Learning Through Social Interactions and Learning to Socially Interact in Multi-Agent Learning

Udari Madhushani Sehwag

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Mechanical and Aerospace Engineering
Adviser: Professor Naomi Ehrich Leonard

May 2023

Abstract

The rapid integration of AI agents into society underscores the need for a deeper understanding of how these agents can benefit from social interactions and develop collective intelligence. Cultural evolution studies have emphasized the importance of cultural transmission of knowledge and intelligence across generations, highlighting that social interactions play a crucial role in a group's ability to solve complex problems or make optimal decisions. Humans are remarkable at learning through social interactions and we posses an innate ability to seamlessly perceive social interactions, acquire and transmit knowledge through social interactions, and transfer cognitive capabilities and knowledge through generations. A natural question is how can we embed these capabilities in AI agents? As a step towards answering this question this dissertation investigates two main research questions: (1) how AI agents can learn to effectively communicate with other agents, and (2) how AI agents can enhance their ability to generalize or adapt to novel partners/opponents through social interactions.

The first section of this dissertation focuses on developing methodologies that facilitate effective communication among AI agents under various communication constraints. We specifically examine communication in sequential decision-making tasks within uncertain environments, where the primary challenge lies in balancing exploration and exploitation to achieve optimal performance. To tackle this challenge, we propose innovative methodologies that enable efficient communication and decision-making among agents, taking into account the intricacies of the problem domain, such as communication costs, different communication networks and agent specific probabilistic communication constraints. Further, we investigate the role agent heterogeneity in individual and group performance and develop methods that can leverage heterogeneity to improve performance.

The second section delves into the topic of generalization in multi-agent AI. Our research investigates how agents can adapt their policies to collaborate with novel agents

they have not previously encountered in tasks that necessitate coordination and cooperation among agents to achieve optimal outcomes. We introduce new techniques, that empower agents to learn and adapt their strategies to novel partners/opponents, fostering improved cooperation and coordination among AI agents. We investigate how heterogeneous social preferences of agents lead to behavioural diversity. Further we investigate how learning a best response to diverse policies can lead to better generalization.

In exploring these research areas, this dissertation aims to enrich our understanding of how AI agents can effectively collaborate in complex social scenarios, thereby contributing to the advancement of the AI field.

Acknowledgements

Completing a PhD is not only an academic achievement, but also a personal and emotional journey that requires the support and encouragement of many individuals. I am deeply grateful to the people who have provided me with unwavering support and encouragement throughout this challenging journey. I would like to take this opportunity to express my gratitude and acknowledge their contributions.

First and foremost, I want to express my deepest appreciation to my advisor, Naomi Ehrich Leonard, who has been an exceptional mentor, teacher, and collaborator throughout my PhD program. Her guidance, expertise, and unwavering support have been instrumental in shaping my research and my development as a scholar. Her mentorship allowed me to explore new ideas, overcome challenges, and succeed as a researcher.

Secondly, I want to thank the members of my thesis committee for their invaluable feedback, time, and expertise. Their insightful comments and constructive criticism have enriched my research and helped me gain a better understanding of the broader implications of my work. I am particularly grateful to Thomas Griffiths, Ryne Beeson, Elad Hazan and Simon Levin for their support and guidance throughout the process.

My colleagues and friends have been a vital part of my life during my PhD program. I am grateful for the wonderful time spent with Xiaohan and Jessica during the first year at Graduate College. I am also thankful to Anastasia, Dan, Chris, Sayantan, Mohan, Mainik and Prakhar for the wonderful time spent at the Lakeside. I enjoyed the company of Mo, Hannah, Pranay, Justin, Shaurya, Pranav, Bhargav, Shashank, and Siddharth during our numerous conversations and meals together.

I am also deeply grateful for the support of my collaborators, as I learned a lot from them over the years. I learned a lot from Maria Santos, Alessia Benevento, Abhimanyu Dubey, Alex Pentland, Udaya Ghai, Elad Hazan, and Justin Lidard in our collaborations. I enjoyed working with Biswadip Dey, Amit Chakraborty, Kalesha

Bullard, and Roberto Calandra during my early internships at Simens and Facebook AI Research. I also want to thank numerous collaborator and mentors at DeepMind, including, John Agapiou, Edgar Duéñez-Guzmán, Joel Leibo, Edward Hughes, Kevin McKee, Thomas Anthony, Richard Everett, DJ Strouse and Karl Tuyls, who were instrumental in making my DeepMind internship a highly rewarding experience. My long time mentor and friend Sanjeeva Maithripala also deserves special recognition for his unwavering support before graduate school.

Furthermore, I appreciate the support of the staff at the MAE department, including Jill (who is currently not in MAE), Katerina, Julia, Melissa, and Caasi, for supporting graduate students across a vast range of administrative issues. I also want to acknowledge the faculty and staff at the graduate school and the Davis International Center for creating a supportive environment.

Lastly, I want to express my deepest gratitude to my family for being with me at every step of my journey. I am deeply grateful to my parents, Nandasena and Suneetha, my brother Dileepa and my extended family back in Sri Lanka for their unwavering love and encouragement. I am glad to have travelled this journey with my partner in science and in life, Vikash, who was extremely supportive at each step. His excitement for research in machine learning is contagious.

In conclusion, I want to express my heartfelt appreciation to everyone who supported and encouraged me throughout my PhD program. I could not have achieved this milestone without your guidance, encouragement, and support. I hope that this thesis will be a valuable contribution to the field of Multi-agent learning, and I am excited for the opportunities that lie ahead.

This dissertation carries T#3448 in the records of the Department of Mechanical and Aerospace Engineering.

To my parents.

Contents

Part I

Social Interactions in Multi-Agent Learning

Chapter 1

Introduction

1.1 Overview and motivation

The exceptional ability of humans to acquire and transfer knowledge and cognitive capabilities through social interactions is a defining characteristic that distinguishes them from other species. The study of cultural evolution highlights the importance of cultural transmission of knowledge and intelligence across generations [88, 32]. This is exemplified by the transfer of knowledge of innovative tools developed by groups of individuals to improve their hunting skills, which is then passed on to succeeding generations, allowing humans to build upon the knowledge of their ancestors and develop complex societies and cultures. In complement, as evidenced by Heider and Simmel's classic experiment [31] in 1944, humans demonstrate remarkable abilities to perceive interactions between agents. This ability allows humans to understand individual differences and reason about how others make decisions and the impact those decisions may have, thus enhancing their capacity to acquire and transfer knowledge leveraging social interactions and developing a better understanding of their environment.

Social interactions play a critical role in the development of collective intelligence, which refers to a group's ability to solve complex problems or make optimal decisions.

This phenomenon is not exclusive to human groups and is observed in other biological groups as well. For instance, red ants transport large pieces of food over vertical surfaces by working together to enhance their physical capabilities, which is impossible for an individual ant [76]. A type of wild birds in the UK have learned to open milk bottles using social learning [24]. The paper [6] provides evidence that the knowledge is transmitted across generations, demonstrating cultural transmission of knowledge among these birds.

Heterogeneity among individuals in a group is a critical factor that can substantially impact social learning and collective intelligence. The presence of diverse skills, knowledge, and perspectives within the group allows for a broader range of ideas and opinions to be considered. This, in turn, can lead to a more comprehensive understanding of the problem at hand and the development of more innovative and effective solutions. In the realm of group dynamics, agent heterogeneity can be strategically leveraged to improve both individual and group performance. By bringing together individuals with different backgrounds, experiences, and cognitive styles, a group can harness the power of collective wisdom, which can result in higher quality decision-making, more creative problem-solving, and increased adaptability in the face of challenges.

As the integration of AI agents into society becomes more widespread, it is increasingly important to investigate how these agents can benefit from social interactions and develop collective intelligence. While AI agents have impressive capabilities at an individual level, they currently lack the ability to work together effectively in groups. The rise of human-AI coordination in various domains, from autonomous cars to human-robot applications and non-embodied AI agents such as chatbots and recommender systems, highlights the necessity of developing effective methods for AI agents to interact with humans. The key aspects of successful human-AI coordination are communication, coordination, and cooperation, along with the ability to generalize

or adapt to novel situations in the presence of humans and tasks involving humans. These elements are critical even in AI-AI interactions. In the context of human-AI coordination, it is also essential to develop AI agents that can comprehend human behavior, values, and preferences. As such, research into enhancing the social skills and abilities of AI agents has become increasingly important.

This dissertation focuses on two main research questions: 1.) how AI agents can learn to effectively communicate with other agents and 2.) how AI agents can improve their ability to generalize or adapt to novel partners/opponents through social interactions. The first section of this dissertation is dedicated to developing methodologies that enable effective communication among AI agents under different communication constraints. Here we specifically focus on communication in sequential decision making tasks in uncertain environments, where the primary challenge is to balance exploration and exploitation to achieve optimal performance. We investigate the role of agent heterogeneity in individual and group performance and how agents can leverage individual differences to improve group performance. The second section of this dissertation explores the topic of generalization in multi-agent AI. Our research work centers on how agents can adapt their policies to collaborate with novel agents they haven't encountered before in tasks that require coordination and cooperation among agents to achieve optimal outcomes. We explore how inherent agent heterogeneity leads to agents with diverse strategies and subsequently how this leads to better generalization. By investigating these research areas, we aim to enhance our understanding of how AI agents can effectively work together in complex social scenarios, contributing to the advancement of the field of AI.

1.1.1 Communication

How agents can learn to effectively communicate in sequential decision-making tasks is an area that has attracted significant attention from both industry and academia

due to its wide range of applications. We study a simple framework where all agents interact with the same environment, and the payoff an agent receives for an action is independent of the actions or past actions of other agents. However, despite the lack of direct dependence, agents can significantly influence the performance of other agents through communication.

Effective communication is crucial for agents to reduce uncertainty and make better-informed choices in sequential decision-making tasks. However, in real-world settings agents can face different communication constraints including costs, potential failures and bandwidth limitations. Our research explores strategies for effective communication among agents under various communication constraints. We study heterogeneity with respect to the agent specific rate of information sharing and how it affects individual and group performance. We also study the role of heterogeneity with respect to where they are in the communication network and how individual differences influence the performance at individual and group level.

1.1.2 Generalization

Despite their impressive performance, machine learning solutions remain predominantly single-skilled and fragile. Developing learning based solutions, each capable of solving only a specified task, to numerous problems can be prohibitively expensive necessitating the methodological development of generalizable solutions. Reinforcement learning seeks to develop policies that generalize to novel tasks and environments. The inherently multi-agent nature of the real world brings developing generalizable solutions to multi-agent problems to the fore. This requires generalization of solution concepts along agent dimension as well as task dimension and environment dimension, exponentially increasing the complexity of the problem, typically rendering the process infeasible. Taking a more pragmatic approach, state-of-the-art research mainly

focuses on generalizing solutions along each dimension decoupled from other dimensions.

A key area of focus in zero-shot generalization in multi-agent reinforcement learning is developing policies that generalize to new agents. To a large extent the existing work in zero-shot generalization focuses on common payoff games or zero-sum games. However, in the real world we often encounter mixed motive games wherein individual goals and socially optimal outcome are misaligned. In this work we investigate how to leverage heterogeneity in social preferences of agents to improve generalization in mixed-motive games.

1.2 Outline of Contributions

We make following contributions in Part I of this dissertation

1. In Chapter 2 we introduce the notations used throughout this dissertation. We also introduce the main computation frameworks multi-agent multi-armed bandit problem and multi-agent reinforcement learning problem. We provide background theoretical results, discuss widely used algorithms and provide brief descriptions of the relevant concepts used in later chapters. We conclude this chapter describing related work in the literature.

2. In Chapter 3 we make contributions to the development of efficient communication protocols in multi-agent multi-armed bandits, where communication among agents is costly. By highlighting the practical relevance of communication cost in real-world applications and measuring it based on the number of messages transmitted, the chapter emphasizes the importance of sharing information about suboptimal options to minimize costs while maximizing the group's cumulative reward. We introduce ComEx communication protocol,

which promotes selective sharing and fosters collaboration among agents, offers a promising solution that balances communication costs, information value, and performance, ultimately resulting in improved decision-making and higher cumulative rewards for the group. We provide theoretical guarantees for the group performance and illustrate the results using numerical simulations.

3. In Chapter 4 we delve into the analysis of probabilistic communication in multi-agent multi-armed bandit problems, specifically examining how individual agents and groups perform under these constraints. The research investigates the role of degree heterogeneity in the communication network, revealing how an agent's connectivity within the network impacts their exploration-exploitation balance and overall performance. The chapter also explores agent-specific communication probabilities and their influence on agent performance. By understanding these dynamics and leveraging individual heterogeneity, adaptive strategies can be developed to improve the performance of multi-agent systems in various applications.

4. In Chapter 5 we explore the impact of behavioral heterogeneity among agents on zero-shot generalization in multi-agent reinforcement learning. By incorporating insights from Social Value Orientation (SVO) research, we investigate the role of diverse social preferences in generating diverse agent behaviors, which can lead to improved generalization when interacting with novel agents in test scenarios. We assess the effects of heterogeneous SVO in a range of incentive structures, such as Prisoner's Dilemma, Chicken, and Stag Hunt, in sequential social dilemmas. We demonstrate that leveraging the resulting diversity through best-response strategies can enhance zero-shot generalization in equilibrium selection sequential social dilemmas. However, we also identify scenarios where training best response may lead to poor generalization, emphasizing the

importance of further research into understanding and exploiting the interplay between diverse social preferences and agent policies.

5. In Chapter 6 we provide conclusions and future research directions for the research presented in this dissertation.

Chapter 2

Background

2.1 Notations

For any positive integer N we denote the set $\{1, 2, \ldots, N\}$ as $[N]$. We define $\mathbf{1}\{x\}$ as an indicator variable that takes value 1 if x is true and 0 otherwise. Further, we use $X \backslash x$ to denote the set X excluding the element x. We use $|X|$ to denote the number of elements in set X. For any general graph G we define $\bar{\chi}(G), \bar{\gamma}(G)$ as clique covering number and dominating number respectively. We use G_γ to denote the γ^{th} power graph of G. We use $\mathsf{P}(\mathcal{A})$ to denote the probability of an event $\mathcal{A}$ and $\mathbb{E}[Z]$ to denote the expectation of a random variable Z. We use $\mathbb{R}$ to denote the real numbers.

2.2 Multi-armed bandits

The multi-armed bandit problem [87, 44] is a mathematical framework for capturing the salient features of sequential decision-making under uncertainty, where an agent is faced with a set of alternatives, each with an unknown reward distribution. In this problem, an agent must decide which alternative to choose to maximize the cumulative reward over time, while balancing the trade-off between exploration and exploitation. Exploration involves trying new alternatives to learn about their reward distributions,

while exploitation involves choosing alternatives that have already shown promising rewards. This problem is named after the *one-armed bandit* slot machines, which have a lever (or *arm*) that a gambler pulls to randomly select one of several possible outcomes.

The multi-armed bandit problem has a wide range of practical applications, such as in clinical trials [22], recommendation systems and user-targeted online advertising [89]. For instance, in clinical trials, a researcher or a medical professional may want to test several treatments, each with an unknown effect on the disease, and must decide how to allocate patients to these treatments to maximize the chances of identifying the most effective one. In recommendation systems, a website may want to recommend which item to show to a user, based on the user's past behavior and preferences, to maximize the user's satisfaction. In online advertising, a company may want to choose which ad to show to a user, based on the user's behavior and preferences, to maximize the click-through rate.

The multi-armed bandit problem has been studied extensively in the literature, and many algorithms have been proposed to solve it efficiently. These algorithms can be broadly classified into two categories: heuristic algorithms and optimal algorithms. Heuristic algorithms are simple and easy to implement, but they may not provide the optimal solution. Optimal algorithms, on the other hand, guarantee the optimal solution, but they may be computationally expensive or require unrealistic assumptions about the reward distributions. The choice of algorithm depends on the specific problem and the available resources.

2.2.1 Single-agent Multi-armed bandit problem

We consider the multi-armed bandit problem in which an agent chooses among K arms. Each arm k has a fixed reward distribution $\mathcal{P}_k$ that is sub-Gaussian with mean μ_k and variance proxy σ_k^2. The reward distributions are not known to the agent.

Definition 1. *A random variable $X \in \mathbb{R}$ is sub-Gaussian with variance proxy σ^2 if $\mathbb{E}[X] = 0$ and its moment generating function satisfies*

$$\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right), \forall s \in \mathbb{R}$$

Let T be the decision making time horizon. At each time step $t \in [T]$ the agent chooses an arm a_t and receives a numerical reward r_t, which is drawn from the probability distribution of the chosen arm. Without loss of generality we assume that the mean rewards of the arms are ordered in descending order, i.e., $\mu_1 \geq \mu_2 \ldots \geq \mu_K$. The expected reward gap between the optimal arm (i.e., the arm with the highest mean reward) and arm k is defined as $\Delta_k = \mu_1 - \mu_k$ for all $k > 1$. We denote the minimum expected reward gap as $\bar{\Delta} = \min_{k \neq 1, k \in [K]} \Delta_k$.

The goal of the agent is maximizing the reward accumulated over the decision making time horizon. The cumulative reward of the agent can be given as $R_T = \sum_{t=1}^{T} r_t$. The performance of the agent is measured using the expected cumulative regret, which is defined as the sum of the expected reward gaps of suboptimal arms times number of times suboptimal arms are chosen,

$$\texttt{Reg} = \sum_{k=2}^{K} \Delta_k \mathbb{E}[n_k(t)]. \tag{2.1}$$

where $n_k(t)$ is a random variable that denotes the number of times arm k has been selected up to time t. This captures the expected regret suffered by the agent when drawing suboptimal arms.

2.2.2 Lower bound

In their seminal paper Lai and Robbins [44] established a lower bound on the expected cumulative regret for the single-agent multi-armed bandit problem. This bound sets

a limit on the maximum expected achievable performance. The lower bound on the number of times a suboptimal arm is chosen up to and including time step T, given a general probability distribution $\mathcal{P}_k$ defining reward for each arm k, is given by:

$$\mathbb{E}[n_k(T)] \geq \left(\frac{1}{\mathcal{D}(\mathcal{P}_k||\mathcal{P}_1)} + o(1)\right) \log T$$

where $\mathcal{D}(\mathcal{P}_k||\mathcal{P}_1)$ represents the Kullback-Leibler divergence between distributions p_k and p_1. For Gaussian rewards with known variance, the above simplifies to:

$$\mathbb{E}[n_k(T)] \geq \left(\frac{2\sigma^2}{\Delta_k^2} + o(1)\right) \log T$$

Then a lower bound for the expected cumulative regret can be given as

$$\texttt{Reg} \geq \sum_{k=1}^{K} \left(\frac{2\sigma^2}{\Delta_k} + o(1)\right) \log T.$$

This lower bound is crucial in setting a benchmark for evaluating the performance of algorithms in the multi-armed bandit problem.

2.2.3 Upper Confidence Bound (UCB) algorithm

In this section we discuss one of the widely used algorithms in multi-armed bandit research. The UCB algorithm [7] is a simple and effective approach that aims to balance exploration and exploitation by using a confidence interval to estimate the upper bound of the true mean reward of each arm. At each time step, the agent selects the arm with the highest upper confidence bound, which balances between choosing the arm with the highest expected reward and exploring other arms. Let $\widehat{\mu}_k(t)$ denote the estimated mean reward of arm k at time step t. Define the upper

confidence bound at time step t as

$$\texttt{UCB}_k(t) = \widehat{\mu}_k(t) + C_k(t) \tag{2.2}$$

where $C_k(t) = \sigma\sqrt{2(\xi+1)\frac{\log t}{n_k(t)}}$. The term $C_k(t)$ captures the uncertainity associated with the estimated mean of the arm k at time step t. $C_k(t)$ increases with the natural logarithm of t and decreases with the number of times the arm k has been sampled until time step t. Then the samplig rule of the agent can be given as

$$a_{t+1} = \arg\max_k \texttt{UCB}_k(t). \tag{2.3}$$

Now we provide a proof of an upper bound for the expected cumulative regret under UCB algorithm.

Lemma 1. (Restatement of results from [7]) *Let* $\eta_k = \left(\frac{8(\xi+1)\sigma^2}{\Delta_k^2}\right)\log T$. *For any suboptimal arm* k *and* $\forall t$ *we have*

$$\mathsf{P}\left(a_{t+1} = k, n_k(t) > \eta_k\right) \leq \mathsf{P}\left(\widehat{\mu}_1(t) \leq \mu_1 - C_1(t)\right) + \mathsf{P}\left(\widehat{\mu}_k(t) \geq \mu_k + C_k(t)\right)$$

Proof. Note that for any $k > 1$ we have

$$\begin{aligned}\{a_{t+1} = k\} &\subset \{\texttt{UCB}_k(t) \geq \texttt{UCB}_1(t)\} \\ &\subset \left\{\{\mu_1 < \mu_k + 2C_k(t)\} \cup \{\widehat{\mu}_1(t) \leq \mu_1 - C_1(t)\} \cup \{\widehat{\mu}_k(t) \geq \mu_k + C_k(t)\}\right\}.\end{aligned}$$

Let $\eta_k = \left(\frac{8(\xi+1)\sigma^2}{\Delta_k^2}\right)\log T$. Since $n_k(t) > \eta_k$ the event $\{\mu_1 < \mu_k + 2C_k(t)\}$ does not occur. Thus we have

$$\mathsf{P}\left(a_{t+1} = k, n_k(t) > \eta_k\right) \leq \mathsf{P}\left(\widehat{\mu}_1(t) \leq \mu_1 - C_1(t)\right) + \mathsf{P}\left(\widehat{\mu}_k(t) \geq \mu_k + C_k(t)\right)$$

This concludes the proof of Lemma 1. □

Theorem 1. (Expected cumulative number of suboptimal option samples) *Let* $\eta_k = \left(\frac{8(\xi+1)\sigma^2}{\Delta_k^2}\right)\log T$. *Then we have*

$$\mathbb{E}\left[n_k(T)\right] \leq \eta_k + \sum_{t=1}^{T} \mathsf{P}\left(\widehat{\mu}_1(t) \leq \mu_1 - C_1(t)\right) + \sum_{t=1}^{T} \mathsf{P}\left(\widehat{\mu}_k(t) \geq \mu_k + C_k(t)\right)$$

Proof. Note that for each suboptimal arm $k > 1$ we have

$$\mathbb{E}\left[n_k(T)\right] = \sum_{t=1}^{T} \mathsf{P}\left(a_t = k\right)$$

Let τ_k be the maximum time step such that the agent has sampled the arm k at most η_k number of times. $\tau_k = \left\{t \in [T] : \sum_{\tau=1}^{t} \mathbf{1}\{a_\tau = k\} \leq \eta_k\right\}$ Then we have

$$\begin{aligned}\mathbb{E}\left[n_k(T)\right] &= \sum_{t=1}^{\tau_k} \mathsf{P}\left(a_t = k\right) + \sum_{t>\tau_k}^{T} \mathsf{P}\left(a_t = k\right) \\ &\leq \eta_k + \sum_{t=1}^{T} \mathsf{P}\left(a_t = k, n_k(t) > \eta_k\right)\end{aligned}$$

From Lemma 1 we have

$$\mathbb{E}\left[n_k(T)\right] \leq \eta_k + \sum_{t=1}^{T} \mathsf{P}\left(\widehat{\mu}_1(t) \leq \mu_1 - C_1(t)\right) + \sum_{t=1}^{T} \mathsf{P}\left(\widehat{\mu}_k(t) \geq \mu_k + C_k(t)\right)$$

□

2.2.4 Multi-agent multi-armed bandits

Multi-agent multi-armed bandit problem considers N agents who are interacting with the same bandits environment. At each time step $t \in [T]$ each agent i chooses an option $a_t^{(i)}$ and receives a numerical reward $r_t^{(i)}$ drawn from the unknown probability distribution associated with the chosen option. The goal of each agent is maximizing the individual cumulative reward. While the rewards received by an agent do

not depend on the actions of other agents, they can share information to enhance their performance. As all agents are choosing from the same set of options, sharing information about the options can reduce uncertainty and improve their performance.

In this thesis we consider that agents are sharing information according to a general communication network. Let $G(V, E)$ be a general graph that encodes the hard communication constraints among agents. The vertex set V is the set of agents $[N]$ and each edge $(i, j) \in E$ indicates that agents i and j are neighbors. We consider that agents directly communicate with their neighbors only. Let $\mathbf{1}\{(i, i) \in E\} = 1, \forall i$. At each time step t we define the communication between agents by $G_t(V, E_t)$ where $E_t \subseteq E$. Let $d^{(i)}$ be the degree of agent i. Let G_γ denote the γ^{th} power graph of G. Denote $d_\gamma^{(i)}$ to be the degree of agent i in graph G_γ, i.e., number of agents within a distance of γ from agent i in graph G. For any γ let $d_\gamma^{(i)^+} = d_\gamma^{(i)} + 1$.

We denote $\mathbf{m}_t^{(i)}$ as the message shared by agent i at time t with its neighbors. This can be either a single message containing information about a particular arm pull, typically the last arm pull of agent i, or a concatenation of information about several arm pulls by more than one agent over several previous time steps. We define $n_k^{(i)}(t) := \sum_{\tau=1}^{t} \mathbf{1}\Big\{a_\tau^{(i)} = k\Big\}$ and $N_k^{(i)}(t) := \sum_{\tau=1}^{t} \sum_{j=1}^{N} \mathbf{1}\Big\{a_\tau^{(j)} = k\Big\}\mathbf{1}\{(i, j) \in E_\tau\}$ to be the number of times until time step t that agent i pulled arm k and observed reward values from arm k, respectively. Note that the number of observations $N_k^{(i)}(t)$ is the sum of the number of pulls drawn by agent i of arm k and the number of times agent i received reward values of arm k from its neighbors. Let $\widehat{\mu}_k^{(i)}(t)$ denote agent i's estimated average reward of arm k at time t.

Similar to the single agent case we measure the performance of the group by total loss suffered by agents due to choosing suboptimal options. The expected cumulative group regret can be given as

$$\mathtt{Reg}_G = \sum_{i=1}^{N} \sum_{k=2}^{K} \Delta_k \mathbb{E}[n_k^{(i)}(t)] \tag{2.4}$$

We define the communication cost as the number of messages shared by agents. We consider the cost of sharing a concatenated message to be the number of single messages included in it. Let $L(t)$ be the cumulative group communication cost at time t. Then, the expected group communication cost can be given as $\mathbb{E}[L(t)] := \sum_{i=1}^{N} \sum_{\tau=1}^{t} \mathbb{E}\left[\left|\mathbf{m}_{\tau}^{(i)}\right|\right]$.

2.2.5 Related work

Decentralized reward sharing. In decentralized reward sharing agents share instantaneous rewards with their neighbors [14, 42, 94]. The paper [42] considered that neighbors are defined according to a fixed communication graph and provide graph structure dependent regret bounds. The paper [14, 61] studied the cooperative bandit problem with time varying communication structures. The papers [13, 10, 19] considered message passing communication rules where each agent initiates a message and send the message to its neighbors. A message received from a neighbor is subsequently forwarded to other neighbors.

Decentralized estimate sharing. In estimate sharing each agent share the estimated average reward and number of arm pulls from each arm with its neighbors defined according to a fixed communication graph. The paper [84] considered a P2P communication where an agent is only allowed to communicate with two other agents at each time step. The papers [47, 46, 66, 49] used a running consensus algorithm to update estimates and provide graph-structure-dependent performance.

Centralized leader-follower setting. A communication strategy where agents observe the rewards and choices of their neighbors according to a leader-follower setting is considered in [48, 42, 94]. In [48, 42], followers pull the last arm pulled by their neighbors. In [94] one leader explores and estimates the mean reward of arms, while all other agents pull the arm with highest estimated mean per the leader.

Communication cost. The paper [86] considered a pure exploration bandit problem and measures the communication by the number of times agents communicate. [94] proposed a leader-follower algorithm with a constant communication cost. The paper [95] proposed an algorithm that achieves near-optimal performance where agents achieve sublinear expected regret. In their work, communication cost is independent of time and measured by the amount of data transmitted.

Distributed Thompson sampling. Recently [92, 45] proposed distributed Thompson sampling rules. The paper [92] studied the problem with sparse communication structures. The paper [45] provided regret guarantees that matches the corresponding centralized regret guarantees.

2.3 Reinforcement learning

Reinforcement learning (RL) is a type of machine learning where an agent learns to make decisions through interactions with its environment. In RL, the agent receives feedback in the form of rewards or penalties based on its actions, and its goal is to maximize the cumulative reward over time. The agent learns to make better decisions by using trial and error, exploring different actions and observing the outcomes.

2.3.1 Multi-agent reinforcement learning

We consider a multi-agent partially observable Markov decision process defined by the tuple $\left\langle N, \mathcal{S}, \mathcal{A}, \mathcal{R}, P, \gamma \right\rangle$, where N is the number of agents, $\mathcal{S}$ is the joint state space, $\mathcal{A} = \times_{i=1}^{N} \mathcal{A}^i$ is the joint action space, P is the state transition probability distribution, $\mathcal{R}$ is the reward function and γ is the discount factor. At each time step t, each agent $i \in 1, \ldots, N$ observes a private (local) observation o_t^i and takes an action a_t^i from a set of actions $\mathcal{A}^i$. The joint action of all agents at time step t is denoted as $a_t = (a_t^1, \ldots, a_t^N)$. The state s_t is unobservable, and the partial

observation o_t^i depends on the current state of the environment s_t and the agent's observation function. The observation function for agent i is denoted as $O^i(o_t^i|s_t)$. Each agent i receives a reward r_t^i which is a function of the joint action a_t and the state s_t of the environment. The state of the environment transitions according to a probability distribution $P(s_{t+1}|s_t, a_t)$.

The objective of each agent i is to maximize their cumulative expected discounted reward, over a given finite time horizon, defined as $J^i = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t^i\right]$, where $\gamma \in [0, 1]$ balances the importance of immediate and future rewards. The agents' policies are defined as the mapping from the agent's observation history to an action, i.e., $\pi^i(a_t^i|o_1^i, \cdots, o_t^i)$. The policies are updated using a multi-agent reinforcement learning algorithm that maximizes the agents' objective functions.

Multi-agent reinforcement learning (MARL) extends RL to the setting where there are multiple agents interacting with each other and the environment. In MARL, each agent has its own observation, action, and reward signals, and its own policy. The agents may have different goals, and their policies may be interdependent, meaning that the actions of one agent affect the rewards and observations of the other agents.

Generalization is an important problem in MARL, as the agents need to be able to adapt to new environments and situations that they have not encountered during training. Generalization in MARL can be achieved through transfer learning, where knowledge learned in one task is transferred to a new task, or through meta-learning, where the agents learn to learn from experience. Another approach to generalization in MARL is to use function approximation, where the agents learn to approximate the value function or the policy using a function approximator, such as a neural network. However, using function approximation can introduce new challenges, such as overfitting and non-stationarity, and require careful regularization and adaptation techniques.

In summary, reinforcement learning is a powerful paradigm for learning to make decisions through trial and error. Multi-agent reinforcement learning extends this paradigm to the setting where there are multiple agents interacting with each other and the environment. Generalization is an important problem in MARL, and can be achieved through transfer learning, meta-learning, or function approximation, although each approach has its own challenges and limitations.

2.3.2 Sequential social dilemmas

Social dilemmas are situations in which individual self-interest conflicts with the collective interest of a group. These dilemmas arise when individuals make choices that are rational for themselves, but collectively lead to a less desirable outcome for everyone. In other words, social dilemmas are situations in which the pursuit of self-interest can lead to a worse outcome for the group as a whole, compared to a scenario in which individuals put aside their own self-interest and cooperate with each other.

Examples of social dilemmas can be found in a wide range of contexts, from environmental issues such as pollution and deforestation to economic issues such as the tragedy of the commons and the prisoner's dilemma in game theory. In each of these cases, individuals must choose between behaving in a way that benefits themselves in the short term but harms the group in the long term, or cooperating with others for the greater good.

Understanding social dilemmas is important because they are pervasive in human societies and can have significant impacts on our collective well-being. By studying social dilemmas, we can develop strategies for promoting cooperation and mitigating the negative consequences of individual self-interest.

Sequential social dilemmas [51] are a class of social dilemma in which the decision-making process of the interacting agents is temporally and spatially extended. Performing well in a sequential social dilemmas tends to require the consideration of

long-term consequences, interdependence, and cooperation among group members. Research on sequential social dilemmas has been widely studied in the context of emergence and maintenance of cooperation [52, 75], inequity aversion [37], partner choices [21, 67] wherein agents have a choice with whom to interact, and generalization [69, 1] wherein agents interact with novel scenarios during test time.

2.3.3 Social Value Orientation

In psychology research, Social Value Orientation (SVO) is a cognitive construct reflecting a person's preference for resource allocation between themselves and others [28, 54, 71]. While some individuals may solipsistically focus on maximizing their personal success, others demonstrate different motivations, including maximizing the difference between their own and others' outcomes (a competitive orientation), maximizing collective welfare (a prosocial orientation), or maximizing other peoples' benefit (an altruistic orientation).

In artificial intelligence research, various algorithms draw inspiration from these insights in their design or implementation [68, 80]. In reinforcement learning, SVO is an intrinsic motivation that transforms an agent's reward based on its particular target distribution between its reward and the reward of others. Recently, there's been research investigating the role of SVO in situations where a group of agents or players interact in ways that involve trade-offs between their self-interest and the collective interest of the group. This research has generated valuable insights into the impact of SVO on the emergence of diverse behaviors and cooperation [68], generalization [69] wherein agents interact with novel scenarios during test time, and partner choices [67] in sequential social dilemmas. SVO research has focused primarily on social dilemmas with underlying incentive structures resembling the *prisoner's dilemma* [77], wherein each player has an incentive to defect, even though both would be better off if they both cooperated.

Omitting the dependence on time t, let r^i be the reward of agent i. Let $\bar{r}^{-i}$ be the average reward of all the agent except agent i. Then we have

$$\bar{r}^{-i} = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} r^j.$$

Let svo^i denote the SVO target angle of agent i. Following the definition given in [67], we define the effective reward $\hat{r}^i$ of agent i as

$$\hat{r}^i = r^i \cos(svo^i) + \bar{r}^{-i} \sin(svo^i).$$

Then agent i optimizes the objective function

$$\hat{J}^i = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \hat{r}_t^i\right].$$

Chapter 3

Efficient communication under communication cost

In this chapter we present our work on developing efficient communication protocols when communication among agents is costly. The chapter summarizes and extends the work presented in papers [63, 60], also included in Chapters 8 and 9.

3.1 Motivation

Studying communication cost in multi-agent multi-armed bandits is important due to its practical relevance in a variety of real-world applications. In multi-agent systems, agents interact with a common environment and make decisions based on the feedback received. Communication among agents can significantly improve their performance; however, it can also be costly, and maintaining a continuous high-bandwidth communication network may not be possible in many real-world scenarios. Thus, developing communication-efficient algorithms that minimize the communication overhead while maintaining performance is essential.

One real-world example where communication-efficient algorithms are crucial is in the field of recommender systems. In such systems, multiple servers are networked

to handle high demands, and the high communication between servers can lead to service latency. Similarly, in a group of robots, communication can increase battery power consumption. Therefore, communication-efficient algorithms can help minimize the communication overhead, enabling the agents to make optimal decisions while conserving resources.

Another example where communication cost is critical is in the field of clinical trials. In clinical trials, groups of patients are often assigned to different treatments, and their responses are recorded over time. By sharing information about the treatments and their outcomes, physicians can make more informed decisions about which treatments to prescribe, leading to better patient outcomes. However, communication among physicians can be time-consuming and costly, making it important to develop communication-efficient algorithms that can reduce the communication overhead while maintaining the same level of performance.

In summary, the importance of studying communication cost in multi-agent multi-armed bandits lies in its practical relevance to real-world applications such as recommender systems and clinical trials. Developing communication-efficient algorithms can significantly improve the performance of agents while minimizing the communication overhead, making it possible to make optimal decisions while conserving resources.

3.2 Communication cost

There are several ways to measure the communication cost, which depends on the specific context and the problem formulation.

One common method of measuring communication cost is to count the total number of messages transmitted between agents. While this method is simple and easy

to implement, it may not account for the size of the messages transmitted, which can vary depending on the amount of information being shared.

Another approach to measuring communication cost is to consider the number of bits transmitted between agents. This method takes into account the size of the messages and is more precise than the previous method. However, it assumes that all messages have the same size, which may not be true in practice.

A more advanced way of measuring communication cost is to consider the energy consumption associated with communication. This method is particularly relevant in wireless communication systems, where energy consumption can be a limiting factor.

Finally, the time required to transmit messages between agents is another way of measuring communication cost. This method is particularly relevant in real-time systems, where delays can have a significant impact on the system's performance.

In the work on multi-agent multi-armed bandits presented in this chapter, we measure communication cost based on the number of messages transmitted between agents. While this method may not be the most sophisticated, it is a reasonable measure of communication cost in many practical scenarios. Moreover, since the message size and energy consumption are not significant factors in our context, counting the number of messages transmitted provides an appropriate measure of communication cost.

3.3 Efficient communication

In scenarios where communication costs are associated with multi-agent multi-armed bandits, it is essential for agents to determine the most effective way to share valuable information. By communicating the most useful information, agents can minimize costs while maximizing the group's cumulative reward. This section highlights

the significance of sharing information about suboptimal options, as it provides the greatest benefit to agents in their decision-making process.

Sharing information about suboptimal options is particularly valuable, as it enables agents to reduce the uncertainty associated with suboptimal arms more rapidly without having to choose them multiple times. By utilizing communicated messages, agents gain a better understanding of suboptimal options, increasing the likelihood of choosing the optimal option.

Furthermore, in multi-agent multi-armed bandits, communication costs can become a significant barrier, especially when the number of agents or options is large. Consequently, striking a balance between minimizing communication and maintaining strong performance is crucial. Efficient stochastic bandit algorithms typically select suboptimal options logarithmically over time. As such, communicating only rewards from suboptimal options can significantly reduce communication costs. Since agents choose suboptimal options a logarithmic number of times under an optimal algorithm, this approach results in a logarithmic communication cost, effectively balancing both communication and performance aspects.

By focusing on sharing information about suboptimal options, agents can successfully minimize communication costs and make better-informed decisions, ultimately increasing the group's cumulative reward. This approach demonstrates the importance of strategically communicating the most useful information to maximize the benefits of communication in multi-agent multi-armed bandits.

3.4 Explore based communication

In multi-agent multi-armed bandits, while sharing information about suboptimal options is highly beneficial, a critical challenge arises due to the fact that agents initially do not know which options are suboptimal. This is because the probability distribu-

tions associated with the options are unknown at the outset, and agents must rely on exploration to gain insights into the reward distributions associated with the options. Therefore, the development of an effective communication protocol that balances the need for exploration and the sharing of valuable information becomes crucial.

A promising heuristic for addressing this challenge is a communication protocol that shares information among agents when they choose an option that is not the one with the highest empirical average reward. This approach provides several advantages, which contribute to the overall effectiveness of the group's decision-making process.

