Social Learning via Multi-Agent Deep Reinforcement Learning

Master Thesis Presented to the
Department of Economics at the
Rheinische Friedrich-Wilhelms-Universität Bonn

In Partial Fulfillment of the Requirements for the Degree of
Master of Science (M.Sc.)

Supervisor: Prof. Dr. Florian Brandl

Submitted in June, 2025 by:
Ege Can Doğaroğlu
Matriculation Number: 3464688

Abstract

This thesis bridges economic social learning theory and multi-agent reinforcement learning by developing Partially Observable Active Markov Games (POAMGs) and the POLARIS algorithm. Our framework addresses fundamental challenges in modeling adaptive behavior under partial observability, where agents must simultaneously learn about uncertain environments and evolving strategies of others. We extend Active Markov Games (Kim et al., 2022) to partially observable settings, derive theoretical convergence results, and develop a practical algorithm combining Transformer-based belief processing, Graph Neural Networks with attention mechanisms, and Multi-Agent Soft Actor-Critic based reinforcement learning optimization. Applied to strategic experimentation and learning without experimentation scenarios, our approach reveals dynamic role assignment as a robust organizing principle where agents naturally differentiate into complementary roles enhancing collective information processing. We find that free-riding behavior does not lead to uniform inefficiencies—for some agents, performance in larger networks exceeds autarky levels. Our ex-post aggregation methodology addresses computational challenges like catastrophic forgetting while providing systematic approaches for extracting insights from multi-agent learning dynamics. The framework contributes theoretical foundations to multi-agent reinforcement learning and computational tools to economic theory, demonstrating how interdisciplinary approaches can illuminate complex social learning phenomena neither field could address alone.

CONTENTS

1 INTRODUCTION

The study of how individuals and groups learn from each other's actions and experiences has been a central focus in economic theory for decades. From the seminal contributions on information cascades by Banerjee (1992) and Bikhchandani, Hirshleifer, and Welch (1992) to more recent explorations of learning in networks (Acemoglu, Dahleh, Lobel, & Ozdaglar, 2011; Golub & Jackson, 2010), economic research has sought to understand how social interactions shape beliefs, decisions, and collective outcomes. Traditional economic models have provided valuable insights into phenomena such as herding behavior, where individuals rationally ignore their private information to follow the observed actions of predecessors. However, these models often rely on simplified assumptions about agents' behavior and reasoning processes, frequently restricting analysis to one-shot sequential decisions or steady-state equilibria rather than capturing the rich dynamics of repeated interactions where agents continuously adapt their strategies based on others' evolving behaviors.

Meanwhile, the field of multi-agent reinforcement learning (MARL) has experienced remarkable progress in developing algorithms that enable autonomous agents to learn effective strategies in complex, interactive environments. Recent advances in reinforcement learning have demonstrated impressive capabilities in strategic games (Silver et al., 2016), cooperative problem-solving (Baker, 2020), and competitive scenarios (OpenAI, 2019). MARL offers powerful tools for modeling adaptive behavior in non-stationary environments where agents must continuously revise their strategies in response to others' changing behaviors. However, much of this work has focused on engineering objectives rather than modeling realistic human learning processes, limiting its applicability to fundamental questions in economic theory. Additionally, many MARL approaches struggle with the challenge of partial observability—a defining feature of social learning contexts where agents cannot directly observe others' private information or belief states.

This thesis bridges these domains by introducing a novel framework that integrates economic social learning theory with multi-agent reinforcement learning under partial observability. Our approach models social learning as a dynamic process where agents with limited information strategically adapt their behaviors while simultaneously learning about their environment and other agents' strategies. Through the development of Partially Observable Active Markov Games (POAMGs) and the POLARIS algorithm, we demonstrate how computational approaches can complement traditional economic theory while revealing new insights into complex social learning dynamics, particularly the emergence of dynamic role assignment as a fundamental organizing principle of multi-agent learning systems.

The first challenge concerns non-stationarity and strategic adaptation. Traditional economic models of social learning typically assume that agents follow fixed, myopic decision rules

or Bayesian updating procedures that do not fully account for strategic adaptation over time. In real world scenarios, however, when agents repeatedly interact and observe each other's actions, they may adjust their strategies in anticipation of others' learning, creating a complex web of interdependent adaptation that fundamentally alters collective learning dynamics. This non-stationarity—where the effective environment an agent faces changes as other agents learn—represents a central challenge that requires new modeling approaches. Standard MARL algorithms often treat this non-stationarity as a technical obstacle to convergence rather than an intrinsic feature of multi-agent systems that agents should explicitly reason about and strategically exploit.

The second challenge involves partial observability, which pervades social learning contexts. In most real-world settings, agents cannot directly observe others' private information, beliefs, or decision processes. Instead, they must infer these hidden states from observable actions and outcomes. This partial observability compounds the complexity of strategic interaction, as agents must simultaneously learn about the underlying state of the world and about the belief formation processes of others. Economic models have often simplified this challenge through strong assumptions about common knowledge or by focusing on one-shot sequential decisions rather than repeated interactions. Similarly, many MARL approaches assume full observability of the state or treat partial observability as a technical challenge to be addressed through belief state tracking, without fully incorporating its strategic implications.

The third challenge concerns long-term strategic considerations in social learning. Agents often face a tension between immediate payoffs and long-term information generation. They may sometimes choose actions that appear suboptimal in the short term to influence the learning trajectories of others or to generate valuable information for future periods. For instance, in strategic experimentation contexts, agents might incur costs to explore unknown options, knowing that the resulting information will benefit both themselves and others in the future. Capturing these farsighted strategic considerations requires models that explicitly account for how current actions shape the future evolution of others' beliefs and behaviors that are usually hard to solve in a closed-form.

To address these challenges, this thesis makes four interconnected contributions that advance both economic theory and multi-agent reinforcement learning. First, we develop Partially Observable Active Markov Games (POAMGs)—a novel formalism that extends the Active Markov Game framework introduced by Kim et al. (2022) to partially observable settings. Unlike standard reinforcement learning frameworks that treat non-stationarity as a challenge to be mitigated, POAMGs incorporate policy evolution as an integral part of the environment dynamics, allowing agents to reason about and strategically influence this evolution process. POAMGs explicitly model how agents' policies evolve over time based on observations and interactions, while accounting for the fundamental constraints imposed by partial observability.

Second, we provide theoretical analysis of convergence and optimization in POAMGs. We establish conditions for joint convergence of states, beliefs, and policy parameters to unique stochastically stable distributions. We derive policy gradient theorems for average and dis-

counted rewards and extend to continuous-time dynamics. Critically, we reveal how game-theoretic equilibrium concepts relate to active equilibrium when agents account for others' learning.

Third, we introduce POLARIS (Partially Observable Learning with Active Reinforcement In Social environments), combining belief processing through Transformers, variational inference with Graph Neural Networks, and Soft Actor-Critic optimization. Our implementation incorporates GNNs with attention mechanisms capturing network topology and temporal dependencies, enabling sophisticated strategies that account for environmental partial observability and strategic adaptation.

Fourth, we validate our framework through strategic experimentation and learning without experimentation applications, building on models of Bolton and Harris (1999), Keller and Rady (2020), Huang, Strack, and Tamuz (2024), and Brandl (2025). Our analysis demonstrates dynamic role assignment as a robust organizing principle, where agents naturally differentiate into complementary roles enhancing collective information processing, challenging theoretical predictions about free-riding inefficiencies.

A key methodological innovation is our discretization approach for continuous-time economic models, treating Lévy processes through Euler-Maruyama schemes and mapping continuous decisions to discrete action probabilities. We also construct an observed reward function enabling reinforcement learning without direct reward signals and develop specialized transformer loss functions for Lévy process observations.

Our finding that dynamic role assignment emerges naturally suggests social learning systems develop specialized roles enhancing collective information processing. This provides guidance for platform design and institutional arrangements in financial markets, organizational learning, and online networks. By demonstrating MARL's value for classical economic questions while revealing computational challenges, we contribute to both fields.

The remainder of this thesis is organized as follows: Chapter 2 reviews literature on social learning and multi-agent reinforcement learning. Chapter 3 presents foundational social learning models. Chapter 4 introduces our theoretical framework, developing POAMGs and presenting the POLARIS algorithm. Chapter 5 applies our framework to strategic experimentation contexts. Chapter 6 concludes with implications and future research directions.

By integrating perspectives from economic theory and multi-agent reinforcement learning, this thesis enhances understanding of social learning processes and provides new tools for modeling strategic interactions in partially observable environments. The insights may encourage further cross-disciplinary work, advance theoretical knowledge, and inform platform design facilitating efficient collective learning in diverse contexts.

In the next chapter, we provide a detailed literature review, tracing development of social learning models and recent MARL advances. This review establishes the conceptual foundation for our theoretical framework and highlights gaps our approach addresses, setting the stage for formal development of Partially Observable Active Markov Games in Chapter 4.

2 RELATED LITERATURE

This section reviews key literature across economic social learning and multi-agent reinforcement learning (MARL), highlighting how our approach bridges these fields to address their respective limitations.

2.1 SOCIAL LEARNING

The economic study of social learning originated with Banerjee (1992) and Bikhchandani et al. (1992), who formalized how rational agents update beliefs using both private information and public information inferred from predecessors' actions, leading to information cascades where public information outweighs private signals and can result in potentially inefficient herding. While foundational, these models rely on one-shot sequential decisions rather than the repeated interactions that characterize many real-world learning contexts. Smith and Sørensen (2000) extended this work by showing how heterogeneous preferences can lead to confounded learning, where private signals remain relevant despite observing others' actions.

Social learning in network settings expands this framework by examining how network structure influences information flow (Acemoglu et al., 2011; Golub & Jackson, 2010). Two principal approaches have emerged: Bayesian models where agents perform rational inference (Gale & Kariv, 2003; Mossel, Sly, & Tamuz, 2015; Rosenberg, Solan, & Vieille, 2009), and non-Bayesian models (DeMarzo, Vayanos, & Zwiebel, 2003; Golub & Jackson, 2010) like the DeGroot framework (DeGroot, 1974) where agents update beliefs through weighted averaging of neighbors' opinions. A critical insight from this literature is that network topology significantly affects learning outcomes. However, most network models still fall short in capturing how agents strategically adapt to others' evolving learning behaviors over time—a key element of our framework. While early models focused on one-shot decisions, more recent work has explored repeated interactions where agents continuously adapt strategies based on observations. These settings more closely align with our MARL approach and will be the focus of our implementation.

2.1.1 Strategic Experimentation

The strategic experimentation literature examines settings where agents balance exploiting current knowledge against generating new information through exploration (Bolton & Harris, 1999; Keller, Rady, & Cripps, 2005). This creates a dynamic tension between individual incentives to free-ride on others' information production and collective benefits from experimentation. Strategic experimentation represents a fundamental departure from classical social learning

models by explicitly accounting for the intertemporal nature of information acquisition. Unlike cascade models where agents make one-shot decisions in sequence, agents in strategic experimentation scenarios face repeated opportunities to learn and adapt their strategies over time. This dynamic perspective connects directly to the reinforcement learning paradigm, where exploration-exploitation tradeoffs are central (Sutton & Barto, 2018). These tradeoffs create the necessary tension for social influence, even in the absence of informational asymmetry among agents that is often central to social learning models (Gale & Kariv, 2003).

The economic foundations of strategic experimentation were established by Rothschild (1974), who analyzed how a monopolist might experiment with different prices to learn about demand. This concept was extended to multi-agent settings by Bolton and Harris (1999), who developed a framework for analyzing experimentation in teams. Their seminal work revealed that when information is a public good, free-riding incentives can significantly reduce aggregate experimentation below socially optimal levels, creating a classic collective action problem.

Several extensions have explored how different information structures affect experimentation incentives. Keller et al. (2005) introduced exponential bandits, where lump-sum rewards arrive according to a Poisson process, demonstrating how the resolution of uncertainty affects the dynamics of experimentation. Their model showed that "encouragement effects" can arise, where agents experiment more intensively to motivate others to join the exploration effort. Klein and Rady (2011) demonstrated how negatively correlated bandits—where success on one experiment decreases the estimated value of others—can encourage more efficient experimentation patterns.

Keller and Rady (2020) developed a particularly relevant model using average reward criteria rather than discounted objectives. Under this framework, the value of information doesn't decay over time, incentivizing different patterns of exploration. This approach aligns with our POAMG framework's emphasis on long-term strategic adaptation in multi-agent systems. Heidhues, Rady, and Strack (2015) further demonstrated how private observations can restore experimentation incentives that fail under public observations, providing insights into how information asymmetry affects collective learning dynamics.

The strategic teaching phenomenon, where agents take seemingly suboptimal actions to influence others' beliefs, emerges naturally in these contexts. Yamamoto (2019) demonstrated how sophisticated agents might deliberately punish or reward others via continuation payoffs due to invariance property of the payoff sets. Similarly, Halac, Kartik, and Liu (2017) showed how optimal incentive structures might intentionally incentivize agents to experiment, highlighting the importance of mechanism design in collective learning environments.

The literature on information design (Bergemann & Morris, 2019; Kamenica & Gentzkow, 2011) provides complementary insights by examining how information revelation mechanisms affect experimentation decisions. Che and Hörner (2018) showed how over-revealing information can sometimes incentivize more experimentation than full transparency. These insights connect directly to the strategic influence aspects of our POAMG framework, which explicitly models how agents reason about and deliberately influence others' learning trajectories.

The formal mathematical structure of strategic experimentation models is detailed in Chapter 3, Section 3.1, which provides the theoretical foundation for understanding multi-agent exploration-exploitation dynamics.

2.1.2 Learning Without Experimentation

A complementary strand of research examines settings where agents learn without directly observing rewards. Huang et al. (2024) studied how long-lived agents learn in networks through repeated interactions, revealing a fundamental inefficiency: regardless of network size, learning speed remains bounded by a constant dependent only on private signal distributions. Brandl (2025) extended these results by showing that this limitation doesn't apply uniformly to all agents, constructing scenarios where some agents learn faster at others' expense.

Unlike strategic experimentation models where agents receive direct payoff feedback, learning without experimentation captures scenarios where agents must form beliefs based primarily on others' observed actions. This distinction is crucial for modeling many real-world social learning contexts, from financial markets (Avery & Zemsky, 1998) to technology adoption (Ellison & Fudenberg, 1993), where payoffs are delayed, noisy, or unobservable.

The theoretical foundations for this approach draw from both Bayesian and non-Bayesian learning traditions. Gale and Kariv (2003) and Smith and Sørensen (2000) developed early models showing how rational agents might become trapped in information cascades when learning from others' actions. Acemoglu et al. (2011) extended this analysis to network settings, demonstrating how network topology influences the aggregation of dispersed information.

A substantial literature has explored learning rates in networked environments. Bala and Goyal (1998) show neighborhood structure influences optimal action adoption, while Golub and Jackson (2012) demonstrate that homophily slows consensus convergence without being affected by network density. Harel, Mossel, Strack, and Tamuz (2014) quantify information loss when observing others' discrete actions, finding only a fraction of private information transmits, approaching zero in large societies due to "groupthink." Jadbabaie, Molavi, and Tahbaz-Salehi (2013) characterize learning rates through agents' signal structures and eigenvector centralities, showing optimal information allocation depends on its distribution—better information should be placed at central nodes when information structures are comparable, but at peripheral nodes when agents possess unique critical information.

The mechanisms behind these learning barriers stem from information loss in the action quantization process. When continuous beliefs are compressed into discrete actions, information is inherently lost (Smith & Sørensen, 2000). Guarino and Jehiel (2013) characterized this as 'coarse inference' where agents make inferences based only on the aggregate distribution of actions across states rather than on the fine details of how actions depend on specific histories, leading to a loss of information.

Strategic considerations emerge naturally in these settings as agents realize their actions influence the future learning of others. Bhattacharya and Mukherjee (2013) demonstrated how forward-looking agents might distort their actions to manipulate the information revealed to

others. Arieli and Babichenko (2019) showed how agents with private information might strategically time their actions to maximize influence on others' beliefs. These strategic dynamics align closely with the active influence mechanisms in our POAMG framework.

The formal mathematical framework for learning without experimentation, including the social learning barrier and coordination benefit theorems, is presented in Chapter 3, Section 3.2. These theoretical results establish fundamental limits and opportunities in social learning networks that directly inform our computational approach.

2.1.3 Non-Bayesian Learning

While most economic models assume fully rational agents, partial rationality perspectives acknowledge cognitive limitations that affect learning. Models like Jadbabaie, Molavi, Sandroni, and Tahbaz-Salehi (2012)'s hybrid learning rule (combining Bayesian updating with naive averaging) and level-k reasoning (Crawford & Iriberri, 2007; Stahl & Wilson, 1994) provide middle grounds between full rationality and simple heuristics.

Evolutionary Game Theory (EGT) also provides a powerful non-Bayesian framework for modeling multi-agent learning dynamics without assuming full rationality. Instead, EGT examines how strategy distributions evolve through selection processes based on relative performance (Weibull, 1997). This perspective aligns naturally with reinforcement learning's trial-and-error approach. Tuyls, Nowe, Lenaerts, and Manderick (2004) demostrates several important connections between EGT and MARL among which, they relate multi-agent Q-learning to replicator dynamics. This mathematical equivalence provides theoretical insights into MARL convergence and equilibrium properties, helping explain empirical observations in complex social dilemmas (Leibo, Zambaldi, Lanctot, Marecki, & Graepel, 2017).

Bridging theoretical frameworks and computational implementations, MARL offers a powerful methodology for operationalizing social learning models. MARL provides computational tools that can simulate the intricate dynamics of social learning environments while implementing the strategic adaptations that economic theories recognize as essential but frequently find difficult to compute in complex, realistic scenarios.

2.2 Multi-Agent Reinforcement Learning

While economic models offer valuable theoretical insights, they often struggle with computational tractability when modeling repeated strategic interactions with partially rational agents. This is where MARL provides complementary tools that address specific limitations of economic approaches. MARL frameworks enable the simulation of complex multi-agent systems where agents learn optimal policies through trial-and-error interactions with their environment and other agents. Unlike traditional economic models that often require closed-form solutions, MARL can handle high-dimensional state spaces, complex agent interactions, and non-stationary dynamics that emerge when multiple agents learn simultaneously.

The core components of MARL include state representations, action spaces, reward functions, and learning algorithms that enable agents to maximize expected cumulative rewards. These components can be tailored to model various aspects of social learning, including partial observability (through belief states), strategic adaptation (through policy gradient methods), and partial rationality (through constrained optimization). Modern MARL approaches incorporate techniques such as centralized training with decentralized execution, value decomposition, and multi-agent actor-critic methods to address coordination challenges that arise in multi-agent settings. We direct the interested reader to Appendix A for a detailed introduction of reinforcement learning techniques, including formal definitions, algorithms, and theoretical properties.

Economic models typically assume either fully rational agents (in Bayesian frameworks) or overly simplistic learning rules (in behavioral models). MARL offers a middle ground by modeling intelligent agents that learn from experience and adapt over time without requiring full rationality. Ndousse, Eck, Levine, and Jaques (2021) demonstrated that even without explicit programming, reinforcement learning agents can develop sophisticated social learning capabilities that mirror human behavior. This addresses the partial rationality problem by providing computational mechanisms for flexible belief updating based on partial information, learning complex strategies through trial and error, and adapting to non-stationary environments created by other learning agents.

A key limitation of economic social learning models is their difficulty in capturing how agents strategically adapt to others' learning processes. The strategic experimentation literature acknowledges these dynamics but often lacks tractable solutions outside of simplified settings. Jaques et al. (2019) addressed this by introducing Social Influence as a mechanism in MARL, where agents receive additional reward for causally influencing others' actions. This creates a computational framework for modeling strategic teaching and information revelation—key dynamics in the economic models of Bolton and Harris (1999) and Heidhues et al. (2015).

More directly relevant to our approach, Kim et al. (2022) developed Active Markov Games, which explicitly model how agents reason about and influence the policy evolution of other agents. This formalism allows us to capture the strategic adaptation that economic models identify as important but struggle to compute in complex environments.

Despite these advantages, standard MARL approaches have their own limitations when applied to social learning. Most MARL algorithms assume full observability of state information, while social learning inherently involves partial observability of others' private information and beliefs. Many MARL approaches treat non-stationarity as a technical obstacle rather than a strategic feature to be exploited. Additionally, MARL often lacks the theoretical foundations that economic models provide for understanding equilibrium behavior.

Our POAMG framework extends Active Markov Games to partially observable settings specifically to address these limitations. By incorporating policy evolution as an integral part of the environment dynamics while accounting for partial observability, we provide a computational approach that preserves the strategic sophistication of economic models.

2.3 BRIDGING ECONOMIC THEORY AND MARL

While economic social learning and MARL have developed largely in parallel, their complementary strengths suggest significant potential for integration. Economic models provide rigorous theoretical foundations for understanding rational behavior, belief formation, and information aggregation in social contexts. However, these models often face computational limitations when addressing complex strategic interactions, especially when agents have heterogeneous beliefs, partial rationality, or operate in environments with partial observability.

Conversely, MARL offers computational frameworks for modeling adaptive agents in complex, high-dimensional environments. These approaches excel at simulating emergent behaviors and can operate effectively without imposing full rationality assumptions. However, MARL approaches frequently lack the theoretical grounding to interpret equilibrium properties and sometimes overlook the strategic sophistication captured in economic models.

Our research bridges these fields by developing a partially observable active Markov game (POAMG) framework that preserves the strategic considerations central to economic theory while leveraging the computational scalability of MARL. This integration addresses three key challenges:

First, we explicitly incorporate policy evolution dynamics and strategic adaptation as fundamental features rather than technical obstacles. Unlike standard MARL approaches that treat non-stationarity as a problem to overcome, our framework models how sophisticated agents reason about and deliberately influence others' learning trajectories (Kim et al., 2022), similar to the strategic teaching phenomena identified in economic experimentation literature (Yamamoto, 2019).

Second, we incorporate partial observability as an intrinsic characteristic of social learning environments. By modeling belief states and observation functions, our approach captures the information asymmetries and strategic uncertainty that economic models identify as crucial determinants of learning outcomes (Heidhues et al., 2015; Rosenberg et al., 2009).

Third, we account for long-horizon strategic planning where agents optimize not just immediate rewards but also their influence on the future learning dynamics of other agents. This aligns with economic perspectives on forward-looking behavior while remaining computationally tractable through reinforcement learning techniques.

The resulting framework enables more realistic modeling of social learning phenomena that resist analysis under either purely economic or purely computational approaches. It combines economic insights on strategic sophistication with MARL's ability to simulate complex adaptive systems, yielding both theoretical insights and practical algorithms for understanding multi-agent learning in partially observable environments.

Having established this theoretical foundation, we now turn to the formal mathematical development of social learning models that will underpin our computational approach.

3 SOCIAL LEARNING MODELS

This chapter presents the formal mathematical foundations underlying our approach, drawing from two key strands of economic literature: strategic experimentation and learning without experimentation. These models provide the theoretical backbone for understanding multi-agent learning dynamics and inform the design of our partially observable active Markov game framework.

3.1 STRATEGIC EXPERIMENTATION

We formalize strategic experimentation through the model developed by Keller and Rady (2020), which captures the fundamental tension between exploration and exploitation in multi-agent settings. In this framework, n agents face a two-armed bandit problem where they continuously allocate resources between a safe arm with known deterministic payoff $r_{safe} > 0$ and a risky arm whose expected payoff depends on an unknown state $\omega \in \{0, 1, \ldots, n_{\text{states}}\}$ with $n_{\text{states}} \geq 1$. The state ω is drawn at the beginning according to a publicly known prior distribution with full support. The risky arm generates payoffs according to a Lévy process:

$$X_t^i = \alpha_\omega t + \sigma Z_t^i + Y_t^i \tag{3.1}$$

where Z^i is a standard Wiener process, Y^i is a compound Poisson process with Lévy measure ν_ω, and α_ω is the drift rate in state ω. The expected payoff per unit of time is $r_\omega = \alpha_\omega + \lambda_\omega h_\omega$, where λ_ω is the expected number of jumps per unit of time and h_ω is the expected jump size.

At each moment, agent i allocates fraction $a_t^i \in [0,1]$ to the risky arm, yielding instantaneous expected payoff $(1-a_t^i)r_{safe} + a_t^i r_\omega$. Each agent observes their own payoff process, the payoff processes of all other agents, and a background information process B_t, which provides free information about the state:

$$B_t = t b_0 (\beta_\omega t + \sigma_B Z_t^B + Y_t^B) \tag{3.2}$$

where $b_0 > 0$ represents the base informativeness of the background signal process, and the informativeness increases linearly with time t, scaling all components of the Lévy process that generates free information about the state.

Under the strong long-run average criterion, agents maximize:

$$\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T \{(1-a_t^i)r_{safe} + a_t^i r_\omega\} dt\right] \tag{3.3}$$

The unique symmetric Markov perfect equilibrium strategy is characterized as:

$$a^*(b) = \begin{cases} 0 & \text{if } I(b) \leq b_0, \\ \frac{I(b)-b_0}{n-1} & \text{if } b_0 < I(b) < b_0 + n - 1, \\ 1 & \text{if } I(b) \geq b_0 + n - 1. \end{cases} \tag{3.4}$$

where $I(b) = \frac{f(b)-r_{safe}}{r_{safe}-m(b)}$ corresponds to the incentive to experiment, when $m(b) < s$, and $I(b) = \infty$ otherwise, where $f(b)$ is the expected flow payoff under full information and $m(b)$ is the expected flow payoff from the risky arm.

This formulation captures several crucial dynamics. When information is a public good, agents have incentives to rely on others' experimentation rather than bearing the cost themselves, creating a free-rider problem. Simultaneously, agents must balance immediate payoffs against long-term information value, embodying the classic exploration-exploitation tradeoff.

3.2 Learning Without Experimentation

The learning without experimentation framework, analyzed by Brandl (2025), examines how agents learn optimal actions through social observation when direct payoff feedback is unavailable. This model is particularly relevant for understanding information aggregation in social networks. A set of agents $I = \{1, \ldots, n\}$ interact over discrete time periods $t \in \{1, 2, \ldots\}$ in a fixed social network. The state of the world ω is drawn from a finite set S according to a prior distribution with full support and remains fixed throughout.

At each period t, each agent i receives a private signal o_t^i from set O, drawn according to a state-dependent signal distribution ξ_s^i, observes actions taken by neighbors $I^i \subset I$ in previous periods, and chooses an action a_t^i from set A. Signals are conditionally independent time periods but not necessarily across agents or states. All agents share the same utility function $u : S \times A \to \mathbb{R}$, which depends on the state and their own action. For each state s, there is a unique optimal action $a_s = \arg\max_{a \in A} u(s, a)$, and no action is optimal in two different states.

Crucially, agents do not observe their realized utilities, eliminating experimentation motives. Agent i's information set at period t consists of private signals up to period t and neighbor actions up to period $t-1$. A (pure) strategy for agent i is a function π^i that maps information sets to actions $\pi^i : \bigcup_{t=1}^{\infty} (O^t \times A^{I^i \times (t-1)}) \to A$. For any given strategy profile $\boldsymbol{\pi} = (\pi^1, \ldots, \pi^n)$, the learning rate of agent i is defined as:

$$r^i(\boldsymbol{\pi}) = \liminf_{t \to \infty} -\frac{1}{t} \log \mathbb{P}(a_t^i \neq a_\omega | \boldsymbol{\pi}) \tag{3.5}$$

This captures how quickly the probability of making mistakes decays exponentially with time, specifically, if the limit exists, we have $\mathbb{P}(a_t^i \neq a_\omega | \boldsymbol{\pi}) \approx e^{-r^i(\boldsymbol{\pi})t}$. The model establishes two key theoretical results. The social learning barrier theorem demonstrates that regardless of

network size, structure or strategies, some agent's learning rate is bounded:

$$\min_{i \in I} r^i(\boldsymbol{\pi}) \leq r_{bdd} = \min_{s \neq s'} \max_{i \in I} \mathbb{E}_s \left[\log \frac{d\xi_s^i}{d\xi_{s'}^i}(o) \right] \tag{3.6}$$

This bound is strictly larger than the autarky rate. The coordination benefit theorem establishes that, if in addition, the signals are also conditionally independent and identically distributed across agents, for large enough strongly connected networks and any $\varepsilon > 0$, all agents can learn at rates above:

$$\min_{i \in I} r^i(\boldsymbol{\pi}) \geq r_{bdd} - \varepsilon \tag{3.7}$$

where agents can achieve learning rates arbitrarily close to the upper bound r_{bdd} through strategic coordination. These results illuminate fundamental challenges in social learning. The learning barrier theorem shows that information aggregation faces inherent limitations regardless of network structure. Despite these barriers, strategic coordination can improve learning outcomes for all agents, as demonstrated by the coordination benefit theorem.

While sharing fundamental social learning dynamics, these models exhibit key structural differences that illuminate complementary aspects of multi-agent learning. The strategic experimentation framework provides agents with direct payoff feedback, whereas learning without experimentation relies exclusively on observing others' actions, representing distinct information revelation mechanisms. In terms of information structure, strategic experimentation treats information as a public good, while learning without experimentation maintains private agent beliefs. The temporal and action space frameworks also differ: strategic experimentation operates in continuous time with long-run average rewards and continuous action spaces, while learning without experimentation uses discrete time with discounted returns and discrete action spaces. The models also diverge in their rationality assumptions - strategic experimentation assumes full rationality with Bayesian updating, while learning without experimentation accommodates a broader class of learning rules. Finally, learning without experimentation explicitly incorporates network topology considerations, while strategic experimentation focuses on symmetric agent interactions.

These formal models inform several key design choices in our POAMG framework. Both models highlight the importance of maintaining probability distributions over unknown states, suggesting the need for explicit belief state tracking in our MARL agents. The distinction between payoff feedback and action observation suggests designing flexible observation functions that can capture different information revelation mechanisms. Both models demonstrate that optimal behavior requires reasoning about others' learning processes, motivating our active influence mechanisms. Finally, the temporal dynamics in both models suggest that agents must optimize over extended time horizons, informing our policy gradient approaches. These models provide the theoretical motivation for our reinforcement learning approach, ensuring that our MARL framework captures the essential strategic dynamics identified in the economic literature while remaining computationally tractable for complex multi-agent scenarios.

4 POAMG FRAMEWORK

We adapt the framework in Kim et al. (2022) to the partially observable setting and formalize the problem of multi-agent learning as a Partially Observable Active Markov Game with periodic policy updates. Throughout the thesis, we will use the bold convention to denote the collection of joint sets, joint variables and joint functions for $i \in I = \{1, \ldots, n\}$, where $\boldsymbol{X} := \times_{i \in I} X^i$ for set X^i, $\boldsymbol{x} := \{x_1, \ldots, x_n\}$ for variable x^i and $\boldsymbol{G}(\cdot) := \prod_{i \in I} G^i(\cdot)$ for function $G^i(\cdot)$.

