s,\cdot)$ by

$$Q_t(s,a^1) - Q_t(s,a^2) = \tau \cdot \log\left(\frac{\underline{\mathrm{br}}(Q_t(s,\cdot))(a^1)}{\underline{\mathrm{br}}(Q_t(s,\cdot))(a^2)}\right) \tag{4.58}$$

$$\leq \tau \cdot \log\left(1 + 2\Lambda(\delta)/\epsilon + \text{a.n.e.}\right), \tag{4.59}$$

where the first line follows from (4.6), and the second from (4.56) and (4.57). Since $\log(\cdot)$ is monotonically increasing and continuous, (4.59) yields that

$$\limsup_{t\to\infty}\{Q_t(s,a^1) - Q_t(s,a^2)\} \leq \Xi(\delta) \coloneqq \tau \cdot \log\left(1 + 2\Lambda(\delta)/\epsilon\right). \tag{4.60}$$

The inequality (4.60) holds for any $a^2 \in A_t(s)$ since $A_t(s)$, as described in (4.30), is the set of maximizers of $\pi_t^{\mathrm{ave}}(s)$. Moreover, by the definition (4.6), $Q_t(s,a^1) = \max_a\{Q_t(s,a)\}$. Since any convex combination of $Q_t(s,a^2)$ over $a^2 \in A_t$, including $\pi_t^{\mathrm{free}}(s) \in \Delta(A)$, also satisfies the inequality (4.60), we obtain

$$\limsup_{t\to\infty} |v_t(s) - \max_a\{Q_t(s,a)\}| \leq \Xi(\delta) \tag{4.61}$$

almost surely, as $v_t(s) = \mathbb{E}_{a\sim\pi_t^{\mathrm{free}}(s)}[Q_t(s,a)]$.

For EL-Q, the right-hand side of (4.55) is 0 by Proposition 4. Hence, (4.61) yields that $|v_t(s) - \max_a\{Q_t(s,a)\} - v_t(s)|$ decay to zero almost surely. □

We can write the Q-update (4.8) as

$$\begin{aligned} Q_{t+1}(s,a) = Q_t(s,a) &+ \mathbb{1}_{\{(s,a)=(s_t,a_t)\}}\beta_{c_t(s,a)} \\ &\times \Big(r(s,a) + \gamma\sum_{s'} p(s' \mid s,a)\max_{\tilde{a}}\{Q_t(s',\tilde{a})\} - Q_t(s,a) \\ &+ \bar{\epsilon}_t(s,a) + \bar{\omega}_{t+1}(s,a)\Big), \end{aligned} \tag{4.62}$$

where the error term and the stochastic approximation noise are, resp., defined by

$$\bar{\epsilon}_t(s,a) = \gamma\sum_{s'} p(s' \mid s,a)\big(v_t(s') - \max_{\tilde{a}}\{Q_t(s',\tilde{a})\}\big) \tag{4.63a}$$

$$\bar{\omega}_{t+1}(s,a) = \gamma\left(v_t(s_{t+1}) - \sum_{s'} p(s' \mid s,a)v_t(s')\right). \tag{4.63b}$$

Corollary 4 and (4.63a) yield that the error term $\bar{\epsilon}_t$ satisfies

$$\limsup_{t\to\infty} |\bar{\epsilon}_t(s,a)| \leq (1-\gamma)e_Q \quad \forall (s,a) \tag{4.64}$$

almost surely. On the other hand, since $s_{t+1} \sim p(\cdot \mid s_t, a_t)$, the stochastic approximation noise $\bar{\omega}_{t+1}$ forms a square-integrable Martingale difference sequence. Due to the exploration inherent to the logit-Q dynamics, the irreducibility assumption in the statement of Theorem 3 yields that each (s,a) pair gets realized infinitely often. Based on the contraction property of the Bellman optimality operator, Theorem 5.1 in the extended version of [35] yields (4.10).[7]

4.4.4 Convergence of the Estimated Plays to Equilibrium

Recall that $v_\pi : S \to \mathbb{R}$ and $Q_\pi : S \times A \to \mathbb{R}$, as described in (4.4), resp., denote the value and Q-function of the Markov stationary policy $\pi : S \to \Delta(A)$. For example, we have $v_\pi(s) = \mathrm{E}_{a\sim\pi(s)}[Q_\pi(s,a)]$. Based on Q_*, as described in (4.11), we define $v_*(s) = \max_a\{Q_*(s,a)\}$ for all s. Then, given the strategy π_t, we have

$$0 \leq v_*(s) - v_{\pi_t}(s) \tag{4.65}$$

$$= v_*(s) - \mathrm{E}_{a\sim\pi_t(s)}[Q_*(s,a)] + \mathrm{E}_{a\sim\pi_t(s)}[Q_*(s,a) - Q_{\pi_t}(s,a)]. \tag{4.66}$$

Since (v_*, Q_*) and (v_{π_t}, Q_{π_t}) pairs satisfy the one-step Bellman equation, we also have

$$\mathrm{E}_{a\sim\pi_t(s)}[Q_*(s,a) - Q_{\pi_t}(s,a)] = \gamma \mathrm{E}_{a\sim\pi_t(s)}\left[\sum_{s'\in S} p(s' \mid s,a)(v_*(s') - v_{\pi_t}(s'))\right] \tag{4.67}$$

$$\leq \gamma \|v_* - v_{\pi_t}\|_\infty \tag{4.68}$$

for both $\pi_t \equiv \pi_t^{\mathrm{ave}}$ and $\pi_t \equiv \pi_t^{\mathrm{free}}$. On the other hand, we can bound the first two terms $v_*(s) - \mathrm{E}_{a\sim\pi_t(s)}[Q_*(s,a)]$ on the right-hand side of (4.66) by

$$|v_*(s) - \mathrm{E}_{a\sim\pi_t(s)}[Q_*(s,a)]| \leq |\max_a\{Q_*(s,a)\} - \max_a\{Q_t(s,a)\}| \tag{4.69}$$

$$+ |\max_a\{Q_t(s,a)\} - v_t(s)| + |\mathrm{E}_{a\sim\pi_t(s)}[Q_*(s,a)] - \mathrm{E}_{a\sim\pi_t(s)}[Q_t(s,a)]|$$

[7] The extended version of [35] is available at `https://arxiv.org/abs/2010.04223`.

Then, we can bound the first and third terms on the right-hand side by (4.10) and bound the second term by Corollary 4, and obtain

$$\limsup_{t\to\infty} |v_*(s) - \mathrm{E}_{a\sim\pi_t(s)}[Q_*(s,a)]| \leq \frac{1+\gamma}{\gamma} \cdot e_Q \tag{4.70}$$

due to the sub-additivity of the limit superior. Since the upper bound in (4.70) does not depend on (s,a), (4.66), (4.68), and (4.70) yields that

$$\limsup_{t\to\infty} \|v_* - v_{\pi_t}\|_\infty \leq \frac{1+\gamma}{\gamma(1-\gamma)} e_Q. \tag{4.71}$$

This completes the proof of Theorem 3.

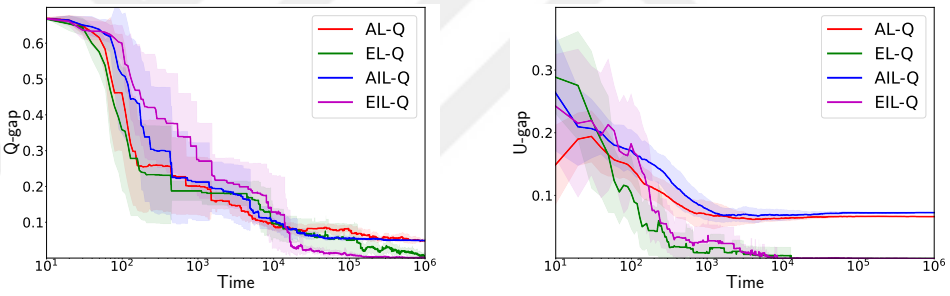
4.5 Illustrative Examples

In this section, we examine the long-run behavior of Algorithm 4 numerically. Consider a stochastic team with three states and three agents. Each agent can take two actions in each state (i.e., $|A^i| = 2$ for each i and $|A| = 8$). The stage-payoffs are randomly selected so that Q-function estimates remain bounded within $[0,1]$. We set $\gamma = 0.6$ and $\tau = 0.05$ (and $\delta = 0.2$ for the independent-revision schemes). For all agents, the initial Q-function values are assigned differently, e.g., $Q_0^i(s,a) = 1 + 0.1 \cdot i$ for all (s,a), to examine the impact of different initializations. We run the experiments for 10^6 time steps over 10 independent trials. Figure 4.2 depicts both the mean error and standard deviation for these trials for the AL-Q, EL-Q, AIL-Q, and EIL-Q algorithms. We use a logarithmic scale for time to illustrate both the transient and steady-state behavior of the dynamics more explicitly.

Figure 4.2a shows the decay of *Q-gap* (defined by $\|Q_t - Q_*\|_\infty$) to (near) zero, implying the convergence of the Q-function estimates Q_t to the (near) optimal ones Q_*. Furthermore, Figure 4.2b shows the decay of the *U-gap* (defined by $\max_\pi\{U(\pi)\} - U(\pi_t)$) to (near) zero, implying the convergence to the (near) efficient equilibrium. Recall the averaging and exploration-free approaches for the estimated play, as described in (4.7). In the AL-Q, and AIL-Q dynamics,

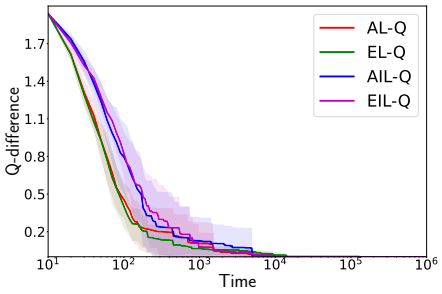
using the averaging scheme, the Q-function estimates Q_t and the estimated plays π_t^{ave} reach *near* efficient equilibrium with an offset error induced by the exploration inherent to the logit dynamics to avoid convergence to inefficient equilibrium. On the other hand, in the EL-Q and EIL-Q dynamics, using the exploration-free scheme, the Q-function estimates Q_t and the estimated plays π_t^{free} reach the *exact* efficient equilibrium, mitigating the offset error.