One significant benefit of this approach is the reduction of exploration redundancy. By sharing information about options that are not the current highest empirical average reward, agents can avoid duplicating exploration efforts on suboptimal options. Instead, they can focus on exploring other potentially more rewarding options, thereby accelerating the discovery of the optimal choice.

Additionally, this heuristic communication protocol speeds up the convergence to the optimal option. As agents communicate their experiences with non-optimal options, they can update their estimates and more quickly eliminate suboptimal choices. This accelerated elimination process increases the likelihood of selecting the optimal option sooner, resulting in improved performance and higher cumulative rewards for the group.

The balance between communication cost and information value is another crucial aspect of this communication protocol. By only sharing information when agents choose options other than the one with the highest empirical average reward, agents communicate less frequently, effectively reducing communication costs. This selective sharing ensures that the shared information has a high value, as it pertains to suboptimal options, which is essential for refining the agents' decision-making process.

Moreover, a communication protocol that shares suboptimal option information fosters collaboration among agents, as it allows them to learn from each other's experiences. This cooperative learning enables agents to adapt their strategies more efficiently, making the entire group more effective at identifying and selecting the optimal option.

Now we define our communication protocol name *comEx* as follows. Let $g(M, x) = M + \sum_{i=1}^{N} \left(12 \log(3(x+1)) + 3 \log\left(x+1\right)\right)$.

Definition 2. (ComEx communication protocol) *Each agent i initiates sharing the message* $m_t^{(i)} := \left\langle i, t, a_t^{(i)}, X_t^{(i)} \right\rangle$ *if* $a_t^{(i)} \neq \arg\max_{k \in [K]} \widehat{\mu}_k^{(i)}(t-1)$

Algorithm 1: ComEx

Input: Bandit environment, algorithm parameters
for each iteration $t \in [T]$ **do**
  for each agent $i \in [N]$ **do**
    // Sampling phase
    Sampling rules: multi-agent UCB
    // Message generating phase
    // Replace full communication with ComEx
    /* ComEx communication protocol */
    if $A_t^{(i)} \neq \arg\max_k \widehat{\mu}_k^{(i)}(t-1)$ **then**
      CREATE $\left(m_t^{(i)} := \left\langle i, t, A_t^{(i)}, X_t^{(i)} \right\rangle\right)$
    end
  end
  for each agent $i \in [N]$ **do**
    // Communication phase
    Communication rule: Decentralized (or centralized) instantaneous reward sharing, Decentralized (or centralized) message passing
    // Estimate updating phase
  end
  for each arm $k \in [K]$ **do**
    CALCULATE $\left(\widehat{\mu}_k^{(i)}(t), N_k^{(i)}(t)\right)$
  end
end

3.4.1 Decentralized instantaneous reward sharing UCB

We present our first algorithm ComEx-UCB by combining the above communication protocol with instantaneous reward sharing. Each agent follows a sampling rule that balances exploiting with exploring. We use a natural extension of Upper Confidence Bound (UCB) algorithm as a sampling rule. In UCB at each time step t for each arm k each agent i constructs an upper confidence bound, i.e., the sum of its estimated expected reward (empirical average of the observed rewards) and the uncertainty associated with the estimate $C_k^{(i)}(t) := \sigma\sqrt{\frac{2(\xi+1)\log t}{N_k^{(i)}(t)}}$ where $\xi > 1.1$, and pull the arm with highest bound. If the pulled arm is instantaneously suboptimal, the agent sends a message $m_t^{(i)} := \left\langle A_t^{(i)}, X_t^{(i)} \right\rangle$ to its neighbors (see Definition 6). Note that under this communication rule agents do not share concatenated messages. Thus passing information about time step and agent id is redundant. Pseudo code for ComEx-UCB is given in Appendix 9.10.12.

Theorem 2. (Group regret of ComEx-UCB) *Consider a group of N agents following ComEx-UCB while sharing instantaneous rewards over a general communication graph G. Then for any $\xi \geq 1.1$ expected cumulative group regret satisfies:*

$$\mathbb{E}\left[R(T)\right] \leq \sum_{k=2}^{K} \frac{8(\xi+1)\sigma}{\Delta_k} \bar{\chi}(G) \log T + \sum_{k=2}^{K} \Delta_k g(4N, d^{(i)})$$

Proof sketch. We follow an approach similar to the standard UCB analysis [7, 19] with a few key modifications. We partition the communication graph into a set of non overlapping cliques and analyze the regret of each clique and take the summation over cliques to obtain the regret of the group. When agents are using full communication group regret can be given as the summation of a $\log T$ term that scales with the clique covering number $\bar{\chi}(G)$ and a term, which is independent of T. The second term depends on the summation of tail probabilities of arms, i.e., $\mathsf{P}\left(\left|\widehat{\mu}_k^{(i)}(t) - \mu_k\right| \geq C_k^{(i)}(t)\right)$. For full communication a similar result can be found in

[19]. Note that full communication is a deterministic communication protocol and ComEx-UCB is a stochastic communication protocol that depends on the decision making process. Two major technical challenges in proving the regret bound for ComEx-UCB are 1.) deriving a tail probability bound for the case in which the communication between agents are stochastic and 2.) bounding the additional regret incurred by not sharing information when pulling the arm with highest estimated average reward, i.e., $a_t^{(i)} = \arg\max_{k\in[K]} \widehat{\mu}_k^{(i)}(t-1)$. We overcome the first challenge by noticing that communication random variables $\mathbf{1}\{(i,j) \in E_t\}, \forall i,j,t$ are previsible, i.e., measurable with respect to the sigma algebra generated by information obtained up to time $t-1$. We address the second challenge by proving that the number of times agents do not share information about any suboptimal arm k can be bounded by tail probabilities of arm k and the optimal arm. A complete proof of Theorem 2 is given in Appendix **??**. □

Remark 1. *By replacing ComEx with full communication in ComEx-UCB algorithm agents obtain an expected cumulative group regret of* $\mathbb{E}[R(T)] = O(K\bar{\chi}(G)\log T + KN)$ *(Appendix H). Thus from Theorem 2 we see that ComEx obtains the same order of performance as full communication.*

Recall that expected communication cost under full communication is $\Theta(T)$. Now we prove that expected communication cost under ComEx is logarithmic in time. In ComEx-UCB algorithm agents are only sending single messages (not concatenated). Thus expected group communication cost at time step t can be given as $\mathbb{E}[L(t)] = \sum_{i=1}^{N}\sum_{\tau=1}^{T} \mathsf{P}\left(A_\tau^{(i)} \neq \arg\max_{k\in[K]} \widehat{\mu}_k^{(i)}(\tau-1)\right)$.

3.4.2 Decentralized message passing UCB

We propose ComEx-MPUCB an improved version of ComEx-UCB by incorporating a message passing method [10, 19] that allows agents to share the messages they

initiated with agents who are within a distance of γ. We call γ *communication density parameter.* We consider that at time t each agent i initiates a message $m_t^{(i)} := \left\langle i, t, A_t^{(i)}, X_t^{(i)} \right\rangle$ according to ComEx given in Definition 6 and sends the messages to its neighbors. Subsequently the agents who receive the message forward it to their neighbors. Messages received at time t are forwarded to neighbors at time $t+1$ resulting that each hop adds a delay of 1 time step. Under this message passing method γ-hop neighbors receive the message after a delay of γ time steps. Agents do not forward the messages that are older than $\gamma - 1$ and discard the messages that are older than γ. Note that for a connected graph maximum number of time step required to pass a message between any two agents equals to the diameter of the graph. Thus we choose γ to be an integer constant which is at most diameter of the communication graph G.

Theorem 3. (Group regret of ComEx-MPUCB) *Consider a group of N agents following ComEx-MPUCB. Then for any $\xi \geq 1.1$ expected cumulative group regret satisfies:*

$$\mathbb{E}\left[R(T)\right] \leq \sum_{k=2}^{K} \frac{8(\xi+1)\sigma}{\Delta_k} \bar{\chi}(G_\gamma) \log T + \sum_{k=2}^{K} \Delta_k \left[(N - \bar{\chi}(G_\gamma))(\gamma - 1) + g\left(4N, d_\gamma^{(i)}\right)\right]$$

Proof sketch. We see that regret under ComEx-MPUCB can be given as the summation of regret of ComEx-UCB when communication graph is G_γ and the regret incurred by the delay in passing messages to agents who are not 1-hop neighbors. We prove that the expected regret due to delay is at most $(N - \bar{\chi}(G_\gamma))(\gamma - 1)$. A detailed proof is provided in Chapter 9. □

3.4.3 Centralized message passing UCB

We propose ComEx-LFUCB by combining ComeEx communication protocol with a leader-follower method [42, 48, 19, 94]. ComEx-LFUCB provides better performance compared to its decentralized counter part ComEx-MPUCB. Let V_γ' be the set of

vertices in minimal dominating set of graph G_γ. We consider each agent $i \in V'_\gamma$ to be a leader and all the other agents to be followers. Note that every follower has at least one leader as a neighbor. We consider that each leader uses ComEx-MPUCB and each follower copies the last action observed from its leader. For each follower j a leader i is assigned such that $d(i,j) = \min_{i'} d(i',j)$ where $d(i,j)$ is the distance between agent i and agent j in graph G. Let $\mathcal{N}^i_\gamma$ be the set of follower of leader i. We consider that each leader sends a message containing the id of the arm it pulls and whether it is instantaneously suboptimal, i.e. for $i \in V'_\gamma$ at time step t, $m^{(i)}_t := \left\langle i, t, A^{(i)}_t, \mathbf{1}\left\{A^{(t)}_i \neq \arg\max_{k\in[K]} \widehat{\mu}^{(i)}_k(t-1)\right\}\right\rangle$ to its neighbors and they subsequently forward it to their neighbors. Note that at time step t follower $j \in \mathcal{N}^{(i)}_\gamma$ pulls the arm $A^{(i)}_{t-d(i,j)}$. Each follower pass a message containing information about the reward and arm id if it pulls an arm that is specified as instantaneously suboptimal by its leader. Thus the followers communicate according to ComEx by initiating a message as follows. Follower $j \in \mathcal{N}^{(i)}_\gamma$ initiates a message $m^{(j)}_t := \left\langle j, t, A^{(j)}_t, X^{(j)}_t\right\rangle$ if $A^{(i)}_{t-d(i,j)} \neq \arg\max_{k\in[K]} \widehat{\mu}^{(i)}_k(t-d(i,j)-1)$. Accordingly under full communication followers share their rewards and arm pulls at every time step. Pseudo code for ComEx-LFUCB is provided in Appendix 9.12.

Theorem 4. (Group regret of ComEx-LFUCB) *Consider a group of N agents following ComEx-LFUCB with communication density parameter γ. Then for any $\xi \geq 1.1$ expected cumulative group regret satisfies:*

$$\mathbb{E}\left[R(T)\right] \leq \sum_{k=2}^{K} \frac{8(\xi+1)\sigma}{\Delta_k} \bar{\gamma}(G_\gamma) \log T + \sum_{k=2}^{K} \Delta_k \left[(N - \bar{\gamma}(G_\gamma))(3\gamma - 1) + \bar{\gamma}(G_\gamma) \cdot g(4N, d^{(i)}_\gamma)\right]$$

Proof sketch. We follow a similar approach to the proof of Theorem 3 with a few key modifications followed by the argument below. Note that number of suboptimal arm pulls by each $j \in \mathcal{N}^{(i)}_\gamma$ can be upper bounded using suboptimal arm pulls by i

and message passing delay. Note that message passing delay can be upper bounded by $d(i,j)$. A detailed proof of Theorem 4 is given in Appendix 9.10.5. □

Remark 2. *Similar to ComEx-MPUCB by replacing ComEx with full communication in ComEx-LFUCB algorithm, i.e. allowing followers to share information about arm pulls at every time step, agents obtain an expected cumulative group regret of* $\mathbb{E}[R(T)] = O\left(K\bar{\gamma}(G_\gamma)\log T + KN\right)$ *(Appendix H). Thus from Theorem 4 we see that ComEx obtains the same order of performance as full communication.*

Now we provide theoretical guarantees that expected group communication cost under ComEx-LFUCB is logarithmically bounded in time.

Theorem 5. (Communication cost of ComEx-LFUCB) *Consider a group of* N *agents following ComEx-LFUCB with communication density parameter* γ*. Then for any* $\xi \geq 1.1$ *expected group communication cost satisfies:*

$$\mathbb{E}[L(T)] \leq \left[8(\xi+1)\sigma\left[\frac{N}{\bar{\Delta}^2}+\sum_{k=2}^{K}\frac{\bar{\gamma}(G_\gamma)}{\Delta_k^2}\right]\log T + K\left[(N-3\bar{\gamma}(G_\gamma)(\gamma-1)\right]\right]\sum_{i=1}^{N} d_{\gamma-1}^{(i)^+}$$
$$+K\sum_{i=1}^{N} d_{\gamma-1}^{(i)^+}\cdot\bar{\gamma}(G_\gamma\cdot g\left(7N, d_\gamma^{(i)}\right)$$

Proof sketch. Note that the expected number of times a leader initiates a message can be upper bounded by twice the expected number of its suboptimal arm pulls. Further the number of times each follower $j \in \mathcal{N}_\gamma^{(i)}$ initiates a message can be bounded by the number of instantaneously suboptimal arms pulled by the leader i. Similar to ComEx-MPUCB in ComEx-LFUCB agents send concatenated messages to their neighbors. Thus each message initiated by any agent i is subsequently forwarded by all agents who are within distance of $\gamma-1$ in graph G. A detailed proof can be found in Chapter 9. □

Remark 3. *Algorithm and results provided in this Section can be specialized to centralized cooperative bandits with instantaneous reward sharing by substituting* $\gamma = 1$.

3.4.4 Estimate sharing

We propose ComEx-EstUCB by combining ComEx with estimate sharing [46, 66, 49], which obtains better performance than instantaneous reward sharing. In estimate sharing, for each arm k, agents maintain estimated sum of rewards and estimated number of pulls from the arm. At each time step, agents average their estimates with their neighbors according to a consensus protocol and update the estimates by incorporating the information of arm pull at that time step. We refer readers to [49] for more details. In ComEx-EstUCB agents only average estimates of instantaneously sub optimal arms. Pseudo code for ComEx-EstUCB is given in Appendix 9.13.

3.4.5 Simulation results

In this section we provide numerical simulations illustrating our results and validating our theoretical claims. All the experiments were run on the first author's personal laptop. We show that ComEx obtains same order of performance, i.e., same order of group regret, as full communication for a significantly smaller communication cost than full communication. We also demonstrate that our algorithms outperform state-of-the-art algorithms in several bandit frameworks.

Experimental setup. We provide simulation results for following cooperative bandit frameworks 1) decentralized instantaneous reward sharing, 2) decentralized message passing, 3) decentralized estimate sharing, 4) centralized leader-follower, and 5) Thompson sampling. We compare performance of our algorithms (ComEx-UCB, ComEx-MPUCB, ComEx-EstUCB, ComEx-LFUCB and ComEx-Thompson) with their corresponding full communication algorithms (Full-UCB, Full-MPUCB, Full-EstUCB and Full-LFUCB) and state-of-the art algorithms in each framework. For all simulations presented in this section we consider 10 arms ($K = 10$), 100 agents ($N = 100$) and 500 time steps ($T = 500$). Communication graph between agents

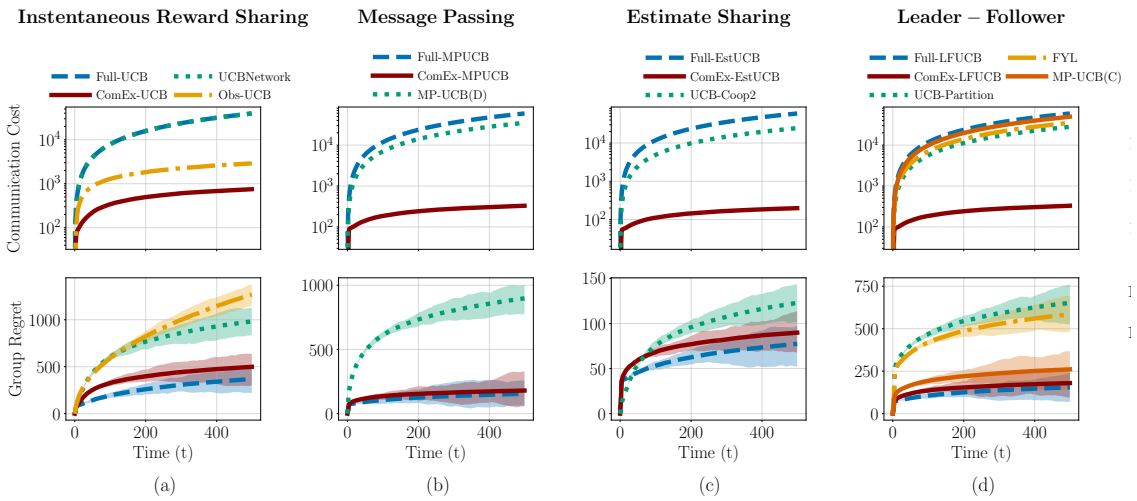


Figure 3.1: A comparison of expected cumulative group regret and communication cost of our algorithms and existing state-of-the-art algorithms in several benchmark cooperative bandit frameworks.

is considered to be a Erdos Renyi random graph with edge probability 0.7. Results are averaged over 100 Monte Carlo simulations. Additional experimental results for different graph structures and parameters (ξ, γ) are provided in Appendix 9.10.11.

Hyper parameters We use tuning parameter $\xi = 1.01$ for UCB based algorithms. For results provided in Figure 3.1(b)-3.1(d) we use communication density parameter $\gamma = 5$. None of the competing algorithms, except UCB-Coop2, MP-UCB(D) and MP-UCB(C) have hyperparameters. We tuned parameters of UCB-Coop2 to get best results of that algorithm and used $\kappa = 0.02, \gamma' = 1.001, \eta = 0.001$ (Equations 9 and 15 in [49]. Here we γ' to avoid confusing with communication parameter γ used in this paper) for final results. Decreasing γ' below 1.001 and η below 0.001 did not offer any significant improvement. MP-UCB(D) and MP-UCB(C) are originally proposed in [19] for heavy-tailed distributions, and we adapt them to sub-Gaussian distributions as directed by the authors. For MP-UCB(D) and MP-UCB(C) we considered the same $C_k^{(i)}(t)$ as in our algorithms. Thus we used the same $\xi = 1.01$ value for a fair comparison.

For results provided in Figures 3.1(a) and 3.1(d), we consider reward distributions to be bounded $[0, 1]$. We consider triangle distributions with mod 1 for the optimal arm and mod 0 for all sub-optimal arms. In simulations provided in Figures 3.1(b) and 3.1(c) we consider Gaussian reward distributions. Expected reward for the optimal arm is $\mu_1 = 11$ and for all sub-optimal arms $k > 1$ is $\mu_k = 10$. We let variance associated with all arms be $\sigma_k^2 = 1, \forall k$. We use the notation Obs-UCB to denote the algorithm presented in [63].

ComEx obtains same order of performance as full communication. Our results in Figure 3.1 illustrate that ComEx obtains the same order of performance, i.e., same order of group regret, as full communication. From Comparing Figures 3.1(a) and 3.1(b) we see that performance difference between full communication and ComEx decrease when communication density γ increase. All results illustrate that our algorithms consistently out preforms state-of-the-art algorithms in all five benchmark cooperative bandit frameworks.

ComEx only incurs a logarithmic communication cost. Our simulation results also illustrate that ComEx only incurs a logarithmic communication cost. In Figure 3.1(a) we observe that Obs-UCB also incurs a logarithmic cost. However ComEx-UCB incurs a smaller cost than Obs-UCB while suffering a smaller group regret. Further, results illustrate that ComEx enabled algorithms incurs a significantly smaller communication cost compared to existing state-of-the-art algorithms.

Additional discussion. State-of-the-art algorithm for leader-follower setting is DPE2 in [94]. DPE2 uses a phased communication protocol, where during the leader selection phase, which lasts at least $2D$ rounds, where D is the diameter of the graph, agents do not pull arms. Thus, this phase accumulates an expected group regret of at least $2DN\mu_1$. In our experimental setup, this alone exceeds the regret accumulated by our algorithms during the entire time horizon. So a meaningful comparison cannot be

provided without modifying DPE2 to allow pulling arms during the leader selection phase.

Chapter 4

Probabilistic communication

In last chapter we discussed the multi-agent multi-armed bandit problem with communication costs. In this chapter we discuss probabilistic communication constraints. The chapter summarizes and extends the work presented in papers [58, 65], also included in Chapters 10 and 11.

4.1 Motivation

The multi-agent multi-armed bandit problem has garnered significant interest due to its applicability in various domains, such as recommendation systems, resource allocation, and online advertising. However, real-world scenarios often involve communication constraints, including the possibility of communication link failures. The motivation for studying multi-agent multi-armed bandit problems with probabilistic communication stems from the need to develop robust and efficient solutions that can address these communication challenges.

In practical applications, communication links can fail randomly due to various factors, such as network congestion, signal interference, or hardware issues. These failures can severely impact the performance of multi-agent systems, leading to suboptimal decision-making and reduced cumulative rewards. By investigating the effects

of probabilistic communication and developing strategies that can cope with communication link failures, we can enhance the performance of multi-agent multi-armed bandits in such settings.

An alternative interpretation of communication probability is to consider it as an agent-dependent information-sharing frequency. In this interpretation, agents share information with a certain probability that is specific to each agent. This probability can be affected by factors like agent priorities, resource limitations, or strategic considerations. Studying agent-dependent information-sharing frequencies can provide valuable insights into the dynamics of multi-agent systems and guide the design of more effective communication protocols.

Furthermore, understanding the impact of probabilistic communication on multi-agent multi-armed bandit problems can enable the development of adaptive communication strategies. These strategies can dynamically adjust the frequency and content of information sharing based on the current state of the environment and agent performance. By incorporating adaptive communication strategies, agents can optimize the use of available communication resources, effectively balancing the trade-offs between exploration, exploitation, and communication overhead.

4.2 Role of degree heterogeneity in the communication network

In the multi-agent multi-armed bandit problem with a general communication graph, the performance of individual agents is significantly influenced by their position within the network. The connectivity of an agent, as determined by its degree (the number of neighbors), plays a crucial role in the information flow and the subsequent decision-making process. In this section, we will discuss the implications of an agent's degree

and connectivity to its neighbors on its exploration-exploitation balance and overall performance.

- Impact of low-degree agents: Agents with a low degree, or a small number of neighbors, generally receive less information from their neighbors. Consequently, these agents have less knowledge about the environment and the performance of other agents. This limited information encourages low-degree agents to engage in more exploration as opposed to exploitation, as they cannot rely as heavily on information from others to make informed decisions.

- Impact of high-degree agents: In contrast, agents with a high degree, or a large number of neighbors, typically receive more information from their neighbors. This abundance of information allows high-degree agents to make better-informed decisions, and as a result, they tend to engage in more exploitation than exploration. However, the performance of high-degree agents is also influenced by the connectivity of their neighbors.

- High-degree agents connected to low-degree neighbors: When high-degree agents are connected to neighbors with low degrees, they often perform better than those connected to high-degree neighbors. This is because low-degree neighbors are more likely to engage in exploration, providing the high-degree agents with valuable information about unexplored options. As a result, high-degree agents can exploit this information to make better decisions and achieve higher rewards.

- High-degree agents connected to high-degree neighbors: On the other hand, high-degree agents connected to high-degree neighbors tend to receive redundant information, as their neighbors are also likely to engage in exploitation. This redundancy limits the exploration of new options, and in turn, the perfor-

mance of these high-degree agents may suffer due to the lack of diversity in the information they receive.

The performance of individual agents in multi-agent multi-armed bandit problems with general communication graphs is highly dependent on their connectivity within the network. Low-degree agents tend to engage in more exploration, while high-degree agents often focus on exploitation. The performance of high-degree agents is further significantly influenced by the connectivity of their neighbors, with those connected to low-degree neighbors generally outperforming those connected to high-degree neighbors. Understanding these dynamics is essential for designing efficient communication protocols and strategies that can enhance the overall performance of multi-agent systems in various applications.

4.3 Role of agent specific communication probabilities

In the multi-agent multi-armed bandit problem, the performance of individual agents can be significantly influenced by the probability with which they receive information from their neighbors. When each agent i receives information from all its neighbors with a probability p_i, the dynamics of the decision-making process are further complicated. In this section, we discuss the implications of agent-specific information reception probability on the exploration-exploitation balance and overall performance of individual agents.

- Impact of low information reception probability: Agents with a low information reception probability (low p_i) are less likely to receive valuable information from their neighbors. This limited information flow might compel these agents to rely more on their own exploration, as the uncertainty about the environment

and the performance of other agents remains high. Consequently, agents with low information reception probability might engage more in exploration than exploitation, which could potentially slow down their learning process and affect their overall performance.

- Impact of high information reception probability: Agents with a high information reception probability (high p_i) have a greater chance of receiving information from their neighbors. This increased information flow enables these agents to make better-informed decisions based on the experiences of their neighbors. As a result, agents with high information reception probability are more likely to engage in exploitation, as they can benefit from the shared knowledge about the environment.

- Balancing exploration and exploitation: The performance of individual agents in the multi-agent multi-armed bandit problem is contingent upon striking the right balance between exploration and exploitation. Agents with low information reception probability may need to adjust their exploration-exploitation strategies to compensate for the lack of information received from their neighbors. Conversely, agents with high information reception probability should be cautious not to over-exploit the available information, as this might lead to suboptimal decisions and reduced cumulative rewards.

- Adapting to the dynamics of information flow: Understanding and adapting to the dynamics of information flow in a multi-agent system with agent-specific information reception probabilities is crucial for optimizing the overall performance. Agents may need to employ adaptive strategies that dynamically adjust their exploration-exploitation balance based on the current state of the environment and the information received from their neighbors. Such adaptive strate-

gies can help agents cope with the uncertainties introduced by the agent-specific probabilities and enhance their decision-making capabilities.

The agent-specific information reception probability significantly influences the performance of individual agents in the multi-agent multi-armed bandit problem. The exploration-exploitation balance and overall performance of agents depend on their ability to adapt to the dynamics of information flow resulting from these probabilities. Developing adaptive strategies that consider agent-specific information reception probabilities can lead to more efficient multi-agent systems that can better cope with the uncertainties and complexities of real-world decision-making problems.

4.4 Group performance under probabilistic communication

The fundamental advantage of cooperative estimation is the ability to leverage observations about suboptimal arms from neighboring agents to reduce exploration. However, when agents are communicating over an arbitrary graph, the amount of information an agent receives varies according to its connectivity in G. For example, agents with a large number of neighbors receive more information, leading them to begin exploitation earlier than agents with fewer neighbors. This means that well-connected agents exhibit better performance early on, but because they quickly do only exploiting, agents that are poorly connected typically only observe exploitative arm pulls, which requires them to explore for longer in order to obtain similarly good estimates for suboptimal arms, increasing their regret. The disparity between performance in well-connected versus poorly connected agents is exacerbated in the presence of random *link failures*, where any message sent by an agent can fail to reach its recipient with a failure probability $1 - p$ (drawn i.i.d. for each message).

Indeed, it is natural to expect the group regret to decrease with decreasing link failure probability, i.e., increasing communication probability p. However, what we observe experimentally (Section 10.7) is that this holds only for graphs G that are *regular* (i.e., each agent has the same degree), or close to regular. When G is irregular, as we increase p from 0 to 1, the group performance oscillates. While, in some cases, the improved performance in the well-connected agents can outweigh the degradation observed in the weakly-connected agents (leading to lower *group* regret), it is prudent to consider an approach that mitigates this disparity by regulating information flow in the network.

Information Regulation in Cooperative Bandits. Our approach to regulate information is straightforward: we direct each agent i to discard any incoming message with an agent-specific probability $1 - p_i$, while always utilizing its own observations. For specific values of p_i, we can obtain various weighted combinations of internal versus group observations. Our first algorithm `RCL-LF` (Link Failures) is built on this regulation strategy, coupled with `UCB1` exploration using all selected observations for each arm. Essentially, each agent runs `UCB1` using the cumulative set of observations it has received from its network. After pulling an arm, it broadcasts its pulled arm and reward through the network, but incorporates each incoming message *only* with a probability p_i. Pseudo code for the algorithm is given in the appendix. We first present a regret bound for `RCL-LF` when run with the *instantaneous reward-sharing* protocol.

Theorem 6 (`RCL-LF` Regret with instantaneous reward-sharing)**.** `RCL-LF` *running with the instantaneous reward-sharing protocol (Figure 10.1, $\gamma = 1$) obtains cumulative group regret of*

$$\mathit{Reg}_G(T) \leq g(\xi, \sigma) \left(\sum_{i=1}^{N} (1 - p_i \cdot p) + \sum_{\mathcal{C} \in \mathcal{C}} (\max_{i \leq \mathcal{C}} p_i) \cdot p \right) \left(\sum_{k>1} \frac{\log T}{\Delta_k} \right) + f(5N, G)$$

where $\mathcal{C}$ is a non-overlapping clique covering of G.

Proof sketch. We follow an approach similar to the analysis of `UCB1` by [7] with several key modifications. First, we partition the communication graph G into a set of non-overlapping cliques and then analyze the regret of each clique. The group regret can be obtained by taking the summation of the regret over each clique. Two major technical challenges in proving the regret bound for `RCL-LF` are (a) deriving a tail probability bound for probabilistic communication, and (b) bounding the regret accumulated by agents by losing information due to communication failures and message discarding. We overcome the first challenge by noticing that communication is independent of the decision making process thus $\mathbb{E}\left(\exp\left(\lambda \sum_{\tau=1}^{t} X_\tau^i \mathbf{1}\{A_\tau^i = k\} - \mu_k N_k^i(t) - \frac{\lambda^2\sigma_k^2}{2} N_k^i(t)\right)\right) \leq 1$ holds under probabilistic communication. We obtain the tail bound by combining this result with the Markov inequality and optimizing over λ using a peeling type argument. We address the second challenge by proving that the number of times agents do not share information about any suboptimal arm k can be bounded by a term that increases logarithmically with time and scales with number of agents, G, and communication probabilities, as $\sum_{i=1}^{N}(1 - p_i \cdot p) + \sum_{\mathcal{C}\in\mathcal{C}}(\max_{i\leq\mathcal{C}} p_i) \cdot p$. □

Remark 4 (Regret bound optimality)**.** *Under perfect communication ($p = 1$) and no message discarding, i.e., $p_i = p = 1, \forall i \in [N]$ the dominant term in our regret bound scales with $\bar{\chi}(G)$, obtaining identical performance to deterministic communication over G [19]. Alternatively, when $p_i = p = 0$, there is no communication, and hence, the regret bound is $\mathcal{O}(N \log T)$. Theorem 17 quantifies the benefit of communication in reducing the group regret under probabilistic link failure and when agents incorporate observations with an agent-specific probability. Note that $\sum_{i=1}^{N}(1 - p_i \cdot p) + \sum_{\mathcal{C}\in\mathcal{C}}(\max_{i\leq\mathcal{C}} p_i) \cdot p = N - p \cdot \left(\sum_{i=1}^{N} p_i - \sum_{\mathcal{C}\in\mathcal{C}}(\max_{i\leq\mathcal{C}} p_i)\right)$. Since the clique covering is non-overlapping, the results show that agents obtain improved group performance for any communication probability $p > 0$ for any nontrivial graph*

as compared to the case with no communication in which each agent learns on its own.

Remark 5 (Controlling information disparity)**.** *In order to regulate the information disparity across the network we set* $p_i = \frac{d_{\min}(G)}{d_i(G)}$*. Thus, the agent(s) with minimum degree* $d_{\min}$ *incorporate each message they receive with probability* 1 *and we have that the* expected *number of messages for each agent is the same, i.e.,* $T \cdot d_{\min}(G)$*. Therefore, every agent receives the same amount of information (in expectation), providing a large performance improvement for irregular graphs (see Section 10.7).*

Message-Passing. Under this communication protocol each agent i communicates with neighbors at distance at most γ, where each hop adds a 1-step delay. Our algorithm `RCL-CF` obtains a similar regret bound in this setting as well, when all agents use the same `UCB1` exploration strategy.

Theorem 7 (`RCL-LF` Regret with message-passing)**.** *Let* $\mathcal{C}$ *be a minimal clique covering of* G_γ*. For any* $\mathcal{C} \in \mathcal{C}$ *and* $i, j \in \mathcal{C}$ *let* $\gamma_i = \max_{j \in \mathcal{C}} d(i,j)$ *be the maximum distance (in graph* G*) between agents* i *and* j*.* `RCL-LF` *running with the message-passing protocol with delay parameter* γ *obtains cumulative group regret of*

$$\texttt{Reg}(T) \leq g(\xi,\sigma)\left(\sum_{i=1}^{N}(1-p_i\cdot p^{\gamma_i})+\bar{\chi}(G_\gamma)\cdot(\max_{i\leq N} p_i\cdot p^{\gamma_i})\right)\left(\sum_{k>1}\frac{\log T}{\Delta_k}\right)+f((\gamma+4)N,G_\gamma).$$

Proof sketch. We partition the graph G_γ into non-overlapping cliques, analyze the regret of each clique and take the summation of regrets over cliques to obtain group regret. In addition to the challenges encountered in Theorem 17 here we are required to account for having different probabilities of failures for messages due to having multiple paths of different length between agents and to account for the delay incurred by each hop when passing messages. We overcome the first challenge by noting that agent i receives each message with at least probability p^{γ_i}. We overcome the second

challenge by identifying that regret incurred by delays can be upper bounded using $\left(\sum_{i=1}^{N} \gamma_i - N\right) \sum_{k>1} \Delta_k$. □

Remark 6. *Finding an optimal observation probability* $\{p_i\}_{1=1}^{N}$ *for* `RCL-LF` *with message-passing is difficult due to the delays added by each hop when forwarding messages. If messages are forwarded without a delay, optimal performance can be obtained by using* $p_i = \frac{d_{\min}(G_\gamma)}{d_i(G_\gamma)}$. *For dense* G_γ, *the above choice of observation probability provides near-optimal performance. When* $\gamma = d_\star(G)$ *we have that* G_γ *is a complete graph,* $p_i = \frac{d_{\min}(G_\gamma)}{d_i(G_\gamma)} = 1$, *and agents do not discard any message. However, when* $\gamma < d_\star(G)$, *the graph* G_γ *is not complete. Therefore agents receive different amounts of information which are approximately proportional to the degree distribution of* G_γ. *As explained earlier this information disparity leads to a performance disparity among agents. As a result group performance decreases. In this case we design the algorithm such that each agent* i *discards messages with* $1 - p_i$ *where* $p_i = \frac{d_{\min}(G_\gamma)}{d_i(G_\gamma)}$. *This regulates the information flow mitigating the bias introduced by information disparity. As a result the group obtains near-optimal performance.*

4.5 Role of sampling rule heterogeneity

In multi-star networks, which are characterized by their irregular and centralized structures, center agents hold a prominent position due to their higher number of connections with peripheral agents. As a result, these center agents receive more information than their peripheral counterparts, creating an imbalance in the exploitation potential across the group. This disparity causes group performance to decline as the number of peripheral agents increases. Our research aims to enhance group performance by capitalizing on the heterogeneity in exploitation potential among the agents. To achieve this, we propose heterogeneous explore-exploit strategies that

require center agents to explore more, subsequently amplifying the exploitation potential of peripheral agents.

The multi-star network is a suitable model for recommender systems, where numerous small servers are assigned to different regions. These servers make sequential recommendations based on user feedback and communicate exclusively with a large central server. By encouraging the central server to provide more exploratory recommendations, the system can gather a broader range of information about user preferences, ultimately improving its performance. Additionally, incorporating probabilistic communication helps to account for potential random communication failures between servers, further enhancing the system's overall reliability and effectiveness.

We focus on *multi-star* graphs defined as follows. Let there be m center agents and $N - m$ peripheral agents. Without loss of generality let each agent i, $i \leq m$, be a *center agent*.

Definition 3. (Heterogeneous Exploration) *Exploration term of agent i at time $t \in [T]$ is*

$$C_k^{(i)}(t) = \sigma_i \sqrt{\frac{2(1+\alpha_i)(\xi+1)\log t}{N_k^{(i)}(t)}} \tag{4.1}$$

where $\xi > 1.1$ and

$$\alpha_i = \begin{cases} \frac{p^{1-p}(d_i - d_i^{avg})}{d_i} & , \quad k \leq m \\ 0 & , \quad k > m. \end{cases} \tag{4.2}$$

4.5.1 Performance of agent in multi-star networks

In this section we provide numerical simulations to illustrate results and validate theoretical bounds. For all simulations, we consider 10 options ($N = 10$) with Gaussian reward distributions. Expected reward for the optimal option is $\mu_1 = 11$ and for all sub-optimal options $k \geq 2$ is $\mu_k = 10$. We let variance associated with all options

k be $\sigma^2 = 1$. Because the expected reward gaps $\Delta_k = 1$, $k \neq 1$, are equal to the variances $\sigma^2 = 1$, it is a challenging problem to distinguish the optimal option from the sub-optimal options. For all simulations, we consider 1000 time steps ($T = 1000$) and use 1000 Monte Carlo simulations with $\xi = 1.01$.

We show simulation results for performance of a group of $K = 36$ agents that communicate over two different symmetric multi-star graphs and use the heterogeneous sampling rules of Definition 10. We compare to the case when agents use the corresponding homogeneous sampling rules of Definition 11. The first multi-star graph has $m = 2$ center agents and $K - m = 34$ peripheral agents, with each center agent communicating with 17 peripheral agents and the other center agent. The second multi-star graph has $m = 3$ center agents and $K - m = 33$ peripheral agents, with each center agent communicating with 11 peripheral agents and the other center agents. In each case, center agents are interchangeable and peripheral agents are interchangeable, so the average performance of a center (peripheral) agent is the same as the individual performance of a center (peripheral) agent.

Figure 4.1 shows how average expected cumulative group regret varies with broadcasting probability p for agents using the heterogeneous rules (dotted) and homogeneous rules (solid). Regret is inversely related to performance: lower group regret implies higher group performance. Results are plotted on the left for the graph with 2 center agents and on the right for the graph with 3 center agents. When $p = 0$ there is no communication at all. So when p becomes even just a little positive and agents learn about options from their neighbors, regret falls, i.e., group performance rises.