4.1 PARTIALLY OBSERVABLE ACTIVE MARKOV GAMES

In social learning environments, agents rarely possess complete information about the underlying state of the world. Instead, they receive private signals—often limited and noisy—that only partially reveal the true state, while also observing the actions of other participants rather than their private information. This fundamental information asymmetry is a cornerstone of economic models of social learning. The challenge of inferring valuable information from others' actions while acknowledging the confounding influence of their own social learning creates rich strategic dynamics that cannot be captured by frameworks assuming full observability. This necessitates extending our theoretical approach to the partially observable setting. Standard approaches to partial observability, such as Partially Observable Markov Decision Processes (POMDPs) and Decentralized POMDPs, address this challenge in single-agent and cooperative multi-agent contexts respectively. However, these frameworks do not adequately capture the strategic nature of policy evolution that characterizes the social learning environments, and neither do they allow for flexible modeling of the reward functions. Building on the Active Markov Game formulation of Kim et al. (2022), we introduce Partially Observable Active Markov Games (POAMGs), which integrate belief state dynamics with evolving policy parameters to model sophisticated social learning dynamics under uncertainty.

Definition 1 (Partially Observable Active Markov Game)**.** *A Partially Observable Active Markov Game is defined as a tuple* $M_n = \langle I, S, \boldsymbol{A}, T, \boldsymbol{O}, \boldsymbol{R}, \boldsymbol{\Theta}, \boldsymbol{U} \rangle$*, where:*

- $I = \{1, \ldots, n\}$ *is the set of n agents;*
- S *is the state space, assumed to be discrete and finite;*
- $\boldsymbol{A} = \times_{i \in I} A^i$ *is the joint action space, where* A^i *is the action space of agent i;*
- $T : S \times \boldsymbol{A} \mapsto \Delta(S)$ *is the Markovian state transition function, with* $T(s'|s, \boldsymbol{a})$ *denoting the probability of transitioning to state* s' *after taking joint action* $\boldsymbol{a}$ *in state s;*

- $\boldsymbol{O} = \times_{i \in I} O^i$ *is the joint observation function, with* $O^i : S \times \Omega^i \mapsto [0,1]$ *denoting the observation function for agent i, where* $O^i(o^i|s)$ *represents the probability of observing* $o^i \in \Omega^i$ *given state s, and* Ω^i *denotes the observation space for agent i;*

- $\boldsymbol{R} = \times_{i \in I} R^i$ *is the joint reward function, with* $R^i : S \times \boldsymbol{A} \mapsto \mathbb{R}$ *denoting the reward function for agent i;*

- $\boldsymbol{\Theta} = \times_{i \in I} \Theta^i$ *is the joint policy parameter space, where* Θ^i *is the policy parameter space for agent i;*

- $\boldsymbol{U} = \times_{i \in I} U^i$ *is the joint Markovian policy update function, with* $U^i : \Theta^i \times \Omega^i \times \boldsymbol{A} \times \mathbb{R} \times \Omega^i \mapsto \Delta(\Theta^i)$ *denoting the policy update function for agent i.*

This formulation extends the Active Markov Game framework by incorporating observation functions that mediate agents' perceptions of the environment state. Next, we define policies of the agents based on their beliefs about the underlying state of the environment.

Definition 2 (Belief-Based Policy)**.** *Under partial observability, each agent i maintains a belief state* $b_t^i \in B^i$ *at time t, representing its probability distribution over states given its observation history. The policy of agent i is defined as a mapping from belief and parameter spaces to distributions over actions:*

$$\pi^i : B^i \times \Theta^i \mapsto \Delta(A^i) \tag{4.1}$$

where $\pi^i(a^i|b^i;\theta^i)$ *represents the probability of agent i taking action* a^i *given belief state* b^i *and policy parameters* θ^i*.*

In partially observable active Markov games, agents form beliefs to infer the underlying state of the environment. Unlike standard partially observable settings, these states evolve through the dynamically changing policies of the agents in addition to the environmental dynamics. The process unfolds as follows: at time step t, each agent i selects an action at state $s_t \in S$ by sampling from its belief-conditioned policy $a_t^i \sim \pi^i(\cdot|b_t^i;\theta_t^i)$. When all agents act collectively through joint action $\boldsymbol{a}_t$, the environment transitions from s_t to s_{t+1} with probability $T(s_{t+1}|s_t,\boldsymbol{a}_t)$. Each agent i then receives an observation o_{t+1}^i with probability $O^i(o_{t+1}^i|s_{t+1})$ and a reward $r_t^i = R^i(s_t,\boldsymbol{a}_t)$, before adjusting its policy parameters via the update function $U^i(\theta_{t+1}^i|\theta_t^i,\tau_t^i)$, where $\tau_t^i = \{o_t^i,\boldsymbol{a_t},r_t^i,o_{t+1}^i\}$ is the trajectory for agent i at time t consisting of the current observation o_t^i, joint action $\boldsymbol{a_t}$, reward received r_t^i, and the next observation o_{t+1}^i. This adaptive cycle continues until non-stationary policies reach convergence. A key insight is that the joint policy update function $\boldsymbol{U}$ depends on a_t^i, which directly impacts state transitions and rewards, thereby enabling agent i to strategically shape future collective policies through its individual decisions. This explicit modeling of strategic influence constitutes the primary advantage of active Markov games over their stationary counterparts.

4.2 CONVERGENCE

A central question in our analysis is whether the joint process of states, beliefs and policy parameters converges to a well-defined stationary distribution, and under what conditions. Following Kim et al. (2022), we establish this connection using the properties of regularly perturbed Markov processes. First, we define the joint process, which operates on the joint space of states, beliefs and policy parameters.

Definition 3 (Joint Process)**.** *In a Partially Observable Active Markov Game, the joint process* $(s_t, \boldsymbol{b}_t, \boldsymbol{\theta}_t)$ *consists of the state* $s_t \in S$*, the joint belief state* $\boldsymbol{b}_t = (b_t^1, \ldots, b_t^n) \in \boldsymbol{B}$ *and joint policy parameters* $\boldsymbol{\theta}_t = (\theta_t^1, \ldots, \theta_t^n) \in \boldsymbol{\Theta}$ *of all agents at time* t*.*

We then make the following assumptions on the subprocesses to ensure the ergodicity of the perturbed joint process.

Assumption 1 (Communicating State Transition)**.** *The state transition* T *is communicating, meaning that for any two states* $s, s' \in S$*, there exists a sequence of joint actions* $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_k$ *and a sequence of states* $s_1, s_2, \ldots, s_{k-1}$ *such that:*

$$T(s_1|s, \boldsymbol{a}_1) > 0, \quad T(s_2|s_1, \boldsymbol{a}_2) > 0, \quad \ldots, \quad T(s'|s_{k-1}, \boldsymbol{a}_k) > 0 \tag{4.2}$$

Assumption 2 (Communicating Belief-State Process)**.** *The belief-state process is communicating, meaning that for any two belief states* $b^i, b'^i \in B^i$*, there exists a sequence of joint actions* $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m$ *and a sequence of observations* $o_1^i, o_2^i, \ldots, o_m^i$ *such that:*

$$\mathbb{P}(b_{t+m}^i = b'^i | b_t^i = b^i, \boldsymbol{a}_t = \boldsymbol{a}_1, o_{t+1}^i = o_1^i, \ldots, \boldsymbol{a}_{t+m-1} = \boldsymbol{a}_m, o_{t+m}^i = o_m^i) > 0 \tag{4.3}$$

Assuming the communicating property of the subprocesses, we can establish convergence of the perturbed joint process to the unique stochastically stable distribution of the unperturbed one, as the perturbation vanishes in the limit.

Theorem 1 (Stochastically Stable Distribution)**.** *For* $\varepsilon > 0$*, define the* ε*-perturbed policy update functions as:*

$$U_i^\varepsilon(\theta_{t+1}^i | \theta_t^i, \tau_t^i) = (1-\varepsilon) U_i(\theta_{t+1}^i | \theta_t^i, \tau_t^i) + \varepsilon \eta_i(\theta_{t+1}^i) \tag{4.4}$$

where η_i *is a baseline distribution over* Θ_i *with full support, and* $\varepsilon > 0$ *is a small constant. Under Assumptions 1 and 2, as* $t \to \infty$ *and* $\varepsilon \to 0$*, the perturbed joint processes defined by these* ε*-perturbed policy update functions converge to the unique stochastically stable distribution* μ^* *of the unperturbed joint process* $(s_t, \boldsymbol{b}_t, \boldsymbol{\theta}_t)$*.*

Proof. See Appendix B.1 for a detailed proof. □

In practice, this ε-perturbation condition is naturally satisfied through entropy regularization in modern multi-agent reinforcement learning algorithms (Kim et al., 2022). Algorithms such

as Multi-Agent Soft Actor-Critic (MASAC) inherently provide the necessary exploration noise through entropy-regularized policy updates, where the entropy term ensures policies maintain positive probability mass across the action space, effectively providing the baseline distribution η_i with full support.

Notably, Assumption 1 is violated in our social learning scenarios where the underlying state of nature remains fixed—agents do not change the fundamental state through their actions. However, as we will see in the next chapter, our results demonstrate convergence even when this assumption is not satisfied, which might be due to our ex-post aggegation mehtodology, where the effective number of states is one per episode. Assumption 2 is more readily satisfied when observation functions provide sufficient information diversity, allowing agents to distinguish between different underlying states through accumulated observations and action sequences.

This convergence result has profound implications for our framework. By establishing the existence of a unique stochastically stable distribution, we provide a solid theoretical foundation for defining optimization objectives and defining equilibrium behavior in partially observable active Markov games. The stochastically stable distribution represents the limiting behavior of the system, independent of initial conditions, allowing us to characterize the long-run outcomes of social learning processes. This property is crucial for developing algorithms that optimize for long-term performance rather than myopic rewards, aligning with our goal of modeling sophisticated social learning dynamics.

4.3 INCENTIVES

We now formalize the optimization objectives for agents in our Partially Observable Active Markov Game framework. In contrast to traditional reinforcement learning settings that typically employ discounted rewards, we first focus on the average reward criterion as our fundamental optimization objective. The average reward formulation provides several compelling advantages for modeling social learning dynamics. As Sutton and Barto (2018) emphasize, this approach is particularly well-suited for continuing tasks without natural episode boundaries—a characteristic that aligns perfectly with the ongoing nature of social learning interactions. While economic theory has predominantly employed discounted reward objectives due to their mathematical tractability and natural correspondence to time preference in utility maximization (Koopmans, 1960; Stokey, Lucas, & Prescott, 1989), the average reward paradigm better captures the strategic considerations in our framework. The key advantage of the average reward approach in our context is its emphasis on the limiting behavior of the multi-agent system. When agents engage in repeated strategic interactions over indefinite horizons, their primary concern becomes the long-run system behavior rather than transient dynamics. Moreover, unlike discounted objectives that can induce myopic behavior depending on the discount factor, the average reward formulation naturally encourages agents to consider the permanent effects of their actions on other agents' learning processes—a crucial aspect for accurately modeling the strategic dimensions of social learning.

Definition 4 (Average Reward Objective under Partial Observability)**.** *Each agent $i \in I$ in a Partially Observable Active Markov Game aims to find policy parameters θ^i and update function U^i that maximize its expected average reward $\rho^i \in \mathbb{R}$ based on the joint beliefs $\boldsymbol{b}$:*

$$\begin{aligned}\max_{\theta^i, U^i} \rho^i(\boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{U}) :&= \max_{\theta^i, U^i} \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T} R^i(s_t, \boldsymbol{a}_t) \,\middle|\, \begin{array}{c} \boldsymbol{b}_0 = \boldsymbol{b}, \boldsymbol{\theta}_0 = \boldsymbol{\theta}, \\ \boldsymbol{a}_t \sim \boldsymbol{\pi}(\cdot|\boldsymbol{b}_t; \boldsymbol{\theta}_t), s_{t+1} \sim T(\cdot|s_t, \boldsymbol{a}_t), \\ \boldsymbol{o}_{t+1} \sim \boldsymbol{O}(\cdot|s_{t+1}), \boldsymbol{\theta}_{t+1} \sim \boldsymbol{U}(\cdot|\boldsymbol{\theta}_t, \boldsymbol{\tau}_t) \end{array}\right] \\ &= \max_{\theta^i, U^i} \sum_{s, \boldsymbol{b}, \boldsymbol{\theta}} b^i(s) \mu^*(s, \boldsymbol{b}, \boldsymbol{\theta}) \sum_{\boldsymbol{a}} \boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{b}; \boldsymbol{\theta}) R^i(s, \boldsymbol{a})\end{aligned} \tag{4.5}$$

This formulation emphasizes that agents aim to maximize their rewards in the limiting behavior of the Markov process, focusing on long-term performance rather than immediate gains. The second equality expresses the objective in terms of the stochastically stable distribution μ^*, weighted by agent's belief, connecting our optimization problem to the theoretical results established earlier. It is worth noting that this definition assumes beliefs, policy parameters, and policy update functions are all publicly observable. Since this assumption rarely holds in practical scenarios, we will implement *variational inference* in our algorithm to enable agents to infer these measures from their partial observations. Next, we derive the policy gradients, which we will be the foundation of our algorithm, allowing the agents to maximize their objectives.

4.4 Belief-Based Policy Gradients

Policy gradient methods constitute a fundamental class of reinforcement learning algorithms that directly optimize policy parameters by ascending the gradient of expected cumulative reward (Sutton & Barto, 2018). Distinguished from value-based approaches, these methods explicitly parametrize the policy function $\pi_\theta(a|s)$ and perform gradient descent on policy parameters θ to maximize expected return. The policy gradient theorem (Sutton, McAllester, Singh, & Mansour, 1999) establishes the theoretical foundation, showing that performance gradients can be estimated without requiring knowledge of underlying state transition dynamics. In partially observable environments, policy gradients adapt by operating on belief states or observation histories (Kaelbling, Littman, & Cassandra, 1998).

Our framework extends these concepts to multi-agent settings with non-stationary policies. The key innovation in our approach is conditioning value functions not just on belief states, but on the joint space of states, belief states, and policy parameters of all agents. This allows us to explicitly model how an agent's actions influence both the environmental dynamics and the learning processes of other agents.

While our theoretical framework proposes maximizing over both policy parameters θ^i and update functions U^i, this joint optimization presents significant computational challenges. As noted in Kim et al. (2022) optimizing over policy update functions essentially constitutes a long-horizon meta-learning problem, which remains computationally intractable for many re-

alistic multi-agent settings. Following their practical approach, we simplify the optimization problem by focusing exclusively on learning optimal fixed-point policies that influence joint policy behavior while using standard update rules:

$$\max_{\theta^i} \rho^i_{\theta^i}(\boldsymbol{b},\boldsymbol{\theta}) := \max_{\theta^i} \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T} R^i(s_t,\boldsymbol{a}_t)\,\middle|\, \begin{array}{c} \boldsymbol{b}_0=\boldsymbol{b},\boldsymbol{\theta}_0=\boldsymbol{\theta}, \\ \boldsymbol{a}_t\sim\boldsymbol{\pi}(\cdot|\boldsymbol{b}_t;\boldsymbol{\theta}_t), s_{t+1}\sim T(\cdot|s_t,\boldsymbol{a}_t), \\ \boldsymbol{o}_{t+1}\sim\boldsymbol{O}(\cdot|s_{t+1}),\boldsymbol{\theta}_{t+1}\sim\boldsymbol{U}(\cdot|\boldsymbol{\theta}_t,\boldsymbol{\tau}_t) \end{array}\right] \tag{4.6}$$

This formulation still preserves the essence of our framework—capturing how agent i's actions influence other agents' learning trajectories—while making the optimization problem tractable. Under the stochastically stable distribution discussed earlier, this simplified objective becomes independent of initial conditions, further supporting the practical viability of our approach. Under the average reward formulation, we define the differential value function for agent i as:

$$v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) = \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T}\left(R^i(s_t,\boldsymbol{a}_t)-\rho^i_{\theta^i}\right)\,\middle|\, \begin{array}{c} s_0=s,\boldsymbol{b}_0=\boldsymbol{b},\boldsymbol{\theta}_0=\boldsymbol{\theta}, \\ \boldsymbol{a}_t\sim\boldsymbol{\pi}(\cdot|\boldsymbol{b}_t;\boldsymbol{\theta}_t), s_{t+1}\sim T(\cdot|s_t,\boldsymbol{a}_t), \\ \boldsymbol{o}_{t+1}\sim\boldsymbol{O}(\cdot|s_{t+1}),\boldsymbol{\theta}_{t+1}\sim\boldsymbol{U}(\cdot|\boldsymbol{\theta}_t,\boldsymbol{\tau}_t) \end{array}\right] \tag{4.7}$$

This function represents the expected total difference between future rewards and the average reward $\rho^i_{\theta^i}$ when starting from state s, belief states $\boldsymbol{b}$, and policy parameters $\boldsymbol{\theta}$. It serves as a crucial component in deriving the policy gradient theorem for our framework.

Theorem 2 (Partially Observable Active Average Reward Policy Gradient Theorem)**.** *The gradient of the active average reward objective with respect to agent i's policy parameters θ^i in a partially observable setting is:*

$$\nabla_{\theta^i} J^i_\pi(\theta^i) = \sum_{s,\boldsymbol{b},\boldsymbol{\theta}} \mu_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) \sum_{a^i} \nabla_{\theta^i}\pi^i(a^i|b^i;\theta^i) \sum_{a^{-i}} \pi^{-i}(a^{-i}|b^{-i};\theta^{-i}) q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{4.8}$$

where the action-value function $q^i_{\theta^i}$ is defined as:

$$\begin{aligned} q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) = &\sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) \\ &\left[R^i(s,\boldsymbol{a}) - \rho^i_{\theta^i} + v^i_{\theta^i}(s', update(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}')\right] \end{aligned} \tag{4.9}$$

with $update(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}')$ representing the belief update function.

Proof. See Appendix B.3 for a detailed proof. □

This theorem extends the policy gradient result in Kim et al. (2022) to partially observable settings by integrating over the joint space of states, beliefs, and policy parameters according to their stochastically stable distribution $\mu_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta})$. The resulting gradient has a form similar

to the standard policy gradient theorem, but with important modifications. First, our policies are conditioned on belief states rather than actual states. Second, the action-value function must account for belief updates and policy parameter evolution. Third, the expectation is taken over the stochastically stable distribution of the joint state space, while also considering the stochasticity of the observations.

Building on the theoretical foundations established for the average reward criterion, we now turn our attention to the discounted return formulation. This alternative objective function is particularly relevant in economic contexts where agents exhibit time preferences, valuing immediate rewards more highly than delayed ones. The following section develops the mathematical framework for optimizing agent behavior under discounted returns, providing a complementary perspective to our earlier analysis. By examining both criteria, we offer a comprehensive theoretical foundation that can accommodate different modeling assumptions about how agents value future consequences of their actions.

4.5 Discounted Returns

While the average reward criterion provides a principled approach for analyzing limiting behaviors in continuing tasks, the discounted return objective remains predominant in economic models of social learning (Bolton & Harris, 1999; Brandl, 2025; Huang et al., 2024; Keller et al., 2005). This formulation incorporates time preference through a discount factor $\gamma \in [0,1)$, giving higher weight to near-term rewards and diminishing importance to those further in the future. The expected effective planning horizon under discounting is approximately $\frac{1}{1-\gamma}$ steps (Mortensen & Talebi, 2025), making it well-suited for scenarios where agents exhibit time preference or where finite planning horizons are appropriate (Frederick, Loewenstein, & O'donoghue, 2002; Sutton & Barto, 2018). We begin by formalizing the discounted return objective in our POAMG framework:

Definition 5 (Discounted Return Objective under Partial Observability)**.** *Each agent $i \in I$ in a Partially Observable Active Markov Game aims to find policy parameters θ^i that maximize its expected discounted return $J^i_{\pi,\gamma}(\theta^i) = v^i_{\theta^i}(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0)$ starting from initial state s_0, joint beliefs $\boldsymbol{b}_0$, and joint policy parameters $\boldsymbol{\theta}_0$:*

$$\max_{\theta^i} v^i_{\theta^i}(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0) := \max_{\theta^i} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^i(s_t, \boldsymbol{a}_t) \,\middle|\, \begin{array}{c} s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0, \\ \boldsymbol{a}_t \sim \boldsymbol{\pi}(\cdot|\boldsymbol{b}_t, \boldsymbol{\theta}_t), s_{t+1} \sim T(\cdot|s_t, \boldsymbol{a}_t), \\ \boldsymbol{o}_{t+1} \sim \boldsymbol{O}(\cdot|s_{t+1}), \boldsymbol{\theta}_{t+1} \sim \boldsymbol{U}(\cdot|\boldsymbol{\theta}_t, \boldsymbol{\tau}_t) \end{array} \right] \tag{4.10}$$

The discounted return objective differs fundamentally from the average reward criterion in its treatment of future consequences. As γ approaches 1, it increasingly resembles the average reward formulation, but with a crucial distinction: even with γ arbitrarily close to 1, the discounted formulation still assigns diminishing weight to distant future rewards. This property has significant implications for strategic behavior in multi-agent settings, particularly in social learning contexts.

4.5.1 Discounted Visitation Measure

To analyze the behavior of agents operating under discounted returns, we introduce the *discounted visitation measure* (also called occupancy measure). This measure represents the normalized expected discounted time spent in each state-belief-policy configuration when following a joint policy (Silver et al., 2014; Sutton et al., 1999).

Definition 6 (Discounted Visitation Measure)**.** *For a Markov process with an initial distribution μ_0 over the joint state-belief-policy space, the discounted visitation measure $d^{\pi}_{\mu_0}$ is defined as:*

$$d^{\pi}_{\mu_0} := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mu_t \tag{4.11}$$

where μ_t is the distribution at time t when starting from μ_0, and $\gamma \in [0,1)$ is the discount factor.

This distribution serves two critical purposes in our framework. First, it provides a well-defined distribution over which expectations can be taken, even in non-stationary environments where policies are constantly changing. Unlike the stochastically stable distribution used in the average reward setting, the discounted distribution automatically exists for any $\gamma < 1$ without requiring additional assumptions on the ergodicity of the underlying processes. Second, it directly links to the optimization objective, allowing us to express the expected discounted return as an expectation of immediate rewards under this distribution.

The discounted visitation measure satisfies important properties that facilitate policy optimization. Most notably, it can be expressed as the unique solution to a functional equation:

$$d^{\pi}_{\mu_0} = (1-\gamma)\mu_0 + \gamma\Psi^* d^{\pi}_{\mu_0} \tag{4.12}$$

where Ψ^* is the adjoint of the transition operator Ψ that describes how probability distributions evolve over one timestep. This relationship can be seen as a fixed-point equation, where $d^{\pi}_{\mu_0}$ is the unique fixed point of the contractive mapping $(1-\gamma)\mu_0 + \gamma\Psi^*(\cdot)$ in the space of finite measures. The existence and uniqueness of this measure is guaranteed for any $\gamma < 1$, providing a solid mathematical foundation for our policy gradient derivations.

4.5.2 Policy Gradient Theorem for Discounted Returns

With the discounted visitation measure established, we now derive the policy gradient theorem that forms the basis for optimization in this framework.

Theorem 3 (Partially Observable Active Discounted Return Policy Gradient Theorem)**.** *The gradient of the discounted return objective $J^i_{\pi,\gamma}(\theta^i) = v^i_{\theta^i}(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0)$ with respect to agent i's*

policy parameters θ^i in a partially observable active Markov game setting can be expressed as:

$$\nabla_{\theta^i} J^i_{\pi,\gamma}(\theta^i) = \frac{1}{1-\gamma} \sum_{s,\boldsymbol{b},\boldsymbol{\theta}} d^{\pi}_{\mu_0}(s,\boldsymbol{b},\boldsymbol{\theta}) \sum_{a^i} \nabla_{\theta^i} \pi^i(a^i|b^i;\theta^i) \sum_{a^{-i}} \pi^{-i}(a^{-i}|\boldsymbol{b}^{-i};\theta^{-i}) q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{4.13}$$

where $d^{\pi}_{\mu_0}(s,\boldsymbol{b},\boldsymbol{\theta})$ is the discounted visitation measure starting from initial distribution μ_0, and $q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a})$ is the action-value function defined as:

$$\begin{aligned} q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) = R^i(s,\boldsymbol{a}) + \gamma \sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \\ \sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) v^i_{\theta^i}(s', update(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \end{aligned} \tag{4.14}$$

with $update(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}')$ representing the belief update function.

Proof. See Appendix C for a detailed proof. □

This theorem connects the policy gradient to expectations with respect to the discounted visitation measure, providing a principled approach to policy optimization in partially observable multi-agent settings. The scaling factor $\frac{1}{1-\gamma}$ accounts for the effective horizon of the discounted objective, with its magnitude increasing as γ approaches 1.

In implementation, the choice between these objectives should be guided by the specific characteristics of the social learning scenario being modeled. The discounted formulation may be appropriate for settings with finite horizons, impatient agents, or where near-term performance is particularly valued. However, for scenarios focused on long-run learning dynamics and asymptotic behavior, the average reward criterion provides a more principled approach by equally valuing rewards across all future time periods.

In practice, computing these gradients exactly is typically infeasible. Instead, sample-based approximations and function approximation techniques using neural networks are employed to estimate the gradient from experience. Nevertheless, in theory, when agents learn to adapt their policies according to these gradients, they eventually converge to an equilibrium concept related to traditional equilibria, which we will introduce in the following section.

4.6 EQUILIBRIUM

Policy gradient optimization naturally leads agents toward stable configurations where no individual agent can unilaterally improve its reward by changing its policy. Given the formalization of our optimization objectives and the policy gradient theorems established above, we now characterize equilibrium concepts in Partially Observable Active Markov Games. By following the gradient directions specified in Theorems 2 and 3, agents can iteratively improve their policies to maximize their respective objectives. This optimization process naturally leads to the concept of Partially Observable Active Equilibrium, which extends the Active Equilibrium from

standard Active Markov Games Kim et al. (2022) to account for partial observability while representing the fixed point of the policy gradient optimization process.

Definition 7 (Partially Observable Active Equilibrium)**.** *A Partially Observable Active Equilibrium is a joint policy parameter* $\boldsymbol{\theta}^* = \{\theta^{i*}, \theta^{-i*}\}$ *with associated joint update function* $\boldsymbol{U}^* = \{U^{i*}, U^{-i*}\}$ *such that for all* $i \in I$*:*

$$J^i_\pi(\boldsymbol{b}, \theta^{i*}, \theta^{-i*}, U^{i*}, U^{-i*}) \geq J^i_\pi(\boldsymbol{b}, \theta^i, \theta^{-i*}, U^i, U^{-i*}) \quad \textit{(average reward case)} \tag{4.15}$$

$$J^i_{\pi,\gamma}(\boldsymbol{b}, \theta^{i*}, \theta^{-i*}, U^{i*}, U^{-i*}) \geq J^i_{\pi,\gamma}(\boldsymbol{b}, \theta^i, \theta^{-i*}, U^i, U^{-i*}) \quad \textit{(discounted reward case)} \tag{4.16}$$

for all $\boldsymbol{b} \in \boldsymbol{B}$, $\theta^i \in \Theta^i$, $U^i \in \mathscr{U}^i$*, where* $\mathscr{U}^i$ *is the space of possible update functions for agent* i*.*

This equilibrium concept captures the idea that rational agents should optimize not just their immediate policies but also their adaptation strategies, taking into account the learning dynamics of the system while operating under partial observability. At equilibrium, no agent can improve its long-term reward (either discounted or average, depending on the formulation) by unilaterally changing either its policy or its update function, overlapping with the definition of the Bayesian Nash equilibrium in non-cooperative games.

Computing partially observable active equilibria exactly is typically intractable due to the complexity of belief spaces and the sophistication of policy update functions. Nevertheless, this equilibrium concept provides a theoretical benchmark against which practical algorithms can be evaluated. In the next section, we discuss the computational challenges inherent in direct implementation of our theoretical framework, before introducing our practical policy optimization method that approximates equilibrium strategies through policy gradient algorithms.

4.7 Challenges with Exact Computation

Theorems 2 and 3 provide mathematically principled approaches for optimizing policies in both average and discounted return settings, but implementing them exactly is computationally infeasible for most real-world applications. In partially observable environments, agents must maintain belief states, whose evolution can be governed by the Bayes' rule as:

$$b^i_{t+1}(s') = \frac{O^i(o^i_{t+1}|s') \sum_{s \in S} T(s'|s, \boldsymbol{a}_t) b^i_t(s)}{\sum_{s'' \in S} O^i(o^i_{t+1}|s'') \sum_{s \in S} T(s''|s, \boldsymbol{a}_t) b^i_t(s)} \tag{4.17}$$

For environments with large or continuous state spaces, representing and updating these distributions becomes prohibitively expensive. Moreover, exact updates require perfect knowledge of observation functions $O^i(o^i|s)$ and transition dynamics T, which may not be readily available to the agents. The complexity compounds in multi-agent settings where agents must reason about others' belief states and policies using Bayesian inference with a function of the form:

$$\mathbb{P}(\boldsymbol{b}_t^{-i}, \boldsymbol{\theta}_t^{-i} | \tau_0^i, \ldots, \tau_t^i) = \mathscr{F}^i_{\text{inference}}(\boldsymbol{b}_{t-1}^{-i}, \boldsymbol{\theta}_{t-1}^{-i}, \tau_t^i) \tag{4.18}$$

that maps from previous beliefs about other agents, their policy parameters, and the current trajectory to updated posterior distributions. This creates nested inference problems—agents maintaining beliefs about others' beliefs—that grow exponentially with the number of agents and the complexity of their policies.

Computing action-value functions adds another layer of intractability, with different formulations for average returns and discounted returns. The triple expectations—over next states, observations, and policy parameters— in equations (4.9) and (4.14) require summing over an exponentially large joint space, where the policy parameter space Θ^i would ideally be modified to be continuous. Furthermore, the policy update functions $\boldsymbol{U}^{-i}$ of other agents are typically unknown, and the recursive nature of the value function creates computational dependencies that quickly become unmanageable.

The policy gradient calculations presented in Equations (4.8) and (4.13) present similarly insurmountable difficulties. Computing the stochastically stable distribution $\mu_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta})$ for average returns or the discounted visitation measure $d^{\pi}_{\mu_0}(s, \boldsymbol{b}, \boldsymbol{\theta})$ for discounted returns requires implementing additional algorithms (Wicks & Greenwald, 2012), while expectations over joint actions create additional combinatorial explosion. These computational barriers—state space explosion, nested belief inference, triple expectations, and distribution computation—collectively render exact computation infeasible except for trivial environments. The curse of dimensionality is especially severe due to the combined complexity of partial observability, multi-agent interactions, and evolving policies.

To address these computational challenges and make our theoretical framework practically applicable, we develop POLARIS, a neural network-based algorithm that approximates the key components of our framework while maintaining its essential properties.