We further examine whether different initializations of Q-function estimates cause non-diminishing differences in their estimates in Figure 4.2c. We observe that the Q-difference (defined by $\sum_{i,j} \|Q_t^i - Q_t^j\|_2$) decays to zero for arbitrary initialization of $Q_0^i(s,a) = 1 + 0.1 \cdot i$ for all (s,a).



(a) Q-gap vs Time

(b) U-gap vs Time



(c) Q-difference vs Time

Figure 4.2: From top to bottom, the solid curves represent the AL-Q, EL-Q, AIL-Q, and EIL-Q dynamics, respectively. The shaded areas show the standard deviation across independent runs. In Figure 4.2a and 4.2b, the gap for Q-functions and game values converge to the (near) optimal ones for all dynamics, corroborating Theorem 3. In Figure 4.2c, the impact of different initializations on the dynamics' long-run behavior diminishes in time, showing robustness to arbitrary initializations.

4.6 Concluding Remarks

This chapter introduced a new family of logit-Q dynamics for efficient learning in *unknown* stochastic teams by combining log-linear learning with Q-learning within the stage-game framework. We proposed four new dynamics: Averaged Logit-Q (AL-Q), Exploration-free Logit-Q (EL-Q), Averaged Independent Logit-Q (AIL-Q), and Exploration-free Independent Logit-Q (EIL-Q). These dynamics differ in terms of how agents revise their actions and model their joint play, offering flexibility for different learning scenarios. We showed that these dynamics converge to a (near) efficient equilibrium in stochastic teams, with quantifiable approximation errors. We also demonstrated the rationality of the dynamics against agents following pure stationary strategies and stochastic games with a single controller where the stage-payoffs specific to each state induces a potential game. We conducted various simulations to illustrate the convergence numerically.

Chapter 5

Conclusion

In this thesis we examined the problem of learning and decision-making in multi-agent systems where the basic unit of strategic interaction is not an individual, but a *team*. Teams coordinate internally while adapting to other learning teams, which makes the environment intrinsically non-stationary. Classical dynamics that treat players independently miss this intra-team coordination layer, and standard equilibrium analyses can fail to predict what learning teams actually do. Motivated by these gaps, we adopted a team-centric lens and developed algorithms and analysis that remain simple, interpretable, and communication-lean while still admitting rigorous guarantees under realistic information constraints.

We first studied learning in multi-team (near) zero-sum settings and networked games, and introduced a novel class of games termed **Zero-Sum Potential Team Games (ZSPTGs)**. This framework serves as a crucial bridge between potential games (typically applicable to single teams) and zero-sum polymatrix games (applicable to individual agents), thereby expanding the scope of Fictitious Play (FP) to complex multi-team interactions. Within this setting, we proposed *Team-Fictitious-Play (Team-FP)*, a dynamic where agents follow a straightforward behavioral rule: responding to the actions of neighbors while incorporating inertia and exploration. We established that under

standard step-size conditions, the empirical averages of team action profiles converge almost surely to a near Team-Nash Equilibrium (TNE). Furthermore, we derived an explicit bound on the approximation error, demonstrating that it diminishes as the exploration in agents' responses decreases.

Secondly, we addressed the problem of efficient learning in *unknown* stochastic teams (Markov games). We introduced the *Logit-Q* family of dynamics, which synthesizes log-linear learning with Q-learning within a stage-game framework. To provide flexibility for different learning scenarios, we proposed four distinct variants: Averaged Logit-Q (AL-Q), Exploration-free Logit-Q (EL-Q), Averaged Independent Logit-Q (AIL-Q), and Exploration-free Independent Logit-Q (EIL-Q). A key contribution here was proving that these dynamics converge to a (near) *efficient* equilibrium—maximizing the team's collective utility rather than getting stuck at suboptimal points—even when the environment dynamics are unknown. We further demonstrated the rationality of these dynamics against opponents employing pure stationary strategies and extended our results to single-controller settings where stage-wise potentials persist.

A unifying technical challenge across both parts is non-stationarity: beliefs about opponents evolve in competitive team games, and Q-values themselves reshape the stage game in stochastic environments. To solve this problem, we developed an epoch-based analysis, tracking the difference between a fictional stationary case and the actual non-stationary dynamics.

Taken together, the thesis advances a practical theory of *team learning in games*. It shows that simple modifications of classical, interpretable behavioral learning rules such as fictitious play or logit updates can yield reliable coordination and competitive performance in team games with minimal or no communication. The algorithms also works in simulated experiments, demonstrating their potential for real-world multi-agent systems.

Building on the results presented in this thesis, future research can expand in

several directions to enhance the applicability of the proposed learning frameworks. First, computational challenges in larger environments arising from tabular settings may be solved through function approximation techniques. Second, bridging the gap between competitive but static games and stochastic and cooperative settings by exploring competitive team games in stochastically evolving environments would provide a more comprehensive understanding of team learning dynamics. Additionally, investigating the robustness of these algorithms against more sophisticated opponents and allowing limited communication among agents could further enhance their practical utility in real-world multi-agent systems.

Bibliography

[1] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, E. Osawa, and H. Matsubara, "Robocup: A challenge problem for AI," *AI magazine*, vol. 18, no. 1, pp. 73–73, 1997.

[2] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, S. Sastry, *et al.*, "Challenges for securing cyber physical systems," in *Workshop on Future Directions in Cyber-physical Systems Security*, vol. 5, p. 7, 2009.

[3] V. N. Silva and L. Chaimowicz, "MOBA: A new arena for game AI," *arXiv preprint arXiv:1705.10443*, 2017.

[4] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, *et al.*, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[5] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castaneda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman, *et al.*, "Human-level performance in 3d multiplayer games with population-based reinforcement learning," *Science*, vol. 364, no. 6443, pp. 859–865, 2019.

[6] D. A. Lazar, E. Bıyık, D. Sadigh, and R. Pedarsani, "Learning how to dynamically route autonomous vehicles on shared roads," *Transportation Research Part C: Emerging Technologies*, vol. 130, p. 103258, 2021.

[7] B. Von Stengel and D. Koller, "Team-maxmin equilibria," *Games Econom. Behav.*, vol. 21, no. 1-2, pp. 309–321, 1997.

[8] G. Farina, A. Celli, N. Gatti, and T. Sandholm, "Ex ante coordination and collusion in zero-sum multi-player extensive-form games," *Advances in Neural Inform. Process. (NeurIPS)*, vol. 31, 2018.

[9] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, "Finite-sample analysis of decentralized batch multiagent reinforcement learning with networked agents," *IEEE Trans. on Automatic Control*, vol. 66, no. 12, pp. 5925–5940, 2021.

[10] D. Fudenberg and D. K. Levine, "Learning and equilibrium," *Ann. Rev. Econ.*, vol. 1, no. 1, pp. 385–420, 2009.

[11] S. Hart and A. Mas-Colell, "Uncoupled dynamics do not lead to Nash equilibrium," *The American Econ. Rev.*, vol. 83, no. 5, pp. 1830–1836, 2003.

[12] J. Hofbauer and W. H. Sandholm, "On the global convergence of stochastic fictitious play," *Econometrica*, vol. 70, no. 6, pp. 2265–2294, 2002.

[13] D. Monderer and L. S. Shapley, "Potential games," *Games Econom. Behav.*, vol. 14, no. 1, pp. 124–143, 1996.

[14] J. R. Marden and J. S. Shamma, "Revisiting log-linear learning: Asynchrony, completeness and payoff-based implementation," *Games Econom. Behav.*, vol. 75, no. 2, pp. 788–808, 2012.

[15] T. Tatarenko, *Game-theoretic Learning and Distributed Optimization in Memoryless Multi-agent Systems*. Springer, 2017.

[16] T. Tatarenko, "Independent log-linear learning in potential games with continuous actions," *IEEE Trans. on Control of Network Systems*, vol. 5, no. 3, pp. 913–923, 2018.

[17] L. M. Bergman and I. N. Fokin, "On separable non-cooperative zero-sum games," *Optimization*, vol. 44, no. 1, pp. 69–84, 1998.

[18] Y. Cai and C. Daskalakis, "On minmax theorems for multiplayer games," in *Proc. of the 22nd Ann. ACM-SIAM Symposium on Discrete Algorithms*, pp. 217–234, SIAM, 2011.

[19] Y. Cai, O. Candogan, C. Daskalakis, and C. Papadimitriou, "Zero-sum polymatrix games: A generalization of minmax," *Math. Oper. Res*, vol. 41, no. 2, pp. 648–655, 2016.

[20] C. Ewerhart and K. Valkanova, "Fictitious play in networks," *Games Econom. Behav.*, vol. 123, pp. 182–206, 2020.

[21] C. Park, K. Zhang, and A. Ozdaglar, "Multi-player zero-sum Markov games with networked separable interactions," in *Advances in Neural Inform. Process. (NeurIPS)*, pp. 37354–37369, 2023.