In the case of the homogeneous rules, as p increases through intermediate values, center agents do less and less exploring and the usefulness of the information received by peripheral agents decreases. This leads to increased regret for peripheral agents, and the group overall, and thus degraded group performance. When p approaches 1,

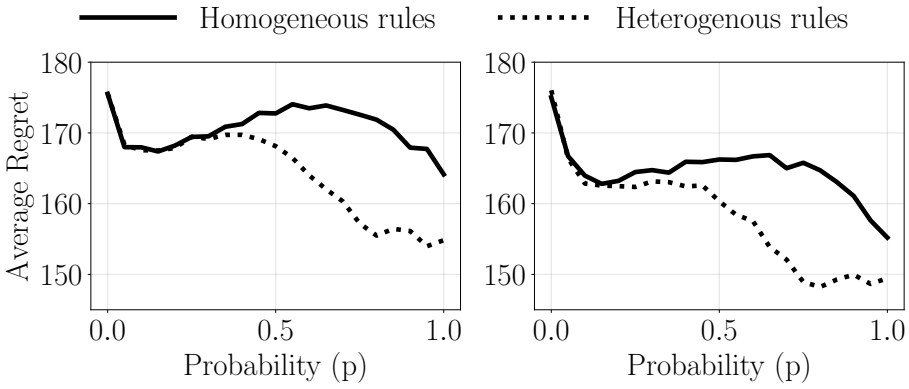


Figure 4.1: Average expected cumulative group regret for $K = 36$ agents at time $t = 1000$ as a function of broadcasting probability p with communication over a symmetric multi-star graph. Left: 2 center and 34 peripheral agents. Right: 3 center and 33 peripheral agents. Dotted lines and solid line shows average regret when agents use heterogeneous and homogeneous sampling rules, respectively.

center agents receive sufficient information from their peripheral neighbors such that their improved performance outweighs the degraded performance of peripheral agents. This leads to a final decrease in group regret and increase in group performance.

The improvement in performance provided by the heterogeneous rules relative to the homogeneous rules, as predicted by Theorem 26 and Remark 21, can be clearly seen in Figure 4.1 by observing how much lower the dotted regret curve is than the solid regret curve. The growth in regret in the homogeneous case, as p increases through intermediate values, is reduced in the heterogeneous case. This is because, by design, center agents are biased toward more exploring, which improves the information that peripheral agents receive. The group performance increase that comes, as p increases further, occurs in the heterogeneous case well before p approaches 1.

The influence of irregularity of the graph can be observed in Figure 4.1 by comparing the left plot (2 center agents and more irregular) to the right plot (3 center agents and less irregular). The results suggest that performance is higher with more center agents, i.e., with greater regularity in the graph.

Figure 4.2 shows expected cumulative regret as a function of time t for center (blue), peripheral (pink), and average (black) agents, when $p = 0.8$ and agents use

the heterogeneous rules (dotted) and homogeneous rules (solid). Results are plotted on the left for the graph with 2 center agents and on the right for the graph with 3 center agents. It can be observed that, as predicted for the heterogeneous rules, the peripheral agent performance increases and the center agent performance decreases, such that group performance (as represented by the average agent) improves. Further, a comparison of left and right plots suggests that group performance improves with more center agents (more regularity).

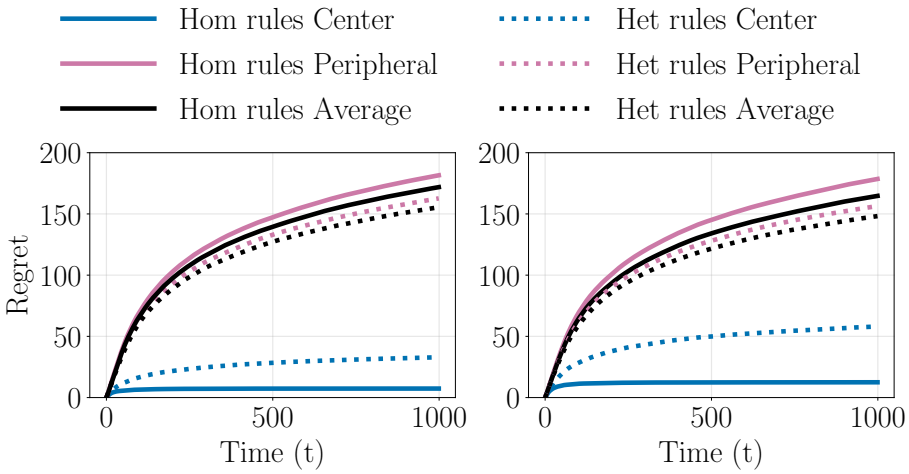


Figure 4.2: Expected cumulative regret of center agent, peripheral agent, and average agent for $K = 36$ agents as a function of time t for $p = 0.8$ and the same two symmetric multi-star graphs as in Figure 4.1: 2 center agents (left) and 3 center agents (right) where agents use heterogeneous (dotted) and homogeneous (solid) sampling rules.

In our study, we developed and examined novel heterogeneous rules for a group of agents sharing information across a network to optimize their collective reward while sampling an uncertain environment. We focused on communication networks characterized by symmetric multi-star graphs, as these are representative of realistic scenarios. By employing the multi-armed bandit problem as the explore-exploit framework, we demonstrated that sampling rules for center agents, which prioritize exploration over exploitation, can enhance the usefulness of information broadcasted to their neighbors, ultimately boosting the group's total reward.

This analysis and design contribute to a deeper understanding of the significance of heterogeneity in collective decision-making processes. The evidence that heterogeneity can be harnessed to improve the performance of a cooperative multi-agent system indicates that further exploration in this area is both valuable and necessary.

Chapter 5

Zero-shot generalization in multi-agent reinforcement learning

In this chapter we present our work on improving generalization in multi-agent reinforcement learning. The chapter summarizes the work presented in the paper [] (the paper will appear on arXiv soon), also included in Chapters 12.

5.1 Role of behavioural heterogeneity among agents

In psychology research, Social Value Orientation (SVO) is a cognitive construct reflecting a person's preference for resource allocation between themselves and others [28, 54, 71]. While some individuals may solipsistically focus on maximizing their personal success, others demonstrate different motivations, including maximizing the difference between their own and others' outcomes (a competitive orientation), maximizing collective welfare (a prosocial orientation), or maximizing other peoples' benefit (an altruistic orientation). In artificial intelligence research, various algorithms draw inspiration from these insights in their design or implementation [68, 80]. In

reinforcement learning, SVO is an intrinsic motivation that transforms an agent's reward based on its particular target distribution between its reward and the reward of others. Recently, there's been research investigating the role of SVO in situations where a group of agents or players interact in ways that involve trade-offs between their self-interest and the collective interest of the group. This research has generated valuable insights into the impact of SVO on the emergence of diverse behaviors and cooperation [68, 69], and partner choice [67]. SVO research has focused primarily on social dilemmas with underlying incentive structures resembling the *Prisoner's dilemma* [77], wherein each player has an incentive to defect, even though both would be better off if they both cooperated.

Sequential social dilemmas are a class of social dilemmas in which the decision-making process of the interacting agents is temporally and spatially extended [51]. Performing well in a sequential social dilemma tends to require the consideration of long-term consequences, interdependence, and cooperation among group members. Research on sequential social dilemmas has been widely studied in the context of emergence and maintenance of cooperation [52, 75], inequity aversion [37], partner choice [21, 67], and generalization [69, 1] wherein agents interact with novel co-players in test scenarios.

While environments provide an *extrinsic reward* that can be used to learn a policy, it is often useful to provide agents with an *intrinsic reward* to shape their behavior towards a policy with desirable properties. Intrinsic reward has be used to capture the social preferences of players, and are typically functions of the vector of all players' reward. In most research in sequential social dilemmas, all players either have no *intrinsic reward*, or they all have the same function (i.e. they have homogeneous social preferences) [52, 93]. However, it has been observed that having a population of agents who differ in their intrinsic reward function (i.e. they have heterogeneous social preferences) can lead to higher levels of cooperation [37]. In [68, 69, 67], the

authors showed that heterogeneity can produce behavioral diversity in fully cooperative games, and in games with incentive structure similar to the Prisoner's dilemma. Other incentive structures have not yet been explored. In addition, the precise interplay between diversity in social preferences and in agent policies, particularly on the mechanisms the enable generalization to novel social partners, remains underexplored.

Diversity in policies has been demonstrated to improve various aspects of agent performance, such as exploration [100], adaptation to environmental changes [18], positive group outcomes [68, 85], generalization to novel co-players [56], and collaboration with humans [82]. One way to quantify diversity is through state-action variation, which measures the distribution of state-action pairs that an agent explores during training. State-action diversity can be assessed by measuring differences in the state visitation frequency [100], action selection frequency in a given state [69], or differences between state-action trajectories starting from a specific state [56]. To complement these methods, diversity can also be quantified by examining the reward an agent obtains when interacting with different co-players (often called *strategic diversity*) [9, 26], which can provide a complementary measure of diversity in behavior. However, defining a universal diversity metric from trajectories can be challenging, and so it is possible instead to use environment-specific measures of diversity.

In this chapter, we assess heterogeneous SVO in a range of incentive structures in sequential social dilemmas. We include temporally and spatially extended environments with an underlying structure that is like: *Prisoner's dilemma*; *Chicken*, where both players have an incentive to choose a risky behavior, but where the worst outcome is if both choose the high risk; and *Stag hunt* wherein players have a safe choice, and an incentive to coordinate on a high-reward strategy that carries a risk of costly miscoordination. Chicken and Stag hunt are equilibrium selection social dilemmas.

5.2 Zero-shot generalization

Zero-shot generalization [36, 35, 82, 50, 69] seeks to develop general agents that are capable of successfully interacting with novel agents during test time (i.e., agents not seen during training). In such situations, the policies of the novel agents encountered at test time can be out-of-distribution for the agents, leading to poor coordination in purely cooperative settings [36, 56], and getting exploited in competitive settings [74]. In mixed-motive games, failure to generalize to novel agents can lead to dead-weight loss by missing an opportunity to cooperate [50]. Learning a best response to partners/opponents with meaningfully diverse policies has emerged as a promising approach to zero-shot generalization [82]. The intuition behind this approach is that training with a set of diverse policies decreases the likelihood of encountering out-of-distribution policies at test time. Despite this promise these best response techniques have not yet been applied in a wide range of incentive structures.

We extend the observation that heterogeneous SVO leads to diverse policies to the incentive structures of Chicken and Stag hunt, and to many players (more than 2). We also show that this diversity, when leveraged via best response, results in better zero-shot generalization in equilibrium selection sequential social dilemmas. We found that best-response agents adapted to partners/opponents with diverse behaviors by learning a conditional policy during training. However, when the test scenario contained conditional policies and the sequential social dilemma was not an equilibrium-selection problem, training best response collapsed to one unconditional policy, leading to poor zero-shot generalization.

5.2.1 Environments

We provide a brief description of the environments considered in this chapter. For all experiments in this paper, we use environments from Melting Pot 2.0 without

modifications [1].

Intertemporal "in the matrix" repeated games: The "in the matrix" repeated games are a family of sequential social dilemmas in Melting Pot 2.0 where two-players interact. In the beginning of each episode the environment is initialized according to a given resource layout, and a set of fixed points where players can spawn. The map consists of two types of resources which can be distinguished by their colour; red corresponds to defection and blue corresponds to cooperation (see Figure 12.3). Players can pick up resources by walking over them, and these resources go into a player inventory. Players spawn with one of each resource type in their inventory. After spawning, each player can move around the map, collect resources, and interact with the co-player by firing an interaction beam. When players interact (by one player hitting the other using their interaction beam), each player gets a reward equal to the expected payoff calculated from the inventory counts and environment-specific payoff matrix. The agent who zaps the other agent is considered as the row player. The inventory count of each player defines a mixed strategy where the probability of playing each pure strategy is equivalent to the percentage of the corresponding resource. Let N_r^i and N_g^i denote the inventory count, number of red resources and green resources respectively, for agent $i \in 1, 2$. For each agent i their mixed strategy is given as

$$p = \Big[\frac{N_r^i}{N_r^i + N_g^i}, \frac{N_g^i}{N_r^i + N_g^i}\Big]$$

Let A be the payoff matrix for both row player and column player. Let r_{row} and r_{col} be the reward of row player and column player respectively. Let p_{row} and p_{col} be the mixed strategy probability vector of row player and column player respectively. Then the rewards can be defined as

$$r_{row} = p_{row}^T A p_{col}, \quad r_{col} = p_{col}^T A^T p_{row}$$

These reward calculations correspond to those used in game theory for matrix games and iterated social dilemmas [97].

Stag hunt

4	**0**
2	**2**

Chicken

3	**2**
5	**0**

Prisoner's dilemma

3	**0**
5	**1**

Figure 5.1: Payoff matrices for Stag hunt, Chicken and Prisoner's dilemma. The values shown correspond to the payoff of the row player. The payoff of the column player is the transpose of the shown matrix (i.e. the games are symmetric games). Cooperation corresponds to the first row and column. Defection corresponds to the send row and column.

The payoff matrices A used are given in Figure 12.2. After interacting, players receive their reward from interaction, have their inventory counts reset (to one of each), and get re-spawned after a delay. Players can have multiple interactions within an episode. Once a resource is picked up, it begins to regenerate after a delay of 10 steps, with a 20% chance of regenerating on each subsequent step. As is standard in Melting Pot 2.0, in each game, there is a 10% chance that the episode will end after every 100 steps, with a minimum of 1000 steps per episode.

Externality mushrooms: Externality mushrooms is sequential social dilemma where players immediately get affected from pro(anti)social behaviors of their co-players. This is a 5-player game where players eat mushrooms in order to receive rewards. Four types of mushrooms grow (in different amounts) on the map: red, green, blue, and orange. Eating a red (fize: full internality zero externality) mushroom gives a reward of 1 to the player who consumed the mushroom. Eating a green (hihe: half internality half externality) mushroom gives a total reward of 2/5 *to all players*. Eating a blue (zife: zero internality full externality) mushroom gives a total

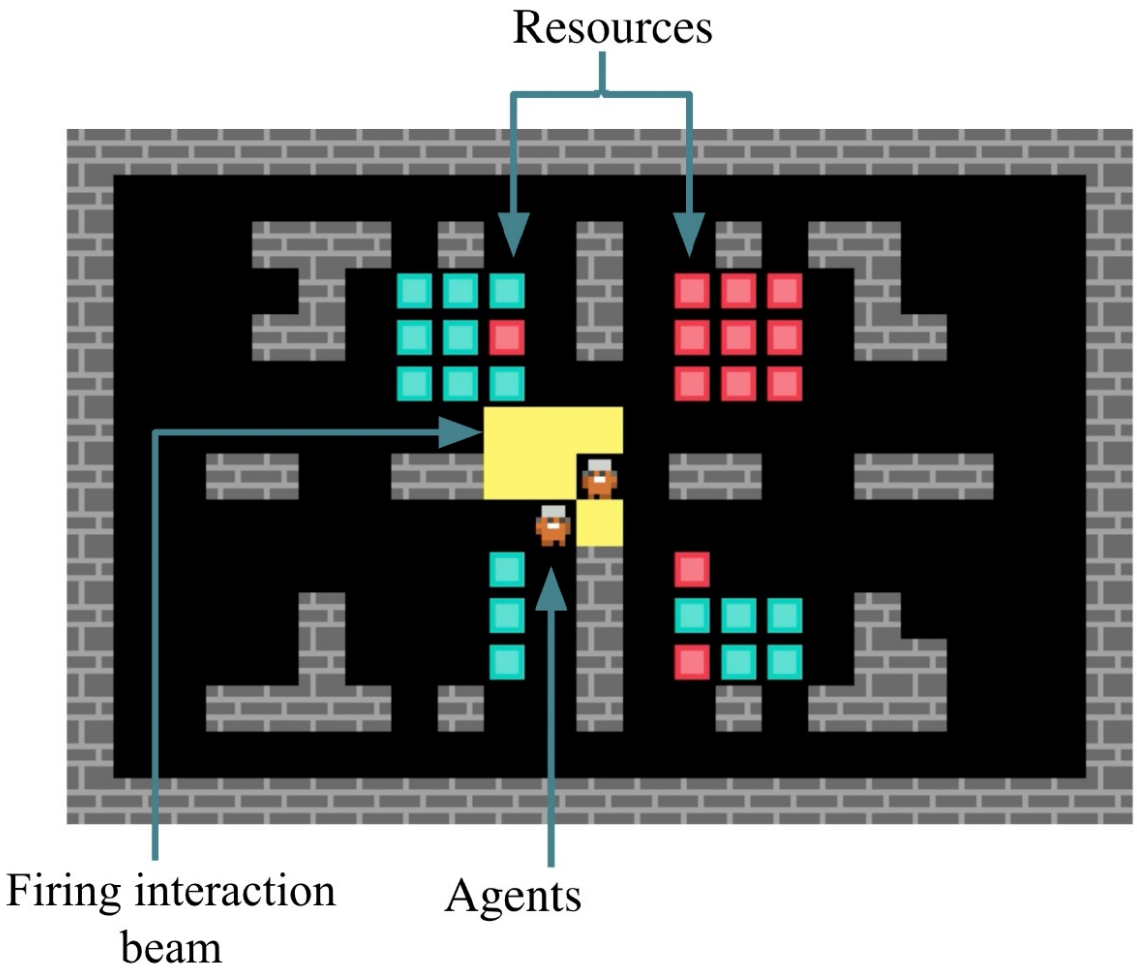


Figure 5.2: "in the matrix" repeated games. This is a 2-player game where agents can gather 2 types of resources (green corresponding to cooperation, red corresponding to defection). When agents interact (using an interaction beam) they get rewards according to their inventory counts and a game specific payoff matrix. The payoff matrix can be Stag hunt, Chicken or Prisoner's dilemma type payoff matrix

reward of 3/4 divided equally among all players *excluding the player who consumed it.* Eating an orange (nize: negative internality zero externality) mushroom causes red mushrooms to be destroyed, each with probability 0.25, and gives a reward of -0.1 to the player who consumed it. After eating a mushroom, the player who consumed it freezes for the mushroom's digestion time: 0 (red), 10 (green), 15 (blue), and 15 steps (orange). After spawning, a mushroom is removed from the map after its perishing time: 200 (red), 100 (green), and 75 steps (blue). Orange mushrooms never perish. Mushrooms respawn from spores depending on consumption of other mush-

rooms. Eating a red, green, or blue mushroom releases 3 spores for red mushrooms, each spore will spawn a mushroom with probability 0.25. Eating a green or blue mushrooms also releases 3 spores for green mushrooms which spawn with probability 0.4. Eating a blue mushroom also releases a blue spore which spawn with probability 0.6. Eating an orange mushroom releases a spore for a new orange mushroom which spawns with probability 1. Similar to "in the matrix" repeated games, in Externality mushrooms each episode runs for at least 1000 steps. Following that the episode terminates with probability 0.2 at every 100 steps.

Externality mushrooms has an incentive structure similar to Chicken, where reward is maximized selfishly by consuming red mushrooms while the others are consuming blue or green mushrooms. But if everyone else is eating red mushrooms, the selfish strategy is to eat green mushrooms, as otherwise all mushrooms would be eventually depleted.

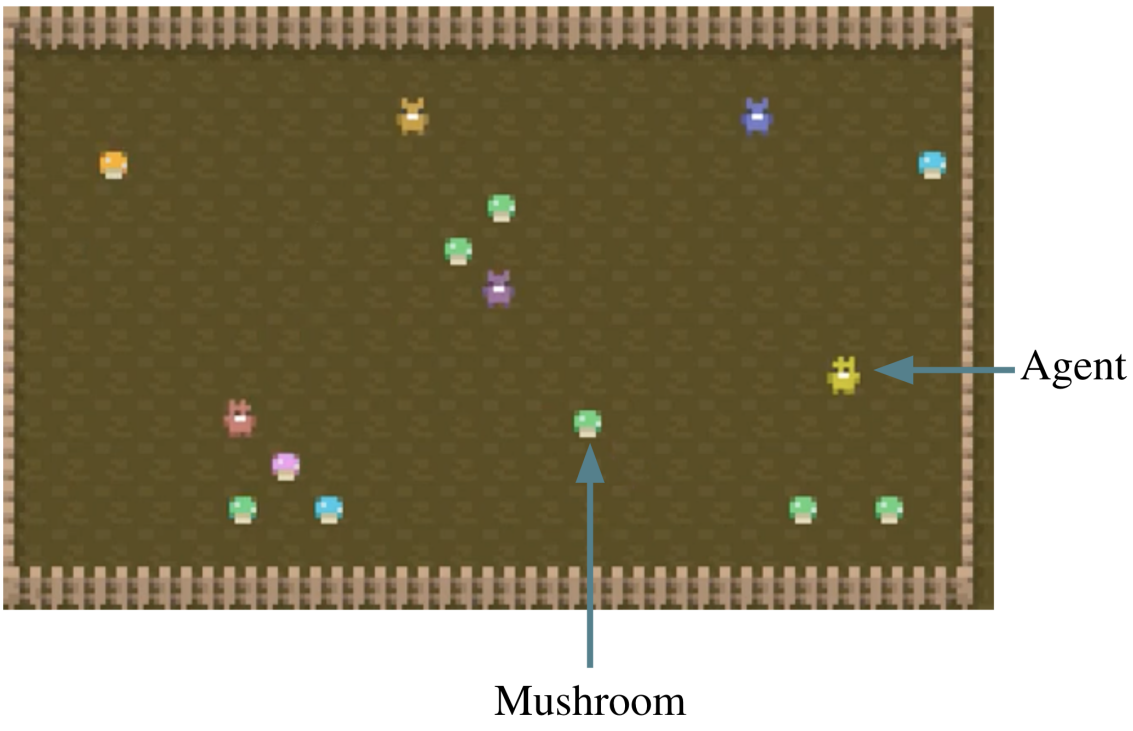


Figure 5.3: Externality Mushrooms. This is a 5-player sequential social dilemma game with immediate feedback. Agents instantaneously share rewards with others depending on the mushroom they are picking.

5.3 Generating diverse policies in sequential social dilemmas

In the beginning of the training process we define distinct SVO angles for each agent. Each environment has a fixed number of players. We train the agents in a distributed asynchronous manner by initializing 'arenas' to train a population of agents. Arenas run in parallel and each arena is a copy of the environment with the number of players specified for that environment. This is a multi-agent version of A3C [70] that is commonly used for multi-agent reinforcement learning [1]. The Melting Pot evaluation protocol requires sampling of policies with replacement. Training in pure self-play introduces skewed reward incentives by playing with copies of oneself. To alleviate this issue, we set players in each arena to play the game for one episode either in self-play or in population-play (with equal probability). During population-play we sample agents without replacement. We train each agent for 10^9 learner frames.

5.4 Training a best-response agent and zero-shot generalization performance evaluation

We train a selfish naive learner without intrinsic reward, to best respond against the policies generated using heterogeneous SVO. In order to avoid confusion we use the term *best-response agent* for the training agent, and *SVO bots* for the pre-trained diverse agents trained with heterogeneous SVO values. In each episode the best-response agent plays with a set of SVO bots sampled without replacement. We train the best-response agent for 10^9 learner frames.

Melting Pot 2.0 [1] provides a protocol for evaluating generalization to novel social partners, which are packaged with the suite as a held-out set of co-players in a suite

of test scenarios. We measure the performance of the best-response agent using the Melting Pot test protocol.

We use the Melting Pot test scenarios for evaluation in Stag hunt, Chicken, Prisoners' dilemma "in the matrix " repeated games and Externality mushrooms. Test scenario details are provided below.

Test scenarios for "in the matrix" repeated.

Focal player (our best response agent) encounters:

S0: *(cooperator + defector)* either a cooperator or a defector with 0.5 probability

S1: *(cooperator)* a cooperator

S2: *(defector)* a defector

S3: *(grim strike 1)* a player who starts by cooperating and defect for the rest the episode when best-response agent defects once

S4: *(grim strike 2)* a player who starts by cooperating and defect for the rest the episode when best-response agent defects twice

S5: *(tit-for-tat)* a player who plays tit-for-tat

S6: *(tit-for-tat tremble)* a player who a player who plays tit-for-tat and occasionally unconditionally defect. (noisy tit-for-tat)

S7: *(flipping)* a player who cooperate during the first 3 interactions and defect for the rest of the episode

S8: *(corrigable tit-for-tat)* a player who starts with defection and switch to tit-for-tat strategy when best-response agent defects

S9: *(corrigable tit-for-tat tremble)* a player who starts with defection and switch to noisy tit-for-tat strategy when best-response agent defects

Test scenarios for Externality mushrooms:

Focal player (our best response agent) encounters:

S0: *(visiting cooperators)* 4 cooperators

S1: *(visiting defectors)* 4 defectors

2 focal players (in our case 2 copies of best response agent) encounter:

S2: *(resident cooperators)* 3 cooperators

S3: *(resident cooperators)* 3 defectors

We provide an overview of the end to end methodological pipeline in Figure 5.4.

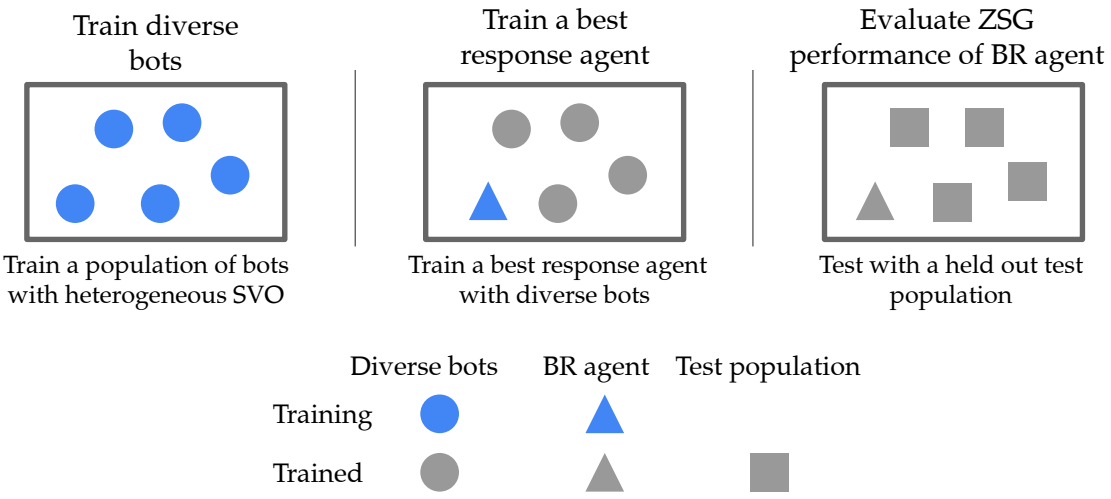


Figure 5.4: Overview of the methodology. Blue shapes show agents that are actively being trained, whereas gray ones denote frozen agents (bots). Circles represent the agents trained with diverse SVO, triangles denote a best response agent, and squares denote a held-out set of co-players. Evaluation is zero-shot, meaning the best response agent is frozen (gray triangle) and is evaluated against the held-out bots.

5.5 Agent architecture

The neural network of the agent consists of two convolutional layers, a two-layer perceptron, and an LSTM—all separated by ReLU activation functions. The convolutional layers have 16 and 32 output channels, kernel shapes of 8 and 4, and strides

of 8 and 1. The perceptron layers are 64 neurons each, and the LSTM layer has 128 units. The policy and baseline for the critic are created by multilayer perceptrons (256 hidden units with ReLU activations) connected to the output of the LSTM.

Representation shaping is achieved through the use of an auxiliary loss and contrastive predictive coding [73], which is used to differentiate between nearby time points via LSTM state representations. PopArt [33] is used to adjust for the different reward scales of the different environments. The optimization method used is RMSProp with a learning rate of 4×10^{-4}, epsilon of 10^{-5}, zero momentum, decay of 0.99, and batch size of 256. The baseline cost for the critic is 0.5, and the entropy regularization cost for the policy is 0.003.

5.6 Results

5.6.1 Experiment 1: Generating diverse policies in "in the matrix" repeated games

Experimental setup: We consider Stag hunt, Chicken and Prisoners' dilemma "in the matrix" repeated games. For each game we average the results over 3 random seeds. We train four agents with SVO values of $-15°, 0°, 60°$, and $75°$, respectively. These values were chosen to cluster around the incentives of competition (-15), selfishness (0) and pro-sociality ($60, 75$), symmetrically. The "in the matrix" repeated games are 2-player games. In addition to SVO bots we also train and freeze a set of selfish-baseline bots (i.e., no intrinsic reward) using the same procedure for comparison.

Finding 1: Heterogeneous SVO bots learn meaningfully diverse policies

We use the inventory count of the bots at the time of interaction as an environment-specific diversity measure. Since the inventory counts define the mixed

Figure 5.5: "in the matrix" repeated. *Diversity of policies of selfish-baseline bots and SVO bots.* Each subfigure shows average inventory counts during evaluation for 4 agents, trained with 50% self-play and 50% population play. The bottom row corresponds to SVO bots with $svo^i \in \{-15°, 0°, 60°, 75°\}$ and the top row corresponds to selfish-baseline bots. Green and red represents cooperative and defective resource counts respectively. Error bars show the standard deviation of results over 3 random seeds.

strategy probability vectors, sufficiently distinct ratios of inventory counts indicate distinct mixed strategies. During evaluation agents play in population-play.

Figure 12.6 shows the inventory counts for the 4 bots averaged over the last 500 interactions during evaluation after the completion of training. Top and bottom rows correspond to resource counts of selfish-baseline bots and SVO bots respectively. Figures 12.6(a), 12.6(b) and 12.6(c) correspond to Stag hunt, Chicken and Prisoners' dilemma respectively. The error bars presented in the figure correspond to the average results of 3 independent runs. The results demonstrate that in each game, all 4 selfish-baseline bots have comparable inventory count ratios, suggesting that their

policies lack diversity. Conversely, the 4 SVO bots exhibit varied inventory count ratios, indicating diverse behaviors. For each "in the matrix" repeated game, resource counts correspond to SVO bots with $svo = [-15°, 0°, 60°, 75°]$, where $svo^i = svo[i]$, $i \in \{1, 2, 3, 4\}$. We denote the cooperative resource counts and defective resource counts using green and red respectively. As the SVO angles increase from $-15°$ to $75°$, the ratio between the red and green resource counts increases, indicating a more cooperative, prosocial or altruistic behavior.

5.6.2 Experiment 2: Generating diverse policies in Externality Mushrooms

Experimental setup: Similar to the training process in "in the matrix" repeated game we average the results from 3 random seeds. For each seed we train 5 agents with SVO values of $-15°, 0°, 60°, 75°$, and $90°$, respectively in 50% self-play and 50% population-play. In addition to SVO bots we also train a set of selfish-baseline bots, using the same procedure for comparison.

Finding 2: The results extends to multi-player games with more than 2 players

We show that our method scales to games with more than 2 players. Figure 12.7 shows that in Externality Mushrooms, agents trained using heterogeneous SVO learn diverse policies. We use the count of mushrooms consumed of each type as the environment-specific diversity metric. The selfish-baseline bots tend to consume mushrooms at similar ratios across different types, whereas the SVO bots consume varying ratios of different mushroom types exhibiting meaningfully diverse behaviors. Agents with low (or negative) SVO consume the selfish mushroom (red), and even the spiteful mushroom (orange), whereas those with high SVO, tend to consume more of the prosocial mushrooms (green and blue).

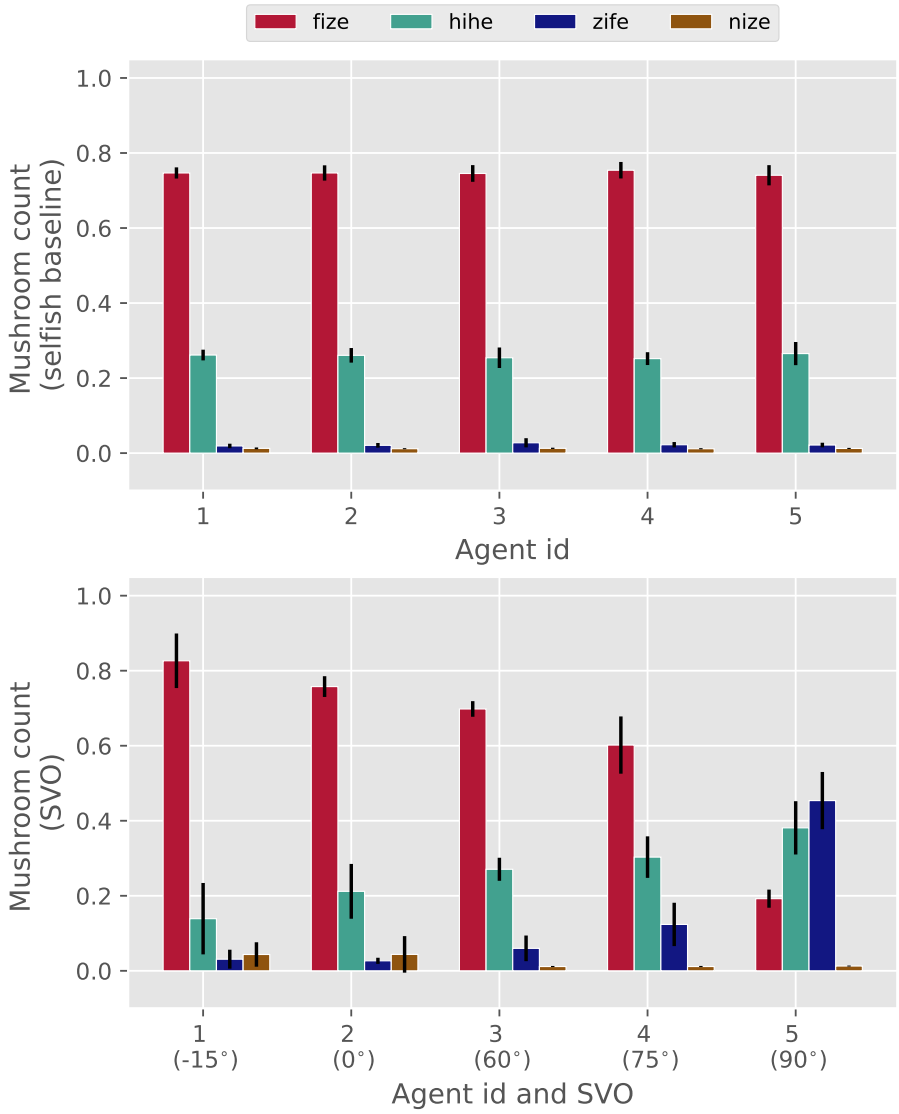


Figure 5.6: Externality mushrooms. *Diversity of policies of selfish-baseline bots and SVO bots.* Each plot shows average fraction of mushrooms consumed by 5 agents during evaluation, trained with 50% self-play and 50% population play in Externality mushrooms dense game. The bottom row corresponds to SVO agents with $svo^i \in \{-15°, 0°, 60°, 75°, 90°\}$ and the top row corresponds to selfish-baseline agents. Error bars show the standard deviation of results over 3 random seeds.

5.6.3 Experiment 3: Zero-shot generalization evaluation

We evaluate the zero-shot generalization performance of a learned best response to the SVO bots trained using heterogeneous SVO.

Baselines: We compare the performance of a learned best response policy for SVO bots with a best response to selfish-baseline bots, Fictitious co-play (FCP, a type of best response that includes also earlier checkpoints of the agents to best respond to) [82] and exploiters (i.e., a best response agent trained on the test scenario directly) [1] We train one exploiter for each test scenario. To train FCP agents we train a naive learning agent with 3 checkpoints for each bot from a bot population. Here we use the first checkpoint, mid checkpoint and last checkpoint. The mid checkpoint is the time during training where the agent first obtains half of its final reward, of the policies of the bots. We report results for FCP applied to the heterogeneous SVO bots FCP(SVO), as well as to selfish baselines FCP(selfish-baseline). We also compare performance of best response agents with zero-shot generalization performance of selfish-baseline agents and random agents.

Experimental setup: We train best-response agents for the selfish-baseline bots and SVO bots. Recall that we trained each type of bots, i.e., selfish-baseline or SVO, for 3 random seeds in this setup. We train a best-response agent for bots from each seed. For each type of test bots we show the average performance evaluation runs correspond to these 3 training runs.

Finding 3: Best-response agents learn a conditional behaviour In order to get a better understanding about the learned policies of the best-response agents we analyze the behaviour of the best-response agents during test time. For each test bot, Figures 12.8 and 12.9 show the fraction of interactions where the best-response agent cooperated with a bot with respect to the fraction of interactions where the bot cooperated with the best-response agent. Figure 12.8 corresponds to Stag hunt "in the matrix" repeated and 12.9 corresponds to Chicken "in the matrix" repeated.

In this analysis we define the best-response agent's interaction as a cooperation when they have higher number of cooperative resources than defective resources in

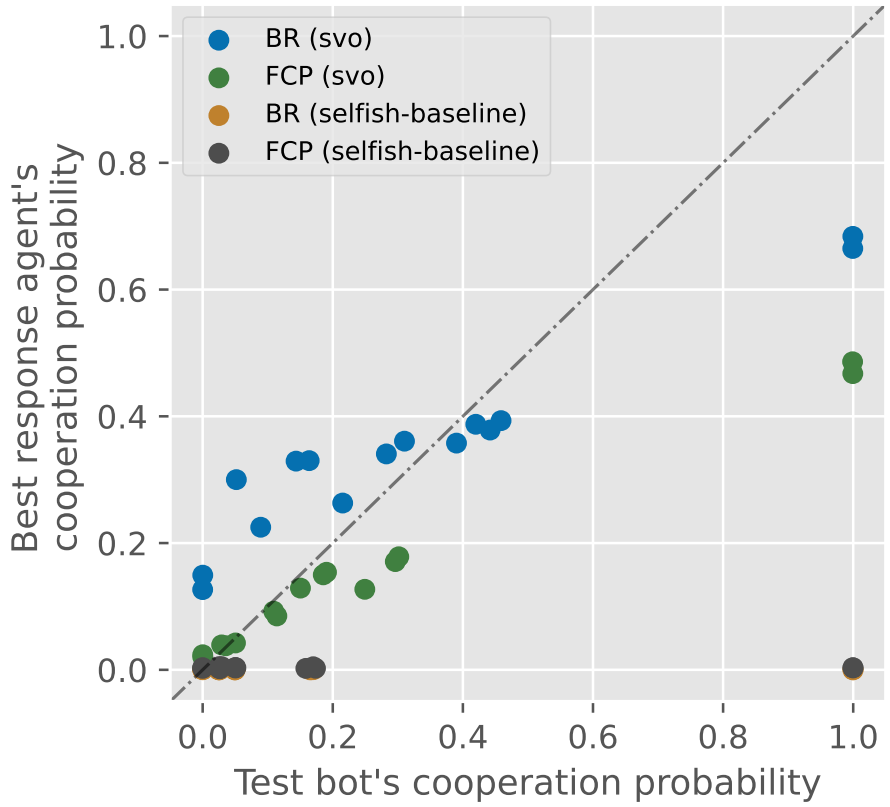


Figure 5.7: *Comparing how well best-response agents learn conditional policies in Stag hunt.*

their inventory at the time of interaction. In Stag hunt both agents cooperating, i.e., both agents playing Stag, yields a higher reward, but it is a riskier strategy. Defecting, yields a secure payoff. Both agents cooperating or both defecting are Nash equilibria, that is, there is no incentive to unilaterally deviate from that strategy. An agent who cooperates with a defector gets 0 reward. When trained in Stag hunt selfish-baseline bots learn to defect. The best response to unconditional defectors is defecting. Hence the best-response agents trained with selfish-baseline bots learn to unconditionally defect. In contrast the heterogeneous SVO bot population consists of both defectors and cooperators with different levels of cooperation and defection. Best-response agents training with SVO bots encounter both cooperators and defectors and subsequently

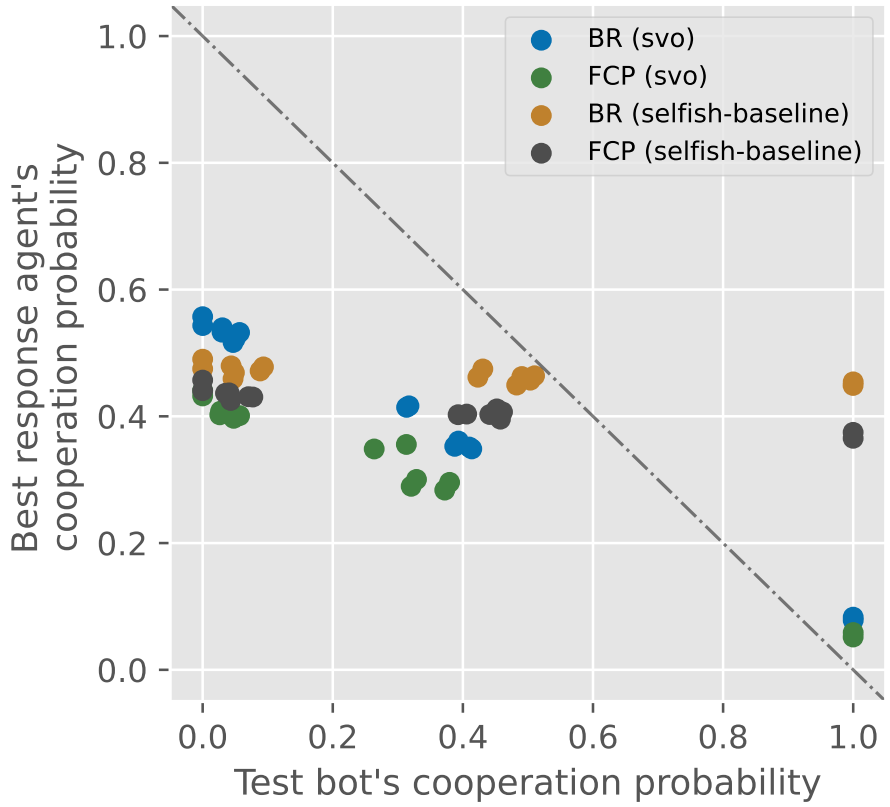


Figure 5.8: *Comparing how well best-response agents learn conditional policies in Chicken.*

learn a conditional policy that tends to cooperate with cooperators and defect with defectors.