4.8 ALGORITHM: POLARIS

In this section, we present POLARIS (Partially Observable Learning with Active Reinforcement In Social environments), our practical implementation of the theoretical framework. POLARIS extends and improves upon the FURTHER algorithm (Kim et al., 2022) with specialized components for belief modeling, information propagation through networks, and advantage-weighted training. The algorithm uses neural network approximations to overcome the computational challenges outlined in the previous section while preserving the key insights of our theoretical framework. It consists of three integrated modules: a belief processing module, an inference learning module, and a reinforcement learning module, which operate in concert to enable sophisticated social learning in partially observable environments. The architecture integrates several neural network components working in tandem: a Transformer-based belief processor (Vaswani et al., 2017) provides sophisticated temporal pattern recognition through its self-

attention mechanisms; a Temporal Graph Neural Network (GNN) with multi-head attention for network-aware representations; and Multi-Agent Soft Actor-Critic (MASAC) (Kim et al., 2022) with dual Q-networks supporting discrete action spaces and both average and discounted reward formulations. We provide the detailed network architectures in Appendix D.

4.8.1 Belief Processing Module

The belief processing module employs a Transformer architecture to encode the agent's belief state based on partial observations and joint actions:

$$b_t^i = \text{Transformer}(b_{t-1}^i, o_t^i; \psi_{\text{Transformer}}^i) \tag{4.19}$$

The Transformer architecture introduces crucial advantages for social learning scenarios. The self-attention mechanism allows the model to weigh the importance of different parts of the observation history dynamically, enabling it to focus on the most informative signals while downplaying noise or irrelevant information. By processing inputs in parallel rather than sequentially, the Transformer enables more efficient training compared to conventional Recurrent Neural Networks, particularly important when dealing with multiple agents and long interaction histories. Perhaps most importantly for social learning, the Transformer architecture captures relationships between temporally distant observations more effectively, allowing agents to recognize patterns that may emerge over extended periods of social interaction. The Transformer-based belief processor also outputs an explicit belief distribution over possible states:

$$\mathbb{P}(s|b_t^i) = \text{Softmax}(f_{\text{belief\_head}}(b_t^i)) \tag{4.20}$$

This explicit representation provides transparency into decision-making processes and facilitates debugging and analysis of social learning dynamics.

The belief processor is trained using different objectives depending on the signal type. For discrete signals, it uses a standard cross-entropy loss to track the underlying state distribution:

$$J_{\text{transformer,discrete}}^i(\psi_{\text{Transformer}}^i) = \mathbb{E}_{D^i}\left[-\log \mathbb{P}(s_t|b_t^i; \psi_{\text{Transformer}}^i)\right] \tag{4.21}$$

For continuous signals (such as Lévy processes), POLARIS employs a specialized transformer loss function that computes the expected signal distribution based on the current belief state and environment parameters:

$$J_{\text{transformer,continuous}}^i(\psi_{\text{Transformer}}^i) = \mathbb{E}_{D^i}\left[|o_t^i - \hat{o}_t^i(b_t^i, \text{env\_params}; \psi_{\text{Transformer}}^i)|^2\right] \tag{4.22}$$

where D^i is agent i's replay buffer containing stored transitions, $\hat{o}_t^i$ is the predicted continuous signal based on the belief state and environment parameters. This ensures that the belief

processor maintains accurate state estimation capabilities for both discrete and continuous observation types throughout training.

4.8.2 Inference Learning Module

POLARIS implements a Temporal Graph Neural Network (GNN) with multi-head attention for network-aware social learning. This inference module represents the multi-agent system as a dynamic graph where nodes correspond to agents and edges represent observational relationships that are automatically inferred from the neighbor actions tensor. The module supports both discrete and continuous signal processing through separate GNN pathways:

$$\bar{z}_t, \log \sigma_t = \text{TemporalGNN}(o_t^i, \boldsymbol{a}_t, r_t^i, o_{t+1}^i; \psi_{\text{GNN}}^i) \tag{4.23}$$

The network topology is dynamically constructed from observed neighbor actions, where non-zero entries indicate actual neighbors (edges exist) and zero entries indicate non-neighbors (no edges). This approach enables automatic adaptation to various social network structures without requiring explicit network specification.

Key architectural features include dual-pathway processing that employs separate Graph Attention Network (GAT) layers for discrete (one-hot encoded) and continuous signals, enabling flexible handling of different observation types. The system incorporates multi-head attention mechanisms that operate both spatially and temporally, differentially weighing connections based on decision relevance. A temporal memory system maintains past node features and edge indices through a sliding window buffer for temporal attention computation. The architecture leverages multiple GAT layers containing attention heads that learn to focus on the most relevant neighboring agents.

The inference module learns a mapping from observable trajectories to latent variables predictive of other agents' behaviors, optimized using the Evidence Lower Bound (ELBO) objective (Kim et al., 2022):

$$J_{\text{elbo}}^i = \mathbb{E}_{\mathbb{P}(\tau_{0:t}^i), \mathbb{P}(\hat{\boldsymbol{z}}_{1:t}^{-i} | \tau_{0:t-1}^i; \psi_{\text{GNN}}^i)} \Bigg[\sum_{t'=1}^{t} \log \mathbb{P}(\boldsymbol{a}_{t'}^{-i} | o_{t'}^i, \hat{\boldsymbol{z}}_{t'}^{-i}; \psi_{\text{GNN}}^i) \tag{4.24}$$

$$- \alpha_{KL} D_{KL}^{\text{temporal}}(\mathbb{P}(\hat{\boldsymbol{z}}_{t'}^{-i} | \tau_{t'-1}^i; \psi_{\text{GNN}}^i) || \mathbb{P}(\hat{\boldsymbol{z}}_{t'-1}^{-i})) \Bigg] \tag{4.25}$$

The ELBO objective balances reconstruction accuracy (predicting other agents' actions) with temporal consistency through a temporal KL divergence term that encourages smooth evolution of latent representations over time:

$$D_{KL}^{\text{temporal}} = \frac{1}{2} \mathbb{E}_t \left[\log \frac{|\boldsymbol{\Sigma}_{t-1}|}{|\boldsymbol{\Sigma}_t|} - d + \text{tr}(\boldsymbol{\Sigma}_{t-1}^{-1} \boldsymbol{\Sigma}_t) + (\bar{\boldsymbol{z}}_{t-1} - \bar{\boldsymbol{z}}_t)^T \boldsymbol{\Sigma}_{t-1}^{-1} (\bar{\boldsymbol{z}}_{t-1} - \bar{\boldsymbol{z}}_t) \right] \tag{4.26}$$

where $\bar{\mathbf{z}}_t$ and $\mathbf{\Sigma}_t$ are the mean and covariance of the latent distribution at time t. This temporal smoothing prevents abrupt changes in opponent models that could destabilize learning dynamics.

4.8.3 Reinforcement Learning Module

The reinforcement learning module uses the Multi-Agent Soft Actor-Critic (MASAC) (Kim et al., 2022) framework with dual Q-networks and entropy regularization. The implementation supports discrete action spaces:

$$\pi^i(a^i|b^i,\hat{\mathbf{z}}^{-i};\theta^i) = \text{Categorical}(a^i|\text{MLP}_{\text{policy}}(b^i,\hat{\mathbf{z}}^{-i};\theta^i)) \tag{4.27}$$

$$q^i_{\theta^i}(b^i,\hat{\mathbf{z}}^{-i},\boldsymbol{a};\psi^i_\beta) = \text{MLP}_{\text{value}}(b^i,\hat{\mathbf{z}}^{-i},\boldsymbol{a};\psi^i_\beta) \tag{4.28}$$

where MLP stands for *Multi Layer Perceptron*. The MASAC framework's entropy regularization promotes exploration and prevents premature convergence to suboptimal equilibria, crucial in partially observable settings where perfect inference is impossible. The RL module optimizes three objectives:

Value Function Objective POLARIS supports both discounted and average reward formulations. For discounted returns:

$$J^i_q(\psi^i_\beta) = \mathbb{E}_{D^i}\left[(y - q^i_{\theta^i}(b^i,\hat{\mathbf{z}}^{-i},\boldsymbol{a};\psi^i_\beta))^2\right] \tag{4.29}$$

$$y = r^i + \gamma \cdot \min_{\beta=1,2} q^i_{\theta^i}(b^i_{t+1},\hat{\mathbf{z}}^{-i}_{t+1},a'^i;\bar{\psi}^i_\beta) \tag{4.30}$$

For average reward (with automatically estimated average reward rate ρ^i):

$$J^i_q(\psi^i_\beta,\rho^i) = \mathbb{E}_{D^i}\left[(y - q^i_{\theta^i}(b^i,\hat{\mathbf{z}}^{-i},\boldsymbol{a};\psi^i_\beta))^2\right] \tag{4.31}$$

$$y = r^i - \rho^i + \min_{\beta=1,2} q^i_{\theta^i}(b^i_{t+1},\hat{\mathbf{z}}^{-i}_{t+1},a'^i;\bar{\psi}^i_\beta) \tag{4.32}$$

Policy Objective Maximizes expected value plus entropy:

$$J^i_\pi(\theta^i) = \mathbb{E}_{D^i}\left[\mathbb{E}_{a^i\sim\pi^i}\left[\min_{\beta=1,2} q^i_{\theta^i}(b^i,\hat{\mathbf{z}}^{-i},a^i,\boldsymbol{a}^{-i};\psi^i_\beta) + \alpha_e H(\pi^i(\cdot|b^i,\hat{\mathbf{z}}^{-i};\theta^i))\right]\right] \tag{4.33}$$

4.8.4 Training Process

The training process interleaves updates across all modules within each iteration. Algorithm 1 outlines the complete POLARIS training procedure.

Key implementation details include sequential batch sampling, where the replay buffer maintains sequential trajectories and samples sequences rather than individual transitions, preserving temporal dependencies crucial for transformer and GNN processing. The system em-

Algorithm 1 POLARIS Training Algorithm for Single Agent

Initialize belief processor, inference GNN, policy and Q-networks with parameters $\psi^i_{\text{Transformer}}$, ψ^i_{GNN}, θ^i, ψ^i_β
Initialize replay buffer D^i, initial belief b^i_0, and latent state $\hat{\boldsymbol{z}}^{-i}_0$
Initialize target networks
for each step t **do**
    Observe signal o^i_t and neighbor actions $\boldsymbol{a}^{-i}_{t-1}$
    Update belief: $b^i_t, \mathbb{P}(s|b^i_t) = \text{Transformer}(b^i_{t-1}, o^i_t, \boldsymbol{a}_{t-1}; \psi^i_{\text{Transformer}})$
    Construct graph and infer latent: $\hat{\boldsymbol{z}}^{-i}_t = \text{TemporalGNN}(o^i_t, \boldsymbol{a}^{-i}_t, r^i_{t-1}, o^i_{t-1}; \psi^i_{\text{GNN}})$
    Select action: $a^i_t \sim \text{Categorical}(\pi^i(\cdot|b^i_t, \hat{\boldsymbol{z}}^{-i}_t; \theta^i))$
    Execute action, observe reward r^i_t and next signal o^i_{t+1}
    Store $(o^i_t, b^i_t, \hat{\boldsymbol{z}}^{-i}_t, a^i_t, r^i_t, o^i_{t+1}, \boldsymbol{a}^{-i}_t)$ in replay buffer D^i
    if update step **then**
        Sample sequential batch B from replay buffer D^i
        Update inference GNN: $\psi^i_{\text{GNN}} \leftarrow \psi^i_{\text{GNN}} - \alpha_\psi \nabla_{\psi^i_{\text{GNN}}} J^i_{\text{elbo}}(B)$
        Update dual Q-networks: $\psi^i_\beta \leftarrow \psi^i_\beta - \alpha_\psi \nabla_{\psi^i_\beta} J^i_q(B)$
        Update policy: $\theta^i \leftarrow \theta^i - \alpha_\theta \nabla_{\theta^i} J^i_\pi(B)$
        Update belief processor: $\psi^i_{\text{Transformer}} \leftarrow \psi^i_{\text{Transformer}} - \alpha_\psi \nabla_{\psi^i_{\text{Transformer}}} J^i_{\text{transformer}}(B)$
        Update target networks: $\bar{\psi}^i_\beta \leftarrow \tau \psi^i_\beta + (1-\tau)\bar{\psi}^i_\beta$
    end if
end for

ploys dual Q-network updates with both Q-networks updated simultaneously using minimum value selection to prevent overestimation bias. All networks utilize gradient clipping with gradient norm clipping to maintain training stability. Target networks implement soft updates using exponential moving averages for stable learning.

In the next chapter, we demonstrate the practical application of our POAMG framework and the POLARIS algorithm to canonical social learning scenarios from economic theory. The versatility of our approach is particularly evident in its ability to handle both average and discounted reward formulations. This capability proves essential when implementing strategic experimentation and learning without experimentation models with discrete action spaces. Through these implementations, we validate theoretical predictions from economic models while uncovering new insights about strategic adaptation in partially observable multi-agent systems that would be difficult to obtain through purely analytical approaches. The experimental results not only demonstrate the effectiveness of our theoretical framework but also highlight how computational methods can complement and extend traditional economic analysis of social learning phenomena.

5 EVALUATION

Having established the theoretical foundation of Partially Observable Active Markov Games, we now turn to the empirical validation of our framework through concrete social learning scenarios. Our implementation of the POLARIS algorithm addresses two canonical problems in economic theory: strategic experimentation and learning without experimentation. For simplicity, we consider binary state environments, but the framework can be extended to more complex state spaces. While our computational approach successfully bridges theoretical economic models with reinforcement learning methods, it also reveals fundamental tensions between theoretical ideals and computational realities.[1]

5.1 METHODOLOGICAL FRAMEWORK AND CHALLENGES

Equilibrium concepts —both in economics and in our POAMG framework— envision agents developing strategies that perform optimally across all possible states of the world. This *ex-ante* perspective requires agents to maintain flexibility and develop sophisticated mixed approaches that can respond appropriately regardless of which state is ultimately realized. Agents should have conditional strategies of the form "if state = A, do X; if state = B, do Y," accounting for uncertainty about the true environmental state.

However, our computational implementation reveals a fundamental departure from this theoretical ideal. Reinforcement learning agents converge to policies that are optimal for specific realized states—adopting an *ex-post* optimization perspective rather than the *ex-ante* conditional strategies that the theory prescribes. This divergence creates a critical incompatibility that proves particularly problematic in the fixed-state structure of social learning environments.

The root of this incompatibility lies in the fundamentally different nature of optimal behavior required across states in social learning settings. In binary state environments, the optimal strategies are not merely different—they are diametrically opposed. Consider the stark contrasts: in strategic experimentation, optimal behavior demands complete allocation to the risky arm when it is good, but complete avoidance when it performs poorly. Similarly, in learning without experimentation, the optimal action is unambiguously action A when the true state is A, but switches entirely to action B when the state is B. These state-dependent strategies represent mutually exclusive policy objectives that cannot be simultaneously optimized within a single fixed-state environment.

This structural constraint creates a learning paradox. Since agents experience only one state realization per episode, they are systematically prevented from developing the flexible, belief-

[1] Code available at `https://github.com/ecdogaroglu/polaris`.

contingent policies that equilibrium concepts require. Rather than learning conditional strategies of the form "if state A, then do X; if state B, then do Y," agents face fundamentally conflicting learning objectives across episodes. They must learn to "do X" during state A episodes, only to encounter the contradictory imperative to "do Y" in subsequent state B episodes. This fundamental incompatibility between ex-ante and ex-post optimal behavior represents a core challenge in evaluating the POAMG framework computationally.

Recognizing this limitation, one might naturally consider training agents across different environmental configurations, then evaluating their behavior under the full spectrum of possible conditions. Unfortunately, this approach consistently fails due to a fundamental constraint of neural network learning. When network parameters are continuously updated across episodes with fixed states, action-value functions inevitably adapt to the most recent state, yielding policies that are only optimal for the last state encountered. This phenomenon—known as *catastrophic forgetting*—fundamentally constrains our ability to conduct ex ante analysis, as it prevents the retention of knowledge about optimal behavior across different states.

5.1.1 Catastrophic Forgetting in Fixed-State Social Learning

Catastrophic forgetting represents a fundamental limitation in neural network learning where the acquisition of new knowledge systematically interferes with previously learned information (van de Ven, Soures, & Kudithipudi, 2024). This phenomenon occurs when neural networks, optimized through gradient descent, overwrite previously learned weights and representations as they adapt to new tasks or environments. While various techniques have been developed to mitigate this issue in continual learning scenarios—including regularization methods, memory replay systems, and architectural approaches—these solutions typically assume some degree of compatibility or transferability between learning objectives (Gong, Yan, Liu, van den Hengel, & Shi, 2022; Kirkpatrick et al., 2017; Rusu et al., 2022; Zenke, Poole, & Ganguli, 2017).

In our social learning environments, this compatibility assumption fails completely. When the optimal policy in state A requires action X while the optimal policy in state B demands the opposite action Y, no meaningful knowledge transfer is possible between episodes with different state realizations. The literature on negative transfer confirms that attempting to force knowledge sharing between fundamentally incompatible tasks degrades performance rather than enhancing it (Wang et al., 2021). Traditional continual learning mechanisms that aim to "consolidate" or "protect" previous learning become counterproductive when such protection would directly undermine the acquisition of conflicting but equally necessary knowledge.

These fundamental constraints force us to reset agent parameters at the beginning of each episode rather than allowing continuous learning across state realizations—a methodological compromise that undermines our ability to directly test the theoretical equilibrium conditions. This parameter reset prevents agents from developing the ex ante conditional strategies that both economic theory and our POAMG framework predict, instead forcing them to learn state-specific policies that cannot generalize across different environmental realizations. Additionally, this methodological compromise prevents the ability to isolate belief formation

mechnanisms from network parameter updates, particularly relevant for the analysis of learning rates. Given these constraints, our experimental methodology necessarily shifts focus to within-episode learning dynamics and cross-episode pattern analysis. This indirect approach still allows us to characterize important features of multi-agent social learning.

5.1.2 Ex-Post Aggregation Methodology

Recognizing that catastrophic forgetting prevents direct measurement of equilibrium strategies, we implement an ex-post aggregation methodology as a workaround to extract off-equilibrium learning patterns. While this approach cannot resolve the fundamental limitations discussed above, it provides a systematic framework for analyzing the learning patterns across different episodes.

Stage 1: Within-Episode Role Identification For each episode e, we identify agents that exhibit extreme performance characteristics. In strategic experimentation, we track cumulative allocation patterns:

$$i^e_{\text{high}} = \arg\max_i \sum_{t=1}^{T} a^i_{t,e}, \quad i^e_{\text{low}} = \arg\min_i \sum_{t=1}^{T} a^i_{t,e} \tag{5.1}$$

In learning without experimentation, we identify the fastest and slowest learning agents by fitting learning rates:

$$i^e_{\text{fast}} = \arg\max_i r^i_e, \quad i^e_{\text{slow}} = \arg\min_i r^i_e \tag{5.2}$$

where r^i_e is agent i's learning rate in episode e.

Stage 2: Cross-Episode Pattern Aggregation We then aggregate the performance of identified extreme agents across episodes to examine systematic patterns:

$$\bar{P}^{\text{extreme}}_t = \frac{1}{n_{\text{episodes}}} \sum_{e=1}^{n_{\text{episodes}}} P^{i^e_{\text{extreme}}}_{t,e} \tag{5.3}$$

where $P^i_{t,e}$ represents the relevant performance metric (cumulative allocation or learning accuracy) for agent i at time t in episode e.

This methodology provides a limited window into multi-agent learning patterns, allowing us to examine whether observed disparities represent structural features that persist across episodes rather than artifacts of specific agent characteristics.

5.2 STRATEGIC EXPERIMENTATION

We reformulate the strategic experimentation model of Keller and Rady (2020) as a Partially Observable Active Markov Game. This framework analyzes undiscounted continuous-time games where a number of symmetric players act non-cooperatively, trying to learn an unknown

state of the world governing the risky arm's expected payoff. The state space $S = \{0, 1\}$ represents the binary states of the world (bad and good states), with a deterministic transition function since the underlying state remains constant throughout the interaction. Each agent's action space in the theoretical model is $A^i = [0, 1]$ representing the fraction of resources allocated to the risky arm at each decision point.

While the original model operates in continuous time with Lévy processes, our POAMG implementation discretizes time using the Euler-Maruyama scheme (Kloeden & Platen, 1992; Platen, 1999) for both the background signal and the agents' payoff processes. However, to enable reinforcement learning with discrete action spaces, we implement a practical discretization where agents choose between two actions: "allocate to risky arm" or "allocate to safe arm". The probability of selecting the risky arm serves as the continuous allocation parameter, allowing us to recover the continuous allocation behavior while maintaining compatibility with discrete policy networks.

The agents receive a public signal produced by the background process, which leads to the observation function:

$$O^i(o|s) = \mathbb{P}(o_t|B_{t-1:t}) \tag{5.4}$$

where o_t represents the observation at discrete time step t and $B_{t-1:t}$ represents the signal increment in discrete time. To address the time-dependent nature of Lévy processes while maintaining compatibility with our POAMG framework, we formulate the reward function for each agent i as:

$$R^i(s, a^i) = (1 - a^i) r_{safe} + a^i \frac{X^i_{t-1:t}}{\Delta t} \tag{5.5}$$

where Δt is the discretization time step,

$$X^i_{t-1:t} = \alpha_s \Delta t + \sigma(W^i_t - W^i_{t-1}) + \Delta Y^i_t, \tag{5.6}$$

$(W^i_t - W^i_{t-1}) \sim \mathcal{N}(0, \Delta t)$, and ΔY^i_t is the increment of the compound Poisson processes over the interval $[t-1, t]$. This normalization converts accumulated rewards to instantaneous reward rates, preserving the incentive structure of the original model. A comprehensive mathematical treatment of this discretization approach, including convergence properties and the preservation of strategic incentives, is provided in Appendix D.4.

Each agent observes the increments in the public background signal, their own payoff process (dependent on their allocation a^i to the risky arm and the true state ω), and the allocations together with the rewards of all other agents. Given continuous signals, POLARIS agents use the specialized transformer loss function that computes the mean squared error of the expected signal based on the current belief state and environment parameters, enabling principled belief updating in response to Lévy process observations.

5.2.1 Experimental Setup

We implement comprehensive experiments using specialized experimental frameworks for both detailed single-configuration analysis and comparative analysis across different group sizes. Our experimental protocol examines group sizes of $n \in \{1,2,4,8\}$ agents; safe payoff of $r_{\text{safe}} = 0.5$; drift rates of $[0,1]$ for bad and good states respectively; jump rates of $[0,0.1]$ with jump sizes of $[1.0,1.0]$; background informativeness of $b_0 = 0.001$; and time discretization step of $\Delta t = 1.0$. We use average return objective for the policy network optimization.

Following our ex-post aggregation methodology, the POLARIS implementation tracks: (1) Convergence of individual agent allocation strategies to state-dependent optimal actions within episodes; (2) Belief fluctuations over time to assess the impact of neighbor action observation on learning; (3) Cumulative allocation disparities between highest and lowest allocators as a function of group size; and (4) Episode-wise analysis of allocation patterns across different true states (good vs. bad). For multiple configurations, we follow the ex-post aggregation methodology in Section 5.1.2 to track the learning dynamics of the extreme agents in each episode.

5.2.2 Results and Analysis

Our experimental results reveal sophisticated strategic learning dynamics while highlighting key challenges in measuring theoretical equilibrium convergence in reinforcement learning settings. We examine three primary aspects of the learning behavior that provide insights into multi-agent strategic experimentation.

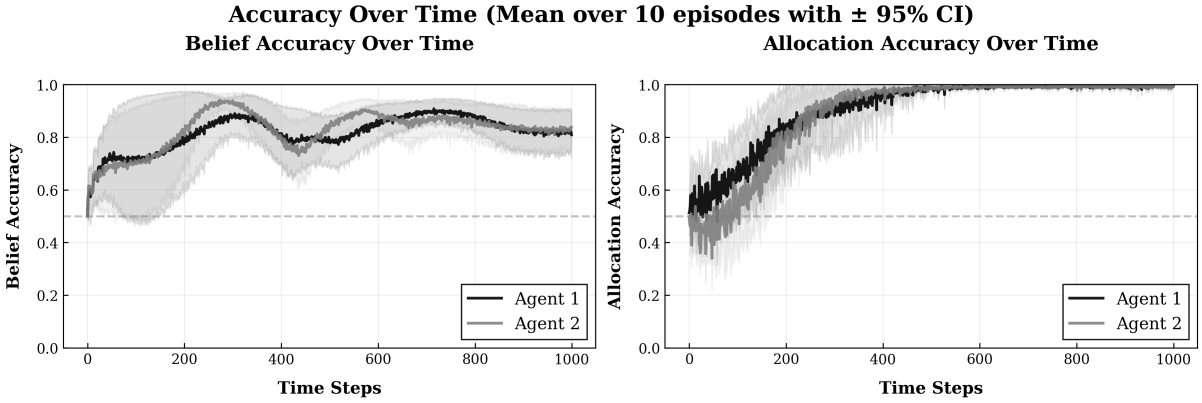

Figure 5.1: Learning dynamics in the strategic experimentation model.

Strategic Action Convergence POLARIS agents successfully learn to converge toward state-dependent optimal allocation strategies. In good states, agents increase their allocation to the risky arm over time, while in bad states, they learn to favor the safe alternative (Figure 5.1). This convergence demonstrates that the discretized action representation effectively captures the continuous allocation dynamics of the theoretical model, enabling agents to discover appropriate exploration-exploitation strategies through reinforcement learning.

Belief Dynamics and Social Learning Effects Figure 5.1 reveals fluctuating belief patterns alongside stable allocation policies. While rewards and signals represent nearly identical stochastic processes, POLARIS agents leverage social information through their inference module to maintain strategic coherence beyond individual signal observation. The observed belief variability reflects ongoing social learning as agents process noisy private signals, yet their allocation policies demonstrate robust convergence toward optimal state-specific strategies despite this underlying uncertainty.

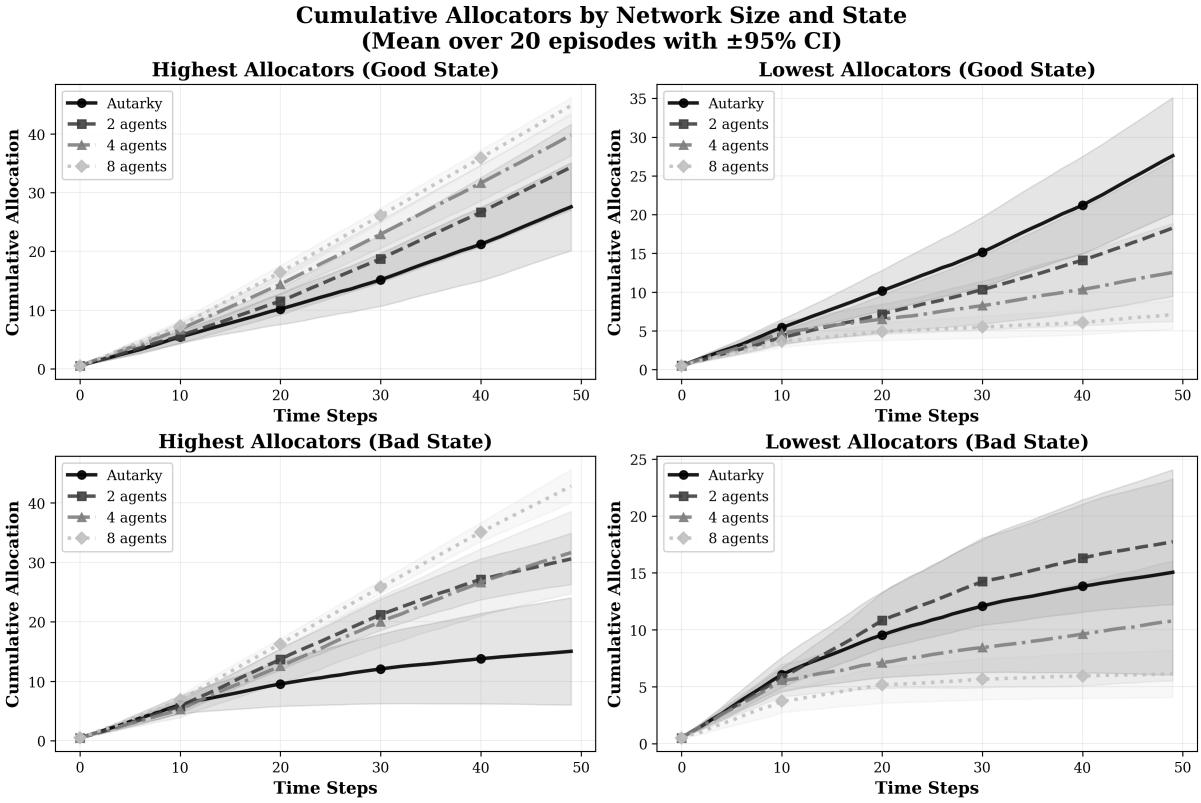

Figure 5.2: Cumulative allocators by network size and state.

Dynamic Role Assignment and Allocation Disparities Figure 5.2 demonstrates a systematic widening of the gap between highest and lowest cumulative allocators as group size increases. This pattern reveals the emergence of dynamic role assignment in multi-agent learning, where agents naturally differentiate into information generators and information exploiters within each episode. In strategic experimentation, this manifests as some agents reducing individual experimentation when benefiting from others' information generation (free-riding effects), while enabling top experimenters to allocate more aggressively when supported by group learning (encouragement effects). Interestingly, despite the fact that the presence of free-riding behavior is typically associated with inefficiency in economic literature, our results show that cumulative allocation in larger networks substantially exceeds autarky levels in both good and bad states, suggesting a more nuanced effiency outcome.

Implications and Validation These results demonstrate that our POAMG framework captures essential features of strategic experimentation dynamics, including state-dependent learning, social information effects, and group size impacts on allocation behavior. The observed patterns of strategic adaptation, belief fluctuations, and allocation disparities provide computational validation for key theoretical insights about multi-agent experimentation and information sharing, even though direct equilibrium convergence measurement is precluded by the catastrophic forgetting constraint discussed in our methodological framework.