[22] A. S. Donmez and M. O. Sayin, "Generalized individual q-learning for polymatrix games with partial observations," in *2024 IEEE 63rd Conf. Decision and Control (CDC)*, pp. 3209–3214, 2024.

[23] A. Celli and N. Gatti, "Computational results for extensive-form adversarial team games," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 32, 2018.

[24] B. Zhang, L. Carminati, F. Cacciamani, G. Farina, P. Olivieri, N. Gatti, and T. Sandholm, "Subgame solving in adversarial team games," *Advances in Neural Inform. Process. (NeurIPS)*, vol. 35, pp. 26686–26697, 2022.

[25] L. Carminati, F. Cacciamani, M. Ciccone, and N. Gatti, "A marriage between adversarial team games and 2-player games: Enabling abstractions, no-regret learning, and subgame solving," in *Int. Conf. Mach. Learn. (ICML)*, pp. 2638–2657, PMLR, 2022.

[26] L. E. Blume, "The statistical mechanics of strategic interaction," *Games Econom. Behav.*, vol. 5, no. 3, pp. 387–424, 1993.

[27] C. Alos-Ferrer and N. Netzer, "The logit-response dynamics," *Games Econom. Behav.*, vol. 68, pp. 413–427, 2010.

[28] G. C. Chasparis, A. Arapostathis, and J. S. Shamma, "Aspiration learning in coordination games," *SIAM J. Control Optim.*, vol. 51, no. 1, pp. 465–490, 2013.

[29] J. R. Marden, H. P. Young, and L. Y. Pao, "Achieving Pareto optimality through distributed learning," *SIAM J. Control Optim.*, vol. 52, pp. 2753–2770, 2014.

[30] B. S. Pradelski and H. P. Young, "Learning efficient Nash equilibria in distributed systems," *Games Econom. Behav.*, vol. 75, pp. 882–897, 2012.

[31] L. S. Shapley, "Stochastic games," *Proc. Natl. Acad. Sci. USA*, vol. 39, no. 10, pp. 1095–1100, 1953.

[32] L. Baudin and R. Laraki, "Fictitious play and best-response dynamics in identical interest and zero-sum stochastic games," in *Int. Conf. Mach. Learn. (ICML)*, pp. 1664–1690, PMLR, 2022.

[33] D. S. Leslie, S. Perkins, and Z. Xu, "Best-response dynamics in zero-sum stochastic games," *J. Econ. Theory*, vol. 189, p. 105095, 2020.

[34] M. O. Sayin, K. Zhang, D. Leslie, T. Basar, and A. Ozdaglar, "Decentralized q-learning in zero-sum Markov games," *Advances in Neural Inform. Process. (NeurIPS)*, vol. 34, pp. 18320–18334, 2021.

[35] M. O. Sayin, F. Parise, and A. Ozdaglar, "Fictitious play in zero-sum stochastic games," *SIAM J. Control Optim.*, vol. 60, no. 4, pp. 2095–2114, 2022.

[36] Zhang, K., Sayin, and M. et al., "(smooth) fictitious-play in identical-interest stochastic games with independent continuation-payoff estimates," *Applied and Computational Mathematics*, vol. 23, no. 3, pp. 366–391, 2024.

[37] M. O. Sayin, "On the global convergence of stochastic fictitious play in stochastic games with turn-based controllers," in *Proc. IEEE Conf. Decision and Control (CDC)*, 2022.

[38] M. O. Sayin and K. Cetiner, "On the heterogeneity of independent learning dynamics in zero-sum stochastic games," in *Learning for Dynamics and Control Conf.*, pp. 994–1005, PMLR, 2022.

[39] A. Ozdaglar, M. O. Sayin, and K. Zhang, "Independent learning in stochastic games," in *Proc. Int. Cong. Math.*, 2021.

[40] Z. Chen, K. Zhang, E. Mazumdar, A. Ozdaglar, and A. Wierman, "A finite-sample analysis of payoff-based independent learning in zero-sum stochastic games," in *Advances in Neural Inform. Process. (NeurIPS)* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 75826–75883, Curran Associates, Inc., 2023.

[41] Y. Arslantas, E. Yuceel, Y. Yalin, and M. O. Sayin, "Convergence of heterogeneous learning dynamics in zero-sum stochastic games," *IEEE Trans. on Automatic Control*, vol. 70, no. 11, pp. 7523–7537, 2025.

[42] Y. Arslantas, "Heterogeneity and strategic sophistication in multi-agent reinforcement learning," master's thesis, Bilkent University, 2024.

[43] I. Bistriz and A. Leshem, "Game of thrones: Fully distributed learning for multiplayer bandits," *Math. Oper. Res.*, vol. 46, no. 1, pp. 159–178, 2021.

[44] B. Yongacoglu, G. Arslan, and S. Yüksel, "Decentralized learning for optimality in stochastic dynamic teams and games with local control and global state information," *IEEE Trans. on Automatic Control*, vol. 67, no. 10, pp. 5230–5245, 2022.

[45] C. Boutilier, "Planning, learning and coordination in multiagent decision processes," in *Conf. Theoretical Aspects of Rationality and Knowledge (TARK)*, pp. 195–210, 1996.

[46] M. Lauer and M. A. Riedmiller, "An algorithm for distributed reinforcement learning in cooperative multi-agent systems," in *Int. Conf. Mach. Learn. (ICML)*, 2000.

[47] M. L. Littman, "Friend-or-foe Q-learning in general-sum games," 2001.

[48] X. Wang and T. Sandholm, "Reinforcement learning to play an optimal Nash equilibrium in team Markov games," in *Proc. Advances in Neural Inform. Process. (NeurIPS)*, 2002.

[49] Y.-W. Cheung and D. Friedman, "Individual learning in normal form games: Some laboratory results," *Games Econom. Behav.*, vol. 19, no. 1, pp. 46–76, 1997.

[50] S. Leonardos, W. Overman, I. Panageas, and G. Piliouras, "Global convergence of multi-agent policy gradient in Markov potential games," *arXiv preprint arXiv:2106.01969*, 2021.

[51] O. Unlu and M. O. Sayin, "Episodic logit-q dynamics for efficient learning in stochastic teams," in *IEEE Conf. Decision and Control (CDC)*, 2023.

[52] V. S. Borkar, "Stochastic approximation with two time scales," *Systems & Control Letters*, vol. 29, no. 5, pp. 291–294, 1997.

[53] M. Benaïm, J. Hofbauer, and S. Sorin, "Stochastic approximations and differential inclusions," *SIAM J. Control Optim.*, vol. 44, no. 1, pp. 328–348, 2005.

[54] S. Perkins and D. S. Leslie, "Asynchronous stochastic approximation with differential inclusions," *Stochastic Systems*, vol. 2, no. 2, pp. 409–446, 2013.

[55] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times*. American Mathematical Society, 2nd ed., 2017.

[56] G. E. Cho and C. D. Meyer, "Comparison of perturbation bounds for the stationary distribution of a Markov chain," *Linear Algebra Appl.*, vol. 335, no. 1-3, pp. 137–150, 2001.

[57] Y. Liu, "Perturbation bounds for the stationary distributions of Markov chains," *SIAM J. Matrix Anal. Appl.*, vol. 33, no. 4, pp. 1057–1074, 2012.

[58] A. S. Donmez, Y. Arslantas, and M. O. Sayin, "Team-fictitious play for reaching team-nash equilibrium in multi-team games," in *Advances in Neural Inform. Process. (NeurIPS)* (A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, eds.), vol. 37, pp. 111515–111543, Curran Associates, Inc., 2024.

[59] G. Arslan, M. F. Demirkol, and Y. Song, "Equilibrium efficiency improvement in MIMO interference systems: A decentralized stream control approach," *IEEE Trans. on Wireless Communications*, vol. 6, no. 8, pp. 2984–2993, 2007.

[60] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y. D. Yao, "Opportunistic spectrum access in cognitive radio networks: Global optimization using local interaction games," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 2, pp. 180–194, 2012.

[61] J. Zheng, Y. Cai, Y. Liu, Y. Xu, B. Duan, and X. Shen, "Optimal power allocation and user scheduling in multicell networks: Base station cooperation using a game-theoretic approach," *IEEE Trans. on Wireless Communications*, vol. 13, no. 12, pp. 6928–6942, 2014.

[62] F. Kalogiannis, I. Panageas, and E.-V. Vlatakis-Gkaragkounis, "Towards convergence to nash equilibria in two-team zero-sum games," in *Int. Conf. Learn. Represent. (ICLR)*, 2022.

[63] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Int. Conf. Mach. Learn. (ICML)*, 1994.

[64] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *J. Mach. Learn. Res.*, pp. 1039–1069, 2003.

[65] S. T. Kiremitci, A. S. Donmez, and M. O. Sayin, "Achieving pareto optimality in games via single-bit feedback," 2025.

[66] M. O. Sayin, "Balancing anarchy and efficiency: partial team formations and learning in potential games," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 33, no. 1, pp. 32–47, 2025.

[67] D. Monderer and L. S. Shapley, "Fictitious play property for games with identical interests," *J. Econ. Theory*, vol. 68, no. 1, pp. 258–265, 1996.

[68] A. S. Donmez, O. Unlu, and M. O. Sayin, "Logit-q dynamics for efficient learning in stochastic teams," *SIAM J. Control Optim.*, vol. 63, no. 5, pp. 3219–3243, 2025.

[69] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, no. 3, pp. 279–292, 1992.