In Chicken the two Nash equilibria are for one agent to cooperate (swerve) and the other agent to defect (straight). In this case selfish-baseline agents learn to do both defection and cooperation. Hence the best-response agents trained with selfish-baseline bots also learn to defect and cooperate. However in Figure 12.9 we see that this behaviour is not conditional. In contrast best-response agents training with SVO bots encounter mostly cooperative and mostly defective bots, leading to best-response agents learning a conditional behavior where they tend to cooperate with defectors and defect against cooperators.

Finding 4: Failure case with Prisoners' dilemma In Prisoner's dilemma the the Nash equilibrium is both agents defecting, as a results selfish-baseline agents learn to defect. Thus, the best response agents that are trained with selfish agents also learn to defect. Moreover, when facing a defector, the best response is to defect, while defecting against a cooperator yields the highest reward. Therefore, agents are incentivized to defect even when faced with an unconditional cooperator. Consequently, the best response to SVO bots (i.e., unconditional cooperators and defectors) is also to unconditionally defect. Figure 12.10 illustrates this showing that all the best-response agents are learning to defect regardless of the level of cooperation of their partners.

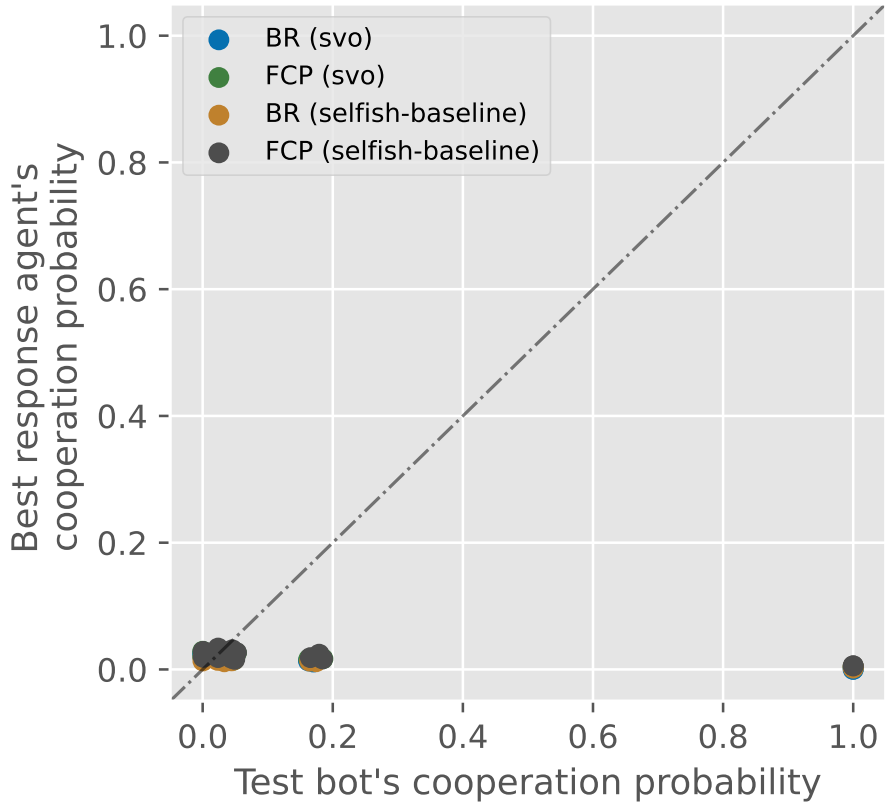


Figure 5.9: *Comparing how well best-response agents learn conditional policies in Prisoners' dilemma.*

	BR(SVO)	FCP(SVO)	BR(selfish-baseline)	FCP(selfish-baseline)	selfish-baseline	random	exploiter
Stag hunt ITMR	**0.876**	0.830	0.856	0.847	0.850	0.000	**0.988**
Chicken ITMR	0.696	0.668	**0.745**	0.723	0.723	0.000	**0.958**
Prisoner's dilemma ITMR	0.738	0.702	0.777	**0.783**	0.754	0.000	**1.000**
Externality mushrooms	0.619	0.764	0.612	**0.846**	0.660	0.000	**0.900**

Table 5.1: Zero-shot generalization performance of best response agents, selfish-baseline agent, random agent and exploiter. The score is calculated by first re scaling the rewards received by each agent such that in each scenario the agent with highest(lowest) reward gets score 1(0) and then averaging over all scenarios for each environment.

Finding 5: Best response agents perform better in zero-shot generalization Zero-shot generalization performance of the best response agents, selfish-baseline agent, random agent, and exploiters are given in Table 5.1. The score is calculated by normalizing the rewards agents receive in an episode across agents for each scenario and then averaging over all scenarios. The exploiters and random agent are intended to provide approximate upper and lower bounds for performance across all environments. As expected the table shows that the exploiters achieve the best performance, while the random agent performs the worst. Across all environments at least one best response agent performs better than the selfish-baseline agent indicating that learning a best response improves zero-shot generalization.

On average in the Stag hunt scenarios, BR(SVO) outperforms other agents. From figure 12.8 we see that BR(SVO) and FCP(SVO) cooperate with unconditional defectors with a small probability. However, in Stag hunt an agent cooperating with a defector or defecting with a defector receives the same reward. Thus when encountering defectors and test bots that are more likely to defect BR(SVO), FCP(SVO) receives comparable rewards to BR(selfish-baseline), FCP(fish-baseline). When encountering more cooperative test bots, best response agents that are able adapt to partner behaviours and cooperate with cooperators receive a higher reward. This leads to the higher score of BR(SVO) agent in Stag hunt.

The table 5.1 shows that in Chicken scenarios, BR(selfish-baseline) outperforms other agents. Note that in Chicken an agent cooperating with a defector

receives a higher reward than an agent defecting against a defector. From results in figure 12.9 we see that when test bots defect with a probability close to 1 all the best response agents cooperate with similar probabilities. Thus in scenarios where test bots are unconditionally defecting all the best response agents obtain comparable performance. Further, note that when test bots are cooperating with nearly 0.4 probability BR(selfish-baseline) and FCP(selfish-baseline) cooperate with a higher matching probability compared to BR(SVO) and FCP(SVO) thus leading to BR(selfish-baseline) and FCP(selfish-baseline) obtaining better performance. In scenarios where best response agents encounter unconditional cooperators BR(SVO) and FCP(SVO) defect with a probability close to 1 obtaining better performance compared to BR(selfish-baseline) and FCP(selfish-baseline). Since most of the test scenarios consists of defectors or test bots that are more likely to defect this leads to BR(selfish-baseline) outperforming BR(SVO) and BR(FCP) agents.

Recall that from Figure 12.10 illustrates that all the best-response agents are defecting against all test bots. Thus we expect the performance score of best response agents for Prisoner's dilemma given in Table 5.1 to be similar. However, surprisingly BR(selfish-baseline) and FCP(selfish-baseline) perform better than BR(SVO) and FCP(SVO). We leave investigating this as future work.

In Externality mushrooms FCP type best response agents perform better than best response agents trained with only final policies of the opponents/ partners. This indicates that best response agents that encounters less proficient agents as well as more proficient agents perform better than the best response agents that only encounters proficient agents during training time.

Chapter 6

Final remarks

6.1 Effective communication in multi-agent multi-armed bandits

6.1.1 Conclusion

In this thesis, we have explored the challenges and intricacies of multi-agent multi-armed bandits in the presence of probabilistic communication failures and communication costs. By investigating the impact of communication constraints on the performance of cooperative bandits, we have gained valuable insights into the development of robust and efficient solutions for a wide range of decision-making problems.

The probabilistic communication failure chapter focused on addressing the issue of unreliable communication links, where each agent's messages can drop with an agent-specific probability. We have proposed effective communication protocols and strategies that are resilient to such failures, ensuring that essential information is shared among agents despite message drops. These adaptive strategies have demonstrated their potential to maintain decision-making efficiency while coping with communication link failures.

In the communication cost chapter, we have examined the trade-offs between information sharing and communication overhead in multi-agent multi-armed bandits. By developing heuristic communication protocols that selectively share information about suboptimal options, we have shown that it is possible to balance communication costs while maintaining good performance. These protocols allow agents to make more informed decisions, reduce exploration redundancy, and ultimately increase the group's cumulative reward.

Overall, our research contributes to the understanding of cooperative bandits under communication constraints, highlighting the importance of robust and efficient communication protocols in the presence of probabilistic communication failures and communication costs. The findings of this thesis can be applied to various domains, including sensor networks, distributed control systems, and multi-robot coordination, among others.

6.1.2 Future work

Building upon the heuristic communication protocol discussed in Chapter 3, several research directions can be explored to further enhance the performance and applicability of multi-agent multi-armed bandits in various domains. The following areas of future work have the potential to deepen our understanding and expand the capabilities of multi-agent multi-armed bandits:

- Adaptive communication protocols: Investigate the development of adaptive communication protocols that dynamically adjust the frequency and content of information sharing based on the evolving state of the environment and the performance of the agents. Such protocols could optimize communication cost while maintaining or improving the group's decision-making efficiency.

- Heterogeneous agents and learning rates: Explore scenarios where agents have different learning rates, capabilities, or access to information. Understanding how to effectively share information in such heterogeneous settings can provide insights into the design of cooperative bandits that cater to diverse agent populations.

- Scalability and distributed learning: Investigate the scalability of the proposed communication protocol for large-scale settings with numerous agents and options. Techniques for distributed learning and communication could be developed to ensure that the performance gains are maintained even in large-scale and complex environments.

- Applications in various domains: Apply the developed heuristic communication protocol to real-world problems in diverse domains, such as recommendation systems, resource allocation, and online advertising. Evaluating the performance of the approach in practical settings can provide valuable feedback for further refinements and improvements.

- Exploration-exploitation trade-offs: Delve deeper into the exploration-exploitation trade-offs in the context of multi-agent multi-armed bandits with communication costs. Develop new techniques that can adaptively balance exploration and exploitation while optimizing communication overhead.

The study of multi-agent multi-armed bandits with communication link failures, where each agent's messages can drop with an agent-specific probability, presents several intriguing research avenues. Addressing the challenges posed by unreliable communication links can further enhance the robustness and applicability of multi-agent multi-armed bandits in real-world scenarios. The following future directions hold promise for expanding our understanding of multi-agent multi-armed bandits under communication constraints:

- Communication link failure models: Investigate different models of communication link failures, such as time-varying, correlated, or burst failures, to understand their impact on the performance of multi-agent multi-armed bandits. Developing strategies that can adapt to various failure models will contribute to the robustness of the proposed solutions.

- Error-resilient communication protocols: Develop communication protocols that are resilient to communication link failures, ensuring that essential information is reliably shared among agents despite message drops. Techniques such as error detection and correction codes or message redundancy could be explored to enhance the robustness of the communication process.

- Adaptive information sharing: Investigate adaptive information-sharing strategies that dynamically adjust the content and frequency of communication based on the current state of communication link failures. Such strategies could optimize the use of available communication resources while maintaining decision-making efficiency.

- Decentralized and cooperative learning: Explore decentralized learning and decision-making algorithms that allow agents to learn and make decisions locally while incorporating shared information from other agents. Developing methods that can effectively incorporate partial or noisy information received from other agents will be crucial for robust performance in the presence of communication link failures.

- Evaluation in real-world scenarios: Apply the developed techniques for multi-agent multi-armed bandits with communication link failures to real-world problems in various domains, such as sensor networks, distributed control systems, and robotic coordination. Evaluating the performance of the approach in prac-

tical settings can provide valuable insights for further refinements and improvements.

By pursuing these research directions, the field of multi-agent multi-armed bandits can continue to grow and advance, contributing to the development of more efficient, robust, and scalable solutions for a wide range of cooperative decision-making problems.

6.2 Generalization in multi-agent reinforcement learning

6.2.1 Conclusion

We investigated the impact of heterogeneous social value orientation on different incentive structures in sequential social dilemmas. We tested whether the presence of heterogeneous SVO leads to diverse policies and if learning a best response to these policies improves zero-shot generalization. The study found that the presence of heterogeneous SVO does indeed lead to measurable diversity in policies, and this diversity often results in better zero-shot generalization for agents that best respond to them.

The best-response agents achieve better performance by learning a conditional policy that adapts to novel agents during test time. The study also revealed that when the sequential social dilemma is not an equilibrium-selection problem, this method still generates meaningful diversity in policies, but it fails to achieve better zero-shot generalization performance. This occurs because the best response to a diverse set of policies collapses to one unconditional policy that performs poorly when encountering conditional policies during test time.

Additionally, the study demonstrated that the results extend to multi-player games with more than two players. Our findings have implications for understanding how heterogeneous SVO impacts incentive structures and policy diversity, and how agents can learn to adapt to diverse policies during test time to achieve better zero-shot generalization performance. Our findings provide new insights into the behavior of agents in sequential social dilemmas and highlights the importance of considering the role of heterogeneity in SVO in the design of incentive structures.

6.2.2 Future work

The exploration of generating heterogeneous policies using heterogeneous prosociality in mixed motive games, and leveraging these policies to improve zero-shot generalization in Multi-Agent Reinforcement Learning (MARL), offers several promising research directions. By developing a better understanding of the interplay between heterogeneous prosociality and policy generation, it may be possible to enhance the adaptability, robustness, and performance of MARL algorithms in various domains. The following areas of future work can contribute to the advancement of this research topic:

- Prosociality models: Investigate different models of prosociality, including varying levels of cooperation and competition, and assess their impact on the generation of heterogeneous policies. Understanding how diverse prosocial behaviors influence policy formation can lead to more nuanced strategies in mixed motive games.

- Adaptive prosociality: Explore adaptive prosociality, where agents can dynamically adjust their prosocial tendencies based on the current state of the environment, the performance of other agents, or their own learning progress. Such adaptability can contribute to more robust and flexible MARL algorithms.

- Heterogeneous agent architectures: Examine the role of heterogeneous agent architectures, such as those with varying learning rates, observation spaces, or action spaces, in the generation of heterogeneous policies. Identifying the interdependencies between agent architecture and prosocial behavior can provide valuable insights for designing more effective MARL solutions.

- Transfer learning and generalization: Investigate the impact of heterogeneous prosociality on transfer learning and generalization across different tasks, environments, and agent populations. Developing methods that can leverage the generated policies for improved zero-shot generalization can significantly enhance the applicability of MARL algorithms.

- Exploration-exploitation trade-offs: Delve deeper into the exploration-exploitation trade-offs in the context of heterogeneous prosociality and mixed motive games. Develop new techniques that can balance exploration and exploitation effectively while taking into account the diverse prosocial tendencies of agents.

- Applications in various domains: Apply the developed techniques for generating heterogeneous policies using heterogeneous prosociality to real-world problems in diverse domains, such as autonomous vehicles, multi-robot coordination, and social dilemmas. Evaluating the performance of the approach in practical settings can provide valuable feedback for further refinements and improvements.

By pursuing these research directions, the field of Multi-Agent Reinforcement Learning can continue to grow and advance, contributing to the development of more efficient, robust, and scalable solutions for a wide range of cooperative and competitive decision-making problems involving heterogeneous prosociality in mixed motive games.

Part II

Published Work

Chapter 7

Overview

Part II of this dissertation consists of five peer-reviewed papers. The notations of the papers are changed to be consistent with the rest of the dissertation. Minor formatting adjustments were made to fit the dissertation format. Only the papers for which the I was the primary contributor are included in this thesis. Other papers where I was a co-author have been mentioned with proper citations.

7.1 Outline

In Chapter 8, we delve into the multi-armed bandit problem in the presence of communication costs. Specifically, we quantify the communication cost by the number of communication rounds required by the agents. To tackle this issue, we propose a novel communication protocol that leverages exploration to minimize communication cost. Under this protocol, agents only communicate during exploration, leading to logarithmic communication cost. Our study sheds light on the potential of using exploration as a means to reduce communication costs in multi-agent systems.

In Chapter 9, we extend the framework introduced in Chapter 8 and propose a new communication protocol. Our novel protocol achieves logarithmic communication cost while maintaining a group cumulative regret that is of the same order as the

regret under full communication. Our study provides insights into the effectiveness of communication protocols in multi-agent systems and contributes to the development of low-cost communication protocols that do not compromise performance. We consider the effect of agent heterogeneity with respect to their position in the communication network.

Chapter 10 proposes decentralized learning algorithms for multi-agent bandit problems under three typical real-world communication scenarios, namely (a) message-passing over stochastic time-varying networks, (b) instantaneous reward-sharing over a network with random delays, and (c) message-passing with adversarially corrupted rewards, including byzantine communication. We demonstrate that our proposed algorithms achieve competitive performance and near-optimal guarantees on the group regret incurred under each of these environments. Additionally, we present an improved delayed-update algorithm for the setting with perfect communication, which outperforms the existing state-of-the-art on various network topologies. We also derive tight network-dependent minimax lower bounds on the group regret.

Chapter 11 aims to explore the potential of leveraging individual differences to enhance group performance. Specifically, we focus on investigating star communication graphs, aiming to identify effective approaches for promoting the center agent to conduct more exploratory actions, which, in turn, can provide more useful information for peripheral agents, ultimately resulting in improved group performance.

Chapter 12 investigates the impact of heterogeneous Social Value Orientation (SVO) on policy diversity and zero-shot generalization in sequential social dilemmas with different incentive structures. It reveals that heterogeneous SVO results in diverse policies, and learning a best response to these policies enhances zero-shot generalization. The observed improvement stems from agents learning to adapt their

behavior based on their partners' or opponents' diverse strategies, enabling more effective responses during training.

7.2 Author contributions

I am the lead authors and lead contribution of materials, including mathematical analysis, illustrations, simulations, presented in the five included papers. My advisor, Professor Naomi Ehrich Leonard, advised me on almost all aspects of my research and my collaborators further helped me improve it. I have described specific contributions in each paper below.

- Chapters 8, 9, 11 is based on the work on Multi-agent multi-armed bandits. Naomi Leonard provided valuable suggestion through out the conception of research problem, executing the research, analysing results and writing the paper. I wrote the initial drafts and Naomi Leonard revised and edited the drafts.

- Chapter 10 started from my initial discussions with Abhimanyu Dubey. Abhimanyu Dubey contributed the sections on reward corruptions and lower bounds. I contributed the sections on probabilistic communication and stochastic delays. Abhimanyu Dubey and I wrote the other sections together. The sections from this chapter that are included in the Chapter 4 are my contributions. Naomi Leonard and Alex Pentland provided valuable suggestion through out the conception of research problem, executing the research, analysing results and writing the paper. Naomi Leonard edited the final draft.

- Chapter 12 was the result of my internship project at DeepMind. My brainstorming sessions with Edgar Duéñez-Guzmán and Joel Leibo helped in conceiving the idea and their help was instrumental in refining my approach. Kevin McKee supplemented the idea on learning adaptive policy, which proved crucial

in the second half of the project. John Agapiou helped me significantly to run large scale experiments on the DeepMind computational resources. I wrote the initial draft. Kevin McKee, Edgar Duéñez-Guzmán, John Agapiou and Joel Leibo revised and edited the draft.

Chapter 8

A Dynamic Observation Strategy for Multi-agent Multi-armed Bandit Problem

Udari Madhushani and Naomi Ehrich Leonard

We define and analyze a multi-agent multi-armed bandit problem in which decision-making agents can observe the choices and rewards of their neighbors under a linear observation cost. Neighbors are defined by a network graph that encodes the inherent observation constraints of the system. We define a cost associated with observations such that at every instance an agent makes an observation it receives a constant observation regret. We design a sampling algorithm and an observation protocol for each agent to maximize its own expected cumulative reward through minimizing expected cumulative sampling regret and expected cumulative observation regret. For our proposed protocol, we prove that total cumulative regret is logarithmically bounded. We verify the accuracy of analytical bounds using numerical simulations.

8.1 Introduction

The effect of communication structure in cooperative and competitive multi-agent systems has been extensively studied in decision theory. Performance of a group of social learners can be improved by the shared information among individuals. In most real-world decision-making processes, however, information sharing between agents can be costly. As a result, directed communication, where each agent only needs to observe its neighbors, has advantages over undirected communication, where each agent sends and receives information. Even when observation costs are high, agents can keep costs to a minimum by choosing when and whom to observe as a function of their own performance. Further, in this setting costs associated with cooperation can be avoided.

Consider the problem of a group of fishermen foraging in an uncertain environment that consists of a distribution of spatial resource (fish). Because of the natural dynamics of fish, environmental conditions, and other external factors, the resource will be distributed stochastically. As a result, a fisherman will receive different reward values (number of fish harvested) at different times, even when sampling from the same patch. Thus, in order to maximize cumulative reward fishermen need to be able to exploit, i.e., forage in well sampled patches known to provide better harvest, and to explore, i.e, forage in poorly sampled patches, which is riskier but may provide even better harvest than well sampled patches. Benefiting from exploitation requires sufficient exploration and identification of the patches that yield highest rewards. More generally, optimal foraging performance comes from balancing the trade-off between exploring and exploiting. This is known as the explore-exploit dilemma.

Multi-armed bandit (MAB) problems are a set of mathematical models that have been proposed to capture the salient features of explore-exploit trade-offs [83, 79]. For the standard MAB problem the reward distributions associated with options are static. An agent estimates the expected reward of each option using the rewards it

receives through sampling. The agent chooses among options by considering a trade-off between estimated expected reward (exploiting) and the uncertainty associated with the estimate (exploring). Therefore, in the frequentist setting, the natural way of estimating the expectation of the reward is to consider the sample average [44, 2, 7]. The papers [41, 78] present how to incorporate prior knowledge about reward expectation in the estimation step by leveraging the theory of conditional expectation in the Bayesian setting.

Multi-agent multi-armed bandit (MAMAB) problems consider a group of individuals facing the same MAB problem simultaneously. For an individual to maximize its own reward, it will naturally seek to observe its neighbors and use those observations to improve its performance. Individual and group performance of agents will vary according to the observation structure, i.e., who is observing whom, and the type of information they observe. For example, if the agents are cooperative and can broadcast signals, they could share their estimates of rewards. When there are constraints, such as communication costs and privacy concerns, they might instead share only their instantaneous rewards and choices. Even without the ability to broadcast, agents may still be able to use sensors to observe the instantaneous rewards and choices of neighbors. A centralized multi-agent setting is considered in [4] and a decentralized setting is considered in [39]. The papers [47, 46] use a running consensus algorithm in which agents observe the reward estimates of their neighbors. In [42, 48] an MAMAB problem is studied in which agents observe instantaneous rewards and choices in a leader-follower setting.

In all of these previous works, communication between agents is assumed to be cost free. However, in real world settings observing neighbors or exchanging information with neighbors is costly. In the present paper, we propose a setting in which agents can decide when and whom to observe in order to receive maximum benefits from observations that incur a cost. An underlying undirected network graph defines

neighbors and models the inherent observation constraints present in the network. Agents receive a fixed observation cost at every instance they observe a neighbor.

To account for the observation cost, we define cumulative regret to be the total cumulative regret agents receive from sampling suboptimal options (sampling regret) and from observing neighbors (observation regret). Deterministic [48] and probabilistic [61] communication strategies proposed in the MAB literature lead to a linear cumulative observation regret. Our main contribution is the design of a new strategy for which we prove a logarithmic total cumulative regret, i.e., order-optimal performance. Our design leverages the intuition that it is most useful to observe neighbors when uncertainty associated with estimations of rewards is high.

In Section **??** we introduce the MAMAB problem and we propose an efficient sampling rule and a communication protocol for an agent to maximize its own total expected cumulative reward. We analyze the performance of the proposed sampling rule in Section **??**. In Section 8.3.1 we analytically upper bound the expected cumulative regret and in Section 8.3.2 we analytically upper bound the expected observation regret. We present the upper bound for the total expected cumulative regret in section 8.3.3. In Section **??** we provide numerical simulation results and computationally validate the analytical results. We conclude in Section **??** and provide additional mathematical details in the Appendix.

8.2 Multi-agent Multi-armed Bandit Problem

In this section we present the mathematical formulation of the MAMAB problem studied here. Let N be the number of options (arms) and K the number of agents. Define X_i as the random variable that denotes reward associated with option $i \in \mathcal{I} = \{1, 2, \ldots, N\}$. In this paper we assume that all the reward distributions are sub-Gaussian. Let σ_i be the variance proxy of X_i, and μ_i the expected reward of option i.

Let i^* be the optimal option with highest expected reward $\mu_{i^*} = \max\{\mu_1, \mu_2, \ldots, \mu_N\}$. Each agent $k \in \{1, \ldots, K\}$ chooses one option at each time step $t \in \{1, 2, \ldots, T\}$ with the goal of minimizing its cumulative regret. In MAB problems, cumulative regret is typically defined as cumulative sampling regret, which is equivalent to expected number of times suboptimal options are selected. We let cumulative regret be the sum of cumulative sampling regret and a cumulative observation regret that accumulates a fixed cost for every observation of a neighbor.

We assume that the expected reward values μ_i are unknown and the variance proxy values σ_i are known to the agents. To improve its own performance, each agent observes its neighbors according to an observation protocol that we define. We use a network graph to encode hard observation constraints and this defines neighbors of agents. Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be an undirected graph. $\mathcal{V}$ is a set of K nodes, such that node k in $\mathcal{V}$ corresponds to agent k for $k \in \{1, \ldots, K\}$. $\mathcal{E}$ is a set of edges between nodes in $\mathcal{V}$. If there is an edge $e(k, j) \in \mathcal{E}$ between node k and node j, then we say that agent k and agent j are neighbors. Since the graph is undirected, $e(k, j) \in \mathcal{E} \iff e(j, k) \in \mathcal{E}$. Let d_k be the number of neighbors of agent k.

Let $\varphi_k^t \in \mathcal{I}$ and X_k^t be random variables that denote the option chosen by agent k and the reward received by agent k at time t, respectively. Let $\mathbb{I}_{\{\varphi_k^t = i\}}$ be a random variable that takes value 1 if option i is chosen by agent k at time t and is 0 otherwise. Let $\mathbb{I}_{\{k,j\}}^t$ be a random variable that takes value 1 if agent k can observe agent j at time t and is 0 otherwise.

In order to maximize the cumulative reward in the long run, agents need to both identify the best options through exploring and sample the best options through exploiting. Observing neighbors allows an agent to receive more information about options and hence obtain better estimates about expected reward values of options. This leads to less exploring and more exploiting, which reduces the regret an agent receives due to sampling suboptimal options. However, since taking observations is

costly, an agent is required to find a trade-off between the information gain and the cost associated with observations. Let $c_{k,j}$ be the cost incurred by agent k when it observes the instantaneous reward and choice of agent j at time step t. In this paper we consider the case in which $c_{k,j} = c, \forall j, k$.

Let the number of times that agent k samples option i until time t be given by the random variable $n_i^k(t) = \sum_{\tau=1}^{t} \mathbb{I}_{\{\varphi_k^\tau = i\}}$. And let the total number of times that agent k observes rewards from option i until time t be given by the random variable $N_i^k(t)$, where

$$N_i^k(t) = \sum_{\tau=1}^{t} \sum_{j=1}^{K} \mathbb{I}_{\{\varphi_j^\tau = i\}} \mathbb{I}_{\{k,j\}}^\tau.$$

We define a sampling rule based on the well known UCB (Upper Confidence Bound) rule for a single agent [7]. The UCB rule chooses the option at time t that maximizes an objective function that is the sum of an exploit term, equal to the estimate of the reward mean at time t, and an explore term, equal to a measure of uncertainty in that estimate at time t. Our sampling rule for agent k in the MAMAB problem accounts for the observations of neighbors by using them to improve its estimate and reduce its uncertainty. Let the estimate by agent k of the expected reward from option i at time t be given by the random variable $\widehat{\mu}_i^k(t)$, where

$$\widehat{\mu}_i^k(t) = \frac{S_i^k(t)}{N_i^k(t)},$$

and $S_i^k(t) = \sum_{\tau=1}^{t} \sum_{j=1}^{K} X_i \mathbb{I}_{\{\varphi_j^\tau = i\}} \mathbb{I}_{\{k,j\}}^\tau$ is the total reward observed by agent k from option i until time t.

Definition 4. *The* sampling rule $\{\varphi_k^t\}_1^T$ for agent k at time $t \in \{1, \ldots, T\}$ *is defined as*

$$\mathbb{I}_{\{\varphi_k^{t+1}=i\}} = \begin{cases} 1 & , \quad Q_i^k(t) = \max\{Q_1^k(t), \cdots, Q_N^k(t)\} \\ 0 & , \quad \text{o.w.} \end{cases} \tag{8.1}$$

with

$$Q_i^k(t) = \widehat{\mu}_i^k(t) + C_i^k(t) \tag{8.2}$$

$$C_i^k(t) = \sigma_i \sqrt{2(\xi+1)\frac{\log t}{N_i^k(t)}}, \tag{8.3}$$

where $\xi > 1$ is a tuning parameter that captures the trade-off between exploring and exploiting.

To find a balance between information gain and observation cost we define an observation rule for agents so that they choose to incur the cost of making observations of neighbors only when observations are most needed, i.e., when their own uncertainty is high. In the following observation rule, an agent observes the instantaneous rewards and choices of all of its neighbors only when it is exploring, since it explores when uncertainty is high. If agent k chooses the option at time t that corresponds to the maximum of its estimates of reward means, $\widehat{\mu}_1^k(t), \ldots, \widehat{\mu}_N^k(t)$, then it is exploiting and it does not observe its neighbors.

Definition 5. *The* observing rule $\mathbb{I}_{\{k,j\}}^t$ for agent k at time $t \in \{1, \ldots, T\}$ *and* $\forall j$ *is defined as*

$$\mathbb{I}_{\{k,j\}}^{t+1} = \begin{cases} 0\,, \varphi_k^t = i,\ s.t.\ \widehat{\mu}_i^k(t) = \max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\} \\ 1\,, \text{o.w.} \end{cases} \tag{8.4}$$

8.3 Performance Analysis

In this section we analyze the cumulative regret of agent k due to sampling suboptimal options and observing neighbors when employing the sampling rule of Definition 4 and observation rule of Definition 5.

8.3.1 Sampling Regret Analysis

Let i be a suboptimal option. The total number of times agent k samples from option i can be upper bounded as

$$n_i^k(T) = \sum_{t=1}^{T} \mathbb{I}_{\{\varphi_k^t = i\}} \leq \sum_{t=1}^{T} \mathbb{I}_{\{Q_i^k(t) \geq Q_{i^*}^k(t)\}}.$$

Here $\mathbb{I}_{\{Q_i^k(t) > Q_{i^*}^k(t)\}}$ is an indicator function such that

$$\mathbb{I}_{\{Q_i^k(t) > Q_{i^*}^k(t)\}} = \begin{cases} 1 & , \quad Q_i^k(t) \geq Q_{i^*}^k(t) \\ 0 & , \quad \text{o.w.} \end{cases}$$

Thus we have

$$\mathbb{E}\left(n_i^k(T)\right) \leq \sum_{t=1}^{T} \mathbb{P}\left(Q_i^k(t) \geq Q_{i^*}^k(t)\right).$$

Let $R_s^k(T)$ be the cumulative sampling regret of agent k from option i until time T. Recall that the cumulative regret is defined as the loss incurred by sampling suboptimal options. Define $\Delta_i = \mu_{i^*} - \mu_i$. Then we have, from [44],

$$\mathbb{E}\left(R_s^k(T)\right) = \sum_{i=1}^{N} \Delta_i \mathbb{E}\left(n_i^k(T)\right). \tag{8.5}$$

To analyze the expected number of samples from suboptimal options until time T, we first note that $\forall i, k, t$ we have

$$\left\{\mathbb{I}_{\{\varphi_k^{t+1}=i\}}\right\} \subseteq \left\{Q_i^k(t) \geq Q_{i^*}^k(t)\right\} \subseteq \left\{\mu_{i^*} < \mu_i + 2C_i^k(t)\right\}$$
$$\cup \left\{\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\right\} \cup \left\{\widehat{\mu}_i^k(t) \geq \mu_i + C_i^k(t)\right\}$$

and so

$$\mathbb{E}\left(n_i^k(T)\right) \leq \sum_{t=1}^{T} \mathbb{P}\left(\mu_{i^*} < \mu_i + 2C_i^k(t)\right) +$$
$$\sum_{t=1}^{T} \mathbb{P}\big(\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\big) + \sum_{t=1}^{T} \mathbb{P}\big(\widehat{\mu}_i^k(t) \geq \mu_i + C_i^k(t)\big)\,. \tag{8.6}$$

Next we analyze concentration probability bounds on the estimates of options.

Theorem 8. *For any $\zeta > 1$ and for $\sigma_i > 0$ there exists a $\vartheta > 0$ such that*

$$\mathbb{P}\left(\widehat{\mu}_i^k(T) - \mu_i > \sqrt{\frac{\vartheta}{N_i^k(T)}}\right) \leq \frac{\nu \log(d_k + 1)T}{\exp(2\kappa\vartheta)}$$

where

$$\nu = \frac{1}{\log \zeta}, \quad \kappa = \frac{1}{\sigma_i^2\left(\zeta^{\frac{1}{4}} + \zeta^{-\frac{1}{4}}\right)^2}.$$

The proof of Theorem 8 can be found in the paper [61]. Using symmetry we conclude that

$$\mathbb{P}\left(\left|\widehat{\mu}_i^k(T) - \mu_i\right| > \sqrt{\frac{\vartheta}{N_i^k(T)}}\right) \leq \frac{\nu \log(d_k + 1)T}{\exp(2\kappa\vartheta)}.$$

Lemma 2. *For $\vartheta = 2\sigma_i^2(\xi+1)\log T$ and $\xi > 1$ there exists a $\zeta > 1$ such that*

$$\mathbb{P}\left(\left|\widehat{\mu}_i^k(T) - \mu_i\right| > \sigma_i\sqrt{\frac{2(\xi+1)\log T}{N_i^k(T)}}\right) \leq \frac{\nu\log(d_k+1)T}{T^{\xi+1}}.$$

The proof of Lemma 2 can be found in the paper [61].

We proceed to upper bound the summation of the probabilities of the events $\left\{\mu_{i^*} < \mu_i + 2C_i^k(t)\right\}$ for $t \in \{1, 2, \ldots, T\}$ as follows. Using equation (11.4) we have that the inequality $\mu_{i^*} < \mu_i + 2C_i^k(t)$ implies

$$\frac{\Delta_i^2}{4\sigma_i^2}\left(N_i^k(t)\right)^2 - 2(\xi+1)\log t\left(N_i^k(t)\right) < 0.$$

This inequality does not hold for $N_i^k(t) > \eta_i(t)$, where

$$\eta_i(t) = \frac{8\sigma_i^2(\xi+1)}{\Delta_i^2}\log t.$$

Thus we have

$$\sum_{t=1}^{T}\mathbb{P}\left(Q_i^k(t) \geq Q_{i^*}^k(t), N_i^k(t) > \eta_i(t)\right) \leq \eta_i(T). \tag{8.7}$$

From the probability bounds given in Lemma 2 and (11.8), the total expected number of times agent k samples suboptimal option i until time T is upper bounded as

$$\begin{aligned}\mathbb{E}\left(n_i^k(T)\right) \leq & \frac{1}{\log\zeta}(1+\log(d_k+1)) + \frac{8\sigma_i^2(\xi+1)}{\Delta_i^2}\log T \\ & + \frac{1}{2^\xi\log\zeta}\left(\frac{\log(d_k+1)}{\xi} + \frac{2}{\xi-1}\right) \\ & + \frac{1}{T^{\xi-1}\log\zeta}\left(\frac{\log(d_k+1)}{T\xi} + \frac{1}{\xi-1}\right)\end{aligned} \tag{8.8}$$

where $\zeta, \xi > 1$.