5.3 Learning Without Experimentation

We reformulate the learning without experimentation model of Brandl (2025) as a Partially Observable Active Markov Game. In this formulation, the state space $S = \{s^1, s^2\}$ represents the binary states of the world, with a deterministic transition function since the state remains constant. Each agent's action space $A^i = \{a^1, a^2\}$ corresponds to the two potential actions, where a^j is the unique optimal action in state s^j. In addition to observing their neighbors' actions, each agent receives a private signal $o_t^i \in \Omega^i = S$ drawn from distribution:

$$O^i(o|s) = \begin{cases} q & \text{if } s = o \\ 1-q & \text{otherwise} \end{cases}$$

where $q > 1/2$ is the signal accuracy. Since agents don't observe real rewards in this model, we construct an observed reward function that facilitates learning:

$$R^i(o_t^i, a_t^i) = \frac{q \cdot \mathbb{I}_{\{a_t^i = \varphi(o_t^i)\}} - (1-q) \cdot \mathbb{I}_{\{a_t^i \neq \varphi(o_t^i)\}}}{2q-1}$$

where φ maps states to their corresponding correct actions.[2] This construction ensures the expected reward matches the true utility function in expectation[3]:

$$u(s,a) := \mathbb{I}_{\{a=\varphi(s)\}} = \mathbb{E}_{o\sim O^i(\cdot|s)}[R^i(s,a)].$$

In this setup, agents are rewarded based on their posterior signal realizations, maintaining compatibility with the original model's incentive structure while providing a learning signal for POLARIS.

5.3.1 Experimental Setup

We implement comprehensive experiments to validate these theoretical predictions and explore the learning dynamics under various network structures. Our experimental protocol examines network sizes of $n \in \{1, 2, 4, 8\}$ agents; network topologies including complete, ring, star, and

[2] Formally, the mapping $\varphi : S \to A$ is a bijective function defined as $\varphi(s^j) = a^j$ for $j \in \{1,2\}$, associating each binary state with the action having the same index.

[3] $u(s,a) = \mathbb{I}_{\{a=\varphi(s)\}} = q \cdot \frac{q\cdot\mathbb{I}_{\{a=\varphi(o)\}} - (1-q)\cdot\mathbb{I}_{\{a\neq\varphi(o)\}}}{2q-1} + (1-q) \cdot \frac{q\cdot\mathbb{I}_{\{a\neq\varphi(o)\}} - (1-q)\cdot\mathbb{I}_{\{a=\varphi(o)\}}}{2q-1} = \mathbb{E}_{o\sim O^i(\cdot|s)}[R^i(o,a)]$.

random networks (with density 0.5); signal accuracy of $q = 0.75$; discount factor of $\gamma = 0.99$; and learning metrics consisting of empirical learning rates extracted via log-linear regression. We use discounted return objective for the policy network optimization.

Following our ex-post aggregation methodology described in Section 5.1.2, we track each agent's incorrect action probability as the probability assignment of the policy network to the incorrect action over time and estimate individual learning rates by fitting log-linear regressions to identify the fastest and slowest learners within each episode. We then aggregate the performance of these extreme learners across episodes to examine structural features of multi-agent learning dynamics.

5.3.2 Results and Analysis

Our analysis reveals three key phenomena that validate and extend the theoretical predictions: the emergence of learning barriers, the realization of coordination benefits, and the dynamic formation of information revelation roles.

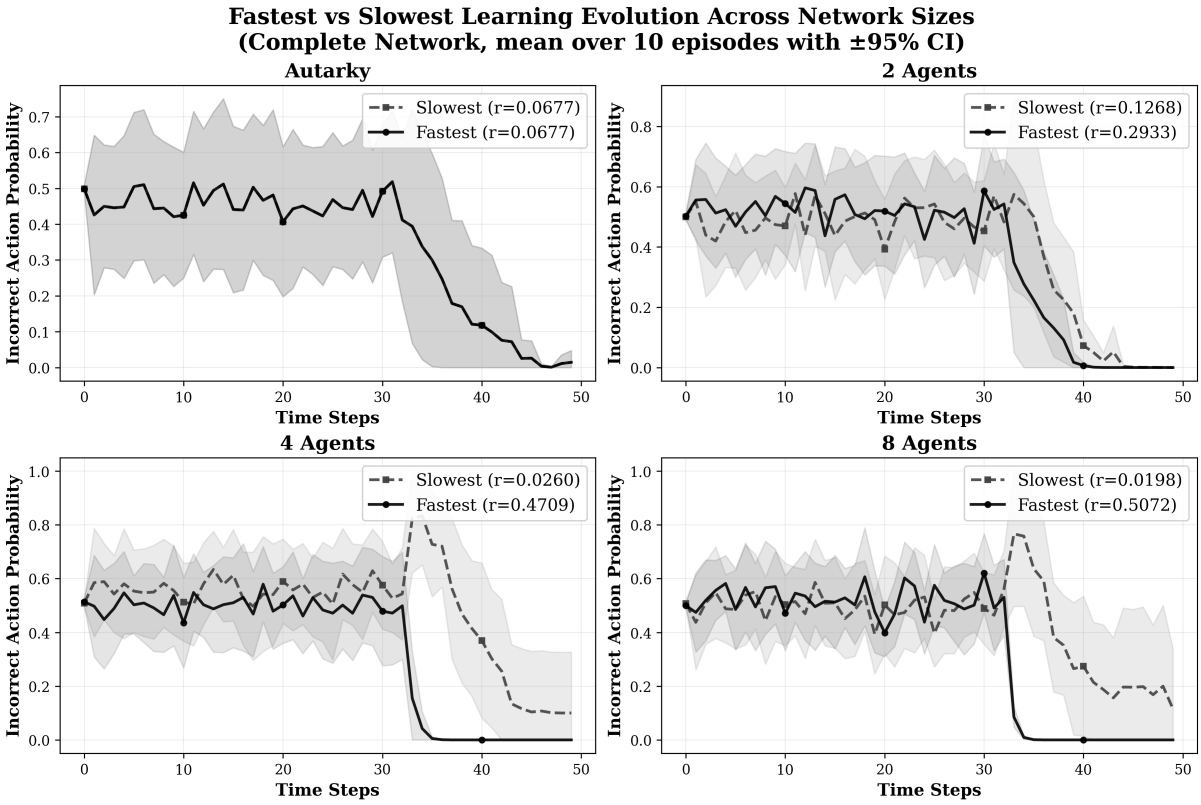

Figure 5.3: Fastest vs slowest agent learning trajectories across network sizes. Learning rates represent the mean trajectories.

Learning Barriers In networks with $n \in \{4, 8\}$, we observe that the slowest-learning agents' learning rates are approximately 0.02 in both network sizes (Figure 5.3). Additionally, Figure 5.4 shows that across different network topologies, the slowest agents consistently achieve rates around or below 0.02. The decreasing trend and the consistent rate across both different network

sizes and topologies provide suggestive evidence for the theoretical learning barrier characterized by the bound r_{bdd}, which predicts that slowest agents' learning rates should stabilize at a fixed value regardless of network size or structure. However, computational constraints prevent us from testing sufficiently large networks to definitively confirm this barrier effect.

Coordination Benefits The theoretical framework predicts that for sufficiently large networks, the slowest-learning agents' rates should approach the r_{bdd} bound, which exceeds autarky performance. Despite being limited by the computational constraints, we observe coordination benefits manifesting through dramatic improvements in fastest agent learning rates across all feasible network configurations. This creates a bifurcated outcome where some agents achieve performance substantially exceeding single-agent capabilities, albeit at the expense of others. The systematic nature of this enhancement indicates that coordination benefits represent robust features of multi-agent social learning systems.

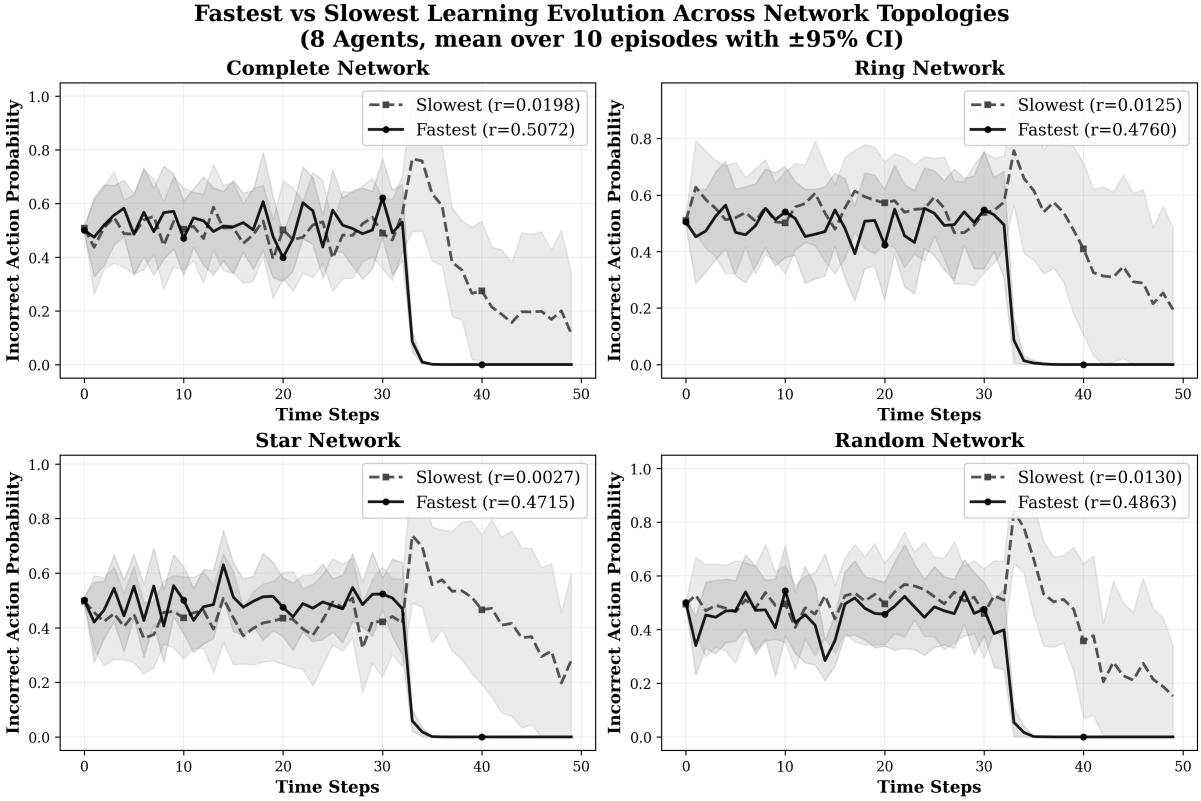

Figure 5.4: Fastest vs slowest agent learning trajectories across network topologies. Learning rates represent the mean trajectories.

Information Revelation Through Dynamic Role Assignment Our results reveal a systematic pattern where agents naturally emerge as either information generators or information exploiters within each episode—a phenomenon that parallels the dynamic role assignment observed in strategic experimentation, where agents differentiate into high and low allocators. In learning without experimentation, some agents engage in more exploratory behavior than they would do in isolation, leading them to make more mistakes, while providing valuable

learning signals that enable other agents to make better-informed decisions and achieve faster convergence. This mirrors the information generation role played by high allocators in strategic experimentation, who provide valuable learning signals through their experimental choices.

Network Structure & Information Flow Figure 5.4 reveals how network topology influences learning dynamics. Complete networks facilitate the most effective information sharing, producing the largest performance gaps between fastest and slowest learners. Ring networks show more uniform learning rates due to limited information flow, while star networks create pronounced differences between central and peripheral agents. These patterns align with the theoretical insights that information aggregation depends critically on the network structure, though the specific mechanisms through which topology affects learning warrant further investigation.

Implications and Validation These results demonstrate that our POAMG framework successfully captures the essential features of social learning without experimentation, validating both the learning barriers that limit information aggregation and the coordination benefits that emerge in well-connected networks. The computational approach complements theoretical analysis by revealing the dynamics of strategy discovery and the distributional properties of learning rates across different network configurations.

Our computational analysis reveals a striking consistency in the emergence of dynamic role assignment across fundamentally different social learning environments. In both strategic experimentation and learning without experimentation, agents naturally differentiate into complementary roles that enhance collective information processing, though the specific mechanisms differ substantially. In strategic experimentation, this differentiation manifests through allocation patterns, where some agents assume higher-risk experimental roles by allocating more resources to uncertain alternatives, while others free-ride on this information generation. In learning without experimentation, the same underlying phenomenon emerges through learning rate disparities and exploration variance, where slower-learning agents exhibit more exploratory behavior that benefits faster-learning agents who can exploit this information revelation. Crucially, these roles are not fixed by agent characteristics but emerge dynamically through the learning process itself, with different agents assuming information generation and exploitation roles across episodes. Our ability to observe these role assignments stems from analyzing asymmetric behavioral outcomes rather than symmetric equilibrium predictions—this reveals that free-riding and encouragement effects coexist simultaneously within the same networks, affecting different agents in opposite ways. While some agents experience substantial performance gains through social learning benefits, others bear the costs of information generation, uncovering systematic inequalities that emerge from multi-agent learning dynamics. This robustness across distinct social learning mechanisms suggests that dynamic role assignment represents a fundamental organizing principle of multi-agent learning systems, enabling efficient information aggregation through asymmetric behavioral specialization that inherently creates performance disparities.

6 CONCLUSION

This thesis bridges economic social learning theory and multi-agent reinforcement learning by developing Partially Observable Active Markov Games (POAMGs) and the POLARIS algorithm. Our framework addresses fundamental challenges in modeling adaptive behavior under partial observability, demonstrating how computational approaches can complement traditional economic theory while providing new insights into complex social learning dynamics. The integration of these traditionally separate domains opens new research directions that neither field could pursue independently, establishing a foundation for understanding how rational agents learn and adapt in environments characterized by both environmental uncertainty and strategic interdependence.

Theoretical and Algorithmic Contributions Our work makes four interconnected contributions that advance both theoretical understanding and computational capabilities. First, we developed the POAMG framework, extending Active Markov Games to partially observable settings where agents reason about both environmental uncertainty and evolving strategies of others. This formalization captures the essential tension between exploration for private information and exploitation of social information that characterizes real-world learning scenarios.

Second, we provided theoretical analysis of convergence and optimization in POAMGs, establishing conditions for stochastic stability and deriving policy gradient theorems for average and discounted reward objectives. We extended the framework to continuous-time dynamics through stochastic differential equations and revealed how traditional game-theoretic equilibrium concepts relate to active equilibrium when agents account for others' learning processes. These theoretical foundations provide rigorous guarantees for algorithm convergence while preserving the strategic structure inherent in social learning problems.

Third, we introduced POLARIS, combining belief processing through Transformer models, variational inference learning with GNNs, and MASAC based reinforcement learning optimization. This enables agents to develop sophisticated strategies accounting for both environmental partial observability and strategic adaptation by other agents. The algorithm's modular architecture allows for principled handling of belief updates, strategic reasoning, and policy optimization in a unified framework.

Fourth, we validated our framework through applications to strategic experimentation and learning without experimentation, demonstrating how our approach captures dynamic role assignment, information generation and exploitation, and the emergence of learning barriers and coordination benefits. These applications illustrate the framework's versatility in addressing diverse social learning phenomena while maintaining computational tractability.

Beyond Symmetric Equilibria These contributions enable a fundamental shift from traditional symmetric equilibrium analysis to understanding asymmetric behavioral outcomes in multi-agent learning. A critical advantage of our computational approach lies in its ability to analyze how free-riding and encouragement effects operate simultaneously within the same networks, but affect a priori identical agents in fundamentally opposing ways. Our analysis reveals systematic inequalities emerging from multi-agent learning dynamics, where some agents capture substantial performance gains through social learning benefits while others shoulder the costs of information generation. These findings demonstrate that the effects predicted by economic theory actually redistribute across individual agents, creating inherent performance disparities that symmetric equilibria cannot capture.

Building on this insight, we observe that dynamic role assignment emerges as a fundamental organizing principle underlying multi-agent learning systems. Consistently across our experimental scenarios, agents spontaneously differentiate into complementary roles that enhance collective information processing capabilities. This spontaneous specialization reveals how asymmetric behavioral patterns can emerge in social systems without central coordination—through the natural development of differentiated agent strategies that redistribute information generation and exploitation across the population.

Methodological Advances Our methodological contributions extend beyond the specific applications studied, providing tools for future research in computational social learning. The ex-post aggregation methodology offers systematic approaches for extracting insights from multi-agent learning dynamics despite catastrophic forgetting in neural networks. Our discretization approach for continuous-time economic models treats Lévy processes through Euler-Maruyama schemes and maps continuous decisions to discrete action probabilities, providing a template for computational implementation while preserving essential incentive structures.

We construct an observed reward function that enables reinforcement learning in environments without direct reward signals and develop a specialized transformer loss function for Lévy process observations. Our implementation incorporates GNNs with temporal and spatial attention mechanisms that capture both network topology and temporal dependencies in multi-agent interactions. Furthermore, we introduce theoretically grounded computational methods for off-equilibrium asymmetric analysis that move beyond the symmetric equilibrium analysis in economic theory, allowing for heterogeneous agent behaviors and potentially asymmetric strategy convergence patterns.

Limitations and Research Challenges Despite these advances, our approach faces important limitations that highlight directions for future research. Catastrophic forgetting prevents direct measurement of equilibrium strategies that economic theory predicts, requiring alternative analytical approaches. Computational complexity limits analysis to small networks and short time horizons, potentially missing asymptotic behaviors that theoretical models emphasize. The discretization of continuous-time models, while necessary for computational tractability, may introduce artifacts that affect the correspondence between theoretical predictions and computa-

tional results.

Future Research Directions Several promising avenues emerge from this work. First, developing sophisticated continual learning techniques specifically for social learning environments could address catastrophic forgetting limitations through architectures maintaining separate policy components for different states. Second, scaling our framework to larger networks and longer horizons would enable validation of asymptotic results and investigation of emergent phenomena in complex social systems. Third, extending to dynamic environments where underlying states change over time would broaden applicability to real-world scenarios characterized by non-stationarity. Fourth, investigating optimal information structures and network topologies could inform platform and institutional design. Fifth, applying our framework to diverse social learning phenomena such as herding behavior, social teaching and mentoring, imitation cascades, and collective decision-making could illuminate the mechanisms underlying these important social processes and inform interventions to promote beneficial outcomes. Sixth, investigating the determinants of role assignment and inequality persistence could reveal what factors decide which agents assume information generation versus exploitation roles and whether agents can escape disadvantageous positions through strategic adaptation.

Broader Implications and Impact This work demonstrates the value of bridging economic theory and computational methods. By preserving theoretical rigor while leveraging modern machine learning techniques, we gain insights into complex social phenomena that neither approach could achieve independently. The practical implications span multiple domains, from financial markets and organizational learning to online social networks and collaborative platforms.

In financial markets and organizational learning, our findings on dynamic role assignment suggest that effective systems naturally develop specialized roles that enhance collective information processing. For online social networks and collaborative platforms, understanding how learning barriers emerge provides guidance for designing systems that promote effective knowledge sharing. Our contributions to multi-agent reinforcement learning include principled approaches for handling partial observability and strategic adaptation, with techniques for belief processing, inference learning, and integrated optimization that may prove valuable for other applications requiring sophisticated coordination under uncertainty.

Conclusion The convergence of economic theory and artificial intelligence presents significant opportunities for advancing our understanding of social learning and strategic behavior. This thesis provides both theoretical foundations and computational tools for future researchers to explore the rich dynamics of learning in complex social environments. While substantial progress has been made, much work remains to fully realize the potential of this interdisciplinary integration—ultimately enabling us to better understand how intelligent agents can learn from each other to make better decisions in an uncertain world.

Appendices

REINFORCEMENT LEARNING BACKGROUND A

This section provides a brief overview of reinforcement learning (RL) and multi-agent reinforcement learning (MARL) to provide a foundation for understanding the theoretical framework presented in this paper.

A.1 REINFORCEMENT LEARNING FOUNDATIONS

Reinforcement learning (RL) provides a mathematical framework for solving sequential decision-making problems under uncertainty. In the classical single-agent setting, an agent interacts with a stationary environment by observing states, taking actions, and receiving rewards, with the objective of maximizing its expected cumulative reward over time Sutton and Barto (2018). This interaction is formalized as a Markov Decision Process (MDP), defined as a tuple $M = \langle S, A, T, R, \gamma \rangle$ where:

- S is the state space, representing all possible configurations of the environment
- A is the action space, representing all possible decisions available to the agent
- $T : S \times A \mapsto \Delta(S)$ is the transition function, specifying the probability distribution over next states given the current state and action
- $R : S \times A \mapsto \mathbb{R}$ is the reward function, specifying the immediate reward received after taking an action in a state
- $\gamma \in [0, 1)$ is the discount factor, balancing immediate versus future rewards

The agent's behavior is characterized by a policy $\pi : S \mapsto \Delta(A)$, which maps states to probability distributions over actions. The policy can be evaluated using value functions: the state-value function $V^{\pi}(s)$ represents the expected return when starting in state s and following policy π thereafter, while the action-value function $Q^{\pi}(s, a)$ represents the expected return after taking action a in state s and following policy π thereafter. The goal of RL is to find an optimal policy π^* that maximizes the expected return from all states.

The Markov property, which states that the future is independent of the past given the present, is a fundamental assumption in MDPs. This property ensures that the current state provides all the necessary information for making optimal decisions, greatly simplifying the learning problem. However, this assumption becomes problematic in multi-agent settings, as we will discuss next.

A.2 Multi-Agent Reinforcement Learning: Concepts and Challenges

Multi-agent reinforcement learning (MARL) extends the single-agent RL framework to environments with multiple autonomous agents that interact simultaneously Albrecht, Christianos, and Schäfer (2024); Busoniu, Babuska, and De Schutter (2010); Huh and Mohapatra (2024); Yang and Wang (2020); Zhang, Yang, and Başar (2021). MARL encompasses a wide spectrum of scenarios, from fully cooperative tasks where agents share a common reward function, to fully competitive zero-sum games, to the general mixed cooperative-competitive case with individual reward functions Nowé, Vrancx, and De Hauwere (2012). These interactions are commonly formalized as Markov games (also known as stochastic games) Littman (1994); Shapley (1953), defined as a tuple $M_n = \langle I, S, \boldsymbol{A}, T, \boldsymbol{R}, \gamma \rangle$, where:

- $I = \{1, \ldots, n\}$ is the set of n agents
- S is the state space, shared among all agents
- $\boldsymbol{A} = \times_{i \in I} A^i$ is the joint action space, where A^i is the action space of agent i
- $T : S \times A \mapsto \Delta(S)$ is the transition function that depends on the joint action
- $\boldsymbol{R} = \times_{i \in I} R^i$ is the joint reward function, where $R^i : S \times A \mapsto \mathbb{R}$ is agent i's individual reward function
- $\gamma \in [0, 1)$ is the discount factor

MARL introduces several fundamental challenges beyond those encountered in single-agent RL. The joint action space grows exponentially with the number of agents, creating a combinatorial explosion that makes exploration and learning increasingly difficult as more agents are added to the system. Coordination challenges further complicate the learning process, as agents must synchronize their actions to achieve effective joint behavior, especially in cooperative settings where team performance depends on complementary actions Du, Leibo, Islam, Willis, and Sunehag (2023), that can also involve explicit communication Zhu, Dastani, and Wang (2024). In scenarios with shared rewards, the credit assignment problem becomes particularly troublesome, as determining individual contributions to team success grows increasingly complex when outcomes result from joint actions rather than individual decisions. MARL also demands sophisticated strategic reasoning, requiring agents to model and reason about other agents' goals, beliefs, and strategies, particularly in competitive or mixed cooperative-competitive settings. Perhaps most critically, when multiple agents learn simultaneously, the environment becomes non-stationary from each agent's perspective, as other agents' changing policies continuously modify the effective environment dynamics, violating a fundamental assumption of traditional reinforcement learning algorithms.

A.3 THE NON-STATIONARITY CHALLENGE

The non-stationarity problem in MARL represents a fundamental departure from single-agent RL assumptions and presents one of the most significant obstacles to effective multi-agent learning Hernandez-Leal, Kaisers, Baarslag, and de Cote (2017); Papoudakis, Christianos, Schäfer, and Albrecht (2019). To understand this challenge more precisely, we can analyze how learning dynamics alter the effective environment for each agent. When multiple agents learn simultaneously, each agent i with policy π^i effectively faces an environment whose dynamics depend on the joint policies of all other agents $\boldsymbol{\pi}^{-i}$. From agent i's perspective, the effective transition function becomes:

$$T^i_{\boldsymbol{\pi}_t^{-i}}(s_{t+1}|s_t, a_t^i) = \sum_{\boldsymbol{a}^{-i} \in \boldsymbol{A}^{-i}} \left(\prod_{j \in I \setminus \{i\}} \pi_t^j(a^j|s_t) \right) \cdot T(s_{t+1}|s_t, (a_t^i, \boldsymbol{a}^{-i})) \tag{A.1}$$

where $\boldsymbol{A}^{-i}$ denotes the joint action space, $\boldsymbol{a}^{-i}$ denotes the joint action and $\boldsymbol{\pi}_t^{-i}$ denotes the joint policies of all agents except i. When other agents update their policies from $\boldsymbol{\pi}_t^{-i}$ to $\boldsymbol{\pi}_{t+1}^{-i}$ through learning, this effective transition function changes, creating a non-stationary environment for agent i. This non-stationarity violates the Markov property assumption underlying most RL algorithms and can lead to several significant challenges that fundamentally undermine the theoretical foundations of single-agent RL. When multiple agents learn simultaneously, standard RL convergence guarantees no longer apply, and learning algorithms may oscillate or diverge as the effective environment continuously shifts Bowling and Veloso (2002). This dynamic environment causes value function approximations to become increasingly inaccurate over time, leading to suboptimal policy decisions as agents base their choices on outdated models of their environment Lauer and Riedmiller (2000). Further complicating matters, agents face a moving target problem where they must simultaneously learn optimal policies while adapting to the evolving strategies of others, creating a complex coupled learning process that resists straightforward optimization approaches Laurent, Matignon, and Fort-Piat (2011). The situation becomes particularly problematic in competitive settings, where learning dynamics may lead to cyclic policy changes rather than convergence to stable strategies, as agents continuously adapt and counter-adapt to each other's evolving behaviors Balduzzi et al. (2018). These interconnected challenges highlight why addressing non-stationarity remains one of the central research questions in multi-agent reinforcement learning.

A.4 TRADITIONAL APPROACHES TO NON-STATIONARITY

Researchers have developed various approaches to address the non-stationarity challenge in MARL. These can be broadly categorized as follows:

A.4.1 Independent Learning

The simplest approach is to ignore non-stationarity entirely, treating other agents as part of the environment and applying standard single-agent RL algorithms independently for each agent (Tan, 1993). This approach, often called Independent Learning (IL), requires no explicit coordination or modeling of other agents. Methods in this category include Independent Q-Learning (Tampuu et al., 2015), where each agent maintains its own Q-function and updates it using only its own experiences. While computationally efficient and naturally scalable to many agents, independent learning lacks theoretical convergence guarantees and can fail in complex multi-agent scenarios due to the violation of the stationarity assumption.

A.4.2 Centralized Training with Decentralized Execution

To mitigate non-stationarity during learning while preserving autonomous execution, many approaches adopt the paradigm of centralized training with decentralized execution (CTDE) (Foerster, Farquhar, Afouras, Nardelli, & Whiteson, 2018; Lowe et al., 2017; Sunehag et al., 2017). In CTDE, agents have access to additional information during training (e.g., joint actions, global state, other agents' policies) but operate based solely on their local observations during execution.

Notable algorithms in this category include Multi-Agent Deep Deterministic Policy Gradient (MADDPG) (Lowe et al., 2017), which uses centralized critics with access to joint actions and states during training but decentralized actors during execution. Another significant approach is Counterfactual Multi-Agent Policy Gradients (COMA) (Foerster et al., 2018), which addresses credit assignment using a centralized critic that computes a counterfactual baseline. The category also features QMIX (Rashid et al., 2018), which learns a centralized value function that factorizes into a mixing of individual agent Q-values, ensuring consistency between centralized and decentralized policies. While CTDE methods help stabilize learning, they still do not fully address the fundamental non-stationarity issue, as they focus on adapting to current policies rather than reasoning about policy evolution over time.

A.4.3 Opponent Modeling and Population-Based Training

Another approach is to explicitly model the behavior and learning processes of other agents, enabling more informed adaptation (Albrecht & Stone, 2018; He, Boyd-Graber, Kwok, & Daumé III, 2016). This can involve predicting other agents' policies, types, or learning dynamics. Methods in this category include Deep Reinforcement Opponent Network (DRON) (He et al., 2016), which integrates opponent modeling into deep Q-learning. Another significant approach is Probabilistic Recursive Reasoning (PR2) (Wen, Yang, Luo, Wang, & Pan, 2019), which models higher-order beliefs about other agents' reasoning processes.

The category also encompasses Population-Based Training (PBT) (Jaderberg et al., 2019), which trains a population of agents simultaneously, creating a naturally changing learning environment that promotes robustness to non-stationarity. While these approaches can better handle

changing policies, they typically focus on adaptation to current policies rather than influencing the learning trajectories of other agents.

A.4.4 Equilibrium Learning and Stability Concepts

Drawing on game theory, some approaches focus on finding policies that constitute game-theoretic solution concepts like Nash equilibria (Bowling, 2005; Hu & Wellman, 2003; Littman, 2001). These methods aim to find stable joint policies where no agent has an incentive to unilaterally deviate. Notable algorithms include Nash Q-Learning (Hu & Wellman, 2003), which converges to Nash equilibria in general-sum Markov games. Another significant approach is Friend-or-Foe Q-Learning (Littman, 2001), which converges to optimal policies in restricted classes of Markov games.

The category also features WoLF-PHC (Win or Learn Fast - Policy Hill Climbing) (Bowling & Veloso, 2002), which adjusts learning rates based on whether an agent is "winning" or "losing" to promote convergence. While these approaches provide theoretical guarantees under certain conditions, they often make strong assumptions about the game structure and other agents' rationality. Moreover, they focus on convergence to static equilibria rather than the dynamic nature of multi-agent learning.

A.4.5 Meta-Learning for Non-Stationarity

Recent work has explored meta-learning as a framework for rapid adaptation to non-stationarity in MARL (Al-Shedivat et al., 2018; Kim et al., 2021). These approaches train agents to quickly adapt to new agents or environments, treating non-stationarity as a meta-learning problem. Key methods include Continuous Adaptation via Meta-Learning (CAML) (Al-Shedivat et al., 2018), which uses meta-learning to quickly adapt to evolving opponents.

Another significant approach is Meta-MAPG (Meta-Multiagent Policy Gradient) (Kim et al., 2021), which extends meta-learning to explicitly account for the influence of an agent's actions on the learning processes of other agents. While meta-learning approaches show promise for rapid adaptation, they often still lack a comprehensive framework for modeling and influencing the long-term learning dynamics of multi-agent systems.

A.5 Active Markov Games

Active Markov Games provide a more sophisticated framework for modeling the dynamic nature of multi-agent learning by explicitly incorporating the policy update processes of all agents (Kim et al., 2022). They extend standard Markov games to capture not just the immediate effects of actions on states and rewards, but also their influence on future policy updates, addressing the non-stationarity challenge at a fundamental level.