[70] M. O. Sayin, K. Zhang, and A. Ozdaglar, "Fictitious play in markov games with single controller," in *Proc. of the 23rd ACM Conf. on Economics and Computation*, EC '22, (New York, NY, USA), pp. 919–936, Association for Computing Machinery, 2022.

[71] P. Guan, M. Raginsky, R. Willett, and D.-S. Zois, "Regret minimization algorithms for single-controller zero-sum stochastic games," in *2016 IEEE 55th Conf. Decision and Control (CDC)*, pp. 7075–7080, 2016.

[72] S. Qiu, X. Wei, J. Ye, Z. Wang, and Z. Yang, "Provably efficient fictitious play policy optimization for zero-sum markov games with structured transitions," in *Proc. of the 38th Int. Conf. on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proc. Mach. Learn. Res.*, pp. 8715–8725, PMLR, 18–24 Jul 2021.

[73] C. Daskalakis, N. Golowich, and K. Zhang, "The complexity of Markov equilibrium in stochastic games," in *36th Ann. Conf. Learn. Theory*, vol. 195, pp. 1–55, 2023.

[74] A. M. Fink, "Equilibrium in stochastic n-person game," *J. Science Hiroshima University Series A-I*, vol. 28, pp. 89–93, 1964.

[75] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, 4th ed., 2017.

[76] S. Perkins and D. S. Leslie, "Asynchronous stochastic approximation with differential inclusions," *Stochastic Systems*, vol. 2, no. 2, pp. 409–446, 2012.

[77] R. I. Brafman and M. Tennenholtz, "R-max: A general polynomial time algorithm for near-optimal reinforcement learning," *J. Mach. Learn. Res.*, vol. 3, pp. 213–231, 2002.

[78] M. Benaim, "Dynamics of stochastic approximation algorithms," in *Le Seminaire de Probabilities, Lecture Notes in Math. 1709*, (Berlin), pp. 1–68, Springer-Verlag, 1999.

[79] B. Gao and L. Pavel, "On the properties of the softmax function with application in game theory and reinforcement learning," *arXiv preprint:1704.00805v4*, 2018.

Appendix A

Proofs of Technical Results in Chapter 2

The followings are the proofs of Lemma 1, and Lemma 2 described in Chapter 2. These lemmas are independently interesting and form the basis of the analysis in this thesis.

A.1 Proof of Lemma 1

By the coupling lemma [55, Proposition 4.2, 4.7], we have:

$$\|\mu_k^w - \widehat{\mu}_k\|_1 \;\leq\; 2\,\Pr(w_k \neq \widehat{w}_k).$$

We construct a coupling of the processes $\{\widehat{w}_k\}$ and $\{(w_k, z_k)\}$ to bound $\Pr(w_k \neq \widehat{w}_k)$. Let us describe the coupling in two cases:

Case (i): $w_k \neq \widehat{w}_k$. The two processes evolve *independently* under their respective transition kernels $\mathbb{P}$ and $\widehat{\mathbb{P}}$.

Case (ii): $w_k = \widehat{w}_k$. They again evolve under $\mathbb{P}$ and $\widehat{\mathbb{P}}$, but now in a *coupled*

manner so that

$$\Pr\big(w_{k+1} \neq \widehat{w}_{k+1} \mid \widehat{w}_k = w_k, \{w_l, z_l\}_{l=0}^{k}\big) = \Big\| \mathbb{P}^w(\cdot \mid \{w_l, z_l\}_{l\leq k}) - \widehat{\mathbb{P}}(\cdot \mid w_k) \Big\|_{\mathrm{TV}}. \tag{A.1}$$

By Condition (ii), the total variation above is at most $1-\lambda$. Summing over any possible history $\{z_l\}_{l\leq k}$ then yields, for all k,

$$\Pr\big(w_{k+1} \neq \widehat{w}_{k+1} \mid \widehat{w}_k = w_k, \{w_l\}_{l=0}^{k}\big) \;\leq\; 1-\lambda. \tag{A.2}$$

Bounding the mismatch probability. Condition (i) states that if $w_k \neq \widehat{w}_k$, then with probability at least $\varepsilon\,\widehat{\varepsilon}$ (over κ steps) both processes "meet" in the same state. Formally, there exist paths $\{\widehat{h}^l\}_{l=0}^{\kappa}$, $\{h^l\}_{l=0}^{\kappa}$ such that $\widehat{h}^l \neq h^l$ for $l<\kappa$ and $\widehat{h}^\kappa = h^\kappa$. Hence,

$$\Pr(w_{k+\kappa} = \widehat{w}_{k+\kappa} \;\mid\; w_k \neq \widehat{w}_k) \;\geq\; \varepsilon\,\widehat{\varepsilon}, \tag{A.3}$$

where independence (Case (i)) ensures that probabilities ε and, $\widehat{\varepsilon}$ can be multiplied for these two paths.

We now partition time in blocks of length κ. From (A.3) we have, whenever $w_{(m-1)\kappa} \neq \widehat{w}_{(m-1)\kappa}$,

$$\Pr\big(w_{m\kappa} = \widehat{w}_{m\kappa} \;\mid\; w_{(m-1)\kappa} \neq \widehat{w}_{(m-1)\kappa}\big) \;\geq\; \varepsilon\,\widehat{\varepsilon}, \tag{A.4}$$

while from (A.2) and Case (ii), if $w_{(m-1)\kappa} = \widehat{w}_{(m-1)\kappa}$ then

$$\Pr\big(w_{m\kappa} = \widehat{w}_{m\kappa} \;\mid\; w_{(m-1)\kappa} = \widehat{w}_{(m-1)\kappa}\big) \;\geq\; \lambda^\kappa. \tag{A.5}$$

Combining (A.4), (A.5), and using induction, one obtains for all $m \geq 0$:

$$\Pr\big(w_{m\kappa} \neq \widehat{w}_{m\kappa}\big) \;\leq\; (1-\varepsilon\,\widehat{\varepsilon})^m \;+\; \frac{1-\lambda^\kappa}{\varepsilon\,\widehat{\varepsilon}}. \tag{A.6}$$

Lastly, for any $k \in [m\kappa, (m+1)\kappa)$ and $m = 0, 1, \ldots$, the mismatch probability is bounded from above by

$$\begin{aligned}
\Pr(w_k \neq \widehat{w}_k) =\; & \Pr(w_k \neq \widehat{w}_k \mid w_{m\kappa} \neq \widehat{w}_{m\kappa}) \times \Pr(w_{m\kappa} \neq \widehat{w}_{m\kappa}) \\
& + (1 - \Pr(w_k = \widehat{w}_k \mid w_{m\kappa} = \widehat{w}_{m\kappa})) \times \Pr(w_{m\kappa} = \widehat{w}_{m\kappa}) \\
\leq\; & \Pr(w_{m\kappa} \neq \widehat{w}_{m\kappa}) + 1 - \lambda^\kappa
\end{aligned} \tag{A.7}$$

since $\Pr(w_k = \widehat{w}_k \mid w_{m\kappa} = \widehat{w}_{m\kappa}) \geq \lambda^{k-m\kappa} \geq \lambda^{\kappa}$ as in (A.5). Combining (A.6) and (A.7), we obtain

$$\Pr\left(w_k \neq \widehat{w}_k\right) \leq (1-\varepsilon\widehat{\varepsilon})^m + (1-\lambda^{\kappa})\frac{1+\varepsilon\widehat{\varepsilon}}{\varepsilon\widehat{\varepsilon}} \quad \forall k \in [m\kappa, (m+1)\kappa). \tag{A.8}$$

for any $m = 0, 1, \ldots$. Since $1 - \varepsilon\widehat{\varepsilon} \in (0,1)$, we have $(1-\varepsilon\widehat{\varepsilon})^m \leq (1-\varepsilon\widehat{\varepsilon})^{k/\kappa - 1}$ for all $k \in [m\kappa, (m+1)\kappa)$. This completes the proof.

A.2 Proof of Lemma 2

Let $a_t \in A$ and $\underline{a}_t \in A$ denote the action profiles, resp., in the classical and independent log-linear dynamics. Both $\{a_t\}_{t\geq 0}$ and $\{\underline{a}_t\}_{t\geq 0}$ form irreducible Markov chains with unique stationary distributions μ^{coor} and μ^{ind}, respectively, and it is known that $\mu^{\text{coor}} = \underline{\text{br}}(\Phi(\cdot))$.

Let P and $\underline{P}$ denote the transition matrices in the Markov chains, resp., associated with the classical and independent log-linear learning dynamics. The ways agents update their actions in both dynamics yield that the transition matrices P and $\underline{P}$ satisfy $\underline{P} = p_0 I + p_1 P + (1 - p_0 - p_1)E$, where $p_0 := (1-\delta)^n \in (0,1)$ and $p_1 := n\delta(1-\delta)^{n-1} \in (0,1)$ correspond, resp., to the probability that none and one of the agents update their actions, and E is the transition matrix corresponding to the case at least two (randomly picked) agents updates their actions. The transition matrix corresponding to the case that at least one (randomly picked) agent updates their actions is given by

$$\overline{P} = \frac{p_1}{1-p_0}P + \frac{1-p_0-p_1}{1-p_0}E. \tag{A.9}$$

Since $\overline{P}$ and $\underline{P}$ share the same unique stationary distribution, we focus on $\overline{a}_t$ evolving by $\overline{P}$ rather than $\underline{P}$, and couple its evolution with a_t based on Lemma 1

To apply Lemma 1, we verify two conditions. First, note that for any two states $a, \overline{a} \in A$, both dynamics allow transitions in which only one agent updates at a time. Thus, there exist N-length paths from a and from $\overline{a}$ to a common state

(say, a') with consecutive profiles differing in at most one coordinate while both process differ until they reach a' at step N. Since each one-agent update occurs with probability uniformly bounded below by a constant $\epsilon > 0$, the probability of following any such N-step path is at least $\left(\frac{\epsilon}{N}\right)^N$ for the process $\{a_t\}$ and $\left(\frac{\epsilon\delta(1-\delta)^{N-1}}{1-(1-\delta)^N}\right)^N$ for the process $\{\overline{a}_t\}$, establishing the required path connectivity.