From equation (8.5) the expected cumulative sampling regret of agent k until time T is upper bounded as

$$\begin{aligned}\mathbb{E}\left(R_s^k(T)\right) \leq & \sum_{i=1}^{N} \frac{\Delta_i}{\log \zeta}(1+\log(d_k+1)) \\ & + \frac{8\sigma_i^2(\xi+1)}{\Delta_i}\log T \\ & + \sum_{i=1}^{N} \frac{\Delta_i}{2^{\xi}\log\zeta}\left(\frac{\log(d_k+1)}{\xi}+\frac{2}{\xi-1}\right) \\ & + \sum_{i=1}^{N} \frac{\Delta_i}{T^{\xi-1}\log\zeta}\left(\frac{\log(d_k+1)}{T\xi}+\frac{1}{\xi-1}\right). \end{aligned} \tag{8.9}$$

8.3.2 Observation Regret Analysis

Recall that c is the constant unit cost associated with observations. Let $R_o^k(T)$ be the cumulative observation regret of agent k at time step T. Then we have

$$R_o^k(T) = c\sum_{t=1}^{T}\sum_{j=1}^{K}\mathbb{I}_{\{k,j\}}^t.$$

This is equivalent to the number of observations taken by agent k until time T. Expected cumulative observation regret can be expressed as

$$\mathbb{E}\left(R_o^k(T)\right) = c\sum_{t=1}^{T}\sum_{j=1}^{K}\mathbb{E}\left(\mathbb{I}_{\{k,j\}}^t\right). \tag{8.10}$$

So expected cumulative observation regret can be upper bounded by upper bounding the expected number of observations until time T:

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{j=1}^{K}\mathbb{E}\left(\mathbb{I}_{\{k,j\}}\right) \\ &\quad = d_k\sum_{t=1}^{T}\mathbb{P}\left(\varphi_k^t = i, \widehat{\mu}_i^k(t) \neq \max\{\widehat{\mu}_1^k(t),\cdots,\widehat{\mu}_N^k(t)\}\right). \end{aligned} \tag{8.11}$$

To analyze the expected number of observation, we use

$$\begin{aligned}
&\mathbb{P}\left(\varphi_k^t = i, \widehat{\mu}_i^k(t) \neq \max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\}\right) = \\
&\mathbb{P}\left(\varphi_k^t = i^*, \widehat{\mu}_{i^*}^k(t) \neq \max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\}\right) \\
&+ \mathbb{P}\left(\varphi_k^t = i, \widehat{\mu}_i^k(t) \neq \max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\}, i \neq i^*\right).
\end{aligned}$$

We first upper bound the expected number of times agent k observes its neighbors until time T when it decides to explore after sampling a suboptimal option.

Lemma 3. *For all suboptimal $i \neq i^*$ we have*

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}\left(\varphi_k^t = i, \widehat{\mu}_i^k(t) \neq \max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\}, i \neq i^*\right) \\
&\leq \frac{N-1}{\log \zeta}(1 + \log(d_k + 1)) + \sum_{\substack{i=1 \\ i \neq i^*}}^{N} \frac{8\sigma_i^2(\xi+1)}{\Delta_i^2} \log T \\
&+ \frac{N-1}{2^{\xi} \log \zeta}\left(\frac{\log(d_k+1)}{\xi} + \frac{2}{\xi - 1}\right) \\
&+ \frac{N-1}{T^{\xi-1} \log \zeta}\left(\frac{\log(d_k+1)}{T\xi} + \frac{1}{\xi - 1}\right).
\end{aligned}$$

The proof of Lemma 3 is given in the Appendix.

Next we analyze the expected number of times agent k observes its neighbors until time T when it decides to explore after sampling the optimal option.

Note that $\forall i, k, t$ we have

$$\begin{aligned}
&\{\varphi_k^t = i^*, \widehat{\mu}_{i^*} \neq \max\{\widehat{\mu}_i^k(t), \cdots, \widehat{\mu}_N^k(t)\}\} \subseteq \\
&\{\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\} \\
&\cup \{\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \exists i, s.t. (\widehat{\mu}_i^k(t) \geq \mu_{i^*} - C_{i^*}^k(t)\}.
\end{aligned}$$

Thus we have

$$\sum_{i=1}^{T} \mathbb{P}\left(\varphi_k^t = i^*, \widehat{\mu}_{i^*} \neq \max\{\widehat{\mu}_i^k(t), \cdots, \widehat{\mu}_N^k(t)\}\right)$$
$$\leq \sum_{i=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\right) +$$
$$\sum_{i=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \exists i, s.t.(\widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t)\right).$$

From Lemma 2 we have

$$\sum_{i=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\right) \leq \frac{1}{\log \zeta}(1 + \log(d_k + 1))$$
$$+ \frac{1}{2^{\xi} \log \zeta}\left(\frac{\log(d_k+1)}{\xi} + \frac{2}{\xi - 1}\right)$$
$$+ \frac{1}{T^{\xi-1} \log \zeta}\left(\frac{\log(d_k+1)}{T\xi} + \frac{1}{\xi - 1}\right). \tag{8.12}$$

Theorem 9. *For all suboptimal options $i \neq i^*$ we have*

$$\sum_{i=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \exists i, s.t.(\widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t)\right) \leq$$
$$\sum_{i=1}^{N} \frac{8\sigma_i(\xi+1)}{\Delta_i^2} \log T + \frac{N-1}{2^{\xi} \log \zeta}\left(\frac{\log(d_k+1)}{\xi} + \frac{2}{\xi-1}\right)$$
$$+ \frac{N-1}{\log \zeta}(1 + \log(d_k+1))$$
$$+ \frac{N-1}{T^{\xi-1} \log \zeta}\left(\frac{\log(d_k+1)}{T\xi} + \frac{1}{\xi-1}\right).$$

The proof of Theorem 9 is given in the Appendix.

Now we proceed to state the main result of this paper, which is that the total expected cumulative observation regret until time T for agent k employing the sampling rule given by Definition 4 and the observation rule given by Definition 5 is upper bounded logarithmically in T.

Theorem 10. *Expected cumulative observation regret until time T for agent k can be upper bounded as*

$$\begin{aligned}\mathbb{E}\left(R_o^k(T)\right) \leq & \sum_{i=1}^{N} \frac{8\sigma_i(\xi+1)}{\Delta_i^2}\log T \\ & + \frac{cd_k(2N-1)}{\log\zeta}(1+\log(d_k+1)) \\ & + \frac{cd_k(2N-1)}{2^{\xi}\log\zeta}\left(\frac{\log(d_k+1)}{\xi}+\frac{2}{\xi-1}\right) \\ & + \frac{cd_k(2N-1)}{T^{\xi-1}\log\zeta}\left(\frac{\log(d_k+1)}{T\xi}+\frac{1}{\xi-1}\right).\end{aligned}$$

Theorem 10 follows from equations (8.10)-(8.12), Lemma 3 and Theorem 9.

Remark 7. *Note that for deterministic communication strategies [47, 48] the expected cumulative observation regret until time T for agent k is linear in T:*

$$\mathbb{E}\left(R_o^k(T)\right) = c\sum_{t=1}^{T}\sum_{j=1}^{K}\mathbb{E}\left(\mathbb{I}_{\{k,j\}}^t\right) = cd_kT.$$

For the probabilistic observation strategy of [61] the expected cumulative observation regret until time T for agent k is linear in T:

$$\mathbb{E}\left(R_o^k(T)\right) = c\sum_{t=1}^{T}\sum_{j=1}^{K}\mathbb{E}\left(\mathbb{I}_{\{k,j\}}^t\right) = cd_kp_kT,$$

where p_k is the observation probability of agent k. Thus, our proposed sampling rule and observation rule outperform these strategies when there are cumulative observation costs.

8.3.3 Total expected cumulative regret

Total expected cumulative regret $\mathbb{E}\left(R^k(T)\right)$ is defined as the summation of expected cumulative sampling regret and expected cumulative observation regret until time T:

$$\mathbb{E}\left(R^k(T)\right) = \sum_{i=1}^{N} \mathbb{E}\left(R_i^k(T)\right) + \mathbb{E}\left(R_o^k(T)\right).$$

Let $\sum_{i=1}^{N} \Delta_i = \widetilde{\Delta}$. Total expected cumulative regret until time T of agent k is upper bounded as

$$\begin{aligned}
\mathbb{E}\left(R_s^k(T)\right) \leq & \sum_{\substack{i=1 \\ i \neq i^*}}^{N} \frac{8\sigma_i^2(\xi+1)}{\Delta_i^2} \log T \\
& \frac{\widetilde{\Delta} + cd_k(2N-1)}{\log \zeta}(1 + \log(d_k+1)) \\
& + \frac{\widetilde{\Delta} + cd_k(2N-1)}{2^{\xi} \log \zeta} \left(\frac{\log(d_k+1)}{\xi} + \frac{2}{\xi - 1}\right) \\
& + \frac{\widetilde{\Delta} + cd_k(2N-1)}{T^{\xi-1} \log \zeta} \left(\frac{\log(d_k+1)}{T\xi} + \frac{1}{\xi - 1}\right).
\end{aligned} \tag{8.13}$$

8.4 Simulation Results

In this section we present numerical simulation results for a network of 6 agents with underlying observation structure defined by the star graph: the center agent observes all other agents and all other agents only observe the center agent. Agents other than the center agent are interchangeable and their average regret and individual regret are the same. We present numerical simulations to evaluate the performance of the sampling rule and observation rule given by Definitions 4 and 5.

The 6 agents play the same MAB problem with 10 options. In all simulations the reward distributions are Gaussian with variance $\sigma_i = 5$, $i = 1, \ldots, 10$, and mean values:

i	1	2	3	4	5	6	7	8	9	10
μ_i	40	50	50	60	70	70	80	90	92	95

.

The communication cost $c = 1$. We set the sampling rule parameter $\xi = 1.01$. We provide results for 1000 time steps with 1000 Monte Carlo simulations.

Figure 8.1 shows simulation results for the expected cumulative sampling regret of a group of 6 agents using the proposed sampling and observation rules. The blue dashed line shows regret of the center agent. The green dash-dot line shows the average regret of the agents not in the center. The red dotted line shows the average expected cumulative sampling regret over all agents. It can be observed that the expected cumulative sampling regret is logarithmic in time. For comparison, we plot the average expected cumulative regret of the agents when they make no observations of neighbors (solid gold line). When agents are not making observations they are interchangeable, and so the average performance and the individual performance are the same. The simulation results illustrate that the performance of every agent improves significantly when it observes neighbors according to the proposed protocol. The simulation results further show that the center agent outperforms the other agents. This is to be expected since the center agent has more neighbors than the other agents.

Figure 8.2 shows simulation results for expected observation regret. It can be seen that the expected observation regret is logarithmic in time, as proved in Theorem 10. Since the center agent has more neighbors than the others agents, its observation regret is the highest. However, the results illustrate that when observation cost is small, a significant performance improvement can be obtained for a small observation regret.

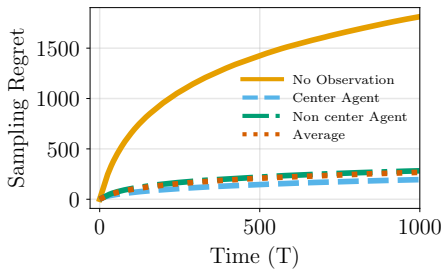


Figure 8.1: Dashed and dotted lines show expected cumulative sampling regret of the agents using the sampling rule and observation rule of Definitions 4 and 5 with underlying star observation structure. The solid line shows the average performance of agents when they are not observing their neighbors.

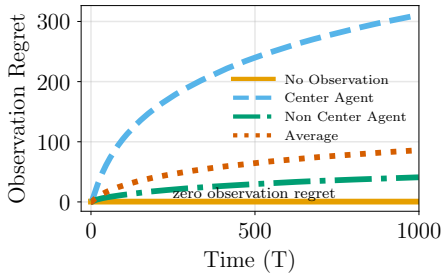


Figure 8.2: Dashed and dotted lines show expected cumulative observation regret of the agents using the sampling rule and observation rule of Definitions 4 and 5 with underlying star observation structure. The solid line shows that agents do not suffer from any observation regret when they do not observe their neighbors.

8.5 Conclusions

We studied an MAMAB problem where agents can observe the instantaneous choices and rewards of their neighbors but incur a cumulative cost each time they make an observation of a neighbor. We proposed a sampling rule and an observation rule in which an agent observes its neighbors only when it has decided to explore. We defined total expected cumulative regret to be the regret agents receive due to sampling sub-optimal options and to observing neighbors. Deterministic and stochastic observation strategies for MAB protocols in the literature yield an expected cumulative observation regret that is linear in time T. We analytically proved that under the proposed sampling and observation rules, expected cumulative regret of each agent is bounded

logarithmically in T. Accuracy of the upper bound has been verified computationally through numerical simulations.

8.6 Appendix

Proof. Note that $\forall i, k, t$ we have

$$\mathbb{P}\left(\varphi_k^t = i, \widehat{\mu}_i^k(t) \neq \max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\}, i \neq i^*\right) \leq \mathbb{E}\left(\mathbb{I}_{\{\varphi_k^t = i\}}\right).$$

Then we have

$$\sum_{t=1}^{T} \mathbb{P}\left(\varphi_k^t = i, \widehat{\mu}_i^k(t) \neq \max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\}, i \neq i^*\right) \leq \sum_{t=1}^{T}\sum_{i=1}^{N} \mathbb{E}\left(\mathbb{I}_{\{\varphi_k^t = i\}}\right).$$

Lemma 3 follows from equation (8.8). □

Proof. Let i be a suboptimal option with highest estimated expected reward for agents k at time t. Then we have $i = \arg\max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\}$ and $i \neq i^*$. If the agent k chooses option i^* at time step $t+1$ we have $Q_{i^*}^k(t) > Q_i^k(t)$. Thus we have $\widehat{\mu}_i^k(t) > \widehat{\mu}_{i^*}^k(t)$ and $C_i^k(t) < C_{i^*}^k(t)$.

Note that for some $\beta_i^k(t) > 0$ we have

$$\mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t)\right) = \beta_i^k(t) + \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t), N_{i^*}^k(t) \geq \beta_i^k(t)\right).$$

Let $\beta_i^k(t) = \frac{8\sigma_i(\xi+1)}{\Delta_i^2}\log t$. Then we have

$$\sum_{t=1}^{T}\mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t)\right) = \beta_i^k(T)$$
$$+\sum_{i=1}^{T}\mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t), N_{i^*}^k(t) \geq \beta_i^k(t)\right).$$

Since $C_i^k(t) < C_{i^*}^k(t)$ we have

$$\sum_{i=1}^{T}\mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t), N_{i^*}^k(t) \geq \beta_i^k(t)\right)$$
$$\leq \sum_{t=1}^{T}\mathbb{P}\left(\widehat{\mu}_i^k(t) \geq \mu_i + C_i^k(t)\right)$$
$$\leq \beta_i^k(T) + \frac{1}{2^\xi \log\zeta}\left(\frac{\log(d_k+1)}{\xi} + \frac{2}{\xi-1}\right)$$
$$+\frac{1}{\log\zeta}(1+\log(d_k+1))$$
$$+\frac{1}{T^{\xi-1}\log\zeta}\left(\frac{\log(d_k+1)}{T\xi} + \frac{1}{\xi-1}\right).$$

Then we have

$$\sum_{i=1}^{T}\mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \exists i, s.t.(\widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t)\right) \leq$$
$$\sum_{i=1}^{N}\frac{8\sigma_i(\xi+1)}{\Delta_i^2}\log T + \frac{N-1}{2^\xi\log\zeta}\left(\frac{\log(d_k+1)}{\xi} + \frac{2}{\xi-1}\right)$$
$$+\frac{N-1}{\log\zeta}(1+\log(d_k+1))$$
$$+\frac{N-1}{T^{\xi-1}\log\zeta}\left(\frac{\log(d_k+1)}{T\xi} + \frac{1}{\xi-1}\right).$$

□

Chapter 9

When to Call Your Neighbor? Strategic Communication in Cooperative Stochastic Bandits

Udari Madhushani and Naomi Ehrich Leonard

In cooperative bandits, a framework that captures essential features of collective sequential decision making, agents can minimize group regret, and thereby improve performance, by leveraging shared information. However, sharing information can be costly, which motivates developing policies that minimize group regret while also reducing the number of messages communicated by agents. Existing cooperative bandit algorithms obtain optimal performance when agents share information with their neighbors at *every time step*, i.e., full communication. This requires $\Theta(T)$ number of messages, where T is the time horizon of the decision making process. We propose *ComEx*, a novel cost-effective communication protocol in which the group achieves the same order of performance as full communication while communicating only $\mathcal{O}(\log T)$ number of messages. Our key step is developing a method to identify and only communicate the information crucial to achieving optimal performance. Further we pro-

pose novel algorithms for several benchmark cooperative bandit frameworks and show that our algorithms obtain *state-of-the-art* performance while consistently incurring a significantly smaller communication cost than existing algorithms.

9.1 Introduction

Sequential decision making in uncertain environments has been extensively studied over the past several decades due to its wide range of real world applications including recommender systems, user-targeted online advertising [89], clinical trials [22] and target searching (e.g. finding nuclear or a temperature source) in robotics. Making optimal decisions under uncertainty requires striking a balance between exploring the environment to identify better decisions and exploiting the decisions that are already known to produce higher outcomes. In collective decision making, i.e., a group of agents making sequential decisions, performance can be greatly improved through cooperative communication by sharing information about the environment. However, often times communication is time consuming and expensive. For example, consider a recommender systems, in which multiple servers networked to handle high demands. In this case high communication between servers can lead to service latency. Similarly, for a group of robots, communication can increase battery power consumption. Thus the cost associated with communication makes it desirable to reduce the amount of shared information. Motivated by this we ask:

Can we minimize communication without sacrificing performance in sequential decision making?

A crucial step in answering this question is, identifying which information is most valuable. We study this problem in bandit framework, which models sequential decision making in uncertain environments [44]. In stochastic bandits, an agent repeatedly pulls an arm from a given set of arms and receives a reward drawn from the

probability distribution associated with the arm. The goal is maximizing cumulative reward. In an uncertain environment, the agent is required to execute a combination of *exploiting actions*, i.e., pulling the arms that are known to provide high rewards, and *exploring actions*, i.e., pulling lesser known arms in order to identify arms that might potentially provide higher rewards [7]. In cooperative bandits a group of agents are faced with the same bandit problem and the goal is maximizing cumulative group reward [47]. Agents can obtain optimal performance by sharing all information they obtained about the arms, i.e., full communication. Thus more specifically we ask how we can minimize communication while obtaining same level of performance as full communication?

In cooperative bandits it is most useful for agents to obtain information about suboptimal arms. Each agent can reduce the number of pulls drawn from suboptimal arms by leveraging communication to reduce the uncertainty associated with the estimates of suboptimal arms. Any efficient stochastic bandit algorithm pulls suboptimal arms logarithmically in time. Thus, when communication is costly, it is desirable to communicate reward values received from suboptimal arms only. Thus our problem effectively reduces to identifying when is it more likely to pull a suboptimal arm?

We solve this problem by proposing ComEx, a new communication protocol, in which agents only communicate the rewards they receive from exploring actions. This is because exploring actions, typically lead to pulling suboptimal arms. Combining ComEx with a cooperative Upper Confidence Bound (UCB) sampling rule [42], we prove that ComEx obtains the same order of performance as full communication, while incurring a significantly smaller communication cost than full communication. We analyze performance of the algorithm using expected group cumulative regret, which is defined as the total expected loss suffered by agents due to pulling suboptimal arms. Measuring the communication cost by the number of messages shared by

agents, we prove that with ComEx agents only suffer a $\mathcal{O}(\log T)$ cost while with full communication they suffer a $\Theta(T)$ cost.

We show that ComEx can be incorporated in a wide range of cooperative bandit algorithms to obtain same order of performance as full communication for a significantly smaller communication cost than full communication. Incorporating ComEx, we propose novel algorithms for bench mark cooperative bandit frameworks: decentralized bandits with 1.) instantaneous rewards sharing, 2.) message passing, 3.) estimate sharing and centralized bandits with 4.) instantaneous rewards sharing 5.) message passing. We propose another algorithm by combining ComEx with message passing and Thompson sampling. Further we provide results illustrating that our algorithms obtain *state-of-the-art* performance while consistently incurring a significantly smaller communication cost than existing algorithms in these benchmark frameworks.

Key contributions. We make following key contributions in this work:

- We propose ComEx, a novel and cost-effective communication protocol for cooperative bandits.
- We provide theoretical guarantees that ComEx obtains the same order group regret as full communication while incurring a $\mathcal{O}(\log T)$ communication cost. In contrast, full communication incurs a $\Theta(T)$ communication cost.
- Incorporating ComEx, we propose novel algorithms in several benchmark cooperative bandit frameworks. We provide both theoretical guarantees and experimental results validating *state-of-the-art* performance of our proposed algorithms.

9.2 Related work

Decentralized reward sharing. In decentralized reward sharing agents share instantaneous rewards with their neighbors [14, 42, 62, 59, 62, 94]. The paper [42] considered that neighbors are defined according to a fixed communication graph and provide graph structure dependent regret bounds. The paper [14, 61, 63, 65] studied the cooperative bandit problem with time varying communication structures. The papers [13, 10, 19] considered message passing communication rules where each agent initiates a message and send the message to its neighbors. A message received from a neighbor is subsequently forwarded to other neighbors.

Decentralized estimate sharing. In estimate sharing each agent share the estimated average reward and number of arm pulls from each arm with its neighbors defined according to a fixed communication graph. The paper [84] considered a P2P communication where an agent is only allowed to communicate with two other agents at each time step. The papers [47, 46, 66, 49] used a running consensus algorithm to update estimates and provide graph-structure-dependent performance.

Centralized leader-follower setting. A communication strategy where agents observe the rewards and choices of their neighbors according to a leader-follower setting is considered in [48, 42, 94]. In [48, 42], followers pull the last arm pulled by their neighbors. In [94] one leader explores and estimates the mean reward of arms, while all other agents pull the arm with highest estimated mean per the leader.

Communication cost. The paper [86] considered a pure exploration bandit problem and measures the communication by the number of times agents communicate. The paper [63] proposed a communication protocol where agents observe their neighbors when they have high uncertainty about arms. [94] proposed a leader-follower algorithm with a constant communication cost. The paper [95] proposed an algorithm

that achieves near-optimal performance where agents achieve sublinear expected regret. In their work, communication cost is independent of time and measured by the amount of data transmitted.

Distributed Thompson sampling. Recently [92, 45] proposed distributed Thompson sampling rules. The paper [92] studied the problem with sparse communication structures. The paper [45] provided regret guarantees that matches the corresponding centralized regret guarantees.

9.3 ComEx: Communicate When Exploring

In this section we provide mathematical formulation and and intuition of our communication protocol.

Notations. For any positive integer M we denote the set $\{1, 2, \ldots, M\}$ as $[M]$. We define $\mathbf{1}\{x\}$ as an indicator variable that takes value 1 if x is true and 0 otherwise. Further, we use $X \backslash x$ to denote the set X excluding the element x. We use $|X|$ to denote the number of elements in set X. For any general graph G we define $\bar{\chi}(G), \bar{\gamma}(G)$ as clique covering number and dominating number respectively. We use G_γ to denote the γ^{th} power graph of G. Let $g(M, x) = M + \sum_{i=1}^{N} \left(12 \log(3(x+1)) + 3 \log\left(x+1\right)\right)$.

Cooperative stochastic bandits. We consider the cooperative bandit problem with K arms and N agents. Reward distributions of each arm $k \in [K]$ is assumed to be sub-Gaussian with mean μ_k and variance proxy σ_k^2. At each time step $t \in [T]$ each agent $i \in [N]$ pulls an arm $A_t^{(i)}$ and receives a numerical reward $X_t^{(i)}$ drawn from the probability distribution associated with the pulled arm. Without loss of generality we assume that $\mu_1 \geq \mu_2 \ldots \geq \mu_K$ and define $\Delta_k := \mu_1 - \mu_k, \forall k > 1$ to be the expected reward gap between optimal arm, i.e., the arm with highest mean reward, and arm k.

Let $\bar{\Delta} := \min_{k \neq 1, k \in [K]} \Delta_k$ be the minimum expected reward gap. We make following assumptions.

Assumptions:

(**A1**) When more than one agent pulls the same arm at the same time they receive rewards independently drawn from the probability distribution associated with the pulled arm.

(**A2**) All the agents know $\sigma^2 \geq \sigma_k^2, \forall k$, an upper bound of the variance proxy associated with arms.

Communication over a general graph. Let $G(V, E)$ be a general graph that encodes the hard communication constraints among agents. The vertex set V is the set of agents $[N]$ and each edge $(i, j) \in E$ indicates that agents i and j are neighbors. We consider that agents directly communicate with their neighbors only. Let $\mathbf{1}\{(i, i) \in E\} = 1, \forall i$. At each time step t we define the communication between agents by $G_t(V, E_t)$ where $E_t \subseteq E$. Let $d^{(i)}$ be the degree of agent i. Let G_γ denote the γ^{th} power graph of G. Denote $d_\gamma^{(i)}$ to be the degree of agent i in graph G_γ, i.e., number of agents within a distance of γ from agent i in graph G. For any γ let $d_\gamma^{(i)^+} = d_\gamma^{(i)} + 1$.

We denote $\mathbf{m}_t^{(i)}$ as the message shared by agent i at time t with its neighbors. This can be either a single message containing information about a particular arm pull, typically the last arm pull of agent i, or a concatenation of information about several arm pulls by more than one agent over several previous time steps. We define $n_k^{(i)}(t) := \sum_{\tau=1}^{t} \mathbf{1}\Big\{A_\tau^{(i)} = k\Big\}$ and $N_k^{(i)}(t) := \sum_{\tau=1}^{t} \sum_{j=1}^{N} \mathbf{1}\Big\{A_\tau^{(j)} = k\Big\}\mathbf{1}\{(i, j) \in E_\tau\}$ to be the number of times until time step t that agent i pulled arm k and observed reward values from arm k, respectively. Note that the number of observations $N_k^{(i)}(t)$ is the sum of the number of pulls drawn by agent i of arm k and the number of times agent i received reward values of arm k from its neighbors. Let $\widehat{\mu}_k^{(i)}(t)$ denote agent i's estimated average reward of arm k at time t.

Regret and communication cost. Following the convention we define regret as the loss suffered by agents due to pulling suboptimal arms. Let $R(t)$ be the cumulative group regret at time t. Then the expected cumulative group regret can be given as $\mathbb{E}\left[R(t)\right] := \sum_{i=1}^{N} \sum_{k=2}^{K} \Delta_k \mathbb{E}[n_k^{(i)}(t)]$. We define the communication cost as the number of messages shared by agents. We consider the cost of sharing a concatenated message to be the number of single messages included in it. Let $L(t)$ be the cumulative group communication cost at time t. Then, the expected group communication cost can be given as $\mathbb{E}[L(t)] := \sum_{i=1}^{N} \sum_{\tau=1}^{t} \mathbb{E}\left[\left|\mathbf{m}_\tau^{(i)}\right|\right]$.

Proposed communication protocol: ComEx. We propose ComEx, a cost-effective partial communication protocol that obtains same order of performance as full communication.

Algorithm 2: ComEx

Input: Bandit environment, algorithm parameters

for each iteration $t \in [T]$ **do**

  for each agent $i \in [N]$ **do**

    `// Sampling phase`

    Sampling rules: Cooperative UCB, Cooperative Thompson

    `// Message generating phase`

    `// Replace full communication with ComEx`

    if $A_t^{(i)} \neq \arg\max_k \widehat{\mu}_k^{(i)}(t-1)$ **then**

      CREATE $\left(m_t^{(i)} := \left\langle i, t, A_t^{(i)}, X_t^{(i)} \right\rangle\right)$

    end

  end

  for each agent $i \in [N]$ **do**

    `// Communication phase`

    Communication rule: Decentralized (or centralized) instantaneous reward sharing, Decentralized (or centralized) message passing

    `// Estimate updating phase`

  end

  for each arm $k \in [K]$ **do**

    CALCULATE $\left(\widehat{\mu}_k^{(i)}(t), N_k^{(i)}(t)\right)$

  end

end

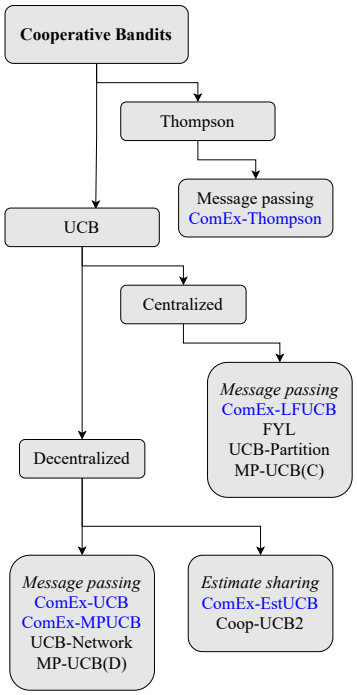


Figure 9.1: A summary of our proposed algorithms and existing state-of-the-art algorithms for different cooperative bandit frameworks.

As motivated above, information about suboptimal arms is most valuable to agents seeking to maximize expected cumulative reward. This is because, with information

from neighbors on a suboptimal arm, an agent can obtain a sufficiently accurate estimate of the expected reward of the suboptimal arm without having to pull the arm by itself. Agents typically pull suboptimal arms when they are exploring. Thus, to provide the means to maintain high performance with low communication costs, we propose a new communication protocol as follows in which agents only share information they obtained through exploring.

Definition 6. (ComEx communication protocol) *Each agent i initiates sharing the message* $m_t^{(i)} := \left\langle i, t, A_t^{(i)}, X_t^{(i)} \right\rangle$ *if* $A_t^{(i)} \neq \arg\max_{k\in[K]} \widehat{\mu}_k^{(i)}(t-1)$

Note that according to the above communication protocol agents initiate sharing messages only about the rewards received from the arms that are instantaneously suboptimal i.e., arm that does not have the maximum estimated expected reward. This maximizes the chance of sharing information about suboptimal arms.

Generalizability of ComEx. As we will demonstrate in next few sections, our communication protocol is an easily implementable general communication protocol that can be incorporated in a wide range of cooperative bandit algorithms. We illustrate the generality by proposing novel algorithms incorporating ComEx in several cooperative bandit frameworks. Figure 9.3 provides a summary of our algorithms and state-of-the-art algorithms in several benchmark cooperative bandit frameworks.

9.4 Decentralized Cooperative Bandits

In this section we propose novel algorithms for decentralized cooperative bandits.

9.4.1 Decentralized instantaneous reward sharing UCB

We present our first algorithm ComEx-UCB by combining the above communication protocol with instantaneous reward sharing. Each agent follows a sampling rule that

balances exploiting with exploring. We use a natural extension of Upper Confidence Bound (UCB) algorithm as a sampling rule. In UCB at each time step t for each arm k each agent i constructs an upper confidence bound, i.e., the sum of its estimated expected reward (empirical average of the observed rewards) and the uncertainty associated with the estimate $C_k^{(i)}(t) := \sigma\sqrt{\frac{2(\xi+1)\log t}{N_k^{(i)}(t)}}$ where $\xi > 1$, and pull the arm with highest bound. If the pulled arm is instantaneously suboptimal, the agent sends a message $m_t^{(i)} := \left\langle A_t^{(i)}, X_t^{(i)} \right\rangle$ to its neighbors (see Definition 6). Note that under this communication rule agents do not share concatenated messages. Thus passing information about time step and agent id is redundant. Pseudo code for ComEx-UCB is given in Appendix 9.10.12.

Theorem 11. (Group regret of ComEx-UCB) *Consider a group of N agents following ComEx-UCB while sharing instantaneous rewards over a general communication graph G. Then for any $\xi \geq 1.1$ expected cumulative group regret satisfies:*

$$\mathbb{E}\left[R(T)\right] \leq \sum_{k=2}^{K} \frac{8(\xi+1)\sigma}{\Delta_k} \bar{\chi}(G) \log T + \sum_{k=2}^{K} \Delta_k g(4N, d^{(i)})$$

Proof sketch. We follow an approach similar to the standard UCB analysis [7, 19] with a few key modifications. We partition the communication graph into a set of non overlapping cliques and analyze the regret of each clique and take the summation over cliques to obtain the regret of the group. When agents are using full communication group regret can be given as the summation of a $\log T$ term that scales with the clique covering number $\bar{\chi}(G)$ and a term, which is independent of T. The second term depends on the summation of tail probabilities of arms, i.e., $\mathsf{P}\left(\left|\widehat{\mu}_k^{(i)}(t) - \mu_k\right| \geq C_k^{(i)}(t)\right)$. For full communication a similar result can be found in [19]. Note that full communication is a deterministic communication protocol and ComEx-UCB is a stochastic communication protocol that depends on the decision making process. Two major technical challenges in proving the regret bound for

ComEx-UCB are 1.) deriving a tail probability bound for the case in which the communication between agents are stochastic and 2.) bounding the additional regret incurred by not sharing information when pulling the arm with highest estimated average reward, i.e., $A_t^{(i)} = \arg\max_{k\in[K]} \widehat{\mu}_k^{(i)}(t-1)$. We overcome the first challenge by noticing that communication random variables $\mathbf{1}\{(i,j) \in E_t\}, \forall i,j,t$ are previsible, i.e., measurable with respect to the sigma algebra generated by information obtained up to time $t-1$. We address the second challenge by proving that the number of times agents do not share information about any suboptimal arm k can be bounded by tail probabilities of arm k and the optimal arm. A complete proof of Theorem 17 is given in Appendix **??**. □

Remark 8. *By replacing ComEx with full communication in ComEx-UCB algorithm agents obtain an expected cumulative group regret of* $\mathbb{E}[R(T)] = O\left(K\bar{\chi}(G)\log T + KN\right)$ *(Appendix H). Thus from Theorem 17 we see that ComEx obtains the same order of performance as full communication.*

Recall that expected communication cost under full communication is $\Theta(T)$. Now we prove that expected communication cost under ComEx is logarithmic in time. In ComEx-UCB algorithm agents are only sending single messages (not concatenated). Thus expected group communication cost at time step t can be given as $\mathbb{E}[L(t)] = \sum_{i=1}^{N}\sum_{\tau=1}^{T} \mathsf{P}\left(A_\tau^{(i)} \neq \arg\max_{k\in[K]} \widehat{\mu}_k^{(i)}(\tau-1)\right)$.

Theorem 12. (Communication cost of ComEx-UCB) *Consider a group of N agents following ComEx-UCB while sharing instantaneous rewards over a general communication graph G. Then for any $\xi \geq 1.1$ expected group communication cost satisfies:*

$$\mathbb{E}[L(T)] \leq 8\sigma(\xi+1)\left[\frac{N}{\bar{\Delta}^2} + \sum_{k=2}^{K}\frac{\bar{\chi}(G)}{\Delta_k^2}\right]\log T + Kg\left(7N, d^{(i)}\right)$$

Proof sketch. Note that expected group communication cost is the sum of 1.) expected number of times agents pull any suboptimal arm when it is instantaneously

suboptimal and 2.) expected number of times agents pull the optimal arm when it is instantaneously suboptimal. We note that the first term can be directly bounded by the expected number of times agents pull suboptimal arms. We prove that the second term can be bounded logarithmically in time. A detailed proof of Theorem 12 is given in Appendix **??**. □

9.4.2 Decentralized message passing UCB

We propose ComEx-MPUCB an improved version of ComEx-UCB by incorporating a message passing method [**?**, 10, 19] that allows agents to share the messages they initiated with agents who are within a distance of γ. We call γ *communication density parameter*. We consider that at time t each agent i initiates a message $m_t^{(i)} := \left\langle i, t, A_t^{(i)}, X_t^{(i)} \right\rangle$ according to ComEx given in Definition 6 and sends the messages to its neighbors. Subsequently the agents who receive the message forward it to their neighbors. Messages received at time t are forwarded to neighbors at time $t+1$ resulting that each hop adds a delay of 1 time step. Under this message passing method γ-hop neighbors receive the message after a delay of γ time steps. Agents do not forward the messages that are older than $\gamma - 1$ and discard the messages that are older than γ. Note that for a connected graph maximum number of time step required to pass a message between any two agents equals to the diameter of the graph. Thus we choose γ to be an integer constant which is at most diameter of the communication graph G. The pseudo code for ComEx-MPUCB is given in Appendix 9.11.

Theorem 13. (Group regret of ComEx-MPUCB) *Consider a group of N agents following ComEx-MPUCB. Then for any $\xi \geq 1.1$ expected cumulative group regret*

satisfies:

$$\mathbb{E}\left[R(T)\right] \leq \sum_{k=2}^{K} \frac{8(\xi+1)\sigma}{\Delta_k} \bar{\chi}(G_\gamma) \log T + \sum_{k=2}^{K} \Delta_k \left[(N - \mathcal{X}(G_\gamma))(\gamma - 1) + g\left(4N, d_\gamma^{(i)}\right)\right]$$

Proof sketch. We see that regret under ComEx-MPUCB can be given as the summation of regret of ComEx-UCB when communication graph is G_γ and the regret incurred by the delay in passing messages to agents who are not 1-hop neighbors. We prove that the expected regret due to delay is at most $(N - \bar{\chi}(G_\gamma))(\gamma - 1)$. A detailed proof is provided in Appendix 9.10.3. □

Remark 9. *Similar to ComEx-UCB by replacing ComEx with full communication in ComEx-MPUCB algorithm agents obtain an expected cumulative group regret of* $\mathbb{E}\left[R(T)\right] = O\left(K\bar{\chi}(G_\gamma) \log T + KN\right)$ *(Appendix H). Thus from Theorem 18 we see that ComEx obtains the same order of performance as full communication.*

Now we proceed to prove that expected group communication cost under ComEx-MPUCB is logarithmic in time.

Theorem 14. (Communication cost of ComEx-MPUCB) *Consider a group of N agents following ComEx-MPUCB with communication density parameter γ. Then for any $\xi \geq 1.1$ expected group communication cost satisfies:*

$$\mathbb{E}\left[L(T)\right] \leq \left[8(\xi+1)\sigma \left[\frac{N}{\bar{\Delta}^2} + \sum_{k=2}^{K} \frac{\bar{\chi}(G_\gamma)}{\Delta_k^2}\right] \log T + K\left[(N - \bar{\chi}(G_\gamma)(\gamma - 1)\right]\right] \sum_{i=1}^{N} d_{\gamma-1}^{(i)^+}$$
$$+ K \sum_{i=1}^{N} d_{\gamma-1}^{(i)^+} \cdot g\left(7N, d_\gamma^{(i)}\right)$$

Proof sketch. Note that under ComEx-MPUCB agents send concatenated messages to their neighbors. Recall that agents do not forward the messages that are older than $\gamma - 1$. Thus each message initiated by agent i is subsequently forwarded by all agents who are within distance of $\gamma - 1$ in graph G. Thus we have

$\mathbb{E}[L(t)] \leq \sum_{i=1}^{N} d_{\gamma-1}^{(i)^+} \sum_{\tau=1}^{t} \mathsf{P}\left(A_\tau^{(i)} \neq \arg\max_{k\in[K]} \widehat{\mu}_k^{(i)}(\tau-1)\right)$. A detailed proof can be found in Appendix 9.10.4. □

9.5 Centralized Cooperative Bandits

We propose ComEx-LFUCB by combining ComeEx communication protocol with a leader-follower method [42, 48, 19, 94]. ComEx-LFUCB provides better performance compared to its decentralized counter part ComEx-MPUCB. Let V'_γ be the set of vertices in minimal dominating set of graph G_γ. We consider each agent $i \in V'_\gamma$ to be a leader and all the other agents to be followers. Note that every follower has at least one leader as a neighbor. We consider that each leader uses ComEx-MPUCB and each follower copies the last action observed from its leader. For each follower j a leader i is assigned such that $d(i,j) = \min_{i'} d(i',j)$ where $d(i,j)$ is the distance between agent i and agent j in graph G. Let $\mathcal{N}_\gamma^i$ be the set of follower of leader i. We consider that each leader sends a message containing the id of the arm it pulls and whether it is instantaneously suboptimal, i.e. for $i \in V'_\gamma$ at time step t, $m_t^{(i)} := \left\langle i, t, A_t^{(i)}, \mathbf{1}\left\{A_i^{(t)} \neq \arg\max_{k\in[K]} \widehat{\mu}_k^{(i)}(t-1)\right\}\right\rangle$ to its neighbors and they subsequently forward it to their neighbors. Note that at time step t follower $j \in \mathcal{N}_\gamma^{(i)}$ pulls the arm $A_{t-d(i,j)}^{(i)}$. Each follower pass a message containing information about the reward and arm id if it pulls an arm that is specified as instantaneously suboptimal by its leader. Thus the followers communicate according to ComEx by initiating a message as follows. Follower $j \in \mathcal{N}_\gamma^{(i)}$ initiates a message $m_t^{(j)} := \left\langle j, t, A_t^{(j)}, X_t^{(j)}\right\rangle$ if $A_{t-d(i,j)}^{(i)} \neq \arg\max_{k\in[K]} \widehat{\mu}_k^{(i)}(t-d(i,j)-1)$. Accordingly under full communication followers share their rewards and arm pulls at every time step. Pseudo code for ComEx-LFUCB is provided in Appendix 9.12.