A.5.1 Formal Definition

An Active Markov Game is defined as a tuple $M_n = \langle I, S, \boldsymbol{A}, T, \boldsymbol{R}, \boldsymbol{\Theta}, \boldsymbol{U} \rangle$, where:

- $I = \{1, \ldots, n\}$ is the set of n agents
- S is the state space
- $\boldsymbol{A} = \times_{i \in I} A^i$ is the joint action space, where A^i is the action space of agent i
- $T : S \times A \mapsto \Delta(S)$ is the transition function
- $\boldsymbol{R} = \times_{i \in I} R^i$ is the joint reward function, where $R^i : S \times A \mapsto \mathbb{R}$ is agent i's individual reward function
- $\boldsymbol{\Theta} = \times_{i \in I} \Theta^i$ is the joint policy parameter space, where Θ^i is the policy parameter space for agent i
- $\boldsymbol{U} = \times_{i \in I} U^i$ is the joint Markovian policy update function, where $U^i : \Theta^i \times T^i \mapsto \Delta(\Theta^i)$ maps current policy parameters and transitions to distributions over next policy parameters

Here, T^i represents the trajectory space for agent i, with a particular element $\tau_t^i = \{s_t, \boldsymbol{a}_t, r_t^i, s_{t+1}\}$ consisting of the current state, joint action, reward received, and the next state. The interaction process in an Active Markov Game unfolds through a sequential procedure of actions and updates. At timestep t, each agent i selects an action $a_t^i \sim \pi^i(\cdot|s_t; \theta_t^i)$ based on its parameterized policy with parameters $\theta_t^i \in \Theta^i$. Following this selection, the environment transitions from state s_t to s_{t+1} according to $T(s_{t+1}|s_t, \boldsymbol{a}_t)$ based on the joint action $\boldsymbol{a}_t = (a_t^1, \ldots, a_t^n)$. Each agent i then receives a reward $r_t^i = R^i(s_t, \boldsymbol{a}_t)$ and subsequently updates its policy parameters according to its update function $\theta_{t+1}^i \sim U^i(\theta_{t+1}^i|\theta_t^i, \tau_t^i)$. This process continues until non-stationary policies converge to a recurrent set of joint policies. The key insight is that agent i's actions not only affect immediate states and rewards but also influence future policy updates of all agents, including itself, through the trajectory τ_t^i.

A.5.2 Augmented Transition Function and Stationarity

The Active Markov Game framework fundamentally transforms the non-stationarity problem by explicitly modeling how policies evolve through Markovian update functions. To understand this transformation, let us contrast the transition functions in standard MARL and Active Markov Games. In standard MARL, the transition function $T : S \times \boldsymbol{A} \mapsto \Delta(S)$ only captures state transitions based on joint actions. However, Active Markov Games introduce a critical innovation: they explicitly model how policies change via Markovian update functions. The augmented transition function becomes:

$$\mathbb{P}(s_{t+1}, \boldsymbol{\theta}_{t+1} | s_t, \boldsymbol{\theta}_t) = \sum_{\boldsymbol{a}_t \in \boldsymbol{A}} \left(\prod_{i \in I} \pi^i(a_t^i | s_t; \theta_t^i) \right) \times T(s_{t+1} | s_t, \boldsymbol{a}_t) \times \boldsymbol{U}(\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t, \boldsymbol{\tau}_t) \tag{A.2}$$

The crucial difference is the inclusion of $\boldsymbol{U}$ - a Markovian update function that specifies exactly how policy parameters $\boldsymbol{\theta}_t$ (and consequently policies $\boldsymbol{\pi}$) evolve based on current parameters and trajectories. By making policy updates an explicit part of the system dynamics, what was previously an unpredictable external process becomes a predictable internal one. This augmented transition function operates over the joint space $(s, \boldsymbol{\theta}) \in S \times \boldsymbol{\Theta}$, creating a joint process that is stationary even though individual policies are changing.

A.5.3 Theoretical Properties and Convergence

Active Markov Games exhibit important theoretical properties that enable rigorous analysis of multi-agent learning dynamics:

Stationary Periodic Distributions. Under mild assumptions, the joint process of states and policies in an Active Markov Game converges to a stationary periodic distribution Kim et al. (2022):

$$\mu_k(s, \boldsymbol{\theta} | s_0, \boldsymbol{\theta}_0, \ell) = \mathbb{P}(s_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta} | s_0, \boldsymbol{\theta}_0, \ell) \tag{A.3}$$

where $\ell = t \% k$ with $\%$ denoting the modulo operation. This distribution represents the limiting behavior of the system after convergence, characterized by potentially periodic patterns of states and policies. The parameter k represents the period length, with $k = 1$ corresponding to fixed-point convergence. To address sensitivity to initial conditions, the framework introduces the concept of *stochastically stable distributions* Kim et al. (2022). These are limiting distributions that emerge when small random perturbations are added to policy updates, providing robustness to initial state and policy configurations.

Long-term Optimization Objective. In Active Markov Games, agents optimize for long-term average reward rather than discounted return Kim et al. (2022); Sutton and Barto (2018):

$$\max_{\theta^i, U^i} \rho^i(s, \boldsymbol{\theta}, \boldsymbol{U}) := \max_{\theta^i, U^i} \lim_{T \to \infty} \mathbb{E} \left[\frac{1}{T} \sum_{t=0}^{T} R^i(s_t, \boldsymbol{a}_t) \middle| \begin{array}{c} s_0 = s, \boldsymbol{\theta}_0 = \boldsymbol{\theta}, \\ \boldsymbol{a}_{0:T} \sim \boldsymbol{\pi}(\cdot | s_{0:T}; \boldsymbol{\theta}_{0:T}), \\ s_{t+1} \sim T(\cdot | s_t, \boldsymbol{a}_t), \boldsymbol{\theta}_{t+1} \sim \boldsymbol{U}(\cdot | \boldsymbol{\theta}_t, \boldsymbol{\tau}_t) \end{array} \right] \tag{A.4}$$

where $0 : T$ denotes the sequence from time 0 to T. This formulation encourages agents to consider how to influence the limiting behavior of the system rather than just short-term performance, addressing the non-stationarity challenge at a fundamental level.

Active Equilibrium. Active Markov Games give rise to a new solution concept called the Active Equilibrium Kim et al. (2022), which generalizes traditional game-theoretic equilibria like Nash equilibrium. An Active Equilibrium is a joint policy parameter $\boldsymbol{\theta}^* = \{\theta^{i*}, \boldsymbol{\theta}^{-i*}\}$ with associated joint update function $\boldsymbol{U}^* = \{U^{i*}, \boldsymbol{U}^{-i*}\}$ such that:

$$\rho^i(s, \theta^{i*}, \boldsymbol{\theta}^{-i*}, U^{i*}, \boldsymbol{U}^{-i*}) \geq \rho^i(s, \theta^i, \boldsymbol{\theta}^{-i*}, U^i, \boldsymbol{U}^{-i*}) \tag{A.5}$$

for all $i \in I$, $s \in S$, $\theta^i \in \Theta^i$, $U^i \in U^i$. This equilibrium concept captures the idea that rational agents should optimize not just their immediate policies but also their adaptation strategies, taking into account the learning dynamics of the system. The active equilibrium generalizes several classic solution concepts in game theory. Stationary Nash equilibria and correlated equilibria are special cases of active equilibria when $k = 1$ (fixed-point convergence) and joint action distributions are independent or correlated, respectively. Similarly, cyclic Nash equilibria and cyclic correlated equilibria can be viewed as special cases of active equilibria when $k > 1$ (periodic behavior), the joint update function is deterministic, and joint action distributions are independent or correlated, respectively. This generality allows the active equilibrium concept to capture a wider range of stable multi-agent behaviors in dynamic learning settings than traditional equilibrium notions.

Active Markov Games address the non-stationarity challenge in several fundamental ways. By incorporating policy parameters and update functions directly into the framework, Active Markov Games make the non-stationarity explicit and amenable to analysis, embracing it as an integral part of the multi-agent learning process rather than treating it as an external challenge to be mitigated. The average reward objective promotes farsighted optimization by encouraging agents to consider the long-term limiting behavior of the system rather than myopically optimizing for immediate or short-term rewards, leading to policies that shape the learning trajectories of other agents in beneficial ways rather than merely reacting to their current behaviors. Furthermore, the framework enables active influence by allowing agents to reason about how their actions affect not just the environment state but also the learning processes of other agents, facilitating sophisticated strategies like teaching, deception, or coordination that explicitly aim to influence other agents' policy updates. Under appropriate conditions, Active Markov Games provide theoretical guarantees about convergence to stationary periodic distributions, offering a more solid foundation for algorithm development than approaches without such guarantees. Finally, the Active Equilibrium concept generalizes traditional game-theoretic solution concepts, providing a more comprehensive framework for understanding stable multi-agent behaviors in learning settings.

A.5.4 Practical Implementations

Recent work has leveraged the Active Markov Game framework to develop practical algorithms for MARL. Notable examples include FURTHER (FUlly Reinforcing acTive influence witH averagE Reward) Kim et al. (2022), which implements a policy gradient approach tailored to

the average reward objective in Active Markov Games, combined with variational inference for estimating other agents' policy dynamics in a decentralized manner. Another significant approach is Meta-MAPG (Meta-Multiagent Policy Gradient) Kim et al. (2021), which integrates meta-learning with explicit modeling of other agents' learning processes, aligning with the influence-aware perspective of Active Markov Games. These algorithms have demonstrated superior performance compared to methods that neglect the learning dynamics of other agents, particularly in environments with high levels of non-stationarity.

A.6 Extension to Partial Observability

While Active Markov Games provide a powerful framework for addressing non-stationarity, they assume full observability of the environment state. In many real-world scenarios, agents have limited perception capabilities and cannot directly observe the complete state of the environment or the internal parameters of other agents. This partial observability introduces additional challenges for multi-agent learning. The Partially Observable Markov Decision Process (POMDP) Kaelbling et al. (1998) extends MDPs to settings with partial observability by introducing observation functions. In the multi-agent context, this leads to Partially Observable Stochastic Games (POSGs) Hansen, Bernstein, and Zilberstein (2004) or Decentralized POMDPs (Dec-POMDPs) for cooperative settings Bernstein, Givan, Immerman, and Zilberstein (2002). Extending Active Markov Games to partial observability settings represents a significant advancement in addressing the combined challenges of non-stationarity and limited information in multi-agent learning. The resulting framework, Partially Observable Active Markov Games, incorporates belief states and observation functions while maintaining the key benefits of Active Markov Games for modeling and influencing learning dynamics.

In partially observable settings, policy gradient methods require careful adaptation to handle the unavailability of the true state. Theoretical formulations, as established by Kaelbling et al. (1998), condition value functions on belief states rather than environmental states. However, this theoretical construction presents a practical challenge: agents don't have access to the true state distribution needed to compute this expectation. Practical implementations resolve this tension through multiple approaches: history-based methods condition policies on observation histories h_t rather than belief states Aberdeen and Baxter (2002); recurrent network formulations implicitly encode history using recurrent architectures Hausknecht and Stone (2015); and belief-explicit methods maintain and update a belief distribution while estimating values from observable quantities Nguyen, Daley, Song, Amato, and Platt (2021). These approaches share a common principle of using trajectory sampling to estimate gradients without requiring knowledge of the true state or transition dynamics. In multi-agent settings with partial observability, the challenge compounds as agents must reason about others' belief states and learning dynamics, requiring sophisticated inference mechanisms to estimate other agents' policies and their evolution over time. For Partially Observable Active Markov Games, policy gradients can be formulated to account for the impact of current actions on both the future environmental states and the future policies of other agents, all while operating from belief states rather than full

state observations.

PROOFS REGARDING AVERAGE RETURNS

B

This section provides the mathematical foundations and detailed proofs for the key theoretical results in our Partially Observable Active Markov Game framework with average return objectives. We establish the Markov properties of belief transitions and joint processes, prove the existence and uniqueness of stochastically stable distributions, and derive the policy gradient theorem for average return optimization.

B.1 MARKOV PROPERTY OF THE JOINT PROCESS

We first establish the Markov properties of belief state transitions and the joint process comprising states, belief states, and policy parameters. These properties form the foundation for our convergence analysis.

Lemma 1 (Markov Property of Belief Transitions)**.** *The sequence of belief states $\{b_t^i\}_{t\geq 0}$ for each agent i forms a Markov process under a fixed policy π, meaning that the distribution of b_{t+1}^i depends only on b_t^i, a_t^i, and o_{t+1}^i, and not on earlier beliefs, actions, or observations.*

Proof. By definition, the belief state b_t^i at time t incorporates all relevant information from the history of observations and actions up to time t. For Bayesian agents, given b_t^i, action a_t^i, and observation o_{t+1}^i, the belief update equation uniquely determines b_{t+1}^i through the rule:

$$b_{t+1}^i(s') = \frac{O^i(o_{t+1}^i|s')\sum_{s\in S} T(s'|s,\boldsymbol{a}_t)b_t^i(s)}{\sum_{s''\in S} O^i(o_{t+1}^i|s'')\sum_{s\in S} T(s''|s,\boldsymbol{a}_t)b_t^i(s)} \tag{B.1}$$

where $O^i(o_{t+1}^i|s')$ represents the probability of agent i observing o_{t+1}^i in state s'. This update depends only on b_t^i, a_t^i, and o_{t+1}^i, and not on the sequence of beliefs, actions, and observations that led to b_t^i. Therefore, the belief state transition satisfies the Markov property. □

Remark 1 (Transformer-Based Belief Representation)**.** *In practice, transformer architectures can be used to represent and update belief states. In such cases, the belief state is updated through an attention-based mechanism of the form:*

$$b_{t+1}^i = Transformer(b_t^i, o_{t+1}^i), \tag{B.2}$$

that is parameterized by $\theta_{Transformer}^i$ and processes the latest observation and previous belief state through self-attention layers, satisfying the Markov property while capturing complex dependencies between observations and beliefs.

Next, we establish the Markov property of the joint process.

Lemma 2 (Markov Property of Joint Process)**.** *The joint process* $(s_t, \boldsymbol{b}_t, \boldsymbol{\theta}_t)$ *forms a Markov process under periodic policy updates, meaning that the distribution of* $(s_{t+1}, \boldsymbol{b}_{t+1}, \boldsymbol{\theta}_{t+1})$ *depends only on* $(s_t, \boldsymbol{b}_t, \boldsymbol{\theta}_t)$ *and not on earlier states.*

Proof. The state transitions are by definition only dependent on current state and current actions. For belief state, as established earlier, $\boldsymbol{b}_{t+1}$ depends only on $\boldsymbol{b}_t$, $\boldsymbol{a}_t$, and $\boldsymbol{o}_{t+1}$. Since actions $\boldsymbol{a}_t$ are drawn from policies $\pi(a_t^i|b_t^i;\theta_t^i)$ that depend only on b_t^i and θ_t^i for each agent i, and observations $\boldsymbol{o}_{t+1}$ depend stochastically on the resulting state, the distribution of $\boldsymbol{b}_{t+1}$ depends only on $\boldsymbol{b}_t$ and $\boldsymbol{\theta}_t$. For policy parameters, at every time step t, policy parameters are updated according to $\boldsymbol{U}(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, \boldsymbol{\tau}_t)$. Under the Markovian policy update assumption, the update depends only on $\boldsymbol{\theta}_t$ and current trajectory $\boldsymbol{\tau}_t$, which are functions of s_t, $\boldsymbol{b}_t$ and $\boldsymbol{\theta}_t$.Therefore, the joint transition probability $\mathbb{P}((s_{t+1}, \boldsymbol{b}_{t+1}, \boldsymbol{\theta}_{t+1})|(s_t, \boldsymbol{b}_t, \boldsymbol{\theta}_t))$ depends only on the current state $(s_t, \boldsymbol{b}_t, \boldsymbol{\theta}_t)$ and not on the history of states, thus establishing the Markov property. □

B.2 CONVERGENCE

Having established the Markov properties, we now analyze the convergence behavior of the joint process to a unique stochastically stable distribution. This convergence result is crucial for developing optimization objectives in our framework.

Theorem 1 (Stochastically Stable Distribution)**.** *For* $\varepsilon > 0$*, define the* ε*-perturbed policy update functions as:*

$$U_i^{\varepsilon}(\theta_{t+1}^i|\theta_t^i, \tau_t^i) = (1-\varepsilon)U_i(\theta_{t+1}^i|\theta_t^i, \tau_t^i) + \varepsilon\eta_i(\theta_{t+1}^i) \tag{B.3}$$

where η_i *is a baseline distribution over* Θ_i *with full support, and* $\varepsilon > 0$ *is a small constant. Under Assumptions 1 and 2, as* $t \to \infty$ *and* $\varepsilon \to 0$*, the perturbed joint processes defined by these* ε*-perturbed policy update functions converge to the unique stochastically stable distribution* μ^* *of the unperturbed joint process* $(s_t, \boldsymbol{b}_t, \boldsymbol{\theta}_t)$*.*

Proof. We establish the convergence to the unique stochastically stable distribution by showing that a perturbed Markov process of a partially observable active Markov game is regular. The ε-perturbation ensures that from any policy parameter configuration $\boldsymbol{\theta}_t$, there is a positive probability of transitioning to any other configuration $\boldsymbol{\theta}_{t+1}$. The perturbed joint process $(s_t, \boldsymbol{b}_t, \boldsymbol{\theta}_t)^{\varepsilon}$ evolves according to the transition probability:

$$\mathbb{P}(s_{t+1}, \boldsymbol{b}_{t+1}, \boldsymbol{\theta}_{t+1}|s_t, \boldsymbol{b}_t, \boldsymbol{\theta}_t) = \sum_{\boldsymbol{a}_t \in \boldsymbol{A}} \boldsymbol{\pi}(\boldsymbol{a}_t|\boldsymbol{b}_t; \boldsymbol{\theta}_t) T(s_{t+1}|s_t, \boldsymbol{a}_t) \tag{B.4}$$

$$\sum_{\boldsymbol{o}_{t+1} \in O} \boldsymbol{O}(\boldsymbol{o}_{t+1}|s_{t+1}) \boldsymbol{U}^{\varepsilon}(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, \boldsymbol{\tau}_t) \tag{B.5}$$

$$\mathbb{I}(\boldsymbol{b}_{t+1} = \text{update}(\boldsymbol{b}_t, \boldsymbol{a}_t, \boldsymbol{o}_{t+1})) \tag{B.6}$$

where $\boldsymbol{\tau}_t = \{\boldsymbol{o}_t, \boldsymbol{a}_t, \boldsymbol{r}_t, \boldsymbol{o}_{t+1}\}$ is the joint trajectory, $\boldsymbol{r}_t = \{R^i(s_t, \boldsymbol{a}_t)\}_{i \in I}$ is the vector of rewards, and $\text{update}(\boldsymbol{b}_t, \boldsymbol{a}_t, \boldsymbol{o}_{t+1})$ represents the belief update function for all agents. Suppose, for

contradiction, that the perturbed Markov process is irregular, meaning it has multiple recurrent classes. Let C_1 and C_2 be two distinct recurrent classes of the joint process. We will show that there must exist a path with positive probability from any state in C_1 to any state in C_2, contradicting the assumption that they are distinct recurrent classes. Consider any $(s^1, \boldsymbol{b}^1, \boldsymbol{\theta}^1) \in C_1$ and $(s^2, \boldsymbol{b}^2, \boldsymbol{\theta}^2) \in C_2$. We construct a path from the first state to the second as follows. First, due to the ε-perturbation in update functions, there is a positive probability of directly transitioning from $\boldsymbol{\theta}_1$ to any $\boldsymbol{\theta}' \in \Theta$ in a single step, regardless of the trajectory. Specifically:

$$\mathbb{P}(\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}' | \boldsymbol{\theta}_t = \boldsymbol{\theta}_1, \boldsymbol{\tau}_t) \geq \prod_{i \in I} \varepsilon \cdot \eta_i(\theta'^i) > 0 \tag{B.7}$$

Thus, there is a positive probability path from $\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_2$. Second, by the communicating state assumption, for any two states $s_1, s_2 \in S$, there exists a sequence of joint actions $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ such that:

$$\mathbb{P}(s_{t+k} = s_2 | s_t = s_1, \boldsymbol{a}_t = \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{t+k-1} = \boldsymbol{a}_k) > 0 \tag{B.8}$$

Given the relationship between policies and actions, this sequence of actions has positive probability under any policy parameters $\boldsymbol{\theta}$ and corresponding belief states $\boldsymbol{b}$ such that $\pi^i(a^i | b^i; \theta^i) > 0$ for all required actions. Third, by the communicating belief-state assumption, for any two belief states $\boldsymbol{b}_1, \boldsymbol{b}_2 \in B$, there exists a sequence of observations $\boldsymbol{o}_1, \ldots, \boldsymbol{o}_m$ such that:

$$\mathbb{P}(\boldsymbol{b}_{t+m} = \boldsymbol{b}_2 | \boldsymbol{b}_t = \boldsymbol{b}_1, \boldsymbol{o}_{t+1} = \boldsymbol{o}_1, \ldots, \boldsymbol{o}_{t+m} = \boldsymbol{o}_m) > 0 \tag{B.9}$$

Since observations depend on states, which can be influenced through actions, and actions are determined by policies, there is a positive probability path from $\boldsymbol{b}_1$ to $\boldsymbol{b}_2$ under appropriate state transitions and policy parameters. Combining these three components, we can construct a path with positive probability from $(s^1, \boldsymbol{b}^1, \boldsymbol{\theta}^1)$ to $(s^2, \boldsymbol{b}^2, \boldsymbol{\theta}^2)$. This contradicts the assumption that C_1 and C_2 are distinct recurrent classes of the perturbed joint process. Since we have shown that the perturbed joint process has only one recurrent class, it is a regular Markov process. Following (Kim et al., 2022; Wicks & Greenwald, 2012), a regular Markov process on a finite state space possesses a unique stationary distribution μ^ε to which it converges regardless of the initial state. Moreover, as $\varepsilon \to 0$, the sequence of stationary distributions μ^ε converges to a limit μ^*, which is the unique stochastically stable distribution of the original, unperturbed process. Therefore, under the given conditions, the joint process $(s_t, \boldsymbol{b}_t, \boldsymbol{\theta}_t)$ in a partially observable active Markov game possesses a unique stochastically stable distribution μ^*. □

B.3 Policy Gradient Theorem

Having established the convergence to a unique stochastically stable distribution, we now derive the policy gradient theorem for optimization in our framework. This theorem provides the

mathematical foundation for gradient-based algorithms to maximize the average return objective.

Theorem 2 (Partially Observable Active Average Reward Policy Gradient Theorem)**.** *The gradient of the active average reward objective with respect to agent i's policy parameters θ^i in a partially observable setting is:*

$$\nabla_{\theta^i} J^i_\pi(\theta^i) = \sum_{s,\boldsymbol{b},\boldsymbol{\theta}} \mu_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) \sum_{a^i} \nabla_{\theta^i} \pi^i(a^i|b^i;\theta^i) \sum_{a^{-i}} \pi^{-i}(a^{-i}|b^{-i};\theta^{-i}) q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{B.10}$$

where the action-value function $q^i_{\theta^i}$ is defined as:

$$\begin{aligned} q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) = &\sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) \\ &\left[R^i(s,\boldsymbol{a}) - \rho^i_{\theta^i} + v^i_{\theta^i}(s', update(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}')\right] \end{aligned} \tag{B.11}$$

with $update(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}')$ representing the belief update function.

Proof. We first define the average reward objective in the partially observable setting:

$$\max_{\theta^i} \rho^i_{\theta^i}(\boldsymbol{b},\boldsymbol{\theta}) := \max_{\theta^i} \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T} R^i(s_t,\boldsymbol{a}_t) \,\middle|\, \begin{array}{c} \boldsymbol{b}_0 = \boldsymbol{b}, \boldsymbol{\theta}_0 = \boldsymbol{\theta}, \\ \boldsymbol{a}_{0:T} \sim \boldsymbol{\pi}(\cdot|\boldsymbol{b}_{0:T};\boldsymbol{\theta}_{0:T}), s_{t+1} \sim T(\cdot|s_t,\boldsymbol{a}_t), \\ \boldsymbol{o}_{t+1} \sim \boldsymbol{O}(\cdot|s_{t+1}), \boldsymbol{\theta}_{t+1} \sim \boldsymbol{U}(\cdot|\boldsymbol{\theta}_t,\boldsymbol{\tau}_t) \end{array}\right] \tag{B.12}$$

$$= \max_{\theta^i} \sum_{s,\boldsymbol{b},\boldsymbol{\theta}} b^i(s) \mu_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) \sum_{\boldsymbol{a}} \boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{b};\boldsymbol{\theta}) R^i(s,\boldsymbol{a}) \tag{B.13}$$

where $\mu_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta})$ is the stationary distribution of the joint process $(s_t,\boldsymbol{b}_t,\boldsymbol{\theta}_t)$ under agent i's policy parameterized by θ^i. The differential value function $v^i_{\theta^i}$ represents the expected total difference between the accumulated rewards from state s, belief states $\boldsymbol{b}$, and policy parameters $\boldsymbol{\theta}$, and the average reward $\rho^i_{\theta^i}$:

$$v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) = \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T} \left(R^i(s_t,\boldsymbol{a}_t) - \rho^i_{\theta^i}\right) \,\middle|\, \begin{array}{c} s_0 = s, \boldsymbol{b}_0 = \boldsymbol{b}, \boldsymbol{\theta}_0 = \boldsymbol{\theta}, \\ \boldsymbol{a}_{0:T} \sim \boldsymbol{\pi}(\cdot|\boldsymbol{b}_{0:T};\boldsymbol{\theta}_{0:T}), s_{t+1} \sim T(\cdot|s_t,\boldsymbol{a}_t), \\ \boldsymbol{o}_{t+1} \sim \boldsymbol{O}(\cdot|s_{t+1}), \boldsymbol{\theta}_{t+1} \sim \boldsymbol{U}(\cdot|\boldsymbol{\theta}_t,\boldsymbol{\tau}_t) \end{array}\right] \tag{B.14}$$

Following the Bellman equation derivation principles, we can express $v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta})$ recur-

sively:

$$v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}(\cdot|\boldsymbol{b};\boldsymbol{\theta}),s'\sim T(\cdot|s,\boldsymbol{a}),\boldsymbol{o}'\sim\boldsymbol{O}(\cdot|s'),\boldsymbol{\theta}'\sim\boldsymbol{U}(\cdot|\boldsymbol{\theta},\boldsymbol{\tau})}\left[R^i(s,\boldsymbol{a}) - \rho^i_{\theta^i} + v^i_{\theta^i}(s',\text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}')\right] \tag{B.15}$$

$$= \sum_{a^i}\pi^i(a^i|b^i;\theta^i)\sum_{\boldsymbol{a}^{-i}}\boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i})\sum_{s'}T(s'|s,\boldsymbol{a})\sum_{\boldsymbol{o}'}\boldsymbol{O}(\boldsymbol{o}'|s')\sum_{\boldsymbol{\theta}'}\boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) \tag{B.16}$$

$$\left[R^i(s,\boldsymbol{a}) - \rho^i_{\theta^i} + v^i_{\theta^i}(s',\text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}')\right] \tag{B.17}$$

where update$(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}')$ represents the belief update function that updates the belief state based on the current belief, action, and new observation. We now define the action-value function $q^i_{\theta^i}$ for the partially observable setting:

$$q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) = \sum_{s'}T(s'|s,\boldsymbol{a})\sum_{\boldsymbol{o}'}\boldsymbol{O}(\boldsymbol{o}'|s')\sum_{\boldsymbol{\theta}'}\boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau})\left[R^i(s,\boldsymbol{a}) - \rho^i_{\theta^i} + v^i_{\theta^i}(s',\text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}')\right] \tag{B.18}$$

Using this definition, we can rewrite the differential value function as:

$$v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) = \sum_{a^i}\pi^i(a^i|b^i;\theta^i)\sum_{\boldsymbol{a}^{-i}}\boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i})q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{B.19}$$

To derive the policy gradient, we begin by computing the gradient of the differential value function with respect to θ^i:

$$\nabla_{\theta^i}v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) = \nabla_{\theta^i}\left[\sum_{a^i}\pi^i(a^i|b^i;\theta^i)\sum_{\boldsymbol{a}^{-i}}\boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i})q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a})\right] \tag{B.20}$$

$$= \sum_{a^i}\nabla_{\theta^i}\pi^i(a^i|b^i;\theta^i)\sum_{\boldsymbol{a}^{-i}}\boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i})q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) + \tag{B.21}$$

$$\sum_{a^i}\pi^i(a^i|b^i;\theta^i)\sum_{\boldsymbol{a}^{-i}}\boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i})\nabla_{\theta^i}q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{B.22}$$

We expand the gradient of the action-value function:

$$\nabla_{\theta^i}q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) = \nabla_{\theta^i}\left[\sum_{s'}T(s'|s,\boldsymbol{a})\sum_{\boldsymbol{o}'}\boldsymbol{O}(\boldsymbol{o}'|s')\right. \tag{B.23}$$

$$\left.\sum_{\boldsymbol{\theta}'}\boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau})\left[R^i(s,\boldsymbol{a}) - \rho^i_{\theta^i} + v^i_{\theta^i}(s',\text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}')\right]\right] \tag{B.24}$$

$$= -\nabla_{\theta^i}\rho^i_{\theta^i} + \sum_{s'}T(s'|s,\boldsymbol{a})\sum_{\boldsymbol{o}'}\boldsymbol{O}(\boldsymbol{o}'|s')\sum_{\boldsymbol{\theta}'}\boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau})\nabla_{\theta^i}v^i_{\theta^i}(s',\text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \tag{B.25}$$

Substituting back into the original expression, we get:

$$\nabla_{\theta^i} v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) = \sum_{a^i} \nabla_{\theta^i}\pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{B.26}$$

$$- \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \nabla_{\theta^i}\rho^i_{\theta^i} \tag{B.27}$$

$$+ \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \tag{B.28}$$

$$\sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) \nabla_{\theta^i} v^i_{\theta^i}(s', \text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \tag{B.29}$$

Rearranging terms:

$$\nabla_{\theta^i}\rho^i_{\theta^i} = \sum_{a^i} \nabla_{\theta^i}\pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{B.30}$$

$$+ \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \tag{B.31}$$

$$\sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) \nabla_{\theta^i} v^i_{\theta^i}(s', \text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \tag{B.32}$$

$$- \nabla_{\theta^i} v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{B.33}$$

Now, we define the transition operator that maps the expectation of a function from one time step to the next:

$$\Psi f(s,\boldsymbol{b},\boldsymbol{\theta}) = \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \tag{B.34}$$

$$\sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) f(s', \text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \tag{B.35}$$

We can rewrite our expression as:

$$\nabla_{\theta^i}\rho^i_{\theta^i} = \sum_{a^i} \nabla_{\theta^i}\pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) + \Psi(\nabla_{\theta^i} v^i_{\theta^i})(s,\boldsymbol{b},\boldsymbol{\theta}) - \nabla_{\theta^i} v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{B.36}$$

Now, we take the expectation with respect to the stationary distribution $\mu_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta})$:

$$\sum_{s,\boldsymbol{b},\boldsymbol{\theta}} \mu_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) \nabla_{\theta^i}\rho^i_{\theta^i} = \sum_{s,\boldsymbol{b},\boldsymbol{\theta}} \mu_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) \sum_{a^i} \nabla_{\theta^i}\pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{B.37}$$

$$+ \sum_{s,\boldsymbol{b},\boldsymbol{\theta}} \mu_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) \Psi(\nabla_{\theta^i} v^i_{\theta^i})(s,\boldsymbol{b},\boldsymbol{\theta}) - \sum_{s,\boldsymbol{b},\boldsymbol{\theta}} \mu_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) \nabla_{\theta^i} v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{B.38}$$

By the definition of the stationary distribution, we have:

$$\sum_{s,\boldsymbol{b},\boldsymbol{\theta}} \mu_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta})\Psi f(s,\boldsymbol{b},\boldsymbol{\theta}) = \sum_{s,\boldsymbol{b},\boldsymbol{\theta}} \mu_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) f(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{B.39}$$

Therefore, the second and third terms in our expression cancel out, giving the final policy gradient theorem:

$$\nabla_{\theta^i} J^i_\pi(\theta^i) := \nabla_{\theta^i} \rho^i_{\theta^i} = \sum_{s,\boldsymbol{b},\boldsymbol{\theta}} \mu_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) \sum_{a^i} \nabla_{\theta^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{B.40}$$

□

PROOFS REGARDING DISCOUNTED RETURNS

We establish the Partially Observable Active Discounted Return Policy Gradient Theorem through a sequence of lemmas that build the necessary mathematical foundation.