On the other hand, (A.9) yields that

$$\left\|P(\cdot \mid h) - \overline{P}(\cdot \mid h)\right\|_{\mathrm{TV}} = \frac{1 - p_0 - p_1}{1 - p_0} \cdot \left\|P(\cdot \mid h) - E(\cdot \mid h)\right\|_{\mathrm{TV}} \tag{A.10}$$

for any $h \in A$. Note that $\left\|P(\cdot \mid h) - E(\cdot \mid h)\right\|_{\mathrm{TV}} \leq 1$.

Therefore, Condition (ii) in Lemma 1 holds for $\lambda = 1-(1-p_0-p_1)/(1-p_0) = p_1/(1-p_0) \leq 1$. Also, Condition (i) holds for $\varepsilon = \left(\frac{\epsilon}{N}\right)^N$ and $\widehat{\varepsilon} = \left(\frac{\epsilon\delta(1-\delta)^{N-1}}{1-(1-\delta)^N}\right)^N$. Since Conditions (i) and (ii) hold, we can invoke Lemma 1 as let $t \to \infty$. Let's define the following variables:

$$\lambda(\delta) := \frac{N\delta(1-\delta)^{N-1}}{1-(1-\delta)^N},$$
$$\overline{\epsilon} := \left(\frac{\epsilon}{N}\right)^N.$$

Then, by Lemma 1, we obtain (2.6), completing the proof.

Appendix B

Proofs of Technical Results in Chapter 3

The followings are the proofs of Lemma 3, Lemma 4, and Propositions 1 and 2 used in §3.4.

B.1 Proof of Lemma 3

We first observe that by some algebraic manipulation of the definition of $\beta_{(n)}$ in (3.16), we have

$$\beta_{(n)} = 1 - \prod_{k=nT}^{(n+1)T-1} (1 - \alpha_k). \tag{B.1}$$

Using the inequality $1 - x \leq e^{-x}$ for $x \geq 0$, we bound the product term:

$$\prod_{k=nT}^{(n+1)T-1} (1 - \alpha_k) \leq \exp\left(- \sum_{k=nT}^{(n+1)T-1} \alpha_k \right). \tag{B.2}$$

Let $S_n = \sum_{k=nT}^{(n+1)T-1} \alpha_k$. Then $\beta_{(n)} \geq 1 - e^{-S_n}$. Since $\alpha_k \to 0$ and T is finite, $S_n \to 0$ as $n \to \infty$. Using the bound $1 - e^{-x} \geq \frac{1}{2}x$ valid for $x \in [0, 1]$, there exists

an N such that for all $n \geq N$:

$$\beta_{(n)} \geq \frac{1}{2} \sum_{k=nT}^{(n+1)T-1} \alpha_k. \tag{B.3}$$

Summing over n, the divergence of $\sum_{k=0}^{\infty} \alpha_k$ implies $\sum_{n=0}^{\infty} \beta_{(n)} = \infty$.

For proving the convergence of the square sum, we use the definition of $\beta_{(n)}$ in (3.16) again:

$$\beta_{(n)} = \alpha_{nT} \prod_{l=nT+1}^{(n+1)T-1} (1 - \alpha_l) + \cdots + \alpha_{(n+1)T-1} \leq \sum_{k=nT}^{(n+1)T-1} \alpha_k, \tag{B.4}$$

because of the fact that $0 < \alpha_k < 1$. Squaring both sides and applying the Cauchy-Schwarz inequality ($(\sum_{i=1}^{T} x_i)^2 \leq T \sum_{i=1}^{T} x_i^2$):

$$\beta_{(n)}^2 \leq \left(\sum_{k=nT}^{(n+1)T-1} \alpha_k \right)^2 \leq T \sum_{k=nT}^{(n+1)T-1} \alpha_k^2. \tag{B.5}$$

Summing over n, we obtain $\sum_{n=0}^{\infty} \beta_{(n)}^2 \leq T \sum_{k=0}^{\infty} \alpha_k^2$. Since $\sum \alpha_k^2 < \infty$, it follows that $\sum \beta_{(n)}^2 < \infty$.

B.2 Proof of Lemma 4

We define two discrete-time processes from the beginning of epoch n. One of them is the joint actions of teams in the reference scenario, defined as $\{\widehat{\omega}_k\}_{k \geq nT} \coloneqq \{(\widehat{\underline{a}}_k^m)_{m \in \mathcal{T}} | \mathcal{F}_{(n)}\}_{k \geq nT}$, and the other one is the joint actions of teams in the actual scenario, with $\{\omega_k\}_{k \geq nT} \coloneqq \{(\underline{a}_k^m)_{m \in \mathcal{T}} | \mathcal{F}_{(n)}\}_{k \geq nT}$.

The transition probabilities between states depend only on the beliefs on the other teams. Therefore, for the reference scenario, during an epoch, this process is a homogeneous MC. Let, $P_{(n),k}$ be the transition probabilities between the states, i.e., joint action profiles of all teams, of the original discrete-time process at step k, and let $\widehat{P}_{(n)}$ be the transition probabilities between the states for the reference scenario. Let $\widehat{\mu}_{(n),k}$ be the distribution of $\widehat{\omega}_k$, and let $\mu_{(n),k}$ be

the distribution of ω_k. In this case, the expected joint actions of a team at step k within an epoch in real and fictional scenarios can be expressed in terms of the distributions as:

$$\mu^m_{(n),k}(a) = \sum_{\omega:\{a^m=a\}} \mu_{(n),k}(\omega), \qquad \widehat{\mu}^m_{(n),k}(a) = \sum_{\widehat{\omega}:\{a^m=a\}} \widehat{\mu}_{(n),k}(\widehat{\omega}). \tag{B.6}$$

Step 1. In classical log-linear learning, at each step, only one agent can change their action and due to the soft-max nature of this action change, any action has positive probability which is bounded from below. This bound only depends on the minimum and maximum values of any agents' reward, which are defined by the game and bounded by definition, and the temperature parameter. If we divide that bound by the number of agents in the team ($|\mathcal{I}^m|$), we can obtain a lower bound on the probability of changing to any joint action which can be reached within a single step. Let's call that bound ξ. Then, for any state ω or $\widehat{\omega}$, there is a $\xi > 0$ such that

$$P_{(n),k}(\omega|\omega_k) > \xi^{|\mathcal{T}|} \iff P_{(n),k}(\omega|\omega_k) > 0, \tag{B.7}$$

$$\widehat{P}_{(n),k}(\widehat{\omega}|\widehat{\omega}_k) > \xi^{|\mathcal{T}|} \iff \widehat{P}_{(n),k}(\widehat{\omega}|\widehat{\omega}_k) > 0. \tag{B.8}$$

Then, we can find a path with positive probability for both of these scenarios, where their state does not match until they reach $\kappa = \max_m |\mathcal{I}^m|$. In other words, $\omega_{nT+\kappa} = \widehat{\omega}_{nT+\kappa}$ and $\omega_k \neq \widehat{\omega}_k$ for all $k = nT, \ldots, nT + \kappa - 1$ with

$$\Pr\left(\bigwedge_{k=nT+1}^{nT+\kappa} \widehat{\omega}_k \mid \widehat{\omega}_{nT}\right) \geq \xi^\kappa \quad \text{and} \quad \Pr\left(\bigwedge_{k=nT+1}^{nT+\kappa} \omega_k \mid \omega_{nT}\right) \geq \xi^\kappa. \tag{B.9}$$

Hence, (B.9) satisfies the first condition for Lemma 1.

Step 2. For this part, we introduce a notation $a^{jm} \in A^{jm}$, where jm represents the j^{th} agent from team m. For example, if all teams have identical agent numbers and agent indexes are ordered for teams, $jm = (m-1)|T_m| + j$. Also, let $\pi_{(n),k} \coloneqq (\pi^m_{(n),k})_{m\in\mathcal{T}}$, and $\widehat{\pi}_{(n),k} \coloneqq (\widehat{\pi}^m_{(n),k})_{m\in\mathcal{T}}$. Now, consider the total variation distance between two transition probabilities. Transition probabilities between state $\omega = \{(a^{im}, a^{-im})\}_{m\in\mathcal{T}}$ and $\widetilde{\omega} = \{(\widetilde{a}^{im}, a^{-im})\}_{m\in\mathcal{T}}$, where im can be any

random agent from team m with a slight abuse of notation, can be written as a function of the belief as

$$P_{\omega\to\widetilde{\omega}}(\pi_{(n),k}) = \prod_{m\in\mathcal{T}} P^m_{\omega\to\widetilde{\omega}}(\pi_{(n),k}), \tag{B.10}$$

where $P^m_{\omega\to\widetilde{\omega}}$ is defined to be

$$P^m_{\omega\to\widetilde{\omega}}(\pi_{(n),k}) = \frac{1}{|\mathcal{I}^m|} \frac{\exp\left(\left(\phi^m\left(\widetilde{a}^{im}, a^{-im}, \pi^{-m}_{(n),k}\right)\right)/\tau\right)}{\sum_{\widetilde{a}'\in A^{im}} \exp\left(\phi^m\left(\widetilde{a}', a^{-im}, \pi^{-m}_{(n),k}\right)/\tau\right)}, \tag{B.11}$$

and $\pi^{-m}_{(n),k} \coloneqq (\pi^{\ell}_{(n),k})_{\ell\neq m}$. Remember that due to the separable structure of the network in ZSPTG we have

$$\phi^m(a^{im}, (a^{-im}), \pi^{-m}_{(n),k}) = \sum_{\ell\neq m} \mathrm{E}_{\underline{a}^\ell\sim\pi^{\ell}_{(n),k}} \phi^{m\ell}(a^{im}, a^{-im}, \underline{a}^\ell) \tag{B.12a}$$

$$= \sum_{\ell\neq m} (\underline{a}^m)^T \Phi^{m\ell} \pi^{\ell}_{(n),k}, \tag{B.12b}$$

where $\Phi^{m\ell}$ is the matrix form of the potential function whose rows are the joint actions of team m and columns are the joint actions of team ℓ.