Theorem 15. (Group regret of ComEx-LFUCB) *Consider a group of N agents following ComEx-LFUCB with communication density parameter γ. Then for any*

$\xi \geq 1.1$ *expected cumulative group regret satisfies:*

$$\mathbb{E}\left[R(T)\right] \leq \sum_{k=2}^{K} \frac{8(\xi+1)\sigma}{\Delta_k} \bar{\gamma}(G_\gamma) \log T + \sum_{k=2}^{K} \Delta_k \left[(N - \bar{\gamma}(G_\gamma))(3\gamma - 1) + \bar{\gamma}(G_\gamma) \cdot g(4N, d_\gamma^{(i)})\right]$$

Proof sketch. We follow a similar approach to the proof of Theorem 18 with a few key modifications followed by the argument below. Note that number of suboptimal arm pulls by each $j \in \mathcal{N}_\gamma^{(i)}$ can be upper bounded using suboptimal arm pulls by i and message passing delay. Note that message passing delay can be upper bounded by $d(i,j)$. A detailed proof of Theorem 15 is given in Appendix 9.10.5. □

Remark 10. *Similar to ComEx-MPUCB by replacing ComEx with full communication in ComEx-LFUCB algorithm, i.e. allowing followers to share information about arm pulls at every time step, agents obtain an expected cumulative group regret of* $\mathbb{E}\left[R(T)\right] = O\left(K\bar{\gamma}(G_\gamma)\log T + KN\right)$ *(Appendix H). Thus from Theorem 15 we see that ComEx obtains the same order of performance as full communication.*

Now we provide theoretical guarantees that expected group communication cost under ComEx-LFUCB is logarithmically bounded in time.

Theorem 16. (Communication cost of ComEx-LFUCB) *Consider a group of N agents following ComEx-LFUCB with communication density parameter γ. Then for any $\xi \geq 1.1$ expected group communication cost satisfies:*

$$\mathbb{E}\left[L(T)\right] \leq \left[8(\xi+1)\sigma\left[\frac{N}{\bar{\Delta}^2} + \sum_{k=2}^{K}\frac{\bar{\gamma}(G_\gamma)}{\Delta_k^2}\right]\log T + K\left[(N - 3\bar{\gamma}(G_\gamma)(\gamma-1)\right]\right]\sum_{i=1}^{N} d_{\gamma-1}^{(i)^+}$$
$$+K\sum_{i=1}^{N} d_{\gamma-1}^{(i)^+} \cdot \bar{\gamma}(G_\gamma \cdot g\left(7N, d_\gamma^{(i)}\right)$$

Proof sketch. Note that the expected number of times a leader initiates a message can be upper bounded by twice the expected number of its suboptimal arm pulls. Further the number of times each follower $j \in \mathcal{N}_\gamma^{(i)}$ initiates a message can be bounded by the number of instantaneously suboptimal arms pulled by the leader i. Similar

to ComEx-MPUCB in ComEx-LFUCB agents send concatenated messages to their neighbors. Thus each message initiated by any agent i is subsequently forwarded by all agents who are within distance of $\gamma - 1$ in graph G. A detailed proof can be found in Appendix 9.10.6. □

Remark 11. *Algorithm and results provided in this Section can be specialized to centralized cooperative bandits with instantaneous reward sharing by substituting* $\gamma = 1$.

Remark 12. (Upper bound on communication cost) *Although smaller* Δ_k *values lead to larger upper bounds for each algorithm (with communication density* γ*) presented in Section 9.4 and 9.5 communication cost is upper bounded by* $T \sum_{i=1}^{N} d_{\gamma-1}^{(i)^+}$.

9.6 Additional Algorithms

We propose two more algorithms, thus extending ComEx to additional cooperative bandit frameworks. We leave providing theoretical guarantees for these as future work.

Estimate sharing. We propose ComEx-EstUCB by combining ComEx with estimate sharing [46, 66, 49], which obtains better performance than instantaneous reward sharing. In estimate sharing, for each arm k, agents maintain estimated sum of rewards and estimated number of pulls from the arm. At each time step, agents average their estimates with their neighbors according to a consensus protocol and update the estimates by incorporating the information of arm pull at that time step. We refer readers to [49] for more details. In ComEx-EstUCB agents only average estimates of instantaneously sub optimal arms. Pseudo code for ComEx-EstUCB is given in Appendix 9.13.

Thompson sampling. We extend our communication protocol to cooperative Thmpson bandits as follows. We propose ComEx-MPThompson, a new algorithm by replacing UCB sampling rule with Thompson sampling rule in ComEx-MPUCB as follows. We combine ComEx with message passing and a natural extension of Thompson sampling to cooperative bandits. Here we provide a brief description of cooperative Thompson sampling rule and refer readers to [45] for more details. Algorithm is initialized by each agent assigning a suitable prior distribution to each arm. Typically Gaussian priors are used for Gaussian reward distributions and Beta priors are used for Bernoulli distributions. At each time step each agent constructs a posterior distribution for each arm using prior distribution and available reward information at that time step. Each agent draws a sample from posterior distributions associated with each arm and pull the arm with highest sampled value. Agents initialize messages according to ComEx and pass the messages to neighbors using a similar protocol given in ComEx-MPUCB. Pseudo code for ComEx-MPThompson is given in Appendix 9.14.

9.7 Experimental Results

In this section we provide numerical simulations illustrating our results and validating our theoretical claims. All the experiments were run on the first author's personal laptop. We show that ComEx obtains same order of performance, i.e., same order of group regret, as full communication for a significantly smaller communication cost than full communication. We also demonstrate that our algorithms outperform state-of-the-art algorithms in several bandit frameworks.

Experimental setup. We provide simulation results for following cooperative bandit frameworks 1) decentralized instantaneous reward sharing, 2) decentralized message passing, 3) decentralized estimate sharing, 4) centralized leader-follower, and 5)

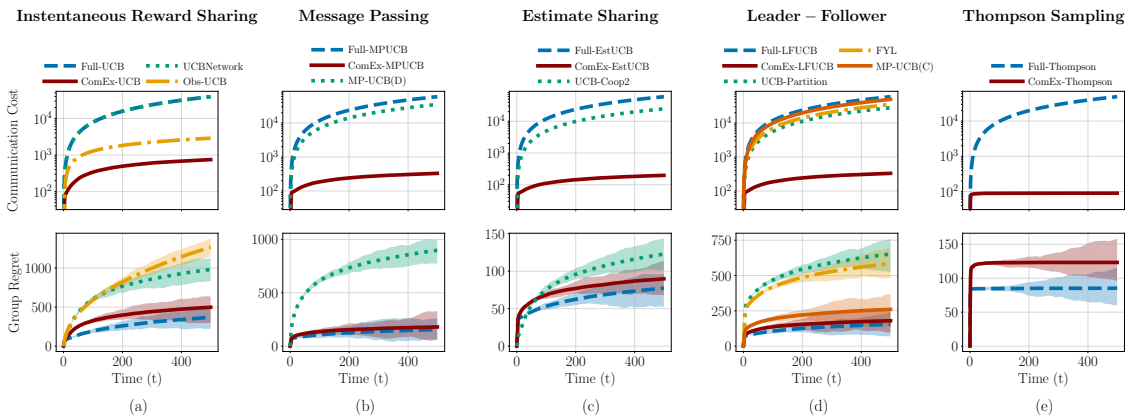


Figure 9.2: A comparison of expected cumulative group regret and communication cost of our algorithms and existing state-of-the-art algorithms in several benchmark cooperative bandit frameworks.

Thompson sampling. We compare performance of our algorithms (ComEx-UCB, ComEx-MPUCB, ComEx-EstUCB, ComEx-LFUCB and ComEx-Thompson) with their corresponding full communication algorithms (Full-UCB, Full-MPUCB, Full-EstUCB, Full-LFUCB and Full-Thompson) and state-of-the art algorithms in each framework. For all simulations presented in this section we consider 10 arms ($K = 10$), 100 agents ($N = 100$) and 500 time steps ($T = 500$). Communication graph between agents is considered to be a Erdos Renyi random graph with edge probability 0.7. Results are averaged over 100 Monte Carlo simulations. Additional experimental results for different graph structures and parameters (ξ, γ) are provided in Appendix 9.10.11.

Hyper parameters We use tuning parameter $\xi = 1.01$ for UCB based algorithms. For results provided in Figure 10.3(b)-10.3(e) we use communication density parameter $\gamma = 5$. None of the competing algorithms, except UCB-Coop2, MP-UCB(D) and MP-UCB(C) have hyperparameters. We tuned parameters of UCB-Coop2 to get best results of that algorithm and used $\kappa = 0.02, \gamma' = 1.001, \eta = 0.001$ (Equations 9 and 15 in [49]. Here we γ' to avoid confusing with communication parameter γ used in this paper) for final results. Decreasing γ' below 1.001 and η below 0.001 did not offer any

significant improvement. MP-UCB(D) and MP-UCB(C) are originally proposed in [19] for heavy-tailed distributions, and we adapt them to sub-Gaussian distributions as directed by the authors. For MP-UCB(D) and MP-UCB(C) we considered the same $C_k^{(i)}(t)$ as in our algorithms. Thus we used the same $\xi = 1.01$ value for a fair comparison.

For results provided in Figures 10.3(a) and 10.3(d), we consider reward distributions to be bounded $[0, 1]$. We consider triangle distributions with mod 1 for the optimal arm and mod 0 for all sub-optimal arms. In simulations provided in Figures 10.3(b), 10.3(c) and 10.3(e) we consider Gaussian reward distributions. Expected reward for the optimal arm is $\mu_1 = 11$ and for all sub-optimal arms $k > 1$ is $\mu_k = 10$. We let variance associated with all arms be $\sigma_k^2 = 1, \forall k$. We use the notation Obs-UCB to denote the algorithm presented in [63].

ComEx obtains same order of performance as full communication. Our results in Figure 10.3 illustrate that ComEx obtains the same order of performance, i.e., same order of group regret, as full communication. From Comparing Figures 10.3(a) and 10.3(b) we see that performance difference between full communication and ComEx decrease when communication density γ increase. Comparing Figure 10.3(e) with others we see that performance difference between full communication and ComEx is smaller when agents are using UCB based sampling rules and Thompson based sampling rules. All results illustrate that our algorithms consistently out preforms state-of-the-art algorithms in all five benchmark cooperative bandit frameworks.

ComEx only incurs a logarithmic communication cost. Our simulation results also illustrate that ComEx only incurs a logarithmic communication cost. In Figure 10.3(a) we observe that Obs-UCB also incurs a logarithmic cost. However ComEx-UCB incurs a smaller cost than Obs-UCB while suffering a smaller group re-

gret. Further, results illustrate that ComEx enabled algorithms incurs a significantly smaller communication cost compared to existing state-of-the-art algorithms.

Additional discussion. State-of-the-art algorithm for leader-follower setting is DPE2 in [94]. DPE2 uses a phased communication protocol, where during the leader selection phase, which lasts at least $2D$ rounds, where D is the diameter of the graph, agents do not pull arms. Thus, this phase accumulates an expected group regret of at least $2DN\mu_1$. In our experimental setup, this alone exceeds the regret accumulated by our algorithms during the entire time horizon. So a meaningful comparison cannot be provided without modifying DPE2 to allow pulling arms during the leader selection phase.

9.8 Discussion

Limitations. Main limitation of this work is that all the theoretical claims are provided using upper bounds. Obtaining lower bounds for cooperative bandits that communicate over general graphs are difficult due to the complex nature of the probability distribution associated with the sampling process of agents. This is an active area of research. We provide a discussion in Appendix B for the optimality of our regret bounds by providing a lower bound when G is a complete graph.

Future extensions. We plan to analyse regret and communication cost for the algorithms provided in Section 9.6. Our intuition can be extended to the collision setting by not allowing agents to share information about the first N instantaneously optimal arms. In the collision setting when more than one agent pulls the same arm at the same time step a collision occurs. This causes agents to either split the reward or completely loose the reward at that time step. Another extension will be proposing similar algorithms for linear bandits and adversarial bandits.

9.9 Conclusion

We proposed ComEx, a general and effective communication protocol which obtains same order of performance as full communication but incurs significantly smaller communication cost than the latter. Next, we proposed novel algorithms for several benchmark bandit frameworks by incorporating ComEx protocol. We provided theoretical guarantees followed by experimental results illustrating the *state-of-the-art* performance of our algorithms.

9.10 Appendix

9.10.1 Proof of Theorem 17

We begin the proof of Theorem 17 by proving a few useful lemmas.

Lemma 4. (Restatement of results from [7]) *Let* $\eta_k = \left(\frac{8(\xi+1)\sigma^2}{\Delta_k^2}\right)\log T$. *For any suboptimal arm* k *and* $\forall i, t$ *we have*

$$\mathsf{P}\left(A_{t+1}^{(i)} = k, N_k^{(i)}(t) > \eta_k\right) \leq \mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right)$$

Proof. Note that for any $k > 1$ we have

$$\begin{aligned}\left\{A_{t+1}^{(i)} = k\right\} &\subset \left\{Q_k^{(i)}(t) \geq Q_1^{(i)}(t)\right\} \\ &\subset \left\{\left\{\mu_1 < \mu_k + 2C_k^{(i)}(t)\right\} \cup \left\{\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right\} \cup \left\{\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right\}\right\}.\end{aligned}$$

Let $\eta_k = \left(\frac{8(\xi+1)\sigma^2}{\Delta_k^2}\right)\log T$. Since $N_k^{(i)}(t) > \eta_k$ the event $\left\{\mu_1 < \mu_k + 2C_k^{(i)}(t)\right\}$ does not occur. Thus we have

$$\mathsf{P}\left(A_{t+1}^{(i)} = k, N_k^{(i)}(t) > \eta_k\right) \leq \mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right)$$

This concludes the proof of Lemma 13. □

Lemma 5. *Let $\bar{\chi}(G)$ is the clique covering number of graph G. Let $\eta_k = \left(\frac{8(\xi+1)\sigma^2}{\Delta_k^2}\right)\log T$. Then we have*

$$\sum_{i=1}^{N}\mathbb{E}[n_k^{(i)}(T)] \leq \bar{\chi}(G)\eta_k + N + \sum_{i=1}^{N}\sum_{t=1}^{T-1}\left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right)\right]$$

Proof. Let $\mathcal{C}$ be a non overlapping clique covering of G. Note that for each suboptimal arm $k > 1$ we have

$$\sum_{i=1}^{N}\mathbb{E}[n_k^{(i)}(T)] = \sum_{i=1}^{N}\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(i)} = k\right) = \sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(i)} = k\right) \tag{9.1}$$

Let $\tau_{k,\mathcal{C}}$ be the maximum time step such that the total number of pulls from arm k shared by agents in the clique $\mathcal{C}$ is at most η_k. This can be stated as $\tau_{k,\mathcal{C}} := \max\left\{t \in [T] : \sum_{i\in\mathcal{C}}\sum_{\tau=1}^{t}\mathbf{1}\left\{A_\tau^{(i)} = k, A_\tau^{(i)} \neq \arg\max_{l\in[K]}\widehat{\mu}_l^{(i)}(\tau-1)\right\} \leq \eta_k\right\}$. Then for all $i \in \mathcal{C}$ we have $N_k^{(i)}(t) > \eta_k, \forall t > \tau_{k,\mathcal{C}}$. We analyse the expected number of times all agents pull suboptimal arm k as follows.

$$\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathbf{1}\left\{A_t^{(i)} = k\right\} = \sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\mathbf{1}\left\{A_t^{(i)} = k\right\} \tag{9.2}$$

$$+\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{T}\mathbf{1}\left\{A_t^{(i)} = k, N_k^{(i)}(t-1) > \eta_k\right\} \tag{9.3}$$

Taking the expectation of (9.3) we have

$$\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(i)} = k\right) = \sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\mathsf{P}\left(A_t^{(i)} = k\right) \tag{9.4}$$

$$+\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{T}\mathsf{P}\left(A_t^{(i)} = k, N_k^{(i)}(t-1) > \eta_k\right) \tag{9.5}$$

Now we proceed to upper bound the first term of right hand side of (9.3) as follows. Note that we have

$$\begin{aligned}\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\mathbf{1}\Big\{A_t^{(i)}=k\Big\} &= \sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\mathbf{1}\left\{A_t^{(i)}=k, A_t^{(i)}\neq\arg\max_{l\in[K]}\widehat{\mu}_l^{(i)}(t-1)\right\}\\ &\quad+\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\mathbf{1}\left\{A_t^{(i)}=k, A_t^{(i)}=\arg\max_{l\in[K]}\widehat{\mu}_l^{(i)}(t-1)\right\}\\ &\leq \eta_k+\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\mathbf{1}\left\{A_t^{(i)}=k, A_t^{(i)}=\arg\max_{l\in[K]}\widehat{\mu}_l^{(i)}(t-1)\right\} \qquad (9.6)\end{aligned}$$

Taking the expectation of (9.6) we have

$$\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\mathsf{P}\left(A_t^{(i)}=k\right)\leq\eta_k+\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\mathsf{P}\left(A_t^{(i)}=k, A_t^{(i)}=\arg\max_{l\in[K]}\widehat{\mu}_l^{(i)}(t-1)\right) \qquad (9.7)$$

Now we proceed to upper bound last term of (9.7) as follows. Note that for any suboptimal arm k we have,

$$\begin{aligned}&\mathsf{P}\left(A_{t+1}^{(i)}=k, A_{t+1}^{(i)}=\arg\max_{l\in[K]}\widehat{\mu}_l^{(i)}(t)\right)\\ &\leq \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t)+C_k^{(i)}(t)\geq\widehat{\mu}_1^{(i)}(t)+C_1^{(i)}(t), \widehat{\mu}_k^{(i)}(t)\geq\widehat{\mu}_1^{(i)}(t),\ \ \widehat{\mu}_1^{(i)}(t)\leq\mu_1-C_1^{(i)}(t)\right)\\ &+\mathsf{P}\left(\widehat{\mu}_k^{(i)}(t)+C_k^{(i)}(t)\geq\widehat{\mu}_1^{(i)}(t)+C_1^{(i)}(t), \widehat{\mu}_k^{(i)}(t)\geq\widehat{\mu}_1^{(i)}(t),\ \ \widehat{\mu}_1^{(i)}(t)>\mu_1-C_1^{(i)}(t)\right)\\ &\leq \mathsf{P}\left(\widehat{\mu}_1^{(i)}(t)\leq\mu_1-C_1^{(i)}(t)\right) \qquad (9.8)\\ &+\mathsf{P}\left(\widehat{\mu}_k^{(i)}(t)+C_k^{(i)}(t)\geq\widehat{\mu}_1^{(i)}(t)+C_1^{(i)}(t), \widehat{\mu}_k^{(i)}(t)\geq\widehat{\mu}_1^{(i)}(t),\ \ \widehat{\mu}_1^{(i)}(t)>\mu_1-C_1^{(i)}(t)\right) \qquad (9.9)\end{aligned}$$

Now we proceed to upper bound the last term of (9.9) as follows. Note that we have

$$\begin{aligned}&\mathsf{P}\left(\widehat{\mu}_k^{(i)}(t)+C_k^{(i)}(t)\geq\widehat{\mu}_1^{(i)}(t)+C_1^{(i)}(t), \widehat{\mu}_k^{(i)}(t)\geq\widehat{\mu}_1^{(i)}(t),\ \ \widehat{\mu}_1^{(i)}(t)>\mu_1-C_1^{(i)}(t)\right)\\ &\leq \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t)+C_k^{(i)}(t)>\mu_1-C_1^{(i)}(t)+C_k^{(i)}(t)\right)\end{aligned}$$

$$\leq \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right). \tag{9.10}$$

From (9.9) and (9.10) we have

$$\mathsf{P}\left(A_t^{(i)} = k, A_t^{(i)} = \arg\max_{l\in[K]} \widehat{\mu}_l^{(i)}(t-1)\right) \leq \mathsf{P}\left(\widehat{\mu}_1^{(i)}(t-1) \leq \mu_1 - C_1^{(i)}(t-1)\right) \tag{9.11}$$

$$+\mathsf{P}\left(\widehat{\mu}_k^{(i)}(t-1) \geq \mu_k + C_k^{(i)}(t-1)\right) \tag{9.12}$$

From (10.9), (9.7) and (9.12) we have

$$\begin{aligned}
\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(i)} = k\right) &\leq \sum_{\mathcal{C}\in\mathcal{C}}\eta_k + \sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{T}\mathsf{P}\left(A_t^{(i)} = k, N_k^{(i)}(t-1) > \eta_k\right) \\
&\quad + \sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right)\right] \\
&\leq \bar{\chi}(G)\eta_k + N + \sum_{i=1}^{N}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right)\right] \\
&\quad + \sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{T-1}\mathsf{P}\left(A_{t+1}^{(i)} = k, N_k^{(i)}(t) > \eta_k\right)
\end{aligned} \tag{9.13}$$

From (9.1), (9.13) and Lemma 13 we have

$$\begin{aligned}
\sum_{i=1}^{N}\mathbb{E}[n_k^{(i)}(T)] &\leq \bar{\chi}(G)\eta_k + N \\
&\quad + \sum_{i=1}^{N}\sum_{t=1}^{T-1}\left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right)\right]
\end{aligned}$$

This concludes the proof of Lemma 5. □

Now we proceed to bound the tail probabilities as follows.

Lemma 6. (Tail probability bound) *Let $d^{(i)}$ be the degree of agent i. For some $\sigma \geq \sigma_k$ and for any $\zeta > 1$*

$$\mathsf{P}\left(\left|\widehat{\mu}_k^{(i)}(t) - \mu_k\right| \geq \sigma\sqrt{\frac{2(\xi+1)\log t}{N_k^{(i)}(t)}}\right) \leq \frac{1}{\log \zeta}\frac{\log\left((d^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}$$

Proof. Let X_k be the sub-Gaussian random variable that models rewards drawn from arm k. Then X_k has mean μ_k and variance proxy σ_k. Then we have

$$\mathbb{E}\left(\exp(\lambda(X_k - \mu_k))\right) \leq \exp\left(\frac{\lambda^2\sigma_k^2}{2}\right).$$

Recall that $\mathbf{1}\Big\{A_\tau^{(i)} = k\Big\}$ is a $\mathcal{F}_{\tau-1}$ measurable random variable. Then we have

$$\mathbb{E}\left(\exp\left(\lambda\left(X_k - \mu_k\right)\mathbf{1}\left\{A_\tau^{(i)} = k\right\}\mathbf{1}\{(i,j) \in E\}\right)\Big|\mathcal{F}_{\tau-1}\right) \leq \exp\left(\frac{\lambda^2\sigma_k^2}{2}\mathbf{1}\left\{A_\tau^{(i)} = k\right\}\mathbf{1}\{(i,j) \in E_\tau\}\right)$$

Define a new random variable such that $\forall \tau > 0$.

$$Y_k^{(i)}(\tau) = (X_k - \mu_k)\sum_{j=1}^{N}\mathbf{1}\left\{A_\tau^{(i)} = k\right\}\mathbf{1}\{(i,j) \in E_\tau\}.$$

Note that $\mathbb{E}\left(Y_k^{(i)}(\tau)\right) = \mathbb{E}\left(Y_k^{(i)}(\tau)|\mathcal{F}_{\tau-1}\right) = 0$. Let $Z_k^{(i)}(t) = \sum_{\tau=1}^{t} Y_k^{(i)}(\tau)$. For any $\lambda > 0$

$$\begin{aligned}
\mathbb{E}\left(\exp(\lambda Y_k^{(i)}(\tau))|\mathcal{F}_{\tau-1}\right) &= \mathbb{E}\left(\exp\left(\lambda\left(X_k - \mu_k\right)\sum_{j=1}^{N}\mathbf{1}\left\{A_\tau^{(i)} = k\right\}\mathbf{1}\{(i,j) \in E_\tau\}\right)\Big|\mathcal{F}_{\tau-1}\right) \\
&= \mathbb{E}\left(\prod_{j=1}^{K}\exp\left(\lambda\left(X_k - \mu_k\right)\mathbf{1}\left\{A_\tau^{(i)} = k\right\}\mathbf{1}\{(i,j) \in E_\tau\}\right)\Big|\mathcal{F}_{\tau-1}\right) \\
&\overset{(a)}{=} \prod_{j=1}^{N}\mathbb{E}\left(\exp\left(\lambda\left(X_k - \mu_k\right)\mathbf{1}\left\{A_\tau^{(i)} = k\right\}\mathbf{1}\{(i,j) \in E_\tau\}\right)\Big|\mathcal{F}_{\tau-1}\right) \\
&\leq \prod_{j=1}^{N}\exp\left(\frac{\lambda^2\sigma_k^2}{2}\mathbf{1}\left\{A_\tau^{(i)} = k\right\}\mathbf{1}\{(i,j) \in E_\tau\}\right)
\end{aligned}$$

$$= \exp\left(\frac{\lambda^2\sigma_k^2}{2}\sum_{j=1}^{N}\mathbf{1}\{A_\tau^{(i)}=k\}\mathbf{1}\{(i,j)\in E_\tau\}\right).$$

Equality (a) follows from the fact that random variables $\left\{\exp\left(\lambda\left(X_k-\mu_k\right)\mathbf{1}\left\{A_\tau^{(i)}=k\right\}\mathbf{1}\{(i,j)\in E_\tau\right.\right.$ are conditionally independent with respect to $\mathcal{F}_{\tau-1}$. Since $\mathbf{1}\left\{A_\tau^{(i)}=k\right\}, \mathbf{1}\{(i,j)\in E_\tau\}$ are $\mathcal{F}_{\tau-1}$ measurable random variable, and so

$$\mathbb{E}\left(\exp\left(\lambda Y_k^{(i)}(\tau)-\frac{\lambda^2\sigma_k^2}{2}\sum_{j=1}^{N}\mathbf{1}\{A_\tau^{(i)}=k\}\mathbf{1}\{(i,j)\in E_\tau\}\right)\Big|\mathcal{F}_{\tau-1}\right)\leq 1.$$

Let $N_k^{(i)}(t)=\sum_{\tau=1}^{t}\sum_{j=1}^{N}\mathbf{1}\left\{A_\tau^{(i)}=k\right\}\mathbf{1}\{(i,j)\in E_\tau\}$. Then we have

Further, using the properties of conditional expectations

$$\mathbb{E}\left(\exp\left(\lambda Z_k^{(i)}(t)-\frac{\lambda^2\sigma_k^2}{2}N_k^{(i)}(t)\right)\Big|\mathcal{F}_{t-1}\right)\leq\exp\left(\lambda Z_k^{(i)}(t-1)-\frac{\lambda^2\sigma_k^2}{2}N_k^{(i)}(t-1)\right).$$

Thus we see that

$$\mathbb{E}\left(\exp\left(\lambda Z_k^{(i)}(t)-\frac{\lambda^2\sigma_k^2}{2}N_k^{(i)}(t)\right)\right)\leq 1.$$

Note that we have

$$\begin{aligned}\mathsf{P}\left(\exp\left(\lambda Z_k^{(i)}(t)-\frac{\lambda^2\sigma_k^2}{2}N_k^{(i)}(t)\right)\geq\exp\left(2\kappa\vartheta\right)\right)&=\mathsf{P}\left(\lambda Z_k^{(i)}(t)-\frac{\lambda^2\sigma_k^2}{2}N_k^{(i)}(t)\geq 2\kappa\vartheta\right)\\&=\mathsf{P}\left(\frac{Z_k^{(i)}(t)}{\sqrt{N_k^{(i)}(t)}}\geq\frac{2\kappa\vartheta}{\lambda}\sqrt{\frac{1}{N_k^{(i)}(t)}}+\frac{\sigma_k^2}{2}\lambda\sqrt{N_k^{(i)}(t)}\right)\end{aligned}$$

Let $\zeta > 1$. Then $1 \leq N_k^{(i)}(t) \leq \zeta^{D_t}$ where $D_t = \frac{\log((d^{(i)}+1)t)}{\log \zeta}$. For $\lambda_l = \frac{2}{\sigma_k}\sqrt{\frac{\kappa\vartheta}{\zeta^{l-1/2}}}$ and $\zeta^{l-1} \leq N_k^{(i)}(t) \leq \zeta^l$ we have

$$\frac{2\kappa\vartheta}{\lambda_l}\sqrt{\frac{1}{N_k^{(i)}(t)}} + \frac{\sigma_k^2}{2}\lambda_l\sqrt{N_i^k(t)} = \sigma_k\sqrt{\kappa\vartheta}\left(\sqrt{\frac{\zeta^{l-1/2}}{N_k^{(i)}(t)}} + \sqrt{\frac{N_k^{(i)}(t)}{\zeta^{l-1/2}}}\right) \leq \sqrt{\vartheta},$$

where $\kappa = \frac{1}{\sigma_k^2\left(\zeta^{\frac{1}{4}}+\zeta^{-\frac{1}{4}}\right)^2}$.

Recall from the Markov inequality that $\mathsf{P}(Y \geq a) \leq \frac{\mathbb{E}(Y)}{a}$ for any positive random variable Y. Thus,

$$\mathsf{P}\left(\frac{Z_k^{(i)}(t)}{\sqrt{N_k^{(i)}(t)}} \geq \sqrt{\vartheta}\right) \leq \sum_{l=1}^{D_T}\exp(-2\kappa\vartheta).$$

Then we have,

$$\mathsf{P}\left(\frac{Z_k^{(i)}(t)}{N_k^{(i)}(t)} \geq \sqrt{\frac{\vartheta}{N_k^{(i)}(t)}}\right) \leq \sum_{l=1}^{D_T}\exp(-2\kappa\vartheta)$$

Substituting $\vartheta = 2\sigma_k^2(\xi+1)\log t$ we get

$$\mathsf{P}\left(\left|\widehat{\mu}_k^{(i)}(t) - \mu_k\right| \geq \sigma_k\sqrt{\frac{2(\xi+1)\log t}{N_k^{(i)}(t)}}\right) \leq \frac{\log((d^{(i)}+1)t)}{\log\zeta}\exp\left(-\frac{4(\xi+1)\log t}{\left(\zeta^{\frac{1}{4}}+\zeta^{-\frac{1}{4}}\right)^2}\right). \tag{9.14}$$

Since $\sigma \geq \sigma_k$ we have

$$\mathsf{P}\left(\left|\widehat{\mu}_k^{(i)}(t) - \mu_k\right| \geq \sigma\sqrt{\frac{2(\xi+1)\log t}{N_k^{(i)}(t)}}\right) \leq \frac{\log((d^{(i)}+1)t)}{\log\zeta}\exp\left(-\frac{4(\xi+1)\log t}{\left(\zeta^{\frac{1}{4}}+\zeta^{-\frac{1}{4}}\right)^2}\right).$$

Note that $\forall \zeta > 1$ we have

$$\frac{4}{\left(\zeta^{\frac{1}{4}} + \zeta^{-\frac{1}{4}}\right)^2} \geq 1 - \frac{(\zeta-1)^2}{16} \tag{9.15}$$

Then we have

$$\mathsf{P}\left(\left|\widehat{\mu}_k^{(i)}(t) - \mu_k\right| \geq \sigma\sqrt{\frac{2(\xi+1)\log t}{N_k^{(i)}(t)}}\right) \leq \frac{1}{\log\zeta}\frac{\log((d^{(i)}+1)t)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}.$$

This concludes the proof of Lemma 15. □

Lemma 7. *Let $\zeta = 1.3, \xi \geq 1.1$, $d^{(i)} \geq 0$ and $t \in [T]$. Then we have*

$$\sum_{t=1}^{T-1}\frac{1}{\log\zeta}\frac{\log\left((d^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}} \leq 12\log(3(d^{(i)}+1)) + 3\left(\log\left(d^{(i)}+1\right)+1\right) \tag{9.16}$$

Proof. For $\zeta = 1.3$ we have $\frac{1}{\log\zeta} < 8.78$. Further $(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right) > 2$ and $\forall t \geq 3$ we see that $\frac{\log\left((d^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}$ is monotonically decreasing. Thus we have

$$\sum_{t=1}^{T-1}\frac{\log\left((d^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}} \leq 1.362\log(3(d^{(i)}+1)) + \int_3^{T-1}\frac{\log\left((d^{(i)}+1)t\right)}{t^2}dt \tag{9.17}$$

Let $z = (d^{(i)}+1)t$. Then we have

$$\int_3^{T-1}\frac{\log\left((d^{(i)}+1)t\right)}{t^2}dt = (d^{(i)}+1)\int_{3(d^{(i)}+1)}^{(d^{(i)}+1)(T-1)}\frac{\log z}{z^2}dz \tag{9.18}$$

$$= (d^{(i)}+1)\left[-\frac{\log z}{z} - \frac{1}{z}\right]_{3((d^{(i)}+1)}^{(d^{(i)}+1)(T-1)} \tag{9.19}$$

Thus we have

$$\int_3^{T-1}\frac{\log\left((d^{(i)}+1)t\right)}{t^2}dt \leq (d^{(i)}+1)\left[\frac{\log(d^{(i)}+1)}{3(d^{(i)}+1)} + \frac{1}{3(d^{(i)}+1)}\right] \tag{9.20}$$

$$= \frac{1}{3}\log(d^{(i)}+1) + \frac{1}{3} \tag{9.21}$$

Recall that For $\zeta = 1.3$ we have $\frac{1}{\log \zeta} < 8.78$. Thus the proof of Lemma 16 follows from (10.59) and (10.63). □

Now we proceed to prove Theorem 17. From definition of expected cumulative group regret and Lemmas 5, 15 and 16 we have

$$\mathbb{E}\left[R(T)\right] \leq \sum_{k=2}^{K} \frac{8(\xi+1)\sigma}{\Delta_k} \bar{\chi}(G) \log T + 4N \sum_{k=2}^{K} \Delta_k \tag{9.22}$$

$$+ \sum_{i=1}^{N} \left(12 \log(3(d^{(i)}+1)) + 3 \log\left(d^{(i)}+1\right)\right) \sum_{k=2}^{K} \Delta_k \tag{9.23}$$

This concludes the proof of Theorem 17.

9.10.2 Proof of Theorem 12

Recall that all the agents communicate their rewards and arm ids at time $t = 1$. Then the expected communication cost can be given as

$$\mathbb{E}\left[L(T)\right] = \sum_{i=1}^{N} \sum_{t=1}^{T-1} \mathsf{P}\left(A_t^{(i)} \neq \arg\max_{k \in [K]} \widehat{\mu}_k^{(i)}(t-1)\right). \tag{9.24}$$

Note that we have

$$\sum_{i=1}^{N} \sum_{t=1}^{T-1} \mathsf{P}\left(A_t^{(i)} \neq \arg\max_{k \in [K]} \widehat{\mu}_k^{(i)}(t-1)\right) = \sum_{i=1}^{N} \sum_{t=1}^{T-1} \mathsf{P}\left(A_t^{(i)} = 1, A_t^{(i)} \neq \arg\max_{k \in [K]} \widehat{\mu}_k^{(i)}(t-1)\right)$$
$$+ \sum_{i=1}^{N} \sum_{t=1}^{T-1} \mathsf{P}\left(A_t^{(i)} \neq 1, A_t^{(i)} \neq \arg\max_{k \in [K]} \widehat{\mu}_k^{(i)}(t-1)\right). \tag{9.25}$$

For all agents we first upper bound the expected number of times they shares rewards and actions with their neighbors until time T when they pull a suboptimal arm:

$$\sum_{i=1}^{N}\sum_{t=1}^{T-1}\mathsf{P}\left(A_t^{(i)}\neq 1, A_t^{(i)}\neq \arg\max_{k\in[K]}\widehat{\mu}_k^{(i)}(t-1)\right)\leq \sum_{i=1}^{N}\sum_{t=1}^{T-1}\mathsf{P}\left(A_t^{(i)}\neq 1\right)\leq \sum_{i=1}^{N}\sum_{k=2}^{K}\mathbb{E}\left[n_k^{(i)}(T)\right]. \tag{9.26}$$

Next for all agents we upper bound the expected number of times they shares rewards and actions with their neighbors until time T when they pull the optimal arm as follows. Let k_t^* be the suboptimal arm with highest estimated expected reward for agents i at time t. This can be stated as $k_t^* = \arg\max_{k\neq 1, k\in[K]}\widehat{\mu}_k^{(i)}(t)$. Note that $\forall i, t$ we have

$$\begin{aligned}\left\{A_{t+1}^{(i)}=1, A_{t+1}^{(i)}\neq \arg\max_{k\in[K]}\widehat{\mu}_k^{(i)}(t)\right\}&\subseteq\left\{\widehat{\mu}_1^{(i)}(t)\leq \mu_1 - C_1^{(i)}(t)\right\}\\ &\cup\left\{A_{t+1}^{(i)}=1, \widehat{\mu}_1^{(i)}(t)\geq \mu_1 - C_1^{(1)}(t), \widehat{\mu}_{k_t^*}^{(i)}(t)\geq \widehat{\mu}_1^{(i)}(t)\right\}.\end{aligned}$$

Thus, we have

$$\begin{aligned}\sum_{t=1}^{T-1}\mathsf{P}\left(A_t^{(i)}=1, A_t^{(i)}\neq \arg\max_{k\in[K]}\widehat{\mu}_k^{(i)}(t-1)\right)&\leq \sum_{t=1}^{T}\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t-1)\leq \mu_1 - C_1^{(i)}(t-1)\right)\\ &+\sum_{t=1}^{T}\mathsf{P}\left(A_{t+1}^{(i)}=1, \widehat{\mu}_1^{(i)}(t-1)\geq \mu_1 - C_1^{(1)}(t-1), \widehat{\mu}_{k_t^*}^{(i)}(t-1)\geq \widehat{\mu}_1^{(i)}(t-1)\right).\end{aligned} \tag{9.27}$$

Note that the first term on the right hand side of the above equation is the summation tail probabilities of the estimate of the optimal arm. Now we proceed to upper bound the second term as follows. Let $\tau_1^{(i)}$ denote the maximum time step when the total number of times agent i pulled the optimal arm and the total number of observations it received from its neighbors about the optimal arm is at most $\bar{\eta}$. This can be stated

as $\tau_1^{(i)} := \max\{t \in [T] : N_1^{(i)}(t) \leq \bar{\eta}\}$. Recall that $N_1^{(i)}(t) \geq n_1^{(i)}(t)$. Thus we have that $n_1^{(i)}(t) \leq \bar{\eta}, \forall t \leq \tau_1^{(i)}$.