C.1 TRANSITION OPERATOR AND ITS ADJOINT

To establish a rigorous framework for analyzing discounted returns in Partially Observable Active Markov Games, we need to formalize how value functions evolve over time and how probability distributions propagate through the system. This dual perspective is captured by two fundamental operators: the transition operator and its adjoint.

C.1.1 Definitions and Duality

We begin by defining the inner product between functions and measures, which forms the foundation for the duality relationship between our operators.

Definition 8 (Inner Product between Functions and Measures)**.** *For a bounded measurable function $f \in \mathscr{B}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$ and a finite measure $\mu \in \mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$, their inner product is defined as the Lebesgue integral:*

$$\langle f, \mu \rangle = \int_{S \times \boldsymbol{B} \times \boldsymbol{\Theta}} f(s, \boldsymbol{b}, \boldsymbol{\theta}) \, d\mu(s, \boldsymbol{b}, \boldsymbol{\theta}) \tag{C.1}$$

This inner product captures the expected value of function f with respect to measure μ, allowing us to express value functions as expectations with respect to appropriate measures.

Definition 9 (Transition Operator and Its Adjoint)**.**

1. *The **transition operator** $\Psi : \mathscr{B}(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) \to \mathscr{B}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$ maps a bounded measurable function to its expected value after one transition:*

$$(\Psi f)(s, \boldsymbol{b}, \boldsymbol{\theta}) = \sum_{a^i} \pi^i(a^i | b^i; \theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i} | \boldsymbol{b}^{-i}; \boldsymbol{\theta}^{-i}) \sum_{s'} T(s' | s, \boldsymbol{a}) \tag{C.2}$$

$$\sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}' | s') \sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}' | \boldsymbol{\theta}, \boldsymbol{\tau}) f(s', \mathit{update}(\boldsymbol{b}, \boldsymbol{a}, \boldsymbol{o}'), \boldsymbol{\theta}') \tag{C.3}$$

   where $\mathit{update}(\boldsymbol{b}, \boldsymbol{a}, \boldsymbol{o}')$ represents the belief update function.

2. *The **adjoint operator** $\Psi^* : \mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) \to \mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$ is defined on the space of finite measures such that:*

$$\langle \Psi f, \mu \rangle = \langle f, \Psi^* \mu \rangle \tag{C.4}$$

for all $f \in \mathscr{B}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$ *and* $\mu \in \mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$.

Intuitively, Ψ describes how function expectations evolve forward in time, meanwhile, Ψ^* describes how probability distributions evolve in time, capturing the propagation of visitation probabilities through the system. The duality relationship ensures that expected values can be computed either by applying Ψ to functions or Ψ^* to measures.

C.1.2 Properties of the Operators

These operators possess several important properties that facilitate our analysis of discounted returns:

Lemma 3 (Properties of the Transition Operator and Its Adjoint)**.** *The transition operator* Ψ *and its adjoint* Ψ^* *satisfy the following properties:*

1. ***Linearity:***
   - *For the operator: For any functions* $f, g \in \mathscr{B}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$ *and constants* $\alpha, \beta \in \mathbb{R}$*:*

$$\Psi(\alpha f + \beta g) = \alpha \Psi f + \beta \Psi g \tag{C.5}$$

   - *For the adjoint: For any measures* $\mu, \nu \in \mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$ *and constants* $\alpha, \beta \in \mathbb{R}$*:*

$$\Psi^*(\alpha \mu + \beta \nu) = \alpha \Psi^* \mu + \beta \Psi^* \nu \tag{C.6}$$

2. ***Positivity Preservation:***
   - *For the operator: If* $f \geq 0$*, then* $\Psi f \geq 0$*.*
   - *For the adjoint: If* μ *is a non-negative measure, then* $\Psi^* \mu$ *is also non-negative.*
3. ***Preservation of Total Measure:***
   - *For the operator: The operator preserves the constant function* $\mathbf{1}$*:*

$$\Psi \mathbf{1} = \mathbf{1} \tag{C.7}$$

     where $\mathbf{1}$ *represents the function that equals 1 everywhere.*
   - *For the adjoint: For any measure* $\mu \in \mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$*:*

$$(\Psi^* \mu)(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) = \mu(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) \tag{C.8}$$

4. ***Bound Preservation:*** *For any bounded function* f*,* Ψf *is bounded with:*

$$\|\Psi f\|_\infty \leq \|f\|_\infty \tag{C.9}$$

where $\|f\|_\infty = \sup_{s, \boldsymbol{b}, \boldsymbol{\theta}} |f(s, \boldsymbol{b}, \boldsymbol{\theta})|$ *is the supremum norm.*

5. ***Explicit Form of the Adjoint:*** *For any measurable set $A \subset S \times \boldsymbol{B} \times \boldsymbol{\Theta}$:*

$$(\Psi^* \mu)(A) = \int_{S \times \boldsymbol{B} \times \boldsymbol{\Theta}} \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \tag{C.10}$$

$$\sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) \mathbf{1}_A(s', update(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}')\, d\mu(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.11}$$

where $\mathbf{1}_A$ is the indicator function for set A.

Proof. We prove each property in turn:

1. Linearity:

For the operator: For any $(s,\boldsymbol{b},\boldsymbol{\theta})$, we have:

$$\Psi(\alpha f + \beta g)(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.12}$$

$$= \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \tag{C.13}$$

$$\sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau})[\alpha f + \beta g](s', \text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \tag{C.14}$$

$$= \alpha \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \tag{C.15}$$

$$\sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) f(s', \text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \tag{C.16}$$

$$+ \beta \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \tag{C.17}$$

$$\sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) g(s', \text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \tag{C.18}$$

$$= \alpha(\Psi f)(s,\boldsymbol{b},\boldsymbol{\theta}) + \beta(\Psi g)(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.19}$$

For the adjoint: For any measurable function $f \in \mathscr{B}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$:

$$\langle f, \Psi^*(\alpha\mu + \beta\nu)\rangle = \langle \Psi f, \alpha\mu + \beta\nu\rangle \tag{C.20}$$

$$= \alpha\langle \Psi f, \mu\rangle + \beta\langle \Psi f, \nu\rangle \tag{C.21}$$

$$= \alpha\langle f, \Psi^*\mu\rangle + \beta\langle f, \Psi^*\nu\rangle \tag{C.22}$$

$$= \langle f, \alpha\Psi^*\mu + \beta\Psi^*\nu\rangle \tag{C.23}$$

Since this holds for any measurable function f, we have $\Psi^*(\alpha\mu + \beta\nu) = \alpha\Psi^*\mu + \beta\Psi^*\nu$ by the uniqueness of the representing measure.

2. Positivity Preservation:

For the operator: If $f \geq 0$, then for any $(s, \boldsymbol{b}, \boldsymbol{\theta})$:

$$(\Psi f)(s, \boldsymbol{b}, \boldsymbol{\theta}) = \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \tag{C.24}$$

$$\sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) f(s', \text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'), \boldsymbol{\theta}') \tag{C.25}$$

Since $f \geq 0$ and all probability distributions π^i, $\boldsymbol{\pi}^{-i}$, T, $\boldsymbol{O}$, and $\boldsymbol{U}$ are non-negative, the entire sum consists of non-negative terms, making $(\Psi f)(s, \boldsymbol{b}, \boldsymbol{\theta}) \geq 0$.

For the adjoint: If μ is non-negative, then for any non-negative measurable function $f \geq 0$:

$$\langle f, \Psi^* \mu \rangle = \langle \Psi f, \mu \rangle \geq 0 \tag{C.26}$$

where the inequality follows from the positivity preservation of Ψ and the fact that μ is non-negative. Since this holds for all non-negative functions f, $\Psi^*\mu$ must be a non-negative measure.

3. Preservation of Total Measure:

For the operator: Let $\mathbf{1}$ be the constant function that equals 1 everywhere. Then:

$$(\Psi \mathbf{1})(s, \boldsymbol{b}, \boldsymbol{\theta}) = \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \tag{C.27}$$

$$\sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) \cdot 1 \tag{C.28}$$

Since each probability distribution sums to 1, applying them sequentially yields:

$$(\Psi \mathbf{1})(s, \boldsymbol{b}, \boldsymbol{\theta}) = \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \cdot 1 \tag{C.29}$$

$$= \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \cdot 1 \tag{C.30}$$

$$= \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \cdot 1 \tag{C.31}$$

$$= \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \cdot 1 = 1 \tag{C.32}$$

For the adjoint: Setting $f = \mathbf{1}$ in the adjoint relationship:

$$\langle \mathbf{1}, \Psi^* \mu \rangle = \langle \Psi \mathbf{1}, \mu \rangle = \langle \mathbf{1}, \mu \rangle \tag{C.33}$$

where we used the unit preservation property of Ψ. This gives:

$$(\Psi^* \mu)(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) = \mu(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) \tag{C.34}$$

4. Bound Preservation:

For any bounded function f with $\|f\|_\infty = \sup_{s,\boldsymbol{b},\boldsymbol{\theta}} |f(s,\boldsymbol{b},\boldsymbol{\theta})|$:

$$|(\Psi f)(s,\boldsymbol{b},\boldsymbol{\theta})| = \Bigg| \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \tag{C.35}$$

$$\sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) f(s', \text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \Bigg| \tag{C.36}$$

$$\leq \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \tag{C.37}$$

$$\sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) \left| f(s', \text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \right| \tag{C.38}$$

$$\leq \|f\|_\infty \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \tag{C.39}$$

$$\sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) = \|f\|_\infty \cdot 1 = \|f\|_\infty \tag{C.40}$$

Taking the supremum over all $(s,\boldsymbol{b},\boldsymbol{\theta})$, we get $\|\Psi f\|_\infty \leq \|f\|_\infty$.

5. Explicit Form of the Adjoint:

For any measurable set $A \subset S \times \boldsymbol{B} \times \boldsymbol{\Theta}$:

$$(\Psi^* \mu)(A) = \langle \mathbf{1}_A, \Psi^* \mu \rangle = \langle \Psi \mathbf{1}_A, \mu \rangle \tag{C.41}$$

Expanding $\Psi \mathbf{1}_A$ using the definition of the transition operator:

$$(\Psi \mathbf{1}_A)(s,\boldsymbol{b},\boldsymbol{\theta}) = \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \tag{C.42}$$

$$\sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) \mathbf{1}_A(s', \text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \tag{C.43}$$

Substituting this back and using the definition of inner product:

$$(\Psi^* \mu)(A) = \int_{S \times \boldsymbol{B} \times \boldsymbol{\Theta}} (\Psi \mathbf{1}_A)(s,\boldsymbol{b},\boldsymbol{\theta})\, d\mu(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.44}$$

$$= \int_{S \times \boldsymbol{B} \times \boldsymbol{\Theta}} \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \sum_{s'} T(s'|s,\boldsymbol{a}) \tag{C.45}$$

$$\sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau}) \mathbf{1}_A(s', \text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}')\, d\mu(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.46}$$

□

C.1.3 Contraction and Propagation

The properties established above lead to two crucial results: the contraction property of the discounted transition operator and the propagation of distributions through the adjoint.

Lemma 4 (Contraction Mapping Property)**.** *For any discount factor $\gamma \in [0,1)$, the operator $\gamma\Psi$ is a contraction mapping on the Banach space of bounded functions on $S \times \boldsymbol{B} \times \boldsymbol{\Theta}$ equipped*

with the supremum norm. Specifically, for any two functions $f, g \in \mathscr{B}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$:

$$\|\gamma\Psi f - \gamma\Psi g\|_\infty \leq \gamma\|f - g\|_\infty \tag{C.47}$$

Proof. Using the linearity of Ψ and its bound preservation property:

$$\|\gamma\Psi f - \gamma\Psi g\|_\infty = \gamma\|\Psi(f - g)\|_\infty \tag{C.48}$$

$$\leq \gamma\|f - g\|_\infty \tag{C.49}$$

Since $\gamma < 1$, $\gamma\Psi$ is a contraction mapping with contraction factor γ. □

This contraction property leads directly to the invertibility of $I - \gamma\Psi$, a result that is fundamental for characterizing value functions in discounted settings.

Lemma 5 (Invertibility of $I - \gamma\Psi$)**.** *For any discount factor $\gamma \in [0, 1)$, the operator $I - \gamma\Psi$ is invertible, and its inverse can be represented as a Neumann series:*

$$(I - \gamma\Psi)^{-1} = \sum_{t=0}^{\infty} \gamma^t \Psi^t \tag{C.50}$$

which converges absolutely in the operator norm.

Proof. By the Banach fixed-point theorem, if T is a contraction mapping on a Banach space, then $I - T$ is invertible, with inverse given by the Neumann series $\sum_{k=0}^{\infty} T^k$. Since $\gamma\Psi$ is a contraction mapping, we have:

$$(I - \gamma\Psi)^{-1} = \sum_{t=0}^{\infty} (\gamma\Psi)^t = \sum_{t=0}^{\infty} \gamma^t \Psi^t \tag{C.51}$$

The absolute convergence follows from:

$$\left\| \sum_{t=0}^{\infty} \gamma^t \Psi^t \right\|_\infty \leq \sum_{t=0}^{\infty} \gamma^t \|\Psi^t\|_\infty \leq \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1 - \gamma} < \infty \tag{C.52}$$

where we use the bound preservation property to establish that $\|\Psi^t\|_\infty \leq 1$ for all t. □

Finally, we establish how distributions evolve in the system through the adjoint operator, formalizing the propagation of probability measures.

Lemma 6 (Distribution Propagation)**.** *If μ_t represents the distribution of the joint process $(s_t, \boldsymbol{b}_t, \boldsymbol{\theta}_t)$ at time t, then the distribution at time $t + 1$ is given by $\mu_{t+1} = \Psi^* \mu_t$.*

Proof. For any measurable function $f \in \mathscr{B}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$:

$$\langle f, \mu_{t+1} \rangle = \mathbb{E}[f(s_{t+1}, \boldsymbol{b}_{t+1}, \boldsymbol{\theta}_{t+1})] \tag{C.53}$$

$$= \mathbb{E}[\mathbb{E}[f(s_{t+1}, \boldsymbol{b}_{t+1}, \boldsymbol{\theta}_{t+1}) \mid s_t, \boldsymbol{b}_t, \boldsymbol{\theta}_t]] \tag{C.54}$$

$$= \mathbb{E}[(\Psi f)(s_t, \boldsymbol{b}_t, \boldsymbol{\theta}_t)] \tag{C.55}$$

$$= \langle \Psi f, \mu_t \rangle = \langle f, \Psi^* \mu_t \rangle \tag{C.56}$$

Since this holds for any measurable function f, we have $\mu_{t+1} = \Psi^* \mu_t$ by the uniqueness of the representing measure. □

The concepts introduced in this section—the transition operator, its adjoint, and their properties—provide the mathematical foundation for analyzing discounted returns in Partially Observable Active Markov Games. The transition operator Ψ captures how function expectations evolve forward in time, while its adjoint Ψ^* describes how probability distributions propagate forward in time. This duality is fundamental to establishing the connection between value-based and distribution-based perspectives in reinforcement learning.

In the next section, we will use these tools to formulate Bellman equations and derive policy gradients for maximizing discounted returns in this complex multi-agent setting.

C.2 Bellman Equations and Value Functions

In this section, we establish the fundamental recursive relationships that characterize the discounted value functions in Partially Observable Active Markov Games. These relationships form the basis for our policy gradient derivation and practical algorithms for optimizing agent behavior. We begin by confirming the well-definedness of value functions, then establish the Bellman equations, and finally derive the gradient expressions necessary for policy optimization.

C.2.1 Well-Definedness of Value Functions

Before deriving recursive relationships, we first establish that the value functions are well-defined mathematical objects under our assumptions of bounded rewards and discount factors less than 1.

Lemma 7 (Well-Defined Value Functions)**.** *Under the assumptions of bounded rewards and $\gamma < 1$, the discounted value functions $v^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta})$ and action-value functions $q^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{a})$ are well-defined and finite for all states, belief states, policy parameters, and actions.*

Proof. Assuming rewards are bounded such that $|R^i(s, \boldsymbol{a})| \leq R_{\max}$ for all $s \in S$ and $\boldsymbol{a} \in A$, we

have:

$$|v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta})| = \left|\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R^i(s_t,\boldsymbol{a}_t)\middle| s_0=s,\boldsymbol{b}_0=\boldsymbol{b},\boldsymbol{\theta}_0=\boldsymbol{\theta},\pi\right]\right| \tag{C.57}$$

$$\leq \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t |R^i(s_t,\boldsymbol{a}_t)|\middle| s_0=s,\boldsymbol{b}_0=\boldsymbol{b},\boldsymbol{\theta}_0=\boldsymbol{\theta},\pi\right] \tag{C.58}$$

$$\leq R_{\max}\sum_{t=0}^{\infty}\gamma^t = \frac{R_{\max}}{1-\gamma} < \infty \tag{C.59}$$

Similarly, for the action-value function:

$$|q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a})| = \left|R^i(s,\boldsymbol{a}) + \gamma\mathbb{E}[v^i_{\theta^i}(s',\boldsymbol{b}',\boldsymbol{\theta}')]\right| \tag{C.60}$$

$$\leq |R^i(s,\boldsymbol{a})| + \gamma\mathbb{E}[|v^i_{\theta^i}(s',\boldsymbol{b}',\boldsymbol{\theta}')|] \tag{C.61}$$

$$\leq R_{\max} + \gamma\frac{R_{\max}}{1-\gamma} = \frac{R_{\max}}{1-\gamma} < \infty \tag{C.62}$$

Thus, both value functions are well-defined and bounded. □

This lemma ensures that our subsequent derivations involving value functions are mathematically sound. The boundedness property is particularly important when we consider expectations and derivatives of these functions.

C.2.2 Bellman Equations for Discounted Value Functions

Next, we establish the recursive relationships between value functions, which are fundamental to dynamic programming approaches.

Lemma 8 (Bellman Equation for Discounted Value Functions)**.** *In a Partially Observable Active Markov Game, the discounted value function $v^i_{\theta^i}$ satisfies the Bellman equation:*

$$v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) = \sum_{a^i}\pi^i(a^i|b^i;\theta^i)\sum_{\boldsymbol{a}^{-i}}\boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i})q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{C.63}$$

where the action-value function $q^i_{\theta^i}$ is defined as:

$$q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) = R^i(s,\boldsymbol{a}) + \gamma\sum_{s'}T(s'|s,\boldsymbol{a})\sum_{\boldsymbol{o}'}\boldsymbol{O}(\boldsymbol{o}'|s')$$
$$\sum_{\boldsymbol{\theta}'}\boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau})v^i_{\theta^i}(s',update(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \tag{C.64}$$

Proof. By definition, the discounted value function is:

$$v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R^i(s_t,\boldsymbol{a}_t)\middle| s_0=s,\boldsymbol{b}_0=\boldsymbol{b},\boldsymbol{\theta}_0=\boldsymbol{\theta},\pi\right] \tag{C.65}$$

We can decompose this into the immediate reward and the future discounted returns:

$$v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) = \mathbb{E}\left[R^i(s_0,\boldsymbol{a}_0) + \gamma\sum_{t=1}^{\infty}\gamma^{t-1}R^i(s_t,\boldsymbol{a}_t)\middle| s_0 = s, \boldsymbol{b}_0 = \boldsymbol{b}, \boldsymbol{\theta}_0 = \boldsymbol{\theta}, \pi\right] \tag{C.66}$$

$$= \mathbb{E}\left[R^i(s_0,\boldsymbol{a}_0)\middle| s_0 = s, \boldsymbol{b}_0 = \boldsymbol{b}, \boldsymbol{\theta}_0 = \boldsymbol{\theta}, \pi\right] + \gamma\mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1}R^i(s_t,\boldsymbol{a}_t)\middle| s_0 = s, \boldsymbol{b}_0 = \boldsymbol{b}, \boldsymbol{\theta}_0 = \boldsymbol{\theta}, \pi\right] \tag{C.67}$$

For the first term, we have:

$$\mathbb{E}\left[R^i(s_0,\boldsymbol{a}_0)\middle| s_0 = s, \boldsymbol{b}_0 = \boldsymbol{b}, \boldsymbol{\theta}_0 = \boldsymbol{\theta}, \pi\right] = \sum_{a^i}\pi^i(a^i|b^i;\theta^i)\sum_{\boldsymbol{a}^{-i}}\boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i})R^i(s,\boldsymbol{a}) \tag{C.68}$$

For the second term, by the law of total expectation:

$$\mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1}R^i(s_t,\boldsymbol{a}_t)\middle| s_0 = s, \boldsymbol{b}_0 = \boldsymbol{b}, \boldsymbol{\theta}_0 = \boldsymbol{\theta}, \pi\right] \tag{C.69}$$

$$= \sum_{a^i}\pi^i(a^i|b^i;\theta^i)\sum_{\boldsymbol{a}^{-i}}\boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i})\sum_{s'}T(s'|s,\boldsymbol{a}) \sum_{\boldsymbol{o}'}\boldsymbol{O}(\boldsymbol{o}'|s')\sum_{\boldsymbol{\theta}'}\boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau})v^i_{\theta^i}(s',\text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \tag{C.70}$$

Combining these two terms and factoring out common terms:

$$v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) = \sum_{a^i}\pi^i(a^i|b^i;\theta^i)\sum_{\boldsymbol{a}^{-i}}\boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i})\Bigg[R^i(s,\boldsymbol{a}) + \gamma\sum_{s'}T(s'|s,\boldsymbol{a})\sum_{\boldsymbol{o}'}\boldsymbol{O}(\boldsymbol{o}'|s') \sum_{\boldsymbol{\theta}'}\boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau})v^i_{\theta^i}(s',\text{update}(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}')\Bigg] \tag{C.71}$$

$$= \sum_{a^i}\pi^i(a^i|b^i;\theta^i)\sum_{\boldsymbol{a}^{-i}}\boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i})q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{C.72}$$

where we define the action-value function $q^i_{\theta^i}$ as specified. □

The Bellman equation provides a recursive characterization of the value function that forms the basis for dynamic programming algorithms. This equation captures how current actions influence both immediate rewards and future value through their effects on the environment state, belief states, and policy parameters of other agents. The key distinction from standard Bellman equations is the inclusion of belief updates and policy parameter dynamics, which reflects the complex dependencies in partially observable multi-agent settings.

C.2.3 Policy Gradient with Respect to Value Function

Having established the recursive relationship for value functions, we now derive how the value function gradient depends on policy parameters, which is essential for policy optimization.

Lemma 9 (Policy Gradient with Respect to Value Function)**.** *The gradient of the value function with respect to policy parameters θ^i can be expressed as:*

$$\nabla_{\theta^i} v^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}) = (I - \gamma\Psi)^{-1} \left[\sum_{a^i} \nabla_{\theta^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{a}) \right] \tag{C.73}$$

where Ψ is the transition operator defined in the previous section.

Proof. Taking the gradient of the Bellman equation with respect to θ^i:

$$\nabla_{\theta^i} v^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}) = \nabla_{\theta^i} \left[\sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{a}) \right] \tag{C.74}$$

Applying the product rule:

$$\begin{aligned} \nabla_{\theta^i} v^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}) = &\sum_{a^i} \nabla_{\theta^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{a}) \\ &+ \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) \nabla_{\theta^i} q^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{a}) \end{aligned} \tag{C.75}$$

For the gradient of the action-value function:

$$\nabla_{\theta^i} q^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{a}) = \nabla_{\theta^i} \left[R^i(s, \boldsymbol{a}) + \gamma \sum_{s'} T(s'|s, \boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta}, \boldsymbol{\tau}) v^i_{\theta^i}(s', \text{update}(\boldsymbol{b}, \boldsymbol{a}, \boldsymbol{o}'), \boldsymbol{\theta}') \right] \tag{C.76}$$

Since $R^i(s, \boldsymbol{a})$, $T(s'|s, \boldsymbol{a})$, $\boldsymbol{O}(\boldsymbol{o}'|s')$, and $\boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta}, \boldsymbol{\tau})$ do not depend directly on θ^i:

$$\begin{aligned} \nabla_{\theta^i} q^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{a}) = &\gamma \sum_{s'} T(s'|s, \boldsymbol{a}) \sum_{\boldsymbol{o}'} \boldsymbol{O}(\boldsymbol{o}'|s') \\ &\sum_{\boldsymbol{\theta}'} \boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta}, \boldsymbol{\tau}) \nabla_{\theta^i} v^i_{\theta^i}(s', \text{update}(\boldsymbol{b}, \boldsymbol{a}, \boldsymbol{o}'), \boldsymbol{\theta}') \end{aligned} \tag{C.77}$$

Using the transition operator Ψ, we can rewrite:

$$\begin{aligned} \nabla_{\theta^i} v^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}) = &\sum_{a^i} \nabla_{\theta^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{a}) \\ &+ \gamma\Psi[\nabla_{\theta^i} v^i_{\theta^i}](s, \boldsymbol{b}, \boldsymbol{\theta}) \end{aligned} \tag{C.78}$$

Rearranging:

$$(I-\gamma\Psi)[\nabla_{\theta^i} v^i_{\theta^i}](s,\boldsymbol{b},\boldsymbol{\theta}) = \sum_{a^i} \nabla_{\theta^i}\pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{C.79}$$

By the lemma on the invertibility of $(I-\gamma\Psi)$, we have:

$$\nabla_{\theta^i} v^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta}) = (I-\gamma\Psi)^{-1}\left[\sum_{a^i} \nabla_{\theta^i}\pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a})\right] \tag{C.80}$$

$$= \sum_{t=0}^{\infty} \gamma^t \Psi^t \left[\sum_{a^i} \nabla_{\theta^i}\pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a})\right] \tag{C.81}$$

□

This lemma expresses the value function gradient in terms of an infinite sum of expected Q-values weighted by policy gradients and propagated through the transition operator. The operator $(I-\gamma\Psi)^{-1}$ captures how immediate changes in policy parameters propagate through future timesteps. This form of the gradient is central to our discounted policy gradient theorem, which we will develop in the next section. The inverse operator can be represented as a Neumann series $\sum_{t=0}^{\infty}\gamma^t\Psi^t$, which has a natural interpretation as the accumulated effect of policy changes over an infinite horizon, weighted by the discount factor.

C.3 Discounted Visitation Measure

The infinite sum formulation derived in the previous sections, while mathematically sound, is not directly amenable to practical computation. In this section, we reformulate these expressions in terms of a probability distribution, which allows for more efficient estimation through sampling-based approaches. This reformulation is central to our policy gradient theorem for discounted returns.

C.3.1 Definition and Properties

Definition 10 (Discounted Visitation Measure)**.** *For a Markov process with an initial distribution μ_0 over the joint state-belief-policy space, the discounted visitation measure $d^{\pi}_{\mu_0}$ is defined as:*

$$d^{\pi}_{\mu_0} := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mu_t \tag{C.82}$$

where μ_t is the distribution at time t when starting from μ_0, and $\gamma \in [0,1)$ is the discount factor.

First, we show that the discounted visitation measure is a proper probability distribution.

Lemma 10 (DVM is a Probability Distribution)**.** *The discounted visitation measure $d^\pi_{\mu_0}$ is a probability distribution, i.e., $d^\pi_{\mu_0}(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) = 1$ when μ_0 is a probability distribution.*

Proof. Since μ_0 is a probability distribution, we have $\mu_0(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) = 1$. For any $t \geq 0$, the measure μ_t is obtained by applying the adjoint operator $(\Psi^*)^t$ to μ_0. Since Ψ^* is a Markov operator (preserves probability mass), we have $\mu_t(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) = 1$ for all $t \geq 0$.

Therefore, the total mass of $d^\pi_{\mu_0}$ is:

$$d^\pi_{\mu_0}(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mu_t(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) \tag{C.83}$$

$$= (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \cdot 1 \tag{C.84}$$

$$= (1-\gamma) \cdot \frac{1}{1-\gamma} \tag{C.85}$$

$$= 1 \tag{C.86}$$

Since $d^\pi_{\mu_0}$ is non-negative by construction (being a convex combination of non-negative measures with positive weights) and has total mass 1, it is indeed a probability distribution. □

This measure represents the expected discounted frequency with which different states, belief states, and policy parameters are visited when following policy π starting from the initial distribution μ_0. The normalization factor $(1-\gamma)$ ensures that $d^\pi_{\mu_0}$ is a proper probability distribution, summing to 1 across the entire state-belief-policy space.

Lemma 11 (Evolution of Discounted Visitation)**.** *The discounted visitation measure $d^\pi_{\mu_0}$ can be expressed in terms of the adjoint operator as:*

$$d^\pi_{\mu_0} = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t (\Psi^*)^t \mu_0 = (1-\gamma)(I - \gamma\Psi^*)^{-1} \mu_0 \tag{C.87}$$

where μ_0 is the initial distribution.