Now, for all ω_{k+1}, which is reachable in one step from the state ω_k, i.e., at most a single agent changes action in each team, the relation between state transition probabilities and $P^m_{\omega\to\widetilde{\omega}}(\cdot)$ can be expressed as

$$P_{(n),k}(\omega_{k+1}|\omega_k, \ldots, \omega_{nT}) = \prod_{m\in\mathcal{T}} P^m_{\omega_k\to\omega_{k+1}}(\pi_{(n),k}), \tag{B.13a}$$

$$\widehat{P}_{(n)}(\omega_{k+1}|\omega_k) = \prod_{m\in\mathcal{T}} P^m_{\omega_k\to\omega_{k+1}}(\widehat{\pi}_{(n),k}). \tag{B.13b}$$

Note that $\pi_{(n),k}$ is a function of history of actions and the initial $\pi_{(n)}$. Therefore, it can be computed given the past states $(\omega_k, \ldots, \omega_{nT})$.

We know that $P^m_{\omega\to\widetilde{\omega}}(\pi_{(n),k})$ is bounded as it is a probability. Now, using the Lipschitz property of the soft-max function, and since $\phi^m(\cdot, \pi^{-m}_{(n),k})$ is a linear function of $\pi_{(n),k}$, we can say that $P^m_{\omega_k\to(\cdot)}(\pi)$ is a Lipschitz continuous function with respect to the input π. Also, using the fact that the product of bounded Lipschitz continuous functions is also Lipschitz continuous, and (B.13), we can

say that there exists an $\mathcal{L} < \infty$ such that the total variation distance between two transition probabilities is bounded as

$$\left\| P_{(n),k}(\cdot|\omega_k, \ldots, \omega_{nT}) - \widehat{P}_{(n)}(\cdot|\omega_k) \right\|_{\mathrm{TV}} \leq \mathcal{L} \sum_{m \in \mathcal{T}} \left\| (\pi^{\ell}_{(n),k} - \widehat{\pi}^{\ell}_{(n),k}) \right\|_1 . \tag{B.14}$$

The distance between the belief of the original scenario and the reference scenario can also be bounded thanks to the small step size. If we bound $\|a^{\ell}_k - \pi^{\ell}_k\|_1 < \|a^{\ell}_k\|_1 + \|\pi^{\ell}_k\|_1 = 2$. Then using triangle inequality

$$\left\| (\pi^{\ell}_{(n),k} - \widehat{\pi}^{\ell}_{(n),k}) \right\|_1 = \left\| (\pi^{\ell}_{(n),k} - \pi^{\ell}_{(n),nT}) \right\|_1 \leq \sum_{t=nT}^{k+NT-1} 2\alpha_t \tag{B.15a}$$

$$\leq \sum_{t=nT}^{(n+1)T-1} 2\alpha_t \tag{B.15b}$$

$$\leq 2T\alpha_{nT}. \tag{B.15c}$$

Let's consider late epochs where $\alpha_{nT} < \dfrac{1}{2T\mathcal{L}|\mathcal{T}|}$, and set $2T\mathcal{L}|\mathcal{T}|\alpha_{nT} = 1 - \lambda_{nT}$ with $0 < \lambda_{nT} \leq 1$. Then,

$$\left\| P_{\omega \to \widetilde{\omega}}(\pi^{-m}_{(n),k}) - P_{\omega \to \widetilde{\omega}}(\widehat{\pi}^{-m}_{(n),k}) \right\|_{\mathrm{TV}} \leq 1 - \lambda_{nT}. \tag{B.16}$$

Hence, the second condition of Lemma 1 is met, and we can invoke Lemma 1 such that given the distributions of the original and reference scenarios, following inequality holds for all $k \geq nT$,

$$\left\| \mu_{(n),k} - \widehat{\mu}_{(n),k} \right\|_1 \leq 2(1-\varepsilon)^{\frac{k-nT}{\kappa}-1} + 2\left(1 - \lambda^{\kappa}_{nT}\right) \frac{1+\varepsilon}{\varepsilon}, \tag{B.17}$$

where $\varepsilon = \xi^{2\kappa}$. If we define constants, $c \coloneqq 2\left(1-\varepsilon\right)^{\frac{1}{k}-1}$, $\rho \coloneqq (1-\varepsilon)^{\frac{1}{\kappa}}$, $d \coloneqq 4\dfrac{1+\varepsilon}{\varepsilon}$, and assume that the reference scenario initial distribution is the stationary distribution of the MC, $\widehat{\mu}_{(n),\star}$, we can rewrite the inequality (B.17) as follows

$$\left\| \mu_{(n),k} - \widehat{\mu}_{(n),\star} \right\|_1 \leq c \cdot \rho^{k-nT} + d \cdot T\alpha_{nT}, \tag{B.18}$$

for all $k \geq nT$. Then, by (B.6), and triangle inequality, we can conclude

$$\left\| \mu^{m}_{(n),k} - \widehat{\mu}^{m}_{(n),\star} \right\|_1 \leq \left\| \mu_{(n),k} - \widehat{\mu}_{(n),\star} \right\|_1 \leq c \cdot \rho^{k-nT} + d \cdot T\alpha_{nT}, \tag{B.19}$$

for all $k \geq nT$.

B.3 Proof of Proposition 1

The set Π is compact set by definition as it is a Cartesian product of probability simplexes. Let's consider a convergent sequence $(\pi_n, \mu_n - \pi_n + e_n)_{n=1,2,\dots}$ in the set $\{(\pi_n, y) \colon y \in F(\pi)\}$, and let $(\pi^\star, \mu^\star - \pi^\star + e^\star)$ be the point that the sequence converges to. Given π_n, any μ_n is a fixed and unique value, and it is an element of the compact set that is generated by mapping probability simplex with the continuous soft-max function. Then, for any $\pi^\star \in \Pi$, $\mu^\star - \pi^\star$ is a fixed value within another compact set. Furthermore, the error term must remain within the compact set $e^\star \in e$. As a result, $(\pi^\star, \mu^\star - \pi^\star + e^\star)$ is also within the set $\{(\pi_n, y) \colon y \in F(\pi)\}$, and $F : \Pi \to A^{\sum_{m\in\mathcal{T}} \|\underline{A}^m\|}$ is a closed-set valued map. Therefore, the condition (i) is satisfied. Given a $\pi \in \Pi$, $\mu_\star \in \Pi$ is a fixed value corresponding to the smoothed best responses to $\pi^{-m}_{\{m\in\mathcal{T}\}}$. Hence, $\mu_\star - \pi$ is a fixed value for a given π. Note that each $e^m \in e$ is a non-empty, bounded, closed and convex subset of $\mathbb{R}^{\sum_{m\in\mathcal{T}} \|\underline{A}^m\|}$. Therefore, for any given $\pi \in \Pi$, $F(\pi) = \mu_\star - \pi + e$ is a non-empty, compact, convex subset of $\mathbb{R}^{\sum_{m\in\mathcal{T}} \|\underline{A}^m\|}$. As a result, (ii) is also satisfied. The function F is bounded such that

$$\sup_{y\in F(x)} \|y\|_1 \leq \sup_{\pi\in\Delta} \|\pi\|_1 + \sup_{\mu\in\Delta} \|\mu\|_1 + \sup \|e\| \leq 2M + M\left(\frac{1}{T}\frac{1}{1-\rho} + K^m(\delta)\right). \tag{B.20}$$

Hence, it satisfies the condition (iii). Since all three conditions are satisfied, F is a Marchaud Map.