Note that we have

$$\begin{aligned}
&\sum_{t=1}^{T-1} \mathsf{P}\left(A_t^{(i)} = 1, \widehat{\mu}_1^{(i)}(t-1) \geq \mu_1 - C_1^{(1)}(t-1), \widehat{\mu}_{k_t^*}^{(i)}(t-1) \geq \widehat{\mu}_1^{(i)}(t-1)\right) \\
&\leq \sum_{t=1}^{\tau_1^{(i)}} \mathsf{P}\left(A_t^{(i)} = 1, \widehat{\mu}_1^{(i)}(t-1) \geq \mu_1 - C_1^{(1)}(t-1), \widehat{\mu}_{k_t^*}^{(i)}(t-1) \geq \widehat{\mu}_1^{(i)}(t-1)\right) \\
&+ \sum_{t>\tau_1^{(i)}}^{T-1} \mathsf{P}\left(A_t^{(i)} = 1, \widehat{\mu}_1^{(i)}(t-1) \geq \mu_1 - C_1^{(1)}(t-1), \widehat{\mu}_{k_{t-1}^*}^{(i)}(t-1) \geq \widehat{\mu}_1^{(i)}(t-1)\right) \\
&\leq \bar{\eta} + 1 + \sum_{t>\tau_1^{(i)}}^{T-2} \mathsf{P}\left(A_{t+1}^{(i)} = 1, \widehat{\mu}_1^{(i)}(t) \geq \mu_1 - C_1^{(1)}(t), \widehat{\mu}_{k_t^*}^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t), N_1^{(i)}(t) > \bar{\eta}\right).
\end{aligned} \tag{9.28}$$

If agent i pulls the optimal arm at time t we have $Q_1^{(i)}(t-1) \geq Q_{k_{t-1}^*}^{(i)}(t-1)$. Further, if $\widehat{\mu}_{k_{t-1}^*}^{(i)}(t-1) \geq \widehat{\mu}_1^{(i)}(t-1)$ then we have $C_{k_{t-1}^*}^{(i)}(t-1) < C_1^{(i)}(t-1)$. Let $\bar{\eta} = \frac{8\sigma(\xi+1)}{\Delta^2} \log T$. Then we have

$$\begin{aligned}
&\sum_{t>\tau_1^{(i)}}^{T-2} \mathsf{P}\left(A_{t+1}^{(i)} = 1, \widehat{\mu}_1^{(i)}(t) \geq \mu_1 - C_1^{(1)}(t), \widehat{\mu}_{k_t^*}^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t), N_1^{(i)}(t) > \bar{\eta}\right) \\
&\quad \leq \sum_{t>\tau_1^{(i)}}^{T-2} \mathsf{P}\left(A_{t+1}^{(i)} = 1, \widehat{\mu}_1^{(i)}(t) \geq \mu_1 - C_1^{(1)}(t), \widehat{\mu}_{k_{t-1}^*}^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t), \mu_1 > \mu_{k_t^*} + 2C_1^{(i)}(t)\right) \\
&\quad \leq \sum_{t>\tau_1^{(i)}}^{T-2} \mathsf{P}\left(\widehat{\mu}_{k_{t-1}^*}^{(i)}(t) \geq \mu_1 - C_1^{(i)}(t), \mu_1 > \mu_{k_t^*} + 2C_1^{(i)}(t)\right) \\
&\quad \leq \sum_{t>\tau_1^{(i)}}^{T-2} \mathsf{P}\left(\widehat{\mu}_{k_t^*}^{(i)}(t) \geq \mu_{k_t^*} + C_{k_t^*}^{(i)}(t)\right).
\end{aligned} \tag{9.29}$$

From (9.27), (9.28) and (9.29) we have

$$\sum_{t=1}^{T} \mathsf{P}\left(A_t^{(i)} = 1, A_t^{(i)} \neq \arg\max_{k\in[K]} \widehat{\mu}_k^{(i)}(t-1)\right) \leq \frac{8\sigma(\xi+1)}{\bar{\Delta}^2}\log T \tag{9.30}$$

$$+\sum_{t=1}^{T-1} \mathsf{P}\left(\widehat{\mu}_1^{(i)}(t-1) \leq \mu_1 - C_1^{(i)}(t-1)\right) + \sum_{t=1}^{T-1} \mathsf{P}\left(\widehat{\mu}_{k_{t-1}^*}^{(i)}(t-1) \geq \mu_{k_{t-1}^*} + C_{k_{t-1}^*}^{(i)}(t-1)\right) \tag{9.31}$$

From (9.31) and Lemma 15 we have

$$\sum_{t=1}^{T} \mathsf{P}\left(A_t^{(i)} = 1, A_t^{(i)} \neq \arg\max_{k\in[K]} \widehat{\mu}_k^{(i)}(t-1)\right) \leq \frac{8\sigma(\xi+1)}{\bar{\Delta}^2}\log T + 2\sum_{t=1}^{T}\frac{1}{\log\zeta}\frac{\log\left((d^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}} \tag{9.32}$$

The proof of Theorem 12 follows from (9.24), (9.25), (9.26), (9.32) and Theorem 17.

9.10.3 Proof of Theorem 18

In section we follow an approach similar to Section **??**. Recall that G_γ is the γ^{th} power graph of G. Thus each pair of vertices in G_γ are adjacent if and only if they distance between them in G is at most γ. We begin the proof of Theorem 18 by proving a lemma similar to Lemma 5.

Lemma 8. *Let $\bar{\chi}(G_\gamma)$ is the clique covering number of graph G_γ. Let $\eta_k = \left(\frac{8(\xi+1)\sigma^2}{\Delta_k^2}\right)\log T$. Then we have*

$$\sum_{i=1}^{N}\mathbb{E}[n_k^{(i)}(T)] \leq \bar{\chi}(G_\gamma)\eta_k + N + (N-\bar{\chi}(G_\gamma))(\gamma-1)$$

$$+\sum_{i=1}^{N}\sum_{t=1}^{T-1}\left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right)\right]$$

Proof. Let $\mathcal{C}_\gamma$ be a non overlapping clique covering of G_γ. Note that for each suboptimal arm $k > 1$ we have

$$\sum_{i=1}^{N} \mathbb{E}[n_k^{(i)}(T)] = \sum_{\mathcal{C}\in\mathcal{C}_\gamma} \sum_{i\in\mathcal{C}} \sum_{t=1}^{T} \mathsf{P}\left(A_t^{(i)} = k\right) \tag{9.33}$$

Let $\tau_{k,\mathcal{C}}$ be the maximum time step such that the total number of messages about pulls from arm k initiated by agents in the clique $\mathcal{C}$ is at most $\eta_k + (|\mathcal{C}| - 1)(\gamma - 1)$. This can be stated as

$$\tau_{k,\mathcal{C}} := \max\left\{ t \in [T] : \sum_{i\in\mathcal{C}} \sum_{\tau=1}^{t} \mathbf{1}\left\{ A_\tau^{(i)} = k, A_\tau^{(i)} \neq \arg\max_{l\in[K]} \widehat{\mu}_l^{(i)}(\tau - 1) \right\} \leq \eta_k + (|\mathcal{C}| - 1)(\gamma - 1) \right\}.$$

Further for all $i \in \mathcal{C}$ we have $N_k^{(i)}(t) > \eta_k, \forall t > \tau_{k,\mathcal{C}}$. We analyse the expected number of times all agents pull suboptimal arm k as follows.

$$\sum_{\mathcal{C}\in\mathcal{C}} \sum_{i\in\mathcal{C}} \sum_{t=1}^{T} \mathbf{1}\left\{A_t^{(i)} = k\right\} = \sum_{\mathcal{C}\in\mathcal{C}} \sum_{i\in\mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathbf{1}\left\{A_t^{(i)} = k\right\} + \sum_{\mathcal{C}\in\mathcal{C}} \sum_{i\in\mathcal{C}} \sum_{t>\tau_{k,\mathcal{C}}}^{T} \mathbf{1}\left\{A_t^{(i)} = k, N_k^{(i)}(t-1) > \eta_k\right\} \tag{9.34}$$

Taking the expectation of (9.34) we have

$$\sum_{\mathcal{C}\in\mathcal{C}} \sum_{i\in\mathcal{C}} \sum_{t=1}^{T} \mathsf{P}\left(A_t^{(i)} = k\right) = \sum_{\mathcal{C}\in\mathcal{C}} \sum_{i\in\mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathsf{P}\left(A_t^{(i)} = k\right) + \sum_{\mathcal{C}\in\mathcal{C}} \sum_{i\in\mathcal{C}} \sum_{t>\tau_{k,\mathcal{C}}}^{T} \mathsf{P}\left(A_t^{(i)} = k, N_k^{(i)}(t-1) > \eta_k\right) \tag{9.35}$$

Now we proceed to upper bound the first term of right hand side of (9.34) as follows. Note that we have

$$\begin{aligned}\sum_{i\in\mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathbf{1}\left\{A_t^{(i)} = k\right\} = &\sum_{i\in\mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathbf{1}\left\{ A_t^{(i)} = k, A_t^{(i)} \neq \arg\max_{l\in[K]} \widehat{\mu}_l^{(i)}(t-1) \right\} \\ &+ \sum_{i\in\mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathbf{1}\left\{ A_t^{(i)} = k, A_t^{(i)} = \arg\max_{l\in[K]} \widehat{\mu}_l^{(i)}(t-1) \right\}\end{aligned}$$

$$\leq \eta_k + (|\mathcal{C}| - 1)(\gamma - 1) + \sum_{i \in \mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathbf{1}\left\{ A_t^{(i)} = k, A_t^{(i)} = \arg\max_{l \in [K]} \widehat{\mu}_l^{(i)}(t-1) \right\} \tag{9.36}$$

Taking the expectation of (9.36) we have

$$\sum_{i \in \mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathsf{P}\left(A_t^{(i)} = k \right) \leq \eta_k + (|\mathcal{C}| - 1)(\gamma - 1) + \sum_{i \in \mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathsf{P}\left(A_t^{(i)} = k, A_t^{(i)} = \arg\max_{l \in [K]} \widehat{\mu}_l^{(i)}(t-1) \right) \tag{9.37}$$

Now we proceed to upper bound last term of (9.37) as follows. Note that for any suboptimal arm k we have,

$$\begin{aligned}
&\mathsf{P}\left(A_{t+1}^{(i)} = k, A_{t+1}^{(i)} = \arg\max_{l \in [K]} \widehat{\mu}_l^{(i)}(t) \right) \\
&\leq \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) + C_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t) + C_1^{(i)}(t), \widehat{\mu}_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t),\ \ \widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t) \right) \\
&+ \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) + C_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t) + C_1^{(i)}(t), \widehat{\mu}_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t),\ \ \widehat{\mu}_1^{(i)}(t) > \mu_1 - C_1^{(i)}(t) \right) \\
&\leq \mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t) \right) \\
&+ \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) + C_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t) + C_1^{(i)}(t), \widehat{\mu}_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t),\ \ \widehat{\mu}_1^{(i)}(t) > \mu_1 - C_1^{(i)}(t) \right)
\end{aligned} \tag{9.38}$$

Now we proceed to upper bound the last term of (9.38) as follows. Note that we have

$$\begin{aligned}
&\mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) + C_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t) + C_1^{(i)}(t), \widehat{\mu}_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t),\ \ \widehat{\mu}_1^{(i)}(t) > \mu_1 - C_1^{(i)}(t) \right) \\
&\leq \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) + C_k^{(i)}(t) > \mu_1 - C_1^{(i)}(t) + C_k^{(i)}(t) \right) \\
&\leq \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t) \right).
\end{aligned} \tag{9.39}$$

From (9.38) and (9.39) we have

$$\mathsf{P}\left(A_t^{(i)}=k, A_t^{(i)}=\arg\max_{l\in[K]}\widehat{\mu}_l^{(i)}(t-1)\right) \leq \mathsf{P}\left(\widehat{\mu}_1^{(i)}(t-1)\leq \mu_1 - C_1^{(i)}(t-1)\right) \quad (9.40)$$

$$+\mathsf{P}\left(\widehat{\mu}_k^{(i)}(t-1)\geq \mu_k + C_k^{(i)}(t-1)\right). \quad (9.41)$$

From (10.103), (9.37) and (9.41) we have

$$\begin{aligned}
&\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(i)}=k\right)\\
&\leq \sum_{\mathcal{C}\in\mathcal{C}}\eta_k + \sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t)\leq \mu_1 - C_1^{(i)}(t)\right)+\mathsf{P}\left(\widehat{\mu}_k^{(i)}(t)\geq \mu_k + C_k^{(i)}(t)\right)\right]\\
&\quad + \sum_{\mathcal{C}\in\mathcal{C}}(|\mathcal{C}|-1)(\gamma-1) + \sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{T}\mathsf{P}\left(A_t^{(i)}=k, N_k^{(i)}(t-1)>\eta_k\right)\\
&\leq \bar{\chi}(G_\gamma)\eta_k + \sum_{i=1}^{N}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t)\leq \mu_1 - C_1^{(i)}(t)\right)+\mathsf{P}\left(\widehat{\mu}_k^{(i)}(t)\geq \mu_k + C_k^{(i)}(t)\right)\right]\\
&\quad + N + (N-\bar{\chi}(G_\gamma))(\gamma-1) + \sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{T-1}\mathsf{P}\left(A_{t+1}^{(i)}=k, N_k^{(i)}(t)>\eta_k\right).
\end{aligned} \quad (9.42)$$

The proof of Lemma 8 follows from (9.33), (9.42) and Lemma 13. □

Now we proceed to prove Theorem 18 as follows. We start by obtaining a modified tail bound similar to the result in Lemma 15. Note that $\forall i,k,t$ we have $1\leq N_k^{(i)}(t) < d_\gamma^{(i)}t$. Thus considering $D_t = \frac{\log\left(\left(d_\gamma^{(i)}+1\right)t\right)}{\log\zeta}$ for any $\zeta>1$ in Lemma 15 we get

$$\mathsf{P}\left(\left|\widehat{\mu}_k^{(i)}(t)-\mu_k\right|\geq\sigma\sqrt{\frac{2(\xi+1)\log t}{N_k^{(i)}(t)}}\right)\leq\frac{1}{\log\zeta}\frac{\log\left((d_\gamma^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}. \quad (9.43)$$

The proof of Theorem 18 follows from Lemmas 16, 8 and (9.43).

9.10.4 Proof of Theorem 14

Following a similar approach to the proof of Theorem 12 we obtain

$$\sum_{i=1}^{N}\sum_{t=1}^{T-1}\mathsf{P}\left(A_t^{(i)}\neq 1, A_t^{(i)}\neq \arg\max_{k\in[K]}\widehat{\mu}_k^{(i)}(t-1)\right)\leq \sum_{i=1}^{N}\sum_{t=1}^{T-1}\mathsf{P}\left(A_t^{(i)}\neq 1\right)\leq \sum_{i=1}^{N}\sum_{k=2}^{K}\mathbb{E}\left[n_k^{(i)}(T)\right]. \tag{9.44}$$

Similarly we get

$$\sum_{t=1}^{T-1}\mathsf{P}\left(A_t^{(i)}=1, A_t^{(i)}\neq \arg\max_{k\in[K]}\widehat{\mu}_k^{(i)}(t-1)\right)\leq \frac{8\sigma(\xi+1)}{\bar{\Delta}^2}\log T+2\sum_{t=1}^{T}\frac{1}{\log\zeta}\frac{\log\left((d_\gamma^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}} \tag{9.45}$$

From (9.44) and (9.45) we have

$$\begin{aligned}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(i)}\neq \arg\max_{k\in[K]}\widehat{\mu}_k^{(i)}(t-1)\right)&\leq \sum_{i=1}^{N}\sum_{k=2}^{K}\mathbb{E}\left[n_k^{(i)}(T)\right]\\&+\sum_{i=1}^{N}\frac{8\sigma(\xi+1)}{\bar{\Delta}^2}\log T+2\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{\log\zeta}\frac{\log\left((d_\gamma^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}\end{aligned} \tag{9.46}$$

Note that (9.46) is the expected number of messages initiated by all the agents. Recall that in ComEx-MPUCB a message initiated by agent i is subsequently passed by agents within a $\gamma-1$ distance in graph G. Thus we have

$$\mathbb{E}\left[L(T)\right]\leq \sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(i)}\neq \arg\max_{k\in[K]}\widehat{\mu}_k^{(i)}(t-1)\right) \tag{9.47}$$

From (9.46) and (9.47) we have

$$\mathbb{E}\left[L(T)\right]\leq \sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\sum_{k=2}^{K}\mathbb{E}\left[n_k^{(i)}(T)\right]+\sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\frac{8\sigma(\xi+1)}{\bar{\Delta}^2}\log T$$

$$+2\sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\sum_{t=1}^{T}\frac{1}{\log\zeta}\frac{\log\left((d_{\gamma}^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}} \tag{9.48}$$

From (9.43), (9.48) and Lemma 8 we have

$$\mathbb{E}\left[L(T)\right]\leq\sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\sum_{k=2}^{K}\left(\bar{\chi}(G_\gamma)\eta_k+N+(N-\bar{\chi}(G_\gamma))(\gamma-1)\right)$$
$$+\sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\frac{8\sigma(\xi+1)}{\bar{\Delta}^2}\log T+2K\sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\sum_{t=1}^{T}\frac{1}{\log\zeta}\frac{\log\left((d_{\gamma}^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}} \tag{9.49}$$

Recall that $\eta_k=\frac{8\sigma(\xi+1)}{\Delta_k^2}\log T$. Thus the proof of Theorem 14 follows from (9.49) and Lemma 16.

9.10.5 Proof of Theorem 15

We follow a similar approach to proof of Theorem 18. We begin the proof by providing a lemma similar to Lemma 8.

Lemma 9. *Let $\bar{\gamma}(G_\gamma)$ is the dominating number of graph G_γ. Let $\eta_k=\left(\frac{8(\xi+1)\sigma^2}{\Delta_k^2}\right)\log T$. Then we have*

$$\sum_{i=1}^{N}\mathbb{E}[n_k^{(i)}(T)]\leq\bar{\gamma}(G_\gamma)\eta_k+N+(N-\bar{\gamma}(G_\gamma))(3\gamma-1)$$
$$+\sum_{i\in V_\gamma'}\left(\left|\mathcal{N}_\gamma^{(i)}\right|+1\right)\sum_{t=1}^{T-1}\left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t)\leq\mu_1-C_1^{(i)}(t)\right)+\mathsf{P}\left(\widehat{\mu}_k^{(i)}(t)\geq\mu_k+C_k^{(i)}(t)\right)\right]$$

where V_γ' is the maximal dominating set of G_γ and $\mathcal{N}_\gamma^{(i)}$ is the set of followers of leader i.

Proof. Recall that V'_γ is the maximal dominating set of G_γ. Let $\mathcal{N}_\gamma^{(i)}$ be the set of followers of leader i. Then for each suboptimal arm $k > 1$ we have

$$\sum_{i=1}^{N} \mathbb{E}[n_k^{(i)}(T)] = \sum_{i \in V'_\gamma} \left(\sum_{t=1}^{T} \mathsf{P}\left(A_t^{(i)} = k\right) + \sum_{j \in \mathcal{N}_\gamma^{(i)}} \sum_{t=1}^{T} \mathsf{P}\left(A_t^{(j)} = k\right) \right) \tag{9.50}$$

Let $\tau_k^{(i)}$ be the maximum time step such that the total number of times agent i pulls arm k and the number of times agents in $\mathcal{N}_\gamma^{(i)}$ initiated messages about pulls from arm k is at most $\eta_k + \mathcal{N}_\gamma^{(i)}(\gamma - 1)$. This can be stated as

$$\tau_k^{(i)} := \max \left\{ t \in [T] : \sum_{\tau=1}^{t} \mathbf{1}\{A_\tau^{(i)} = k\} + \sum_{j \in \mathcal{N}_\gamma^{(i)}} \sum_{\tau=1}^{t} \mathbf{1}\left\{ A_\tau^{(i)} = k, A_\tau^{(i)} \neq \arg\max_{l \in [K]} \widehat{\mu}_l^{(i)}(\tau - 1) \right\} \right.$$
$$\left. \leq \eta_k + \mathcal{N}_\gamma^{(i)}(\gamma - 1) \right\}.$$

Then we have $N_k^{(i)}(t) > \eta_k, \forall t > \tau_k^{(i)}$. We analyse the expected number of times all agents pull suboptimal arm k as follows. Let $d(i,j)$ be the distance between agents i and j in graph G. Then note that for any $j \in \mathcal{N}_\gamma^{(i)}$ we have $A_t^{(j)} = A_{t-d(i,j)}^{(i)}$ and $d(i,j) \leq \gamma$.

$$\sum_{i \in V'_\gamma} \left\{ \sum_{t=1}^{T} \mathbf{1}\left\{A_t^{(i)} = k\right\} + \sum_{j \in \mathcal{N}_\gamma^{(i)}} \sum_{t=1}^{T} \mathbf{1}\left\{A_t^{(j)} = k\right\} \right\} \leq \sum_{i \in V'_\gamma} \left\{ \sum_{t=1}^{\tau_k^{(i)}} \mathbf{1}\left\{A_t^{(i)} = k\right\} \right.$$
$$\left. + \sum_{j \in \mathcal{N}_\gamma^{(i)}} \sum_{t=d(i,j)}^{\tau_k^{(i)}} \mathbf{1}\left\{A_t^{(j)} = k\right\} \right\}$$
$$+ \sum_{i \in V'_\gamma} \left\{ \sum_{t > \tau_k^{(i)}}^{T} \mathbf{1}\left\{A_t^{(i)} = k\right\} + \sum_{j \in \mathcal{N}_\gamma^{(i)}} \sum_{t > \tau_k^{(i)}}^{T-d(i,j)} \mathbf{1}\left\{A_t^{(i)} = k\right\} \right\} + \sum_{i \in V'_\gamma} \sum_{j \in \mathcal{N}_\gamma^{(i)}} 2d(i,j)$$
$$\leq \sum_{i \in V'_\gamma} \left\{ \sum_{t=1}^{\tau_k^{(i)}} \mathbf{1}\left\{A_t^{(i)} = k\right\} + \sum_{j \in \mathcal{N}_\gamma^{(i)}} \sum_{t=d(i,j)}^{\tau_k^{(i)}} \mathbf{1}\left\{A_t^{(j)} = k\right\} \right\} + \sum_{i \in V'_\gamma} \left(\left| \mathcal{N}_\gamma^{(i)} \right| + 1 \right) \sum_{t > \tau_k^{(i)}}^{T} \mathbf{1}\left\{A_t^{(i)} = k\right\}$$

$$+2(N-\bar{\gamma}(G_\gamma))\gamma \tag{9.51}$$

Now we proceed to upper bound the first two terms of right hand side of (9.51) as follows. Note that we have

$$\begin{aligned}
&\sum_{i\in V'_\gamma}\left\{\sum_{t=1}^{\tau_k^{(i)}}\mathbf{1}\Big\{A_t^{(i)}=k\Big\}+\sum_{j\in\mathcal{N}_\gamma^{(i)}}\sum_{t=d(i,j)}^{\tau_k^{(i)}}\mathbf{1}\Big\{A_t^{(j)}=k\Big\}\right\}\\
&=\sum_{i\in V'_\gamma}\left\{\sum_{t=1}^{\tau_k^{(i)}}\mathbf{1}\Big\{A_t^{(i)}=k\Big\}+\sum_{j\in\mathcal{N}_\gamma^{(i)}}\sum_{t=d(i,j)}^{\tau_k^{(i)}}\mathbf{1}\Big\{A_t^{(j)}=k, A_{t-d(i,j)}^{(i)}\neq\arg\max_{l\in[K]}\widehat{\mu}_l^{(i)}(t-d(i,j)-1)\Big\}\right.\\
&\qquad\left.+\sum_{j\in\mathcal{N}_\gamma^{(i)}}\sum_{t=d(i,j)}^{\tau_k^{(i)}}\mathbf{1}\Big\{A_t^{(j)}=k, A_{t-d(i,j)}^{(i)}=\arg\max_{l\in[K]}\widehat{\mu}_l^{(i)}(t-d(i,j)-1)\Big\}\right\}\\
&\leq\sum_{i\in V'_\gamma}\left(\eta_k+\mathcal{N}_\gamma^{(i)}(\gamma-1)\right)+\sum_{i\in V'_\gamma}\left|\mathcal{N}_\gamma^{(i)}\right|\sum_{t=1}^{\tau_k^{(i)}}\mathbf{1}\Big\{A_t^{(i)}=k, A_t^{(i)}=\arg\max_{l\in[K]}\widehat{\mu}_l^{(i)}(t-1)\Big\}
\end{aligned} \tag{9.52}$$

Taking the expectation of (9.51) and (9.52) we have

$$\begin{aligned}
&\sum_{i\in V'_\gamma}\left(\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(i)}=k\right)+\sum_{j\in\mathcal{N}_\gamma^{(i)}}\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(j)}=k\right)\right)\leq\bar{\gamma}(G_\gamma)\eta_k+(N-\bar{\gamma}(G))(3\gamma-1)\\
&+\sum_{i\in V'_\gamma}\left(\left|\mathcal{N}_\gamma^{(i)}\right|+1\right)\sum_{t>\tau_k^{(i)}}^{T}\mathsf{P}\left(A_t^{(i)}=k\right)+\sum_{i\in V'_\gamma}\left|\mathcal{N}_\gamma^{(i)}\right|\sum_{t=1}^{\tau_k^{(i)}}\mathsf{P}\left(A_t^{(i)}=k, A_t^{(i)}=\arg\max_{l\in[K]}\widehat{\mu}_l^{(i)}(t-1)\right)
\end{aligned} \tag{9.53}$$

Now we proceed to upper bound last term of (9.53) as follows. Note that for any suboptimal arm k we have,

$$\mathsf{P}\left(A_{t+1}^{(i)}=k, A_{t+1}^{(i)}=\arg\max_{l\in[K]}\widehat{\mu}_l^{(i)}(t)\right)$$

$$
\begin{aligned}
&\leq \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) + C_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t) + C_1^{(i)}(t), \widehat{\mu}_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t),\ \ \widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right) \\
&+ \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) + C_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t) + C_1^{(i)}(t), \widehat{\mu}_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t),\ \ \widehat{\mu}_1^{(i)}(t) > \mu_1 - C_1^{(i)}(t)\right) \\
&\leq \mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right) \\
&+ \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) + C_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t) + C_1^{(i)}(t), \widehat{\mu}_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t),\ \ \widehat{\mu}_1^{(i)}(t) > \mu_1 - C_1^{(i)}(t)\right)
\end{aligned} \tag{9.54}
$$

Now we proceed to upper bound the last term of (9.54) as follows. Note that we have

$$
\begin{aligned}
&\mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) + C_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t) + C_1^{(i)}(t), \widehat{\mu}_k^{(i)}(t) \geq \widehat{\mu}_1^{(i)}(t),\ \ \widehat{\mu}_1^{(i)}(t) > \mu_1 - C_1^{(i)}(t)\right) \\
&\quad \leq \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) + C_k^{(i)}(t) > \mu_1 - C_1^{(i)}(t) + C_k^{(i)}(t)\right) \\
&\quad \leq \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right).
\end{aligned} \tag{9.55}
$$

From (9.54) and (9.55) we have

$$
\begin{aligned}
\mathsf{P}\left(A_t^{(i)} = k, A_t^{(i)} = \arg\max_{l \in [K]} \widehat{\mu}_l^{(i)}(t-1)\right) &\leq \mathsf{P}\left(\widehat{\mu}_1^{(i)}(t-1) \leq \mu_1 - C_1^{(i)}(t-1)\right) \\
&+ \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t-1) \geq \mu_k + C_k^{(i)}(t-1)\right).
\end{aligned} \tag{9.56}
$$

Recall that $N_k^{(i)}(t) > \eta_k, \forall t > \tau_k^{(i)}$. Thus from (9.53), (9.56) and Lemma 13 we have

$$
\begin{aligned}
\sum_{i \in V_\gamma'} \left(\sum_{t=1}^{T} \mathsf{P}\left(A_t^{(i)} = k\right) + \sum_{j \in \mathcal{N}_\gamma^{(i)}} \sum_{t=1}^{T} \mathsf{P}\left(A_t^{(j)} = k\right) \right) &\leq \bar{\gamma}(G_\gamma)\eta_k + N + (N - \bar{\gamma}(G))(3\gamma - 1) \\
&+ \sum_{i \in V_\gamma'} \left(\left|\mathcal{N}_\gamma^{(i)}\right| + 1\right) \sum_{t=1}^{T-1} \left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right)\right].
\end{aligned} \tag{9.57}
$$

The proof of Lemma 9 follows from (9.50) and (9.57). □

Now we proceed to prove Theorem 15 as follows. We start by obtaining a modified tail bound similar to the result in Lemma 15. Note that $\forall i \in V'_\gamma$ we have $1 \leq N_k^{(i)}(t) < d_\gamma^{(i)} t$. Thus considering $D_t = \frac{\log\left(\left(d_\gamma^{(i)}+1\right)t\right)}{\log \zeta}$ for any $\zeta > 1$ in Lemma 15 we get

$$\mathsf{P}\left(\left|\widehat{\mu}_k^{(i)}(t) - \mu_k\right| \geq \sigma\sqrt{\frac{2(\xi+1)\log t}{N_k^{(i)}(t)}}\right) \leq \frac{1}{\log \zeta}\frac{\log\left((d_\gamma^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}. \tag{9.58}$$

The proof of Theorem 15 follows from Lemmas 16, 9 and (9.58).

9.10.6 Proof of Theorem 16

Following a similar approach to the proof of Theorem 14 we obtain

$$\sum_{i=1}^{N}\sum_{t=1}^{T-1}\mathsf{P}\left(A_t^{(i)} \neq 1, A_t^{(i)} \neq \arg\max_{k\in[K]} \widehat{\mu}_k^{(i)}(t-1)\right) \leq \sum_{i=1}^{N}\sum_{t=1}^{T-1}\mathsf{P}\left(A_t^{(i)} \neq 1\right) \leq \sum_{i=1}^{N}\sum_{k=2}^{K}\mathbb{E}\left[n_k^{(i)}(T)\right]. \tag{9.59}$$

Similarly we get

$$\sum_{t=1}^{T-1}\mathsf{P}\left(A_t^{(i)} = 1, A_t^{(i)} \neq \arg\max_{k\in[K]} \widehat{\mu}_k^{(i)}(t-1)\right) \leq \frac{8\sigma(\xi+1)}{\bar{\Delta}^2}\log T + 2\sum_{t=1}^{T}\frac{1}{\log\zeta}\frac{\log\left((d_\gamma^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}} \tag{9.60}$$

From (9.59) and (9.60) we have

$$\begin{aligned}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\mathsf{P}\left(A_t^{(i)} \neq \arg\max_{k\in[K]} \widehat{\mu}_k^{(i)}(t-1)\right) &\leq \sum_{i=1}^{N}\sum_{k=2}^{K}\mathbb{E}\left[n_k^{(i)}(T)\right] \\ &+ \sum_{i=1}^{N}\frac{8\sigma(\xi+1)}{\bar{\Delta}^2}\log T + 2\sum_{i=1}^{N}\sum_{t=1}^{T-1}\frac{1}{\log\zeta}\frac{\log\left((d_\gamma^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}\end{aligned} \tag{9.61}$$

Note that (9.61) is the expected number of messages initiated by all the agents. Recall that in ComEx-LFUCB a message initiated by agent i is subsequently passed

by agents within a $\gamma-1$ distance in graph G. Thus we have

$$\mathbb{E}\left[L(T)\right] \leq \sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\sum_{t=1}^{T-1}\mathsf{P}\left(A_t^{(i)} \neq \arg\max_{k\in[K]}\widehat{\mu}_k^{(i)}(t-1)\right) \tag{9.62}$$

From (9.61) and (9.62) we have

$$\begin{aligned}\mathbb{E}\left[L(T)\right] \leq \sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\sum_{k=2}^{K}\mathbb{E}\left[n_k^{(i)}(T)\right] + \sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\frac{8\sigma(\xi+1)}{\bar{\Delta}^2}\log T \\ +2\sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\sum_{t=1}^{T}\frac{1}{\log\zeta}\frac{\log\left((d_{\gamma}^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}\end{aligned} \tag{9.63}$$

From (9.58), (9.63) and Lemma 9 we have

$$\begin{aligned}\mathbb{E}\left[L(T)\right] \leq \sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\sum_{k=2}^{K}\bar{\gamma}(G_\gamma)\eta_k + N + (N-\bar{\gamma}(G_\gamma))(3\gamma-1) \\ +\sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\frac{8\sigma(\xi+1)}{\bar{\Delta}^2}\log T + 2K\sum_{i=1}^{N}(d_{\gamma-1}^{(i)}+1)\sum_{t=1}^{T}\frac{1}{\log\zeta}\frac{\log\left((d_{\gamma}^{(i)}+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}\end{aligned} \tag{9.64}$$

Recall that $\eta_k = \frac{8\sigma(\xi+1)}{\Delta_k^2}\log T$. Thus the proof of Theorem 16 follows from (9.64) and Lemma 16.

9.10.7 Regret Under Full Communication

In this section we provide theoretical bounds for group regret of Full-UCB, Full-MPUCB and Full-LFUCB as follows.

9.10.8 Group Regret for Full-UCB

We start by proving a Lemma similar to Lemma 5.

Lemma 10. *Let* $\eta_k = \left(\frac{8(\xi+1)\sigma^2}{\Delta_k^2}\right)\log T$. *Let* $\mathcal{C}$ *be a non overlapping clique covering and* $\bar{\chi}(G)$ *be the clique covering number of the graph* G. *Let* $\tau_{k,\mathcal{C}}$ *be the maximum time*

step such that the total number of pulls from arm k by agents in the clique $\mathcal{C} \in \mathcal{C}$ is at most η_k. Define $\tau_k := \min_{\mathcal{C}} \tau_{k,\mathcal{C}}$. Then we have

$$\sum_{i=1}^{N} \mathbb{E}[n_k^{(i)}(T)] \leq \bar{\chi}(G)\eta_k + N + \sum_{i=1}^{N} \sum_{t>\tau_k}^{T-1} \left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t) \right) + \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t) \right) \right]$$

Proof. Let $\mathcal{C}$ be a non overlapping clique covering of the graph G. Then we have

$$\sum_{i=1}^{N} \mathbb{E}[n_k^{(i)}(T)] = \sum_{\mathcal{C} \in \mathcal{C}} \sum_{i \in \mathcal{C}} \sum_{t=1}^{T} \mathsf{P}\left(A_t^{(i)} = k \right) \tag{9.65}$$

Let $\tau_{k,\mathcal{C}}$ be the maximum time step such that the total number of pulls from arm k by agents in the clique $\mathcal{C}$ is at most η_k. This can be stated as $\tau_{k,\mathcal{C}} := \max\left\{ t \in [T] : \sum_{i \in \mathcal{C}} \sum_{\tau=1}^{t} \mathbf{1}\left\{ A_\tau^{(i)} = k \right\} \leq \eta_k \right\}$. Further for all $i \in \mathcal{C}$ we have $N_k^{(i)}(t) > \eta_k, \forall t > \tau_{k,\mathcal{C}}$. We analyse the expected number of times all agents pull suboptimal arm k as follows.

$$\sum_{\mathcal{C} \in \mathcal{C}} \sum_{i \in \mathcal{C}} \sum_{t=1}^{T} \mathbf{1}\left\{ A_t^{(i)} = k \right\} = \sum_{\mathcal{C} \in \mathcal{C}} \sum_{i \in \mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathbf{1}\left\{ A_t^{(i)} = k \right\} + \sum_{\mathcal{C} \in \mathcal{C}} \sum_{i \in \mathcal{C}} \sum_{t>\tau_{k,\mathcal{C}}}^{T} \mathbf{1}\left\{ A_t^{(i)} = k, N_k^{(i)}(t-1) > \eta_k \right\} \tag{9.66}$$

Taking the expectation of (9.66) we have

$$\begin{aligned} \sum_{\mathcal{C} \in \mathcal{C}} \sum_{i \in \mathcal{C}} \sum_{t=1}^{T} \mathsf{P}\left(A_t^{(i)} = k \right) &= \sum_{\mathcal{C} \in \mathcal{C}} \sum_{i \in \mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathsf{P}\left(A_t^{(i)} = k \right) + \sum_{\mathcal{C} \in \mathcal{C}} \sum_{i \in \mathcal{C}} \sum_{t>\tau_{k,\mathcal{C}}}^{T} \mathsf{P}\left(A_t^{(i)} = k, N_k^{(i)}(t-1) > \eta_k \right) \\ &\leq \bar{\chi}(G)\eta_k + N + \sum_{\mathcal{C} \in \mathcal{C}} \sum_{i \in \mathcal{C}} \sum_{t>\tau_{k,\mathcal{C}}}^{T-1} \mathsf{P}\left(A_{t+1}^{(i)} = k, N_k^{(i)}(t) > \eta_k \right) \end{aligned} \tag{9.67}$$

Let $\tau_k := \min_{\mathcal{C}} \tau_{k,\mathcal{C}\in\mathcal{C}}$. Similarly to Lemma 5 from (9.65), (9.67) and Lemma 13 we have

$$\sum_{i=1}^{N} \mathbb{E}[n_k^{(i)}(T)] \leq \bar{\chi}(G)\eta_k + N + \sum_{i=1}^{N}\sum_{t>\tau_k}^{T-1} \left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right)\right]$$

This concludes the proof of Lemma 10. □

Then from Lemmas 15, 16 and 10 it follows that

$$\mathbb{E}\left[R(T)\right] = O\left(K\bar{\chi}(G)\log T + KN\right).$$

9.10.9 Group Regret for Full-MPUCB

We start by proving a Lemma similar to Lemma 8.