Proof. By the propagation lemma established earlier, $\mu_t = (\Psi^*)^t \mu_0$ represents the distribution at time t when starting from μ_0. The discounted visitation measure weights these distributions by γ^t and normalizes by $(1-\gamma)$ to ensure it's a proper probability measure:

$$d^\pi_{\mu_0} = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mu_t \tag{C.88}$$

$$= (1-\gamma) \sum_{t=0}^{\infty} \gamma^t (\Psi^*)^t \mu_0 \tag{C.89}$$

The second equality follows from the Neumann series expansion:

$$(I - \gamma\Psi^*)^{-1} = \sum_{t=0}^{\infty} (\gamma\Psi^*)^t = \sum_{t=0}^{\infty} \gamma^t (\Psi^*)^t \tag{C.90}$$

which converges for $\gamma < 1$ since Ψ^* is a non-expansion operator in the total variation norm. □

This establishes the essential connection between the adjoint operator and the discounted visitation measure, which is central to the policy gradient theorem. The measure $d^{\pi}_{\mu_0}$ represents the normalized expected discounted time spent in each state-belief-policy configuration, weighted according to the discount factor γ. This connection allows us to transform expressions involving $(I - \gamma\Psi)^{-1}$ in the value function to expectations with respect to $d^{\pi}_{\mu_0}$, which is key for deriving practical policy gradient algorithms.In practical implementations, the discounted visitation measure is rarely computed explicitly due to the prohibitive computational cost. Instead, sampling-based approaches are used to estimate expectations with respect to this measure. The policy gradient theorem leverages this measure to derive update rules that can be efficiently implemented using trajectory samples collected from the environment.

C.3.2 Existence and Uniqueness

To provide a rigorous mathematical foundation for our policy gradient theorem, we establish the existence and uniqueness of the discounted visitation measure.

Lemma 12 (Existence of the Discounted Visitation Measure)**.** *Let $\mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$ be the space of finite measures on the joint state-belief-policy space, equipped with the total variation norm. For any initial measure $\mu_0 \in \mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$ and discount factor $\gamma \in [0,1)$, the discounted state-visitation measure*

$$d^{\pi}_{\mu_0} = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t (\Psi^*)^t \mu_0 \tag{C.91}$$

exists and is a well-defined measure in $\mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$.

Proof. To establish the existence of $d^{\pi}_{\mu_0}$, we need to prove that the infinite sum $(1-\gamma)\sum_{t=0}^{\infty} \gamma^t (\Psi^*)^t \mu_0$ converges to a well-defined measure. First, let's recall that $\mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$ equipped with the total variation norm $\|\mu\|_{TV} = \sup_{|f| \leq 1} |\int f d\mu|$ is a Banach space. Now, we need to show that the series $\sum_{t=0}^{\infty} \gamma^t (\Psi^*)^t \mu_0$ converges in this space. For any $t \geq 0$, $(\Psi^*)^t \mu_0$ represents the measure over the joint state-belief-policy space after t transitions, starting from the initial measure μ_0. Since Ψ^* is a Markov operator (the adjoint of a Markov transition kernel), it preserves the total mass of a measure: if μ_0 is a probability measure (i.e., $\mu_0(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) = 1$), then so is $(\Psi^*)^t \mu_0$ for all $t \geq 0$. Therefore, $\|(\Psi^*)^t \mu_0\|_{TV} = \mu_0(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$ for all $t \geq 0$. For a general finite measure μ_0 with total mass $M = \mu_0(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$, we have $\|(\Psi^*)^t \mu_0\|_{TV} = M$ for all $t \geq 0$.

Now, consider the partial sum:

$$S_n = (1-\gamma)\sum_{t=0}^{n} \gamma^t (\Psi^*)^t \mu_0 \tag{C.92}$$

For $n < m$, the difference between two partial sums is:

$$\|S_m - S_n\|_{TV} = \left\| (1-\gamma) \sum_{t=n+1}^{m} \gamma^t (\Psi^*)^t \mu_0 \right\|_{TV} \tag{C.93}$$

$$\leq (1-\gamma) \sum_{t=n+1}^{m} \gamma^t \|(\Psi^*)^t \mu_0\|_{TV} \tag{C.94}$$

$$= (1-\gamma) M \sum_{t=n+1}^{m} \gamma^t \tag{C.95}$$

$$= (1-\gamma) M (\gamma^{n+1} + \gamma^{n+2} + \ldots + \gamma^m) \tag{C.96}$$

$$= (1-\gamma) M \gamma^{n+1} \frac{1-\gamma^{m-n}}{1-\gamma} \tag{C.97}$$

$$= M \gamma^{n+1} (1-\gamma^{m-n}) \tag{C.98}$$

As $n \to \infty$, $\gamma^{n+1} \to 0$ (since $\gamma < 1$), which means that the sequence of partial sums $\{S_n\}$ is Cauchy in the Banach space $\mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$. By the completeness of this space, the sequence converges to a measure which we denote as $d^{\pi}_{\mu_0}$. Moreover, the total mass of $d^{\pi}_{\mu_0}$ is:

$$d^{\pi}_{\mu_0}(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t (\Psi^*)^t \mu_0 (S \times \boldsymbol{B} \times \boldsymbol{\Theta}) \tag{C.99}$$

$$= (1-\gamma) M \sum_{t=0}^{\infty} \gamma^t \tag{C.100}$$

$$= (1-\gamma) M \frac{1}{1-\gamma} \tag{C.101}$$

$$= M \tag{C.102}$$

Thus, $d^{\pi}_{\mu_0}$ is a finite measure with the same total mass as μ_0. □

Having established existence, we now turn to the uniqueness of the discounted visitation measure.

Lemma 13 (Uniqueness of the Discounted Visitation Measure)**.** *The discounted state-visitation measure $d^{\pi}_{\mu_0}$ is the unique solution to the functional equation:*

$$d^{\pi}_{\mu_0} = (1-\gamma)\mu_0 + \gamma \Psi^* d^{\pi}_{\mu_0} \tag{C.103}$$

in the space of finite measures $\mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$.

Proof. First, we verify that $d^{\pi}_{\mu_0} = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t (\Psi^*)^t \mu_0$ satisfies the functional equation:

$$(1-\gamma)\mu_0 + \gamma\Psi^* d^\pi_{\mu_0} = (1-\gamma)\mu_0 + \gamma\Psi^* \left((1-\gamma) \sum_{t=0}^{\infty} \gamma^t (\Psi^*)^t \mu_0 \right) \tag{C.104}$$

$$= (1-\gamma)\mu_0 + (1-\gamma) \sum_{t=1}^{\infty} \gamma^t (\Psi^*)^t \mu_0 \tag{C.105}$$

$$= (1-\gamma) \left(\mu_0 + \sum_{t=1}^{\infty} \gamma^t (\Psi^*)^t \mu_0 \right) \tag{C.106}$$

$$= (1-\gamma) \sum_{t=0}^{\infty} \gamma^t (\Psi^*)^t \mu_0 \tag{C.107}$$

$$= d^\pi_{\mu_0} \tag{C.108}$$

To prove uniqueness, suppose there exists another measure $\tilde{d} \in \mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$ that satisfies the functional equation:

$$\tilde{d} = (1-\gamma)\mu_0 + \gamma\Psi^* \tilde{d} \tag{C.109}$$

Let's consider the operator $T : \mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta}) \to \mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$ defined as:

$$T(\mu) = (1-\gamma)\mu_0 + \gamma\Psi^* \mu \tag{C.110}$$

For any two measures $\mu, \nu \in \mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$:

$$\|T(\mu) - T(\nu)\|_{TV} = \|\gamma\Psi^*\mu - \gamma\Psi^*\nu\|_{TV} \tag{C.111}$$

$$= \gamma\|\Psi^*(\mu - \nu)\|_{TV} \tag{C.112}$$

$$\leq \gamma\|\mu - \nu\|_{TV} \tag{C.113}$$

Since $\gamma < 1$, T is a contraction mapping on the Banach space $\mathscr{M}(S \times \boldsymbol{B} \times \boldsymbol{\Theta})$ with contraction factor γ. By the Banach fixed-point theorem, T has a unique fixed point. We've already shown that $d^\pi_{\mu_0}$ is a fixed point of T, and by assumption, $\tilde{d}$ is also a fixed point. Therefore, $d^\pi_{\mu_0} = \tilde{d}$, proving the uniqueness of the discounted state-visitation measure. □

C.3.3 Connection to Value Function

Having established the theoretical properties of the discounted visitation measure, we now connect it to the value function, which will provide important insights later on.

Lemma 14 (Value Function as Inner Product)**.** *For any discount factor $\gamma \in [0,1)$ and initial state s_0, belief state $\boldsymbol{b}_0$, and policy parameter $\boldsymbol{\theta}_0$, the value function can be expressed as:*

$$v^i_{\theta^i}(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0) = \frac{1}{1-\gamma} \int_{S \times \boldsymbol{B} \times \boldsymbol{\Theta}} r^i(s, \boldsymbol{b}, \boldsymbol{\theta}) d^\pi_{\mu_0}(s, \boldsymbol{b}, \boldsymbol{\theta}) \tag{C.114}$$

where $\mu_0 = \delta_{(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0)}$ is the Dirac measure concentrated at $(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0)$, and r^i is the expected immediate reward function:

$$r^i(s,\boldsymbol{b},\boldsymbol{\theta}) = \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) R^i(s,\boldsymbol{a}) \tag{C.115}$$

Proof. Let $\mu_0 = \delta_{(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0)}$ be the Dirac measure concentrated at the point $(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0)$. By definition of the value function:

$$v^i_{\theta^i}(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^i(s_t,\boldsymbol{a}_t) \middle| s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0,\pi\right] \tag{C.116}$$

$$= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[R^i(s_t,\boldsymbol{a}_t) \middle| s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0,\pi\right] \tag{C.117}$$

Define the expected immediate reward function $r^i : S \times \boldsymbol{B} \times \boldsymbol{\Theta} \to \mathbb{R}$ as:

$$r^i(s,\boldsymbol{b},\boldsymbol{\theta}) = \sum_{a^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) R^i(s,\boldsymbol{a}) \tag{C.118}$$

Then:

$$v^i_{\theta^i}(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[r^i(s_t,\boldsymbol{b}_t,\boldsymbol{\theta}_t) \middle| s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0,\pi\right] \tag{C.119}$$

$$= \sum_{t=0}^{\infty} \gamma^t \int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}} r^i(s,\boldsymbol{b},\boldsymbol{\theta}) d\mu_t(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.120}$$

where μ_t is the distribution over the joint state-belief-policy space at time t, starting from μ_0. From previous lemmas, we know that $\mu_t = (\Psi^*)^t \mu_0$. Using the duality relationship between functions and measures:

$$\int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}} r^i(s,\boldsymbol{b},\boldsymbol{\theta}) d\mu_t(s,\boldsymbol{b},\boldsymbol{\theta}) = \int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}} r^i(s,\boldsymbol{b},\boldsymbol{\theta}) d((\Psi^*)^t \mu_0)(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.121}$$

$$= \int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}} (\Psi^t r^i)(s,\boldsymbol{b},\boldsymbol{\theta}) d\mu_0(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.122}$$

Therefore:

$$v^i_{\theta^i}(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0) = \sum_{t=0}^{\infty} \gamma^t \int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}} (\Psi^t r^i)(s,\boldsymbol{b},\boldsymbol{\theta}) d\mu_0(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.123}$$

$$= \int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}} \left(\sum_{t=0}^{\infty} \gamma^t (\Psi^t r^i)(s,\boldsymbol{b},\boldsymbol{\theta})\right) d\mu_0(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.124}$$

$$= \int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}} ((I-\gamma\Psi)^{-1} r^i)(s,\boldsymbol{b},\boldsymbol{\theta}) d\mu_0(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.125}$$

Since $\mu_0 = \delta_{(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0)}$, we have:

$$v^i_{\theta^i}(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0) = ((I-\gamma\Psi)^{-1} r^i)(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0) \tag{C.126}$$

Alternatively, using the duality between $(I-\gamma\Psi)^{-1}$ and $(I-\gamma\Psi^*)^{-1}$:

$$\int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}}((I-\gamma\Psi)^{-1}r^i)(s,\boldsymbol{b},\boldsymbol{\theta})d\mu_0(s,\boldsymbol{b},\boldsymbol{\theta})=\int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}}r^i(s,\boldsymbol{b},\boldsymbol{\theta})d((I-\gamma\Psi^*)^{-1}\mu_0)(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.127}$$

From previous lemmas, we know that $d^\pi_{\mu_0}=(1-\gamma)(I-\gamma\Psi^*)^{-1}\mu_0$. Therefore:

$$\begin{aligned} v^i_{\theta^i}(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0) &= \int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}} r^i(s,\boldsymbol{b},\boldsymbol{\theta})d((I-\gamma\Psi^*)^{-1}\mu_0)(s,\boldsymbol{b},\boldsymbol{\theta}) && \text{(C.128)}\\ &= \int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}} r^i(s,\boldsymbol{b},\boldsymbol{\theta})\frac{d^\pi_{\mu_0}(s,\boldsymbol{b},\boldsymbol{\theta})}{1-\gamma} && \text{(C.129)}\\ &= \frac{1}{1-\gamma}\int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}} r^i(s,\boldsymbol{b},\boldsymbol{\theta})d^\pi_{\mu_0}(s,\boldsymbol{b},\boldsymbol{\theta}) && \text{(C.130)} \end{aligned}$$

□

This lemma establishes that the value function can be expressed as an expectation with respect to the discounted visitation measure, with an appropriate scaling factor. This formulation provides a direct link between the value function and the distribution of states.

C.4 Policy Gradient Theorem

Having established the mathematical foundations for discounted returns in partially observable multi-agent settings, we now derive the policy gradient theorem that forms the basis for practical optimization algorithms. This theorem provides a principled expression for computing gradients of the expected discounted return with respect to policy parameters, enabling efficient policy improvement. The policy gradient theorem connects an agent's policy parameters to its expected long-term performance through the discounted visitation measure, providing a mathematically rigorous foundation for optimization.

Theorem 3 (Partially Observable Active Discounted Return Policy Gradient Theorem)**.** *The gradient of the discounted return objective $J^i_{\pi,\gamma}(\theta^i)=v^i_{\theta^i}(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0)$ with respect to agent i's policy parameters θ^i in a partially observable active Markov game setting can be expressed as:*

$$\nabla_{\theta^i}J^i_{\pi,\gamma}(\theta^i)=\frac{1}{1-\gamma}\sum_{s,\boldsymbol{b},\boldsymbol{\theta}}d^\pi_{\mu_0}(s,\boldsymbol{b},\boldsymbol{\theta})\sum_{a^i}\nabla_{\theta^i}\pi^i(a^i|b^i;\theta^i)\sum_{a^{-i}}\pi^{-i}(a^{-i}|\boldsymbol{b}^{-i};\theta^{-i})q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{C.131}$$

where $d^\pi_{\mu_0}(s,\boldsymbol{b},\boldsymbol{\theta})$ is the discounted visitation measure starting from initial distribution μ_0, and $q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a})$ is the action-value function defined as:

$$\begin{aligned} q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) = R^i(s,\boldsymbol{a})+\gamma\sum_{s'}T(s'|s,\boldsymbol{a})\sum_{\boldsymbol{o}'}\boldsymbol{O}(\boldsymbol{o}'|s')\\ \sum_{\boldsymbol{\theta}'}\boldsymbol{U}(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\tau})v^i_{\theta^i}(s',update(\boldsymbol{b},\boldsymbol{a},\boldsymbol{o}'),\boldsymbol{\theta}') \end{aligned} \tag{C.132}$$

with update$(\boldsymbol{b}, \boldsymbol{a}, \boldsymbol{o}')$ *representing the belief update function.*

We now provide a detailed proof of the policy gradient theorem, building on the lemmas established in previous sections.

Proof. From our previous lemma, we know that the gradient of the value function can be expressed as:

$$\nabla_{\theta^i} v^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}) = (I - \gamma\Psi)^{-1} \left[\sum_{a^i} \nabla_{\theta^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{a}) \right] \tag{C.133}$$

For notational clarity, we define the term inside the brackets as:

$$g(s, \boldsymbol{b}, \boldsymbol{\theta}) = \sum_{a^i} \nabla_{\theta^i} \pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{a}) \tag{C.134}$$

Now we can write:

$$\nabla_{\theta^i} v^i_{\theta^i}(s, \boldsymbol{b}, \boldsymbol{\theta}) = (I - \gamma\Psi)^{-1} g(s, \boldsymbol{b}, \boldsymbol{\theta}) \tag{C.135}$$

For our specific initial state configuration $(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0)$:

$$\nabla_{\theta^i} v^i_{\theta^i}(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0) = ((I - \gamma\Psi)^{-1} g)(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0) \tag{C.136}$$

Let's define $\mu_0 = \delta_{(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0)}$, which is the Dirac measure concentrated at the initial state. We can view $((I - \gamma\Psi)^{-1} g)(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0)$ as the integration of the function $(I - \gamma\Psi)^{-1} g$ against this Dirac measure:

$$\nabla_{\theta^i} v^i_{\theta^i}(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0) = ((I - \gamma\Psi)^{-1} g)(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0) \tag{C.137}$$

$$= \int ((I - \gamma\Psi)^{-1} g)(s, \boldsymbol{b}, \boldsymbol{\theta}) d\delta_{(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0)}(s, \boldsymbol{b}, \boldsymbol{\theta}) \tag{C.138}$$

Recall the duality principle we established earlier, which states that:

$$\int ((I - \gamma\Psi)^{-1} f)(x) d\mu(x) = \int f(x) d((I - \gamma\Psi^*)^{-1} \mu)(x) \tag{C.139}$$

Allowing us to move the operator from acting on the function g to acting on the measure μ_0:

$$\int ((I - \gamma\Psi)^{-1} g)(s, \boldsymbol{b}, \boldsymbol{\theta}) d\mu_0(s, \boldsymbol{b}, \boldsymbol{\theta}) = \int g(s, \boldsymbol{b}, \boldsymbol{\theta}) d((I - \gamma\Psi^*)^{-1} \mu_0)(s, \boldsymbol{b}, \boldsymbol{\theta}) \tag{C.140}$$

Therefore:

$$\nabla_{\theta^i} v^i_{\theta^i}(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0) = \int g(s, \boldsymbol{b}, \boldsymbol{\theta}) d((I - \gamma\Psi^*)^{-1} \delta_{(s_0, \boldsymbol{b}_0, \boldsymbol{\theta}_0)})(s, \boldsymbol{b}, \boldsymbol{\theta}) \tag{C.141}$$

From our earlier derivation of the discounted visitation measure, we established that:

$$d^{\pi}_{\mu_0} = (1-\gamma)(I-\gamma\Psi^*)^{-1}\mu_0 = (1-\gamma)(I-\gamma\Psi^*)^{-1}\delta_{(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0)} \tag{C.142}$$

We can now substitute this expression:

$$\nabla_{\theta^i} v^i_{\theta^i}(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0) = \int g(s,\boldsymbol{b},\boldsymbol{\theta}) d\left(\frac{d^{\pi}_{\mu_0}}{1-\gamma}\right)(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.143}$$

$$= \frac{1}{1-\gamma}\int g(s,\boldsymbol{b},\boldsymbol{\theta}) d^{\pi}_{\mu_0}(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.144}$$

Substituting the definition of $g(s,\boldsymbol{b},\boldsymbol{\theta})$ back:

$$\nabla_{\theta^i} J^i_{\pi,\gamma}(\theta^i) = \nabla_{\theta^i} v^i_{\theta^i}(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0) \tag{C.145}$$

$$= \frac{1}{1-\gamma}\int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}} \sum_{a^i} \nabla_{\theta^i}\pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) d^{\pi}_{\mu_0}(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.146}$$

For finite state, belief, and policy parameter spaces, this integral can be written as a sum:

$$\nabla_{\theta^i} J^i_{\pi,\gamma}(\theta^i) = \frac{1}{1-\gamma}\sum_{s,\boldsymbol{b},\boldsymbol{\theta}} d^{\pi}_{\mu_0}(s,\boldsymbol{b},\boldsymbol{\theta}) \sum_{a^i} \nabla_{\theta^i}\pi^i(a^i|b^i;\theta^i) \sum_{\boldsymbol{a}^{-i}} \boldsymbol{\pi}^{-i}(\boldsymbol{a}^{-i}|\boldsymbol{b}^{-i};\boldsymbol{\theta}^{-i}) q^i_{\theta^i}(s,\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{a}) \tag{C.147}$$

This completes the proof of the Partially Observable Active Discounted Return Policy Gradient Theorem. □

The proof reveals a symmetry in the mathematical structure of our results.

Remark 2 (Symmetry)**.** *From the Value Function as Inner Product lemma, the value function*

$$v^i_{\theta^i}(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0) = \frac{1}{1-\gamma}\int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}} r^i(s,\boldsymbol{b},\boldsymbol{\theta}) d^{\pi}_{\mu_0}(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.148}$$

is an expectation of immediate rewards r^i under the discounted visitation measure $d^{\pi}_{\mu_0}$. While its gradient

$$\nabla_{\theta^i} v^i_{\theta^i}(s_0,\boldsymbol{b}_0,\boldsymbol{\theta}_0) = \frac{1}{1-\gamma}\int_{S\times\boldsymbol{B}\times\boldsymbol{\Theta}} g(s,\boldsymbol{b},\boldsymbol{\theta}) d^{\pi}_{\mu_0}(s,\boldsymbol{b},\boldsymbol{\theta}) \tag{C.149}$$

is an expectation of policy-gradient-weighted action values g under the same discounted visitation measure $d^{\pi}_{\mu_0}$.

This symmetric structure provides not only mathematical elegance but also practical advantages. It means that the same samples from the discounted visitation distribution can be used to estimate both values and policy gradients.

The policy gradient theorem provides the mathematical foundation for agents to optimize their policies in ways that account for both the partial observability of the environment and

the learning dynamics of other agents. This enables sophisticated strategic behaviors such as information revelation, teaching, and influence that are essential in social learning contexts.

POLARIS ARCHITECTURE

D

This section provides detailed technical specifications of all neural network architectures used in our POLARIS implementation. We describe each component's structure, activation functions, and design considerations to enable reproducibility and thorough understanding of our approach. POLARIS represents a significant advancement over traditional multi-agent reinforcement learning frameworks by employing Graph Neural Networks (GNNs) as the primary architecture for modeling agent interactions and temporal dependencies in partially observable social learning environments, with novel integration of Lévy process discretization for handling stochastic policy updates and jump-diffusion dynamics in multi-agent learning.

POLARIS employs a modular architecture consisting of four primary components that work in tandem to enable effective learning in partially observable multi-agent environments. The belief processing module maintains belief states using transformers with attention mechanisms, while the inference learning module models other agents' behavior using Temporal Graph Neural Networks with Lévy process discretization to capture sudden behavioral shifts and non-Gaussian policy evolution patterns. The policy network maps beliefs and latent states to action decisions through a stochastic framework that incorporates discretized Lévy jumps for exploration and adaptation, and dual value networks estimate expected returns using a critic architecture enhanced with jump-diffusion value function approximation. All components operate over the joint space of states, belief states, and policy parameters as defined in our Partially Observable Active Markov Game framework, explicitly modeling how actions influence both environmental dynamics and other agents' learning processes through continuous-time stochastic processes discretized via Euler-Maruyama schemes with Lévy increment compensation.

D.1 TRANSFORMER NETWORK

The belief processing module maintains and updates agents' belief states about the underlying environment state using a transformer-based architecture (Vaswani et al., 2017). This component processes sequences of observations and actions to form beliefs about the partially observable environment. The belief processor employs a transformer encoder with an input dimension configurable based on the observation space, using dimension $|S|$ for discrete observation spaces and dimension 1 for continuous observation spaces. The hidden dimension defaults to 256, with 4 attention heads across 2 layers and a dropout rate of 0.1.

For discrete observation spaces, the input signal $\mathbf{o}_t \in \mathbb{R}^{|S|}$ is projected to the hidden dimension through $\mathbf{x}_{\text{proj}} = W_{\text{proj}}\mathbf{o}_t + \mathbf{bias}_{\text{proj}}$. For continuous observation spaces with $\mathbf{o}_t \in \mathbb{R}$, a separate projection layer handles the scalar input: $\mathbf{x}_{\text{proj}} = W_{\text{cont}}\mathbf{o}_t + \mathbf{bias}_{\text{cont}}$. Positional encodings are added to capture temporal ordering: $\mathbf{x}_{\text{input}} = \mathbf{x}_{\text{proj}} + \mathbf{PE}$, where $\mathbf{PE}$ is a learned positional

encoding parameter. The transformer processes the sequence through multi-head self-attention layers, where each attention head computes:

$$\text{head}_i = \text{Attention}(\mathbf{X}W_i^Q, \mathbf{X}W_i^K, \mathbf{X}W_i^V) \tag{D.1}$$

where $\mathbf{X}$ is the input sequence of observations and actions. The attention mechanism operates through a query-key-value framework where the queries $\mathbf{Q}_i = \mathbf{X}W_i^Q \in \mathbb{R}^{n \times d_k}$, keys $\mathbf{K}_i = \mathbf{X}W_i^K \in \mathbb{R}^{n \times d_k}$, and values $\mathbf{V}_i = \mathbf{X}W_i^V \in \mathbb{R}^{n \times d_v}$ are computed via learned linear transformations, where n is the sequence length, d_k is the key/query dimension, and d_v is the value dimension. The attention computation proceeds through the following steps:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d_k}}\right) \tag{D.2}$$

$$\text{head}_i = \mathbf{A}\mathbf{V}_i \in \mathbb{R}^{n \times d_v} \tag{D.3}$$

where $\mathbf{q}_j^T\mathbf{k}_k$ represents the unnormalized attention score between position j and position k, and $\mathbf{A}_{jk}$ represents the normalized attention weight. The scaling factor $1/\sqrt{d_k}$ prevents the dot products from growing too large, which would push the softmax function into regions with extremely small gradients.

The attention mechanism creates belief states $\mathbf{b}_t$ by computing weighted combinations of historical information, where attention weights determine the relevance of past observations for current state estimation. Specifically, the self-attention layers compute queries, keys, and values from the observation sequence, allowing the model to selectively focus on informative historical patterns while updating the belief distribution over possible world states. This approach enables the transformer to maintain uncertainty estimates and capture long-range dependencies that are crucial for effective belief state representation in partially observable environments.

The final belief distribution is computed using appropriate activations based on the observation space: $\mathbf{b}_{\text{dist}} = \text{softmax}(W_{\text{discrete}}\mathbf{b}_t + \mathbf{bias}_{\text{discrete}})$ for discrete observation spaces and $\mathbf{b}_{\text{dist}} = \text{sigmoid}(W_{\text{continuous}}\mathbf{b}_t + \mathbf{bias}_{\text{continuous}})$ for continuous observation spaces, where $\mathbf{b}_t$ is the output from the transformer layers.

D.2 Temporal Graph Neural Network

The core innovation of POLARIS lies in its use of Graph Neural Networks (Hamilton, 2020) to explicitly model the network structure of agent interactions. Unlike traditional approaches that treat other agents as part of the environment, our GNN-based architecture captures the relational structure between agents and enables sophisticated reasoning about network effects in social learning.

The Temporal GNN serves as the inference learning module, enabling agents to predict and model the behavior of other agents over time. This component represents agents and their interactions as a dynamic graph where nodes correspond to agents and edges represent obser-

vational relationships. For each agent i, the GNN constructs a graph representation where each agent becomes a node with features encoding their current observation and action, connections represent observational relationships based on network topology, and node features consist of the concatenation of belief state and action: $[\mathbf{b}_t^i; a_t^i]$ for the ego agent, $[\mathbf{0}; \boldsymbol{a}_t^{-i}]$ for neighbors. allocation values are directly included in the feature representation.

The GNN employs Graph Attention Networks (GAT) (Veličković et al., 2018) that compute spatial attention coefficients between connected nodes through a multi-step process. For each node i with feature vector $\mathbf{h}_i \in \mathbb{R}^{d_{in}}$ and each neighbor $j \in \mathscr{N}^i$, the spatial attention mechanism computes:

$$\mathbf{z}_i = W\mathbf{h}_i \in \mathbb{R}^{d_{out}} \tag{D.4}$$

$$\mathbf{z}_j = W\mathbf{h}_j \in \mathbb{R}^{d_{out}} \tag{D.5}$$

$$\mathbf{c}_{ij} = [\mathbf{z}_i; \mathbf{z}_j] \in \mathbb{R}^{2d_{out}} \tag{D.6}$$

where $W \in \mathbb{R}^{d_{out} \times d_{in}}$ is a learned transformation matrix and $[\cdot;\cdot]$ denotes concatenation. The attention scores are computed using a learned attention function:

$$e_{ij} = \text{LeakyReLU}\left(\mathbb{A}^T \mathbf{c}_{ij}\right) \tag{D.7}$$

where $\mathbb{A} \in \mathbb{R}^{2d_{out}}$ is a learned attention vector that parameterizes the importance function. The spatial attention coefficients are then normalized across the local neighborhood:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in I^i} \exp(e_{ik})} = \frac{\exp(\text{LeakyReLU}(\mathbb{A}^T[\mathbf{z}_i; \mathbf{z}_j]))}{\sum_{k \in I^i} \exp(\text{LeakyReLU}(\mathbb{A}^T[\mathbf{z}_i; \mathbf{z}_k]))} \tag{D.8}$$

where $I^i = \{j : (i,j) \in \mathscr{E}\} \cup \{i\}$ represents the neighborhood of node i including self-connections, and $\mathscr{E}$ denotes the edge set of the graph. The spatial attention weights α_{ij} satisfy the normalization constraint $\sum_{j \in I^i} \alpha_{ij} = 1$, ensuring that the attention mechanism produces a valid probability distribution over the neighborhood.

The node representations are updated through an attention-weighted aggregation of neighbor features:

$$\mathbf{h}_i' = \text{ReLU}\left(\sum_{j \in I^i} \alpha_{ij} \mathbf{z}_j\right) = \text{ReLU}\left(\sum_{j \in I^i} \alpha_{ij} W \mathbf{h}_j\right) \tag{D.9}$$

This spatial attention mechanism enables each node to adaptively focus on the most relevant neighbors, with attention weights $\alpha_{ij} \approx 1$ for highly relevant neighbors and $\alpha_{ij} \approx 0$ for less important connections, allowing the model to learn which agent interactions are most informative for predicting behavior.