B.4 Proof of Proposition 2

The smoothness of the entropy regularization in (3.1) yields that we can invoke the envelope theorem to compute the time derivative of $L(\pi)$ as

$$\frac{d}{dt}L(\pi) = \sum_{m\in\mathcal{T}} \phi^m(\mu^m_\star, \dot{\pi}^{-m}) \tag{B.21}$$

$$\overset{(a)}{=} \sum_{m\in\mathcal{T}} \sum_{\ell\neq m} \phi^{m\ell}(\mu^m_\star, \dot{\pi}^\ell), \tag{B.22}$$

where $\mu_\star^m := \mathrm{br}_\tau(\phi^m(\cdot, \pi^{-m}))$ and (a) follows from (3.4). By (3.34) and (3.32), we have $\dot{\pi}^\ell = \mu_\star^\ell - \pi^\ell + e^\ell$. Therefore, we can write (B.22) as

$$\frac{d}{dt}L(\pi) = \sum_{m\in\mathcal{T}}\sum_{\ell\neq m}\left(\phi^{m\ell}(\mu_\star^m, \mu_\star^\ell) - \phi^{m\ell}(\mu_\star^m, \pi^\ell) + \sum_{\underline{a}^\ell\in\underline{A}^\ell} e^\ell(\underline{a}^\ell)\phi^{m\ell}(\mu_\star^m, \underline{a}^\ell)\right) \tag{B.23}$$

For the first two terms on the right-hand side, we have

$$\sum_{m\in\mathcal{T}}\sum_{\ell\neq m}\left(\phi^{m\ell}(\mu_\star^m, \mu_\star^\ell) - \phi^{m\ell}(\mu_\star^m, \pi^\ell)\right) \overset{(a)}{=} -\sum_{m\in\mathcal{T}}\phi^m(\mu_\star^m, \pi^{-m}), \tag{B.24}$$

$$\overset{(b)}{\leq} -L(\pi) + \tau\log|A|, \tag{B.25}$$

where (a) is due to (3.3) and (3.4), and (b) follows from the definition (3.36a) as $\mathcal{H}(\mu_\star^m) \leq \log|\underline{A}^m|$. On the other hand, we have

$$\sum_{m\in\mathcal{T}}\sum_{\ell\neq m}\sum_{\underline{a}^\ell\in\underline{A}^\ell} e^\ell(\underline{a}^\ell)\phi^{m\ell}(\mu_\star^m, \underline{a}^\ell) \leq \sum_{m\in\mathcal{T}}\sum_{\ell\neq m}\|e^\ell\|_1\cdot\overline{\phi} \leq |\mathcal{T}|^2\overline{\phi}\cdot C(\varepsilon, T), \tag{B.26}$$

due to the bound on the errors. By the bounds (B.25) and (B.26), we can bound the time derivative of $L(\pi)$ as

$$\frac{d}{dt}L(\pi) < -L(\pi) + \Xi(\lambda, \varepsilon, T) \Leftrightarrow \frac{d}{dt}\left(L(\pi) - \Xi(\lambda, \varepsilon, T)\right) < -L(\pi) + \Xi(\lambda, \varepsilon, T), \tag{B.27}$$

where the constant $\Xi(\lambda, \varepsilon, T) > 0$ is as described in (3.36b). Since we have $V(\pi) = \min(0, L(\pi) - \Xi(\lambda, \varepsilon, T))$, the strict inequality in (B.27), yields that $V(\cdot)$ is a Lyapunov function.

Appendix C

Proof of Technical Results in Chapter 4

The followings are the proofs of Lemma 5 and Lemma 6 used in §4.4.

C.1 Proof of Lemma 5

We focus on the timesteps at which the state s is visited. Let t_k be the time index of the k-th visit to s, and define $w_k := (s, a_{t_k})$ and $\widehat{w}_k := (s, \widehat{a}_{t_k})$. Let μ_k^w, $\widehat{\mu}_k \in \Delta(W)$ be their distributions at stage t_k, conditioned on $\mathcal{F}_{(n)}$.

The probability of transition from the action $\widehat{w}$ to $\widehat{w}'$ depends on $\widehat{w}$ and $Q_{(n)}$. Since $Q_{(n)}$ is $\mathcal{F}_{(n)}$-measurable, $\{\widehat{w}_k\}_k$ form a time homogeneous Markov chain conditioned on $\mathcal{F}_{(n)}$. Let $\widehat{\mathbb{P}}$ denote the transition matrix of this time homogeneous Markov chain. In contrast, $\{w_k\}_k$ is not time-homogeneous since w_{k+1} depends on w_k and on $Q_{t_{k+1}}(s, \cdot)$, which is not $\mathcal{F}_{(n)}$-measurable. Recall that $z_k \in \mathbb{Z}$ denotes the trajectory of state-action pairs realized between the k-th and $(k+1)$-th visits to state s, where $\mathbb{Z}$ is the countable set of all possible such trajectories. By the Q-update (4.8), each estimate $Q_{t_{k+1}}(s, a)$ is

$\sigma(\{w_l, z_l\}_{l\leq k})$-measurable. Hence, the estimate $Q_{(n)}$ (which is $\mathcal{F}_{(n)}$-measurable) and the history $\{w_l, z_l\}_{l<k}$ from the $\big(k_{(n)}-1\big)$-th visitation to state s together determine the transition probability from $w_{k-1} = (s, a_{t_{k-1}})$ to $w_k = (s, a_{t_k})$. We denote these transition probabilities by $\mathbb{P}\big(\cdot \mid \{w_l, z_l\}_{l\leq k}\big)$. This implies that the discrete-time processes $\{\widehat{w}_k\}$ and $\{(w_k, z_k)\}$ are in the framework of Lemma 1. Correspondingly, in the following, we prove the claims that the conditions listed in Lemma 1 hold so that we can invoke the lemma to address the distance between μ_k^w and $\widehat{\mu}_k$.

Claim 1. The processes $\{\widehat{w}_k\}$ and $\{(w_k, z_k)\}$ satisfy Condition (i) in Lemma 1 for $\kappa = |A|$, $\varepsilon = (\epsilon\delta(1-\delta)^{n-1})^{2\kappa}$, and any $w' \in W$.

Proof. Due to the update rule of log-linear learning $\widehat{w} = (s, \widehat{a}) \in W$ can transit to $\widehat{w} = (s, \widehat{a}') \in W$ only if $\widehat{a}$, and $\widehat{a}'$ are

- arbitrarily different at most for one agent's action in coordinated scheme,
- arbitrarily different for every agent's actions in independent scheme.

This yields that the Markov chain $\{\widehat{w}_k\}_{k\geq k_{(n)}}$ is irreducible in both cases. The Markov chain also has self-loops for every extended state.

Since the Markov chain $\{\widehat{w}_k\}$ is irreducible and has self loops at every state, there exists a κ-length path $\widehat{w} = \widehat{h}_0, \ldots, \widehat{h}_\kappa = \widehat{w}'$ for any $(\widehat{w}, \widehat{w}') \in W \times W$ in the fictional scenario, where $\kappa = |A|$. Even though $\{(w_k, z_k)\}$ do not form a homogeneous Markov chain, $\widehat{\mathbb{P}}$ and $\mathbb{P}$ have the same pattern. In other words, given any trajectory $\{w_l, z_l\}_{l<k}$, we have $\mathbb{P}^w(w' \mid, \{w_l, z_l\}_{l\leq k}) > 0$ if and only if $\widehat{\mathbb{P}}(w' \mid w_k) > 0$ since agents updating their actions can take any action with some positive probability irrespective of the Q-function estimate due to the logit response (4.6) and the main and fictional scenarios differ only in terms of the Q-function estimates used. Therefore, there also exists a κ-length path $w = h^0, \ldots, h^\kappa = w'$ for any $(w, w') \in W \times W$ in the main scenario.

In the coordinated scheme, every feasible transition (in both main and fictional scenarios) has probability at least $\frac{\epsilon}{n}$ because exactly one agent updates

with probability $\frac{1}{n}$ and then chooses an action with probability at least ϵ (see (4.57)). These transitions are also valid in the independent scheme, where each such update occurs with probability at least $\epsilon\,\delta\,(1-\delta)^{n-1}$. Note that $\delta\,(1-\delta)^{n-1} < \frac{1}{n}$. Hence, in both settings, any κ-length path has probability at least $\left(\epsilon\,\delta\,(1-\delta)^{n-1}\right)^{\kappa}$. Therefore,

$$\Pr\left(\bigwedge_{l=1}^{\kappa} \widehat{w}_{k+l} = \widehat{h}_l \mid \widehat{a}_{t_k} = \widehat{h}_0\right) \geq (\epsilon\delta(1-\delta)^{n-1})^{\kappa}, \tag{C.1a}$$

$$\Pr\left(\bigwedge_{l=1}^{\kappa} w_{k+l} = h_l \mid (w_k, z_k) = (h_0, z)\right) \geq (\epsilon\delta(1-\delta)^{n-1})^{\kappa} \quad \forall z \in \mathbb{Z}. \tag{C.1b}$$

Given any $w' \in W$ and any pair $(\widehat{w}, w)$ with $\widehat{w} \neq w$, there exist shortest paths of length at most κ from $\widehat{w}$ to w' and from w to w', each changing at most one agent's action at a time. If these paths share a common state, we align them from that point onward, inserting self-loops to ensure both paths have total length κ. If one path is shorter ($\ell < \widehat{\ell}$), we let $\{w_k\}$ follow its ℓ-step path and then do self-loops at w' for $\kappa - \ell$ steps, while $\{\widehat{w}_k\}$ waits in self-loops at $\widehat{w}$ for $\kappa - \widehat{\ell}$ steps and then follows its $\widehat{\ell}$-step path (and vice versa). If $\ell = \widehat{\ell}$, each path still remains of length less than κ, so we add self-loops at w' after traversing the path and add self-loops at $\widehat{w}$ before traversing the path. In all cases, the two processes avoid occupying the same state simultaneously for the first κ steps, thanks to the artificially inserted delays, and each path has probability at least $\left(\epsilon\,\delta\,(1-\delta)^{n-1}\right)^{\kappa}$ by (C.1). This establishes Condition (i) of Lemma 1 and completes the proof of Claim 1.

□

Claim 2. The processes $\{\widehat{w}_k\}$ and $\{(w_k, z_k)\}$ meet Condition (ii) in Lemma 1 with $\lambda_{(n)} = 1 - CHK_{(n)} \times \beta_{(n)}$, for some constant $C > 0$, and sufficiently large n.

Proof. The transition kernels $\mathbb{P}^w(\cdot \mid \{w_l, z_l\}_{l \leq k})$ and $\widehat{\mathbb{P}}(\cdot \mid \widehat{w}_k)$ both depend on agents' softmax responses to the Q-function estimates $Q_{t_{k+1}}$ and $Q_{(n)}$. Since the softmax map (4.6) is $1/\tau$-Lipschitz with respect to $\|\cdot\|_2$ [79, Proposition 4], there

exists $C > 0$ such that

$$\left\| \widehat{\mathbb{P}}(\cdot \mid w_k) - \mathbb{P}^w(\cdot \mid \{w_l, z_l\}_{l \leq k}) \right\|_{\mathrm{TV}} \leq C \|Q_{t_{k+1}} - Q_{(n)}\|_{\max} \leq CHK_{(n)}\beta_{(n)}, \quad \text{(C.2)}$$

for all $k \in [k_{(n)}, k_{(n+1)})$, where $\beta_{(n)} := \max_a\{\beta_{c_{(n)}(s,a)}\}$. Hence we set $\lambda_{(n)} := 1 - C\,H\,K_{(n)}\,\beta_{(n)}$, which stays in $(0, 1]$ for sufficiently large n since $K_{(n)} \times \beta_{(n)}$ is decaying to zero as $n \to \infty$ by Lemma 7 whereas C, and H are time-invariant. Hence, Condition (ii) in Lemma 1 also holds for sufficiently large n. This completes the proof of Claim 2. □

Claims 1, 2, and Lemma 1 yields that

$$\|\mu_k^w - \widehat{\mu}_k\|_1 \leq 2(1-\varepsilon)^{\frac{(k-k_{(n)})}{\kappa}-1} + 2\left(1 - \lambda_{(n)}^{\kappa}\right)\frac{1+\varepsilon}{\varepsilon} \quad \text{(C.3)}$$

for all $k \geq k_{(n)}$. Recall that $\{\widehat{w}_k\}$ is irreducible and aperiodic, so it has a unique stationary distribution $\widehat{\mu}_{(n)}$. if we had $\widehat{\mu}_{k_{(n)}-1} = \widehat{\mu}_{(n)}$, i.e., $\widehat{a}_{k_{(n)}} \sim \widehat{\mu}_{(n)}$, then we would have $\widehat{\mu}_k = \widehat{\mu}_{(n)}$, i.e., $\widehat{w}_k \sim \widehat{\mu}_{(n)}$, for all $k \geq k_{(n)}$. Therefore, (C.3) yields that

$$\|\mu_k^w - \widehat{\mu}_{(n)}\|_1 \leq 2(1-\varepsilon)^{\frac{(k-k_{(n)})}{\kappa}-1} + 2\left(1 - \lambda_{(n)}^{\kappa}\right)\frac{1+\varepsilon}{\varepsilon} \quad \text{(C.4)}$$

for all $k \geq k_{(n)}$ which completes the proof of Lemma 5.

C.2 Proof of Lemma 6

The proof is by construction. First of all, by (4.37), we have

$$\alpha_{k_{(n+1)}-1} \leq \overline{\alpha}_{(n)} = \alpha_{k_{(n)}} \prod_{l=k_{(n)}+1}^{k_{(n+1)}-1} (1-\alpha_l) + \ldots + \alpha_{k_{(n+1)}-1} \leq K_{(n)}\alpha_{k_{(n)}} \quad \text{(C.5)}$$

since α_k's are monotonically non-increasing and take values in $[0, 1]$. As α_k decays to zero, the bounds in (C.5) would imply that $\overline{\alpha}_{(n)}$ also decays to zero if we had $\lim_{n\to\infty} K_{(n)}\alpha_{k_{(n)}} = 0$ which is true if $K_{(n)}$ grows at a sub-linear rate, and $k_{(n)} \geq n$. Some algebra would also show that $\overline{\alpha}_{(n)} = 1 - \prod_{k=k_{(n)}}^{k_{(n+1)}-1}(1-\alpha_k)$, and therefore, $\overline{\alpha}_{(n)} \in [0, 1]$. Next we show that such a realization of $K_{(n)}$ is possible such that $\overline{\alpha}_{(n)}$ satisfies standard conditions.

Define the following:

$$\overline{A}_N \coloneqq \sum_{k=1}^{N} \alpha_k, \qquad \underline{A}_N \coloneqq \sum_{k=N}^{\infty} \alpha_k^2. \tag{C.6}$$

By definition and the assumptions stated in the Lemma 6, $\lim_{N\to\infty} \overline{A}_N = \infty$, and $\lim_{N\to\infty} \underline{A}_N = 0$. Then, for any $m \in \mathbb{N}$,

$$\overline{\nu}(m) \coloneqq \min\{N : \overline{A}_N \geq 2^m, N \in \mathbb{N}\}, \quad \underline{\nu}(m) \coloneqq \min\{N : \underline{A}_N \leq 2^{-m}, N \in \mathbb{N}\} \tag{C.7}$$

are well defined functions. $\overline{\nu}(m)$ and $\underline{\nu}(m)$ are monotonically non-decreasing by definition. Also, define the following:

$$\overline{K}_{(n)} = \max(1, \max\{m : \overline{A}_n \geq 2^m, m \in \mathbb{N}\}), \tag{C.8}$$

$$\underline{K}_{(n)} = \max(1, \max\{m : \underline{A}_n \leq 2^{-m}, m \in \mathbb{N}\}). \tag{C.9}$$

Note that, $\overline{K}_{(n)} = m$ if and only if $\overline{\nu}(m) \leq n < \overline{\nu}(m+1)$, and $\underline{K}_{(n)} = m$ if and only if $\underline{\nu}(m) \leq n < \underline{\nu}(m+1)$. Based on these definitons, we let $K_{(n)} = \min(\overline{K}_{(n)}, \underline{K}_{(n)})$. Notice that, by this definition $K_{(n)}$ grows at a sub-linear rate.

Due to the epoch length having a minimum length of 1, i.e. $K_{(n)} \geq 1$, and $k_{(1)} = 1$, we have $k_{(n)} \geq n$. Morever, since $\alpha_k \in [0, 1)$, we have $\overline{\nu}(m) \geq m$ for all m.

By the left inequality in (C.5), we have

$$\sum_{n=1}^{\infty} \overline{\alpha}_{(n)} \geq \sum_{n=1}^{\infty} \alpha_{k_{(n+1)}-1} \geq \sum_{n=1}^{\infty} \sum_{k=k_{(n+1)}-1}^{k_{(n+2)}-1} \frac{\alpha_k}{K_{(n+1)}} \tag{C.10}$$

$$\geq \sum_{k=k_{(2)}-1}^{\infty} \frac{\alpha_k}{K_{(k+1)}} \tag{C.11}$$

$$\geq \sum_{k=k_{(2)}-1}^{\infty} \frac{\alpha_k}{\overline{K}_{(k+1)}} \tag{C.12}$$

$$\geq \sum_{m=k_{(2)}}^{\infty} \sum_{k=\overline{\nu}(m)}^{\overline{\nu}(m+1)-1} \frac{\alpha_k}{\overline{K}_{(k+1)}} \tag{C.13}$$

$$\geq \sum_{m=k_{(2)}}^{\infty} \frac{1}{m+1} \left(\sum_{k=1}^{\overline{\nu}(m+1)} \alpha_k - \sum_{k=1}^{\overline{\nu}(m)-1} \alpha_k - 1 \right) \tag{C.14}$$

$$\geq \sum_{m=k_{(2)}}^{\infty} \frac{1}{m+1} (2^{m+1} - 2^m - 1) \tag{C.15}$$

$$\geq \sum_{m=k_{(2)}}^{\infty} \frac{2^m - 1}{m+1} = \infty. \tag{C.16}$$

By the right inequality in (C.5), we have

$$\sum_{n=\underline{\nu}(1)}^{\infty} \overline{\alpha}_{(n)}^2 \leq \sum_{n=\underline{\nu}(1)}^{\infty} K_{(n)}^2 \alpha_{k_{(n)}}^2 \leq \sum_{n=\underline{\nu}(1)}^{\infty} \underline{K}_{(n)}^2 \alpha_n^2 \leq \sum_{m=1}^{\infty} \sum_{n=\underline{\nu}(m)}^{\underline{\nu}(m+1)-1} \underline{K}_{(n)}^2 \alpha_n^2 \tag{C.17}$$

$$\leq \sum_{m=1}^{\infty} (m)^2 \sum_{n=\underline{\nu}(m)}^{\underline{\nu}(m+1)-1} \alpha_n^2 \tag{C.18}$$

$$\leq \sum_{m=1}^{\infty} (m)^2 \sum_{n=\underline{\nu}(m)}^{\infty} \alpha_n^2 \tag{C.19}$$

$$\leq \sum_{m=1}^{\infty} \frac{(m)^2}{2^m} \leq \infty. \tag{C.20}$$

For the claim that $K_{(n)}/k_{(n)}$ decaying to zero, we can simply use the definition of $K_{(n)}$. Since $\alpha_k < 1$ for all k, we can say that $\overline{A}_{2^m} < 2^m$. Then, $\overline{K}_{(2^m)} < m$. Since $\{\overline{K}_{(n)}\}_{n \geq 1}$ is a monotonically non-decreasing sequence, $\overline{K}_{(n)} \leq \lceil \log_2(n) \rceil$.

Also, we know that $k_{(n)} \geq n$. Then, the following is true

$$0 \leq \lim_{n\to\infty} \frac{K_{(n)}}{k_{(n)}} \leq \lim_{n\to\infty} \frac{\overline{K}_{(n)}}{n} \leq \lim_{n\to\infty} \frac{\lceil \log_2(n) \rceil}{n} = 0. \tag{C.21}$$

This completes the proof.