To capture different types of relationships between agents simultaneously, the GNN employs multi-head spatial attention with $K = 4$ attention heads. Each head $k \in \{1, 2, \ldots, K\}$ maintains its own parameter set $\{W^k, \mathbb{A}^k\}$ where $W^k \in \mathbb{R}^{d_{out} \times d_{in}}$ and $\mathbb{A}^k \in \mathbb{R}^{2d_{out}}$, computing head-specific

representations:

$$\mathbf{z}_i^k = W^k \mathbf{h}_i \in \mathbb{R}^{d_{out}} \tag{D.10}$$

$$e_{ij}^k = \text{LeakyReLU}\left((\mathbb{A}^k)^T \left[\mathbf{z}_i^k; \mathbf{z}_j^k\right]\right) \tag{D.11}$$

$$\alpha_{ij}^k = \frac{\exp(e_{ij}^k)}{\sum_{l \in \mathcal{N}^i} \exp(e_{il}^k)} \tag{D.12}$$

$$\mathbf{h}_i^k = \text{ReLU}\left(\sum_{j \in \mathcal{N}^i} \alpha_{ij}^k \mathbf{z}_j^k\right) \tag{D.13}$$

The final node representation is obtained by concatenating outputs from all attention heads:

$$\mathbf{h}_i' = [\mathbf{h}_i^1; \mathbf{h}_i^2; \ldots; \mathbf{h}_i^K] \in \mathbb{R}^{Kd_{out}} \tag{D.14}$$

This multi-head architecture enables the model to attend to different aspects of agent relationships simultaneously, such as one head focusing on recent behavioral patterns with α_{ij}^1 emphasizing temporal proximity, another head capturing strategic complementarity with α_{ij}^2 emphasizing action coordination patterns, and additional heads modeling competitive dynamics or information flow patterns within the agent network.

A critical innovation in our Temporal GNN is the incorporation of temporal dependencies through a sliding window memory mechanism. The system maintains a buffer of previous graph representations over a temporal window of size $T = 5$:

$$\mathcal{M}_t = \{\mathbf{H}_{t-T+1}, \mathbf{H}_{t-T+2}, \ldots, \mathbf{H}_t\} \tag{D.15}$$

where $\mathbf{H}_\tau$ represents the node features at time step τ. Temporal attention is applied across this sequence using a standard transformer attention mechanism:

$$\text{TemporalAttention}(\mathbf{Q}_t, \mathbf{K}_{t-T+1:t}, \mathbf{V}_{t-T+1:t}) = \text{softmax}\left(\frac{\mathbf{Q}_t \mathbf{K}_{t-T+1:t}^T}{\sqrt{d}}\right)\mathbf{V}_{t-T+1:t} \tag{D.16}$$

This temporal attention mechanism enables the model to capture how agent behaviors evolve over time and identify patterns in their learning dynamics.

The Temporal GNN outputs parameters for a variational posterior distribution over latent states representing other agents' hidden characteristics:

$$\boldsymbol{m}_t = W_\mu \mathbf{h}_i^{\text{final}} + \mathbf{bias}_m \tag{D.17}$$

$$\log \boldsymbol{\sigma}_t = W_\sigma \mathbf{h}_i^{\text{final}} + \mathbf{bias}_\sigma \tag{D.18}$$

where $\mathbf{h}_i^{\text{final}}$ is the final node representation for agent i after all GNN layers and temporal

attention. The latent variables are sampled using the reparameterization trick:

$$\hat{\mathbf{z}}_{t+1} = \boldsymbol{m}_t + \boldsymbol{\sigma}_t \odot \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{D.19}$$

D.3 Policy and Value Networks

The policy networks map belief states and inferred latent variables to action decisions, implementing the agent's strategy for maximizing long-term rewards. For discrete action spaces, the policy network outputs categorical distributions:

$$\mathbf{h}^1 = \text{ReLU}(W^1[\mathbf{b}_t^i; \hat{\mathbf{z}}_t] + \mathbf{bias}_1) \tag{D.20}$$

$$\mathbf{h}^2 = \text{ReLU}(W^2\mathbf{h}^1 + \mathbf{bias}_2) \tag{D.21}$$

$$\boldsymbol{\pi}^i(\cdot|\mathbf{b}_t^i, \hat{\mathbf{z}}_t; \boldsymbol{\theta}^i) = \text{softmax}(W^{\text{out}}\mathbf{h}^2 + \mathbf{bias}_{\text{out}}) \tag{D.22}$$

where $[\cdot;\cdot]$ denotes concatenation.

POLARIS employs dual Q-networks (Hasselt, Guez, & Silver, 2016) to mitigate overestimation bias, a critical consideration in multi-agent settings where overconfident value estimates can lead to suboptimal strategic behaviors. For discrete actions, the Q-networks take belief states, latent variables, and one-hot encoded neighbor actions:

$$\mathbf{x}_q = [\mathbf{b}_t^i; \hat{\mathbf{z}}_t; \boldsymbol{a}_{\text{onehot}}^{\text{neighbors}}] \tag{D.23}$$

$$\mathbf{h}_q^1 = \text{ReLU}(W_q^1\mathbf{x}_q + \mathbf{b}_q^1) \tag{D.24}$$

$$\mathbf{h}_q^2 = \text{ReLU}(W_q^2\mathbf{h}_q^1 + \mathbf{b}_q^2) \tag{D.25}$$

$$\mathbf{Q}^i = W_q^{\text{out}}\mathbf{h}_q^2 + \mathbf{b}_q^{\text{out}} \tag{D.26}$$

Target networks are maintained for stability, updated using exponential moving averages:

$$\boldsymbol{\theta}_{\text{target}} \leftarrow \tau\boldsymbol{\theta} + (1-\tau)\boldsymbol{\theta}_{\text{target}} \tag{D.27}$$

with $\tau = 0.005$ in our implementation.

D.4 Lévy Process Discretization

This appendix provides a comprehensive mathematical treatment of the discretization of Lévy processes for implementing strategic experimentation models within our Partially Observable Active Markov Game framework. We establish rigorous theoretical foundations for the numerical approximation schemes used in our implementation and analyze their convergence properties and preservation of strategic incentives.

D.4.1 Mathematical Foundations of Lévy Processes

Lévy processes form a fundamental class of stochastic processes that include Brownian motion and Poisson processes as special cases. They are characterized by stationary and independent increments, serving as the natural continuous-time generalization of random walks.

Definition 11 (Lévy Process)**.** *A stochastic process $X = \{X_t : t \geq 0\}$ on $\mathbb{R}^d$ with $X_0 = 0$ almost surely is a Lévy process if:*

1. *It has independent increments: for any $0 \leq t_1 < t_2 < \cdots < t_n < \infty$, the random variables $X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \ldots, X_{t_n} - X_{t_{n-1}}$ are mutually independent.*
2. *It has stationary increments: for any $s < t$, the distribution of $X_t - X_s$ depends only on $t - s$.*
3. *It is stochastically continuous: for any $t \geq 0$ and $\varepsilon > 0$, $\lim_{h \to 0} \mathbb{P}(|X_{t+h} - X_t| > \varepsilon) = 0$.*

The celebrated Lévy-Khintchine formula provides a complete characterization of Lévy processes through their characteristic functions:

Theorem 4 (Lévy-Khintchine Formula Applebaum (2009); Sato (1999))**.** *If $X = \{X_t : t \geq 0\}$ is a Lévy process, then its characteristic function has the form:*

$$\mathbb{E}[e^{i\theta X_t}] = e^{t\psi(\theta)} \tag{D.28}$$

where

$$\psi(\theta) = i\alpha\theta - \frac{1}{2}\sigma^2\theta^2 + \int_{\mathbb{R}\setminus\{0\}} (e^{i\theta x} - 1 - i\theta x \mathbf{1}_{|x|<1})\nu(dx) \tag{D.29}$$

for some $\alpha \in \mathbb{R}$, $\sigma \geq 0$, and a measure ν on $\mathbb{R} \setminus \{0\}$ satisfying $\int_{\mathbb{R}\setminus\{0\}} \min(1, x^2)\nu(dx) < \infty$.

The triplet (α, σ^2, ν) is called the Lévy-Khintchine triplet or the characteristics of the Lévy process. Here, α represents a drift term, σ^2 parametrizes the continuous Gaussian component, and ν is the Lévy measure characterizing the jump behavior.

Theorem 5 (Lévy-Itô Decomposition Protter (2005); Sato (1999))**.** *Any Lévy process X_t can be decomposed as:*

$$X_t = \alpha t + \sigma W_t + \int_{|x|<1} x(\tilde{Y}(t, dx) - t\nu(dx)) + \int_{|x|\geq 1} xY(t, dx) \tag{D.30}$$

where W_t is a standard Brownian motion, $Y(t, A)$ counts the number of jumps of size in set A occurring up to time t, and $\tilde{Y}(t, dx) = Y(t, dx) - t\nu(dx)$ is the compensated Poisson random measure.

The Lévy-Itô decomposition provides a pathwise representation of a Lévy process as the sum of a drift term, a Brownian motion, and potentially infinitely many jumps, both small and large.

D.4.2 Time Discretization of Lévy Processes

To implement continuous-time Lévy processes within discrete computational frameworks, we must employ appropriate numerical approximation schemes. For our strategic experimentation model, we adopt the Euler-Maruyama scheme, extended to accommodate the jump components of general Lévy processes.

Definition 12 (Euler-Maruyama Scheme for Lévy Processes)**.** *Given a Lévy process X_t with characteristics (α, σ^2, ν) and a discretization time step Δt, the Euler-Maruyama approximation constructs a discrete-time process $\{X_{t_n}\}_{n=0}^{N}$ where $t_n = n\Delta t$ through the recursive relation:*

$$X_{t_{n+1}} = X_{t_n} + \alpha\Delta t + \sigma\sqrt{\Delta t}Z_n + \Delta J_n \tag{D.31}$$

where $Z_n \sim \mathcal{N}(0,1)$ are independent standard normal random variables and ΔJ_n represents the jump increment over $[t_n, t_{n+1}]$.

Theorem 6 (Convergence of Euler-Maruyama Scheme (Higham & Kloeden, 2005; Platen & Bruti-Liberati, 2010))**.** *Let X_t be a Lévy process and $\hat{X}_t$ be its Euler-Maruyama approximation with time step Δt. Then for any fixed $T > 0$:*

1. *(Weak Convergence) For any smooth function f with polynomial growth:*

$$|\mathbb{E}[f(X_T)] - \mathbb{E}[f(\hat{X}_T)]| \leq C\Delta t \tag{D.32}$$

2. *(Strong Convergence) If $\int_{|x|>1} |x|^2 \nu(dx) < \infty$, then:*

$$\mathbb{E}[\sup_{0\leq t\leq T} |X_t - \hat{X}_t|^2] \leq C\Delta t \tag{D.33}$$

where C is a constant depending on T and the characteristics of the Lévy process.

The weak and strong convergence properties ensure that our numerical scheme accurately approximates both the distributional properties and pathwise behavior of the continuous-time process as the time step decreases.

D.4.3 Implementing Strategic Experimentation Models

In our implementation of the strategic experimentation model from Keller and Rady (2020), we must discretize both the background signal process B_t and the individual payoff processes X_t^i, while preserving the strategic incentives that drive experimentation decisions.

Discretization of Diffusion-Poisson Processes

In the original model, both B_t and X_t^i follow Lévy processes that combine continuous diffusion and discrete jumps:

$$dB_t = \beta_s dt + \sigma_B dW_t^B + dY_t^B \tag{D.34}$$
$$dX_t^i = \alpha_s dt + \sigma dW_t^i + dY_t^i \tag{D.35}$$

where W_t^B and W_t^i are standard Brownian motions, and Y_t^B and Y_t^i are compound Poisson processes. Using the Euler-Maruyama scheme (Platen & Bruti-Liberati, 2010), we discretize these continuous-time stochastic differential equations:

$$B_{t+\Delta t} - B_t = \beta_s \Delta t + \sigma_B (W_{t+\Delta t}^B - W_t^B) + (Y_{t+\Delta t}^B - Y_t^B) \tag{D.36}$$
$$X_{t+\Delta t}^i - X_t^i = \alpha_s \Delta t + \sigma (W_{t+\Delta t}^i - W_t^i) + (Y_{t+\Delta t}^i - Y_t^i) \tag{D.37}$$

For implementation, we denote these increments as:

$$B_{t-1:t} = \beta_s \Delta t + \sigma_B (W_t^B - W_{t-1}^B) + \Delta Y_t^B \tag{D.38}$$
$$X_{t-1:t}^i = \alpha_s \Delta t + \sigma (W_t^i - W_{t-1}^i) + \Delta Y_t^i \tag{D.39}$$

where $(W_t^B - W_{t-1}^B) \sim \mathcal{N}(0, \Delta t)$, $(W_t^i - W_{t-1}^i) \sim \mathcal{N}(0, \Delta t)$, and ΔY_t^B, ΔY_t^i are the increments of the compound Poisson processes over the interval $[t-1, t]$.

Reward Function Equivalence

A critical challenge in our implementation is reconciling the time-dependent nature of continuous-time rewards with the time-independent reward structure required by the POAMG framework. In the original continuous-time model, agents' instantaneous rewards are:

$$dR_t^i = (1 - a_t^i) r_{safe} dt + a_t^i dX_t^i \tag{D.40}$$

where $a_t^i \in [0,1]$ is the allocation to the risky arm and r_{safe} is the safe arm's deterministic flow payoff and the Levy process increment is a function of the state s. To translate this structure into the POAMG framework, we need a reward function $R^i(s, a^i)$ that depends only on the state and action, not explicitly on time. We achieve this through a normalization approach:

$$R^i(s, a^i) = (1 - a^i) r_{safe} + a^i \frac{X_{t-1:t}^i}{\Delta t} \tag{D.41}$$

This transformation preserves the incentive structure of the original model while eliminating explicit time dependence. The following proposition establishes this equivalence formally:

Proposition 1 (Reward Equivalence)**.** *The expected value of the discrete-time reward function $R^i(s, a^i)$ exactly equals the expected instantaneous flow payoff in the continuous-time model.*

Specifically:

$$\mathbb{E}[R^i(s,a^i)] = (1-a^i)r_{safe} + a^i(\alpha_s + \lambda_s h_s) \tag{D.42}$$

where α_s is the drift of the Levy process, λ_s is the jump intensity and h_s is the mean jump size of the compound Poisson process component of X_t^i in state s.

Proof. The expected value of the discrete-time reward function is:

$$\mathbb{E}[R^i(s,a^i)] = (1-a^i)r_{safe} + a^i\mathbb{E}\left[\frac{X^i_{t-1:t}}{\Delta t}\right] \tag{D.43}$$

$$= (1-a^i)r_{safe} + a^i\frac{\mathbb{E}[\alpha_s\Delta t + \sigma(W_t^i - W_{t-1}^i) + \Delta Y_t^i]}{\Delta t} \tag{D.44}$$

$$= (1-a^i)r_{safe} + a^i\left(\alpha_s + \frac{\mathbb{E}[\Delta Y_t^i]}{\Delta t}\right) \tag{D.45}$$

Since $\mathbb{E}[W_t^i - W_{t-1}^i] = 0$ and $\mathbb{E}[\Delta Y_t^i] = \lambda_s h_s \Delta t$, where λ_s is the jump intensity and h_s is the mean jump size, we have:

$$\mathbb{E}[R^i(s,a^i)] = (1-a^i)r_{safe} + a^i\left(\alpha_s + \frac{\lambda_s h_s \Delta t}{\Delta t}\right) \tag{D.46}$$

$$= (1-a^i)r_{safe} + a^i(\alpha_s + \lambda_s h_s) \tag{D.47}$$

This exactly matches the expected instantaneous flow payoff in the continuous-time model. □

REFERENCES

Aberdeen, D., & Baxter, J. (2002). Policy-gradient algorithms for partially observable Markov decision processes. *Advances in Neural Information Processing Systems*, *15*.

Acemoglu, D., Dahleh, M. A., Lobel, I., & Ozdaglar, A. (2011). Bayesian learning in social networks. *The Review of Economic Studies*, *78*(4), 1201–1236.

Albrecht, S. V., Christianos, F., & Schäfer, L. (2024). *Multi-agent reinforcement learning: Foundations and modern approaches*. MIT Press. Retrieved from https://www.marl-book.com

Albrecht, S. V., & Stone, P. (2018). Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, *258*, 66–95.

Al-Shedivat, M., Bansal, T., Burda, Y., Sutskever, I., Mordatch, I., & Abbeel, P. (2018). Continuous adaptation via meta-learning in nonstationary and competitive environments. *Proceedings of the 6th International Conference on Learning Representations*.

Applebaum, D. (2009). *Lévy processes and stochastic calculus* (2nd ed., Vol. 116). Cambridge University Press.

Arieli, I., & Babichenko, Y. (2019). Private bayesian persuasion. *Journal of Economic Theory*, *182*, 185-217. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022053118302217 doi: https://doi.org/10.1016/j.jet.2019.04.008

Avery, C., & Zemsky, P. (1998). Multidimensional uncertainty and herd behavior in financial markets. *American Economic Review*, *88*(4), 724–748.

Baker, B. (2020). *Emergent reciprocity and team formation from randomized uncertain social preferences.* Retrieved from https://arxiv.org/abs/2011.05373

Bala, V., & Goyal, S. (1998). Learning from neighbours. *Review of Economic Studies*, *65*(3), 595–621.

Balduzzi, D., Racaniere, S., Martens, J., Foerster, J., Tuyls, K., & Graepel, T. (2018). The mechanics of n-player differentiable games. *Proceedings of the 35th International Conference on Machine Learning*, 354–363.

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, *107*(3), 797–817.

Bergemann, D., & Morris, S. (2019, March). Information design: A unified perspective. *Journal of Economic Literature*, *57*(1), 44–95. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jel.20181489 doi: 10.1257/jel.20181489

Bernstein, D. S., Givan, R., Immerman, N., & Zilberstein, S. (2002). The complexity of decentralized control of markov decision processes. *Mathematics of Operations Research*, *27*(4), 819–840.

Bhattacharya, S., & Mukherjee, A. (2013). Strategic information revelation when experts compete to influence. *The RAND Journal of Economics*, *44*(3), 522–544. Retrieved 2025-05-14, from http://www.jstor.org/stable/43186430

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, *100*(5), 992–1026.

Bolton, P., & Harris, C. (1999). Strategic experimentation. *Econometrica*, *67*(2), 349–374.

Bowling, M. (2005). Convergence and no-regret in multiagent learning. *Advances in Neural Information Processing Systems*, *17*.

Bowling, M., & Veloso, M. (2002). Multiagent learning using a variable learning rate. *Artificial Intelligence*, *136*(2), 215–250.

Brandl, F. (2025). *The social learning barrier.* Retrieved from https://arxiv.org/abs/2504.12136

Busoniu, L., Babuska, R., & De Schutter, B. (2010). Multi-agent reinforcement learning: An overview. *Innovations in Multi-Agent Systems and Applications-1*, 183–221.

Che, Y.-K., & Hörner, J. (2018). Optimal design for social learning. *Quarterly Journal of Economics*, *133*(2), 871–925.

Crawford, V. P., & Iriberri, N. (2007). Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica*, *75*(6), 1721–1770.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, *69*(345), 118–121.

DeMarzo, P. M., Vayanos, D., & Zwiebel, J. (2003, 08). Persuasion bias, social influence, and unidimensional opinions*. *The Quarterly Journal of Economics*, *118*(3), 909-968. Retrieved from https://doi.org/10.1162/00335530360698469 doi: 10.1162/00335530360698469

Du, Y., Leibo, J. Z., Islam, U., Willis, R., & Sunehag, P. (2023). *A review of cooperation in multi-agent learning.* Retrieved from https://arxiv.org/abs/2312.05162

Ellison, G., & Fudenberg, D. (1993). Rules of thumb for social learning. *Journal of Political Economy*, *101*(4), 612-43. Retrieved from https://EconPapers.repec.org/RePEc:ucp:jpolec:v:101:y:1993:i:4:p:612-43

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2018). Counterfactual multi-agent policy gradients. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

Frederick, S., Loewenstein, G., & O'donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of economic literature*, *40*(2), 351–401.

Gale, D., & Kariv, S. (2003). Bayesian learning in social networks. *Games and Economic Behavior*, *45*(2), 329–346.

Golub, B., & Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, *2*(1), 112–149.

Golub, B., & Jackson, M. O. (2012). How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, *127*(3), 1287–1338. Retrieved 2025-05-14, from http://www.jstor.org/stable/23251986

Gong, D., Yan, Q., Liu, Y., van den Hengel, A., & Shi, J. Q. (2022). *Learning bayesian sparse networks with full experience replay for continual learning.* Retrieved from https://arxiv.org/abs/2202.10203

Guarino, A., & Jehiel, P. (2013, February). Social learning with coarse inference. *American Economic Journal: Microeconomics*, *5*(1), 147–74. Retrieved from https://www.aeaweb.org/articles?id=10.1257/mic.5.1.147 doi: 10.1257/mic.5.1.147

Halac, M., Kartik, N., & Liu, Q. (2017). Contests for experimentation. *Journal of Political Economy*, *125*(5), 1523-1569. Retrieved from https://doi.org/10.1086/693040 doi: 10.1086/693040

Hamilton, W. L. (2020). Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *14*(3), 1–159.

Hansen, E. A., Bernstein, D. S., & Zilberstein, S. (2004). Dynamic programming for partially observable stochastic games. *Proceedings of the 19th National Conference on Artificial Intelligence*, 709–715.

Harel, M., Mossel, E., Strack, P., & Tamuz, O. (2014). The speed of social learning.. Retrieved from https://api.semanticscholar.org/CorpusID:1483403

Hasselt, H. v., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double q-learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1).

Hausknecht, M., & Stone, P. (2015). Deep recurrent Q-learning for partially observable MDPs. In *Aaai fall symposium on sequential decision making for intelligent agents.*

He, H., Boyd-Graber, J., Kwok, K., & Daumé III, H. (2016). Opponent modeling in deep reinforcement learning. *Proceedings of the 33rd International Conference on Machine Learning*, 1804–1813.

Heidhues, P., Rady, S., & Strack, P. (2015). Strategic experimentation with private payoffs. *Journal of Economic Theory*, *159*, 531-551. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022053115001404 doi: https://doi.org/10.1016/j.jet.2015.07.017

Hernandez-Leal, P., Kaisers, M., Baarslag, T., & de Cote, E. M. (2017). A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*.

Higham, D., & Kloeden, P. (2005, July). Numerical methods for nonlinear stochastic differential equations with jumps. *Numerische Mathematik*, *101*(1), 101–119. doi: 10.1007/s00211-005-0611-8

Hu, J., & Wellman, M. P. (2003). Nash q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, *4*, 1039–1069.

Huang, A., Strack, P., & Tamuz, O. (2024). Learning in networks: Social learning, multi-agent

learning and diffusion of information. *Annual Review of Economics*, *16*, 1–28.

Huh, D., & Mohapatra, P. (2024). *Multi-agent reinforcement learning: A comprehensive survey.* Retrieved from https://arxiv.org/abs/2312.10256

Jadbabaie, A., Molavi, P., Sandroni, A., & Tahbaz-Salehi, A. (2012). Non-bayesian social learning. *Games and Economic Behavior*, *76*(1), 210–225.

Jadbabaie, A., Molavi, P., & Tahbaz-Salehi, A. (2013). Information heterogeneity and the speed of learning in social networks. *Writing Technologies eJournal*. Retrieved from https://api.semanticscholar.org/CorpusID:17438004

Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castañeda, A. G., ... Graepel, T. (2019, May). Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, *364*(6443), 859–865. Retrieved from http://dx.doi.org/10.1126/science.aau6249 doi: 10.1126/science.aau6249

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., ... De Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. *Proceedings of the 36th International Conference on Machine Learning*, *97*, 3040–3049.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1-2), 99–134.

Kamenica, E., & Gentzkow, M. (2011, October). Bayesian persuasion. *American Economic Review*, *101*(6), 2590–2615. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590 doi: 10.1257/aer.101.6.2590

Keller, G., & Rady, S. (2020). Undiscounted bandit games. *Games and Economic Behavior*, *124*, 43–61.

Keller, G., Rady, S., & Cripps, M. (2005). Strategic experimentation with exponential bandits. *Econometrica*, *73*(1), 39–68.

Kim, D.-K., Liu, M., Riemer, M., Sun, C., Abdulhai, M., Habibi, G., ... How, J. P. (2021). *A policy gradient algorithm for learning to learn in multiagent reinforcement learning.* Retrieved from https://arxiv.org/abs/2011.00382

Kim, D. K., Yuan, L., Gong, X., Ullman, T. D., Spelke, E., Tenenbaum, J. B., & Chuang, I. (2022). Influencing long-term behavior in multiagent reinforcement learning. *Advances in Neural Information Processing Systems*, *35*, 41417–41432.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... others (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13), 3521–3526.

Klein, N., & Rady, S. (2011). Negatively correlated bandits. *The Review of Economic Studies*, *78*(2), 693–732.

Kloeden, P. E., & Platen, E. (1992). *Numerical solution of stochastic differential equations*. Springer-Verlag, Berlin.

Koopmans, T. C. (1960). Stationary ordinal utility and impatience. *Econometrica*, *28*(2), 287–309.

Lauer, M., & Riedmiller, M. (2000). An algorithm for distributed reinforcement learning in cooperative multi-agent systems. *Proceedings of the 17th International Conference on Machine Learning*, 535–542.

Laurent, G. J., Matignon, L., & Fort-Piat, N. L. (2011). The world of independent learners is not markovian. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, *15*(1), 55-64. doi: 10.3233/KES-2010-0206

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 464–473.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. *Proceedings of the 11th International Conference on Machine Learning*, 157–163.

Littman, M. L. (2001). Friend-or-foe q-learning in general-sum games. *Proceedings of the 18th International Conference on Machine Learning*, 322–328.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, *30*.

Mortensen, O., & Talebi, M. S. (2025). *Entropic risk optimization in discounted mdps: Sample complexity bounds with a generative model.* Retrieved from `https://arxiv.org/abs/2506.00286`

Mossel, E., Sly, A., & Tamuz, O. (2015). Strategic learning and the topology of social networks. *Econometrica*, *83*(5), 1755–1794.

Ndousse, K., Eck, D., Levine, S., & Jaques, N. (2021). Emergent social learning via multi-agent reinforcement learning. *Proceedings of the 38th International Conference on Machine Learning*, *139*, 7991–8004.

Nguyen, H., Daley, B., Song, X., Amato, C., & Platt, R. (2021). *Belief-grounded networks for accelerated robot learning under partial observability.* Retrieved from `https://arxiv.org/abs/2010.09170`

Nowé, A., Vrancx, P., & De Hauwere, Y.-M. (2012). Game theory and multi-agent reinforcement learning. *Reinforcement Learning*, 441–470.

OpenAI. (2019). *Dota 2 with large scale deep reinforcement learning.* Retrieved from `https://arxiv.org/abs/1912.06680`

Papoudakis, G., Christianos, F., Schäfer, L., & Albrecht, S. V. (2019). Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*.

Platen, E. (1999). *An introduction to numerical methods for stochastic differential equations*. Springer-Verlag, Berlin.

Platen, E., & Bruti-Liberati, N. (2010). *Numerical solution of stochastic differential equations with jumps in finance* (Vol. 64). Springer Science & Business Media.

Protter, P. E. (2005). *Stochastic integration and differential equations* (2nd ed., Vol. 21). Springer.

Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018). Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. *Proceedings of the 35th International Conference on Machine Learning*, 4295–4304.

Rosenberg, D., Solan, E., & Vieille, N. (2009). Informational externalities and emergence of consensus. *Games and Economic Behavior*, *66*(2), 979-994. Retrieved from https://www.sciencedirect.com/science/article/pii/S0899825608001772 (Special Section In Honor of David Gale) doi: https://doi.org/10.1016/j.geb.2008.09.027

Rothschild, M. (1974). Two-armed bandit theory of market pricing. *Journal of Economic Theory*, *9*(2), 185–202.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., ... Hadsell, R. (2022). *Progressive neural networks.* Retrieved from https://arxiv.org/abs/1606.04671

Sato, K.-i. (1999). *Lévy processes and infinitely divisible distributions* (Vol. 68). Cambridge University Press.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences*, *39*(10), 1095–1100.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Silver, D., Lever, G., Heess, N. M. O., Degris, T., Wierstra, D., & Riedmiller, M. A. (2014). Deterministic policy gradient algorithms. In *International conference on machine learning.* Retrieved from https://api.semanticscholar.org/CorpusID:13928442

Smith, L., & Sørensen, P. (2000). Pathological outcomes of observational learning. *Econometrica*, *68*(2), 371–398.

Stahl, D. O., & Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, *25*(3), 309–327.

Stokey, N. L., Lucas, R. E., & Prescott, E. C. (1989). *Recursive methods in economic dynamics*. Cambridge, MA: Harvard University Press.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., ... Graepel, T. (2017). *Value-decomposition networks for cooperative multi-agent learning.* Retrieved from https://arxiv.org/abs/1706.05296

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). Cambridge, MA: MIT Press.

Sutton, R. S., McAllester, D. A., Singh, S. P., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, *12*.

Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., ... Vicente, R. (2015). *Multiagent cooperation and competition with deep reinforcement learning.* Retrieved

from https://arxiv.org/abs/1511.08779

Tan, M. (1993). Multi-agent reinforcement learning: Independent vs. cooperative agents. *Proceedings of the 10th International Conference on Machine Learning*, 330–337.

Tuyls, K., Nowe, A., Lenaerts, T., & Manderick, B. (2004). An evolutionary game theoretic perspective on learning in multi-agent systems. *Synthese*, *139*(2), 297–330. Retrieved 2025-05-14, from http://www.jstor.org/stable/20118420

van de Ven, G. M., Soures, N., & Kudithipudi, D. (2024). *Continual learning and catastrophic forgetting.* Retrieved from https://arxiv.org/abs/2403.05175

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. *6th International Conference on Learning Representations*.

Wang, L., Zhang, M., Jia, Z., Li, Q., Bao, C., Ma, K., ... Zhong, Y. (2021). *Afec: Active forgetting of negative transfer in continual learning.* Retrieved from https://arxiv.org/abs/2110.12187

Weibull, J. W. (1997). Evolutionary game theory.

Wen, Y., Yang, Y., Luo, R., Wang, J., & Pan, W. (2019). Probabilistic recursive reasoning for multi-agent reinforcement learning. *Proceedings of the 7th International Conference on Learning Representations*.

Wicks, J., & Greenwald, A. (2012). An algorithm for computing stochastically stable distributions with applications to multiagent learning in repeated games. *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, 623–632.

Yamamoto, Y. (2019). Stochastic games with hidden states. *Theoretical Economics*, *14*(3), 1115-1167. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.3982/TE3068 doi: https://doi.org/10.3982/TE3068

Yang, Y., & Wang, J. (2020). An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*.

Zenke, F., Poole, B., & Ganguli, S. (2017). *Continual learning through synaptic intelligence.* Retrieved from https://arxiv.org/abs/1703.04200

Zhang, K., Yang, Z., & Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, 321–384.

Zhu, C., Dastani, M., & Wang, S. (2024). *A survey of multi-agent deep reinforcement learning with communication.* Retrieved from https://arxiv.org/abs/2203.08975

4.4 STATEMENT OF AUTHORSHIP

I hereby confirm that the work presented has been performed and interpreted solely by myself except for where I explicitly identified the contrary. I assure that this work has not been presented in any other form for the fulfillment of any other degree or qualification. Ideas taken from other works in letter and in spirit are identified in every single case.

Date: _____ Signature: _____