Lemma 11. *Let $\eta_k = \left(\frac{8(\xi+1)\sigma^2}{\Delta_k^2}\right)\log T$. Let $\mathcal{C}_\gamma$ be a non overlapping clique covering and $\bar{\chi}(G_\gamma)$ be the clique covering number of the graph G_γ, which is the γ^{th} power graph of G. Let $\tau_{k,\mathcal{C}}$ be the maximum time step such that the total number of pulls from arm k by agents in the clique $\mathcal{C} \in \mathcal{C}_\gamma$ is at most $\eta_k + (|\mathcal{C}-1|)(\gamma-1)$. Define $\tau_k := \min_{\mathcal{C}} \tau_{k,\mathcal{C}}$. Then we have*

$$\sum_{i=1}^{N} \mathbb{E}[n_k^{(i)}(T)] \leq \bar{\chi}(G_\gamma)\eta_k + N + (N - \bar{\chi}(G_\gamma))(\gamma-1)$$
$$+ \sum_{i=1}^{N}\sum_{t>\tau_k}^{T-1} \left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right)\right]$$

Proof. Let $\mathcal{C}_\gamma$ be a non overlapping clique covering of the graph G_γ. Then we have

$$\sum_{i=1}^{N} \mathbb{E}[n_k^{(i)}(T)] = \sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T} \mathsf{P}\left(A_t^{(i)} = k\right) \tag{9.68}$$

Let $\tau_{k,\mathcal{C}_\gamma}$ be the maximum time step such that the total number of pulls from arm k by agents in the clique $\mathcal{C}$ is at most η_k. This can be stated as $\tau_{k,\mathcal{C}} := \max\left\{t \in [T] : \sum_{i\in\mathcal{C}} \sum_{\tau=1}^{t} \mathbf{1}\left\{A_\tau^{(i)} = k\right\} \leq \eta_k + (|\mathcal{C}|-1)(\gamma-1)\right\}$. Further for all $i \in \mathcal{C}$ we have $N_k^{(i)}(t) > \eta_k, \forall t > \tau_{k,\mathcal{C}}$. We analyse the expected number of times all agents pull suboptimal arm k as follows.

$$\sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathbf{1}\left\{A_t^{(i)}=k\right\} = \sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\mathbf{1}\left\{A_t^{(i)}=k\right\} + \sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{T}\mathbf{1}\left\{A_t^{(i)}=k, N_k^{(i)}(t-1)>\eta_k\right\} \tag{9.69}$$

Taking the expectation of (9.69) we have

$$\begin{aligned}\sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(i)}=k\right) &= \sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\mathsf{P}\left(A_t^{(i)}=k\right) + \sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{T}\mathsf{P}\left(A_t^{(i)}=k, N_k^{(i)}(t-1)>\eta_k\right)\\ &\leq \bar{\chi}(G_\gamma)\eta_k + N + (N-\bar{\chi}(G_\gamma))(\gamma-1) \qquad (9.70)\\ &+ \sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{T-1}\mathsf{P}\left(A_{t+1}^{(i)}=k, N_k^{(i)}(t)>\eta_k\right) \qquad (9.71)\end{aligned}$$

Let $\tau_k := \min_{\mathcal{C}\in\mathcal{C}_\gamma}\tau_{k,\mathcal{C}}$. Similarly to Lemma 8 from (9.68), (9.71) and Lemma 13 we have

$$\begin{aligned}\sum_{i=1}^{N}\mathbb{E}[n_k^{(i)}(T)] &\leq \bar{\chi}(G_\gamma)\eta_k + N + (N-\bar{\chi}(G_\gamma))(\gamma-1)\\ + \sum_{i=1}^{N}\sum_{t>\tau_k}^{T-1}&\left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t) \leq \mu_1 - C_1^{(i)}(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t) \geq \mu_k + C_k^{(i)}(t)\right)\right]\end{aligned}$$

This concludes the proof of Lemma 11. □

Then from Lemmas 15, 16 and 11 it follows that

$$\mathbb{E}\left[R(T)\right] = O\left(K\bar{\chi}(G_\gamma)\log T + KN\right).$$

9.10.10 Group Regret for Full-LFUCB

We begin the proof by providing a lemma similar to Lemma 9.

Lemma 12. *Let $\bar{\gamma}(G_\gamma)$ is the clique covering number of graph G_γ. Let $\eta_k = \left(\frac{8(\xi+1)\sigma^2}{\Delta_k^2}\right)\log T$. Then we have*

$$\sum_{i=1}^{N}\mathbb{E}[n_k^{(i)}(T)] \leq \bar{\gamma}(G_\gamma)\eta_k + N + (N-\bar{\gamma}(G_\gamma))(3\gamma-1) + \sum_{i\in V_\gamma'}\left(\left|\mathcal{N}_\gamma^{(i)}\right|+1\right)\sum_{t>\tau_k^{(i)}}^{T-1}\left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t)\leq \mu_1 - C_1^{(i)}(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^{(i)}(t)\geq \mu_k + C_k^{(i)}(t)\right)\right]$$

where V_γ' is the maximal dominating set of G_γ and $\mathcal{N}_\gamma^{(i)}$ is the set of followers of leader i. Here $\tau_k^{(i)}$ be the maximum time step such that the total number of times agent i pulls arm k and the number of times agents in $\mathcal{N}_\gamma^{(i)}$ pull from arm k is at most $\eta_k + \mathcal{N}_\gamma^{(i)}(\gamma-1)$.

Proof. Recall that V_γ' is the maximal dominating set of G_γ. Let $\mathcal{N}_\gamma^{(i)}$ be the set of followers of leader i. Then for each suboptimal arm $k>1$ we have

$$\sum_{i=1}^{N}\mathbb{E}[n_k^{(i)}(T)] = \sum_{i\in V_\gamma'}\left(\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(i)}=k\right) + \sum_{j\in\mathcal{N}_\gamma^{(i)}}\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(j)}=k\right)\right) \tag{9.72}$$

Let $\tau_k^{(i)}$ be the maximum time step such that the total number of times agent i pulls arm k and the number of times agents in $\mathcal{N}_\gamma^{(i)}$ pull from arm k is at most $\eta_k + \mathcal{N}_\gamma^{(i)}(\gamma-1)$. This can be stated as

$$\tau_k^{(i)} := \max\left\{t\in[T] : \sum_{\tau=1}^{t}\mathbf{1}\{A_\tau^{(i)}=k\} + \sum_{j\in\mathcal{N}_\gamma^{(i)}}\sum_{\tau=1}^{t}\mathbf{1}\{A_\tau^{(i)}=k\} \leq \eta_k + \mathcal{N}_\gamma^{(i)}(\gamma-1)\right\}.$$

Then we have $N_k^{(i)}(t) > \eta_k, \forall t > \tau_k^{(i)}$. We analyse the expected number of times all agents pull suboptimal arm k as follows. Let $d(i,j)$ be the distance between agents i and j in graph G. Then note that for any $j \in \mathcal{N}_\gamma^{(i)}$ we have $A_t^{(j)} = A_{t-d(i,j)}^{(i)}$ and $d(i,j) \leq \gamma$.

$$\sum_{i \in V'_\gamma} \left\{ \sum_{t=1}^{T} \mathbf{1}\Big\{A_t^{(i)} = k\Big\} + \sum_{j \in \mathcal{N}_\gamma^{(i)}} \sum_{t=1}^{T} \mathbf{1}\Big\{A_t^{(j)} = k\Big\} \right\} \leq \sum_{i \in V'_\gamma} \left\{ \sum_{t=1}^{\tau_k^{(i)}} \mathbf{1}\Big\{A_t^{(i)} = k\Big\} + \sum_{j \in \mathcal{N}_\gamma^{(i)}} \sum_{t=d(i,j)}^{\tau_k^{(i)}} \mathbf{1}\Big\{A_t^{(j)} = k\Big\} \right\}$$
$$+ \sum_{i \in V'_\gamma} \left\{ \sum_{t > \tau_k^{(i)}}^{T} \mathbf{1}\Big\{A_t^{(i)} = k\Big\} + \sum_{j \in \mathcal{N}_\gamma^{(i)}} \sum_{t > \tau_k^{(i)}}^{T-d(i,j)} \mathbf{1}\Big\{A_t^{(i)} = k\Big\} \right\} + \sum_{i \in V'_\gamma} \sum_{j \in \mathcal{N}_\gamma^{(i)}} 2d(i,j)$$
$$\leq \sum_{i \in V'_\gamma} \left\{ \sum_{t=1}^{\tau_k^{(i)}} \mathbf{1}\Big\{A_t^{(i)} = k\Big\} + \sum_{j \in \mathcal{N}_\gamma^{(i)}} \sum_{t=d(i,j)}^{\tau_k^{(i)}} \mathbf{1}\Big\{A_t^{(j)} = k\Big\} \right\}$$
$$+ \sum_{i \in V'_\gamma} \left(\left| \mathcal{N}_\gamma^{(i)} \right| + 1 \right) \sum_{t > \tau_k^{(i)}}^{T} \mathbf{1}\Big\{A_t^{(i)} = k\Big\} + 2(N - \bar{\gamma}(G_\gamma))\gamma \quad (9.73)$$

Now we proceed to upper bound the first two terms of right hand side of (9.73) as follows. Note that we have

$$\sum_{i \in V'_\gamma} \left\{ \sum_{t=1}^{\tau_k^{(i)}} \mathbf{1}\Big\{A_t^{(i)} = k\Big\} + \sum_{j \in \mathcal{N}_\gamma^{(i)}} \sum_{t=d(i,j)}^{\tau_k^{(i)}} \mathbf{1}\Big\{A_t^{(j)} = k\Big\} \right\}$$
$$\leq \sum_{i \in V'_\gamma} \left(\eta_k + \mathcal{N}_\gamma^{(i)}(\gamma - 1) \right) + \sum_{i \in V'_\gamma} \left| \mathcal{N}_\gamma^{(i)} \right| \sum_{t=1}^{\tau_k^{(i)}} \mathbf{1}\Big\{A_t^{(i)} = k\Big\} \quad (9.74)$$

Taking the expectation of (9.73) and (9.74) we have

$$\sum_{i \in V'_\gamma} \left(\sum_{t=1}^{T} \mathsf{P}\left(A_t^{(i)} = k\right) + \sum_{j \in \mathcal{N}_\gamma^{(i)}} \sum_{t=1}^{T} \mathsf{P}\left(A_t^{(j)} = k\right) \right) \leq \bar{\gamma}(G_\gamma)\eta_k + N + (N - \bar{\gamma}(G_\gamma))(3\gamma - 1)$$

$$+\sum_{i\in V'_\gamma}\left(\left|\mathcal{N}_\gamma^{(i)}\right|+1\right)\sum_{t>\tau_k^{(i)}}^{T}\mathsf{P}\left(A_t^{(i)}=k\right)+\sum_{i\in V'_\gamma}\left|\mathcal{N}_\gamma^{(i)}\right|\sum_{t>\tau_k^{(i)}}^{\tau_k^{(i)}}\mathsf{P}\left(A_t^{(i)}=k\right) \tag{9.75}$$

Recall that $N_k^{(i)}(t)>\eta_k, \forall t>\tau_k^{(i)}$. Thus from (9.75) and Lemma 13 we have

$$\sum_{i\in V'_\gamma}\left(\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(i)}=k\right)+\sum_{j\in\mathcal{N}_\gamma^{(i)}}\sum_{t=1}^{T}\mathsf{P}\left(A_t^{(j)}=k\right)\right)\leq\bar{\gamma}(G_\gamma)\eta_k+(N-\bar{\gamma}(G_\gamma))(3\gamma-1)$$
$$+\sum_{i\in V'_\gamma}\left(\left|\mathcal{N}_\gamma^{(i)}\right|+1\right)\sum_{t>\tau_k^{(i)}}^{T-1}\left[\mathsf{P}\left(\widehat{\mu}_1^{(i)}(t)\leq\mu_1-C_1^{(i)}(t)\right)+\mathsf{P}\left(\widehat{\mu}_k^{(i)}(t)\geq\mu_k+C_k^{(i)}(t)\right)\right]. \tag{9.76}$$

The proof of Lemma 12 follows from (9.72) and (9.76). □

Then from Lemmas 15, 16 and 12 it follows that

$$\mathbb{E}\left[R(T)\right]=O\left(K\bar{\gamma}(G_\gamma)\log T+KN\right).$$

9.10.11 Additional Experimental Results

In this section we provide additional simulation results. We observe that performance of the algorithms improve when we decrease ξ. Thus for simulations provided in this section we use $\xi=1.001$. Further when γ is increased communication density increases and performance improve. For simulations provided in this section we consider $\gamma=7$. We use the same graph structure and reward structures used in the results provided in the main paper.

Additional details on estimate sharing Note that in estimate sharing agents average their estimates of instantaneously suboptimal arms at every time step. Thus at each time step each agent creates $2K$ number of messages (estimated sum of re-

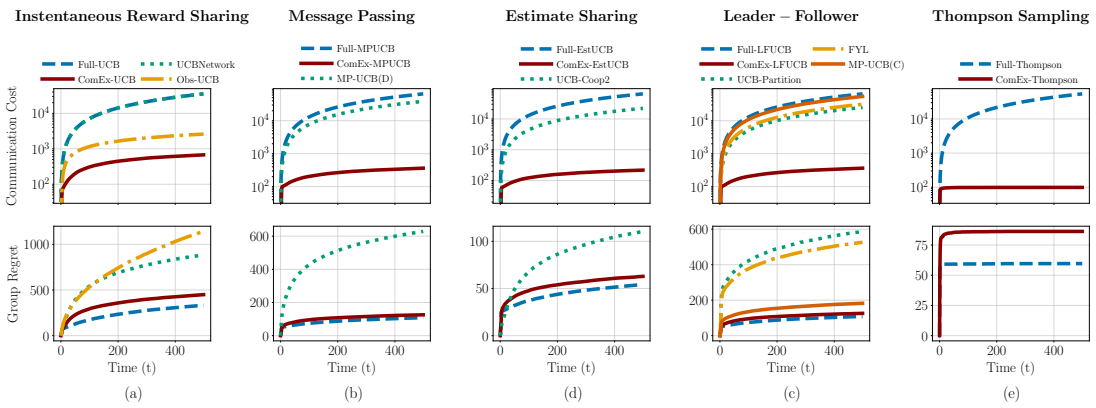


Figure 9.3: A comparison of expected cumulative group regret and communication cost of our algorithms and existing state-of-the-art algorithms in several benchmark cooperative bandit frameworks.

wards for each arm and estimated number of pulls from each arm). If the number of arms are of same order as time horizon this leads to $O(T^2)$ cost for Full-EstUCB and $O(T \log T)$ cost for ComEx-EstUCB. However we consider that nummber of arms are fixed and $K << T$ for large T and when providing simulation results for the communication cost we only considered the communication cost associated with initiating messages and passing them through network neglecting the dependence on number of arms. This leads to $O(T)$ cost for Full-EstUCB and $O(\log T)$ cost for ComEx-EstUCB.

9.10.12 Pseudo code of ComEx-UCB

9.11 Pseudo code of ComEx-MPUCB

9.12 Pseudo code of ComEx-LFUCB

For all $i \in V'_\gamma$ the indicator variable $I_t^{(i)}$ takes value 1 if $A_t^{(i)}$ is instantaneously suboptimal.

Algorithm 3: ComEx-UCB

Input: Arms $k \in [K]$, variance proxy upper bound σ^2, parameter ξ
Initialize: $N_k^{(i)}(0) = \widehat{\mu}_k^{(i)}(0) = C_k^{(i)}(0) = 0, \forall k, i$
for each iteration $t \in [T]$ **do**
  $E_t \leftarrow \emptyset$
  for each agent $i \in [N]$ **do**
    /* Sampling phase */
    if $t = 1$ **then**
      $A_t^{(i)} \leftarrow \textsc{RandomArm}([K])$
    end
    else
      $A_t^{(i)} \leftarrow \arg\max_k \widehat{\mu}_k^{(i)}(t-1) + C_k^{(i)}(t-1)$
    end
    /* Send messages */
    if $A_t^{(i)} \neq \arg\max_k \widehat{\mu}_k^{(i)}(t-1)$ **then**
      $\textsc{Send}\left(m_t^{(i)} := \left\langle A_t^{(i)}, X_t^{(i)} \right\rangle\right)$
    end
    $\mathbf{m}_t^{(i)} \leftarrow m_t^{(i)}$
  end
  for each agent $i \in [N]$ **do**
    /* Receive messages */
    for each neighbor j s. t. $\{(j \to i) \in E\}$ **do**
      $\mathbf{m}_t^{(i)} \leftarrow \mathbf{m}_t^{(i)} \cup m_t^{(j)}$
    end
    for each agent $j \in [N]$ **do**
      if $m_t^{(j)} \in \mathbf{m}_t^{(i)}$ **then**
        $E_t \leftarrow E_t \cup \{(j \to i)\}$
      end
    end
    /* Update estimates */
    for each arm $k \in [K]$ **do**
      $\textsc{Calculate}\left(N_k^{(i)}(t), \widehat{\mu}_k^{(i)}(t), C_k^{(i)}(t)\right)$
    end
  end
end

9.13 Pseudo code of ComEx-EstUCB

Let $\widehat{N}_k^{(i)}(t)$ be the estimated number of pulls from arm k for agent i up to time t.

9.14 Pseudo code of ComEx-MPThompson

In Thompson sampling for each arm k each agent i maintains a posterior distribution $\phi_k^{(i)}$ and updates the distribution according to the available information. Then draw samples from the posterior distribution and pull the arm with highest sample value.

Algorithm 4: ComEx-MPUCB

Input: Arms $k \in [K]$, variance proxy upper bound σ_k^2, parameter ξ, γ
Initialize: $N_k^{(i)}(0) = \widehat{\mu}_k^{(i)}(0) = C_k^{(i)}(0) = 0, \forall k, i$
for each iteration $t \in [T]$ **do**
  $E_t \leftarrow \emptyset$
  for each agent $i \in [N]$ **do**
    /* Sampling phase */
    if $t = 1$ **then**
      $A_t^{(i)} \leftarrow \textsc{RandomArm}([K])$
    end
    else
      $A_t^{(i)} \leftarrow \arg\max_k \widehat{\mu}_k^{(i)}(t-1) + C_k^{(i)}(t-1)$
    end
    /* Send messages */
    if $A_t^{(i)} \neq \arg\max_k \widehat{\mu}_k^{(i)}(t-1)$ **then**
      $\textsc{Create}\left(m_t^{(i)} := \left\langle i, t, A_t^{(i)}, X_t^{(i)} \right\rangle\right)$
      $\mathbf{m}_t^{(i)} \leftarrow m_t^{(i)}$
    end
    $\textsc{Send}\left(\mathbf{m}_t^{(i)}\right)$
  end
  for each agent $i \in [N]$ **do**
    /* Receive messages */
    for each neighbor j s. t. $\{(j \to i) \in E\}$ **do**
      $\mathbf{m}_t^{(i)} \leftarrow \mathbf{m}_t^{(i)} \cup m_t^{(j)}$
    end
    /* Discard messages older than γ */
    for each neighbor $j \in [N]$ **do**
      $\mathbf{m}_t^{(i)} \leftarrow \mathbf{m}_t^{(i)} \backslash m_\tau^{(j)}, \forall \tau$ s. t. $\tau < t - \gamma$
      for each time step $\tau \in \{t - \gamma + 1, \ldots, t\}$ **do**
        if $m_\tau^{(j)} \in \mathbf{m}_t^{(i)}$ **then**
          $E_\tau \leftarrow E_\tau \cup \{(j \to i)\}$
        end
      end
    end
    /* Update estimates */
    for each arm $k \in [K]$ **do**
      $\textsc{Calculate}\left(N_k^{(i)}(t), \widehat{\mu}_k^{(i)}(t), C_k^{(i)}(t)\right)$
    end
    $\mathbf{m}_{t+1}^{(i)} \leftarrow \mathbf{m}_t^{(i)}$
  end
end

Algorithm 5: ComEx-LFUCB

Input: Arms $k \in [K]$, variance proxy upper bound σ_k^2, parameter ξ, γ

Initialize: $N_k^{(i)}(0) = \widehat{\mu}_k^{(i)}(0) = C_k^{(i)}(0) = 0, \forall k, i$

for each iteration $t \in [T]$ **do**

$E_t \leftarrow \emptyset$

for each agent $i \in V_\gamma'$ **do**

/* Sampling phase */

Same as ComEx-MPUCB

/* Send messages */

$\textsc{Create}\left(m_t^{(i)} := \left\langle i, t, A_t^{(i)}, I_t^{(i)} \right\rangle\right)$

$\mathbf{m}_t^{(i)} \leftarrow m_t^{(i)}$

$\textsc{Send}\left(\mathbf{m}_t^{(i)}\right)$

for each agent $j \in \mathcal{N}_\gamma^{(i)}$ **do**

/* Sampling phase */

if $t < d(i,j)$ **then**

$A_t^{(i)} \leftarrow \textsc{RandomArm}([K])$

end

else

$A_t^{(j)} \leftarrow A_{t-d(i,j)}^{(i)}$

end

if $I_{t-d(i,j)}^{(i)} = 1$ **then**

$\textsc{Create}\left(m_t^{(j)} := \left\langle j, t, A_t^{(j)}, X_t^{(j)} \right\rangle\right)$

$\mathbf{m}_t^{(j)} \leftarrow m_t^{(j)}$

end

end

end

for each agent $i \in V_\gamma'$ **do**

/* Receive messages */

for each neighbor j s. t. $\{(j \to i) \in E\}$ **do**

$\mathbf{m}_t^{(i)} \leftarrow \mathbf{m}_t^{(i)} \cup m_t^{(j)}$

end

/* Discard messages older than γ */

for each neighbor $j \in [N]$ **do**

$\mathbf{m}_t^{(i)} \leftarrow \mathbf{m}_t^{(i)} \backslash m_\tau^{(j)}, \forall \tau$ s. t. $\tau < t - \gamma$

for each time step $\tau \in \{t - \gamma + 1, \ldots, t\}$ **do**

if $m_\tau^{(j)} \in \mathbf{m}_t^{(i)}$ **then**

$E_\tau \leftarrow E_\tau \cup \{(j \to i)\}$

end

end

end

/* Update estimates */

for each arm $k \in [K]$ **do**

$\textsc{Calculate}\left(N_k^{(i)}(t), \widehat{\mu}_k^{(i)}(t), C_k^{(i)}(t)\right)$

end

$\mathbf{m}_{t+1}^{(i)} \leftarrow \mathbf{m}_t^{(i)}$

end

Algorithm 6: ComEx-EstUCB

Input: Arms $k \in [K]$, variance proxy upper bound σ_k^2, parameter ξ, γ
Initialize: $\widehat{N}_k^{(i)}(0) = \widehat{\mu}_k^{(i)}(0) = C_k^{(i)}(0) = 0, \forall k, i$
for each iteration $t \in [T]$ **do**
  $E_t \leftarrow \emptyset$
  for each agent $i \in [N]$ **do**
    /* Sampling phase */
    if $t = 1$ **then**
      $A_t^{(i)} \leftarrow \textsc{RandomArm}([K])$
    end
    else
      $A_t^{(i)} \leftarrow \arg\max_k \widehat{\mu}_k^{(i)}(t-1) + C_k^{(i)}(t-1)$
    end
    /* Send messages */
    if $A_t^{(i)} \neq \arg\max_k \widehat{\mu}_k^{(i)}(t-1)$ **then**
      $\textsc{Create}\left(m_t^{(i)} := \left\langle i, t, \widehat{N}_t^{(i)}, \widehat{\mu}_k^{(i)}(t-1) \right\rangle\right)$
      $\mathbf{m}_t^{(i)} \leftarrow m_t^{(i)}$
    end
    $\textsc{Send}\left(\mathbf{m}_t^{(i)}\right)$
  end
  for each agent $i \in [N]$ **do**
    /* Receive messages */
    for each neighbor j s. t. $\{(j \to i) \in E\}$ **do**
      $\mathbf{m}_t^{(i)} \leftarrow \mathbf{m}_t^{(i)} \cup m_t^{(j)}$
    end
    /* Discard messages older than γ */
    for each neighbor $j \in [N]$ **do**
      $\mathbf{m}_t^{(i)} \leftarrow \mathbf{m}_t^{(i)} \backslash m_\tau^{(j)}, \forall \tau$ s. t. $\tau < t - \gamma$
      for each time step $\tau \in \{t - \gamma + 1, \ldots, t\}$ **do**
        if $m_\tau^{(j)} \in \mathbf{m}_t^{(i)}$ **then**
          $E_\tau \leftarrow E_\tau \cup \{(j \to i)\}$
        end
      end
    end
    /* Update estimates */
    for each arm $k \in [K]$ **do**
      $\textsc{Calculate}\left(\widehat{N}_k^{(i)}(t), \widehat{\mu}_k^{(i)}(t), C_k^{(i)}(t)\right)$
      according to consensus algorithm
    end
    $\mathbf{m}_{t+1}^{(i)} \leftarrow \mathbf{m}_t^{(i)}$
  end
end

Algorithm 7: ComEx-MPThompson

Input: Arms $k \in [K]$, parameter γ

Initialize: $\phi_k^{(i)}(0), \forall k, i$

for each iteration $t \in [T]$ **do**

  $E_t \leftarrow \emptyset$

  for each agent $i \in [N]$ **do**

    /* Sampling phase */

    for each arm $k \in [K]$ **do**

      $y_k^{(i)}(t) \sim \phi_k^{(i)}(t-1)$

      $A_t^{(i)} \leftarrow \arg\max_k y_k^{(i)}(t)$

    end

    /* Send messages */

    $\textsc{Create}\left(m_t^{(i)} := \left\langle i, t, A_t^{(i)}, X_t^{(i)} \right\rangle\right)$

    $\mathbf{m}_t^{(i)} \leftarrow m_t^{(i)}$

    $\textsc{Send}\left(\mathbf{m}_t^{(i)}\right)$

  end

  for each agent $i \in [N]$ **do**

    /* Receive messages */

    for each neighbor j s. t. $\{(j \to i) \in E\}$ **do**

      $\mathbf{m}_t^{(i)} \leftarrow \mathbf{m}_t^{(i)} \cup m_t^{(j)}$

    end

    /* Discard messages older than γ */

    for each neighbor $j \in [N]$ **do**

      $\mathbf{m}_t^{(i)} \leftarrow \mathbf{m}_t^{(i)} \backslash m_\tau^{(j)}, \forall \tau$ s. t. $\tau < t - \gamma$

      for each time step $\tau \in \{t-\gamma+1, \ldots, t\}$ **do**

        if $m_\tau^{(j)} \in \mathbf{m}_t^{(i)}$ **then**

          $E_\tau \leftarrow E_\tau \cup \{(j \to i)\}$

        end

      end

    end

    /* Update estimates */

    for each arm $k \in [K]$ **do**

      $\textsc{Calculate}\left(\phi_k^{(i)}(t)\right)$

    end

    $\mathbf{m}_{t+1}^{(i)} \leftarrow \mathbf{m}_t^{(i)}$

  end

end

Chapter 10

One More Step Towards Reality: Cooperative Bandits with Imperfect Communication

Udari Madhushani, Abhimanyu Dubey, Naomi Ehrich Leonard and Alex Pentland

The cooperative bandit problem is increasingly becoming relevant due to its applications in large-scale decision-making. However, most research for this problem focuses exclusively on the setting with perfect communication, whereas in most real-world distributed settings, communication is often over stochastic networks, with arbitrary corruptions and delays. In this paper, we study cooperative bandit learning under three typical real-world communication scenarios, namely, (a) message-passing over stochastic time-varying networks, (b) instantaneous reward-sharing over a network with random delays, and (c) message-passing with adversarially corrupted rewards, including byzantine communication. For each of these environments, we propose decentralized algorithms that achieve competitive performance, along with near-optimal guarantees on the incurred group regret as well. Furthermore, in the set-

ting with perfect communication, we present an improved delayed-update algorithm that outperforms the existing state-of-the-art on various network topologies. Finally, we present tight network-dependent minimax lower bounds on the group regret. Our proposed algorithms are straightforward to implement and obtain competitive empirical performance.

10.1 Introduction

The cooperative multi-armed bandit problem involves a group of N agents collectively solving a multi-armed bandit while communicating with one another. This problem is relevant for a variety of applications that involve decentralized decision-making, for example, in distributed controls and robotics [81] and communication [43]. In the typical formulation of this problem, a group of agents are arranged in a network $G = (\mathcal{V}, \mathcal{E})$, wherein each agent interacts with the bandit, and communicates with its neighbors in G, to maximize the cumulative reward.

A large body of recent work on this problem assumes the communication network G to be fixed [42, 49]. Furthermore, these algorithms inherently require precise communication, as they construct careful confidence intervals for cumulative arm statistics across agents, e.g., for stochastic bandits, it has been shown that the standard `UCB1` algorithm [7] with a $n$eighborhood confidence interval is close to optimal [19, 42], and correspondingly, for adversarial bandits, a neighborhood-weighted loss estimator can be utilized with the `EXP3` algorithm to provide competitive regret [13]. Such approaches are indeed feasible when communication is perfect, e.g., the network G is fixed, and messages are not lost or corrupted. In real-world environments, however, this is rarely true: messages can be lost, agents can be byzantine, and communication networks are rarely static [53]. This aspect has hence received much attention in the distributed optimization literature [99]. However, contrary to network optimization

where dynamics in communication can behave synergistically [34], bandit problems additionally bring a decision-making component requiring an explore-exploit trade-off. As a result, external randomness and corruption are incompatible with the default optimal approaches, and require careful consideration [91, 57]. This motivates us to study the multi-agent bandit problem under real-world communication, which regularly exhibits external randomness, delays and corruptions. Our key contributions include the following.

Contributions . We provide a set of algorithms titled Robust Communication Learning (`RCL`) for the cooperative stochastic bandit under three real-world communication scenarios.

First, we study stochastic communication, where the communication network G is time-varying, with each edge being present in G with an unknown probability p. For this setting, we present a `UCB`-like algorithm, `RCL-LF` (Link Failures), that directs agent i to discard messages with an additional probability of $1 - p_i$ in order to control the bias in the (stochastic) reward estimates. `RCL-LF` obtains a group regret of $\mathcal{O}\left(\left(\sum_{i=1}^{N}(1 - p \cdot p_i) + \sum_{\mathcal{C}\in\mathcal{C}}(\max_{i\leq\mathcal{C}} p_i) \cdot p\right)\left(\sum_{k=1}^{K} \frac{\log T}{\Delta_k}\right)\right)$, where $\mathcal{C}$ is a non overlapping clique covering of G, T is time horizon, and Δ_k is the difference in reward mean between the optimal and kth arm. The regret exhibits a smooth interpolation between known rates for no communication ($p = 0$) and perfect communication ($p = 1$).

Second, we study the case where messages from any agent can be delayed by a random (but bounded) number of trials τ with expectation $\mathbb{E}[\tau]$. For this setting, simple reward-sharing with a natural extension of the `UCB` algorithm (`RCL-SD` (Stochastic Delays)) obtains a regret of

$$\mathcal{O}\left(\bar{\chi}(G) \cdot \left(\sum_{k>1} \frac{\log T}{\Delta_k}\right) + \left(N \cdot \mathbb{E}[\tau] + \log(T) + \sqrt{N \cdot \mathbb{E}[\tau] \log(T)}\right) \cdot \sum_{k>1} \Delta_k\right)$$

, which is reminiscent of that of single-agent bandits with delays [38] (Remark 16). Here $\bar{\chi}(G)$ is the clique covering number of G.

Third, we study the *c*orrupted setting, where any message can be (perhaps in a byzantine manner) corrupted by an unknown (but bounded) amount ϵ. This setting presents the two-fold challenge of receiving feedback after (variable) delays as well as adversarial corruptions, making existing arm elimination [57, 15, 30] or cooperative estimation [19] methods inapplicable. We present algorithm `RCL-AC` (Adversarial Corruptions) that overcomes this issue by limiting exploration only to well-positioned agents in G, who explore using a *h*ybrid robust arm elimination and local confidence bound approach. `RCL-AC` obtains a regret of $\mathcal{O}\left(\psi(G_\gamma)\cdot\sum_{k=1}^{K}\frac{\log T}{\Delta_k}+N\sum_{k=1}^{K}\frac{\log\log T}{\Delta_k}+NTK\gamma\epsilon\right)$, where $\psi(G_\gamma)$ denotes the domination number of the γ graph power of G, which matches the rates obtained for corrupted single-agent bandits without knowledge of ϵ.

Finally, for perfect communication, we present a simple modification of cooperative `UCB1` that provides significant empirical improvements, and also provides minimax lower bounds on the group regret of algorithms based on message-passing.

Related Work. A variant of the networked adversarial bandit problem without communication constraints (e.g., delay, corruption) was studied first in the work of [8], who demonstrated an average regret bound of order $\sqrt{(1+{}^{K}/_{N})T}$. This line of inquiry was generalized to networked communication with at most γ rounds of delays in the work of [13], that demonstrate an average regret of order $\sqrt{(\gamma+{}^{\alpha(G_\gamma)}/_{N})KT}$ where $\alpha(G_\gamma)$ denotes the independence number of G_γ, the γ-power of network graph G. This line of inquiry has been complemented for the stochastic setting with problem-dependent analyses in the work of [42] and [19]. The former presents a `UCB1`-style algorithm with instantaneous reward-sharing that obtains a regret bound of $\mathcal{O}(\alpha(G)\cdot$

$\sum_{k=1}^{K} \frac{\log T}{\Delta_k}$) that was generalized to message-passing communication with delays in the latter.

Alternatively, [49] consider the multi-agent bandit where communication is done instead using a $r$unning consensus protocol, where neighboring agents average their reward estimates using the DeGroot consensus model [17]. This algorithm was refined in the work of [66] by a delayed mixing scheme that reduces the bias in the consensus reward estimates. A specific setting of Huber contaminated communication was explored in the work of [20]; however, in contrast to our algorithms, that work assumes that the total contamination likelihood is known a priori. Additionally, multi-agent networked bandits with stochastic communication was considered in [61, 64, 65], however, only for regular networks and multi-star networks.

Our work also relates to aspects of stochastic delayed feedback and corruptions in the context of single-agent multi-armed bandits. There has been considerable research in these areas, beginning from the early work of [98] that proposes running multiple bandit algorithms in parallel to account for (fixed) delayed feedback. [91] discuss the multi-armed bandit with stochastic delays, and provide algorithms using optimism indices based on the `UCB1` [7] and `KL-UCB` [25] approaches. Stochastic bandits with adversarial corruptions have also received significant attention recently. [57] present an arm elimination algorithm that provides a regret that scales linearly with the total amount of corruption, and present lower bounds demonstrating that the linear dependence is inevitable. This was followed up by [29] who introduce the algorithm `BARBAR` that improves the dependence on the corruption level by a better sampling of worse arms. Alternatively, [3] discuss best-arm identification under contamination, which is a weaker adversary compared to the one discussed in this paper. The corrupted setting discussed in our paper combines both issues of (variable) delayed feedback along with adversarial corruptions, and hence requires a novel approach.

Table 10.1: Quantity (with notation) for any graph G.

Average degree ($\bar{d}$)	Maximum degree ($d_{\max}$)	Degree of i (d_i)
Message life (γ)	Minimum degree ($d_{\min}$)	Neighborhood of i ($\mathcal{N}_i$)
k-power of G (G_k)	Diameter ($d_\star$)	$\mathcal{N}_i \cup \{i\}$ ($\mathcal{N}_i^+$)
Independence number (α)	Domination number (ψ)	Clique covering number ($\bar{\chi}$)

In another line of related work, Chawla *et al.*[15] discuss gossip-based communication protocols for cooperative multi-armed bandits. While the paper provides similar results, there are several differences in the setup considered in Chawla et al compared to our setup. First, we can see that Chawla *et al.*do not provide a uniform $\mathcal{O}(\frac{1}{N})$ speedup, but in fact, their regret depends on the difficulty of the first $\frac{K}{N}$ arms, which is a $\mathcal{O}(\frac{1}{N})$ speed up only when all arms are "uniformly" suboptimal, i.e., $\Delta_i \approx \Delta_j \forall i, j \in [K]$. In contrast, our algorithm will always provide a speed up of order $\frac{\alpha(G_\gamma)}{N}$ regardless of the arms themselves, and when we run our algorithm by setting the delay parameter $\gamma = d_\star(G)$ (diameter of the graph G), we obtain an $\mathcal{O}(\frac{1}{N})$ speedup regardless of the sparsity of G. Additionally, our constants (per-agent) scale as $\mathcal{O}(K)$ in the worst case, whereas Chawla et al obtain a constant between $\mathcal{O}(K + (\log N)^\beta)$ and $\mathcal{O}(K + N^\beta)$ for some $\beta \gg 1$, based on the graph structure, which can dominate the $\log T$ term when we have a large number of agents present.

10.2 Preliminaries

Notation (Table 10.1). We denote the set $a, ..., b$ as $[a, b]$, and as $[b]$ when $a = 1$. We define the indicator of a Boolean predicate x as $\mathbf{1}\{x\}$. For any graph G with diameter $d_\star(G)$, and any $1 \leq \gamma \leq d_\star(G)$, we define G_γ as the γ-power of G, i.e., the graph with edge (i, j) if i, j are at most a distance γ.

Problem Setting. We consider the cooperative stochastic multi-armed bandit problem with K arms and a group $\mathcal{V}$ of N agents. In each round $t \in [T]$, each agent

$i \in \mathcal{V}$ pulls an arm $A_i(t) \in [K]$ and receives a random reward $X_i(t)$ (realized as $r_i(t)$) drawn i.i.d. from the corresponding arm's distribution. We assume that each reward distribution is sub-Gaussian with an unknown mean μ_k and unknown variance proxy σ_k^2 upper bounded by a known constant σ^2. Without loss of generality we assume that $\mu_1 \geq \mu_2 \ldots \geq \mu_K$ and define $\Delta_k := \mu_1 - \mu_k, \forall k > 1$, to be the reward gap (in expectation) of arm k. Let $\overline{\Delta} := \min_{k>1} \Delta_k$ be the minimum expected reward gap. For brevity in our theoretical results, we define $g(\xi, \sigma) := 8(\xi + 1)\sigma^2 = o(1)$ and $f(M, G) := M \sum_{k>1} \Delta_k + 4 \sum_{i=1}^{N} (3 \log(3(d_i(G) + 1)) + (\log(d_i(G) + 1))) \cdot \sum_{k>1} \Delta_k = o((M + N \log N) \cdot \sum_{k>1} \Delta_k)$.

Networked Communication (Figure 10.1). Let $G = (\mathcal{V}, \mathcal{E})$ be a connected, undirected graph encoding the communication network, where $\mathcal{E}$ contains an edge (i, j) if agents i and j can communicate directly via messages with each other. After each round t, each agent j broadcasts a message $\mathbf{m}_j(t)$ to all their neighbors. Each message is forwarded at most γ times through G, after which it is discarded. For any value of $\gamma > 1$, the protocol is called *m*essage-passing [55], but for $\gamma = 1$ it is called *i*nstantaneous reward sharing, as this setting has no delays in communication.

Exploration Strategy (Figure 10.2). For Sections 10.3 and 10.4 we use a natural extension of the `UCB1` algorithm for exploration. Thus we modify `UCB1` [7] such that at each time step t for each arm k each agent i constructs an upper confidence bound, i.e., the sum of its estimated expected reward $\widehat{\mu}_k^i(t-1)$ (empirical average of all the observed rewards) and the uncertainty associated with the estimate $C_k^i(t-1) := \sigma \sqrt{\frac{2(\xi+1)\log t}{N_k^i(t-1)}}$ where $\xi > 1$, and pulls the arm with the highest bound.

Regret. The performance measure we consider, *g*roup regret, is a straightforward extension of *p*seudo regret for a single agent. Group regret is the regret (in expectation) incurred by the group $\mathcal{V}$ by pulling suboptimal arms. The group regret is given

by $\mathsf{Reg}_G(T) = \sum_{i=1}^{N} \sum_{k>1} \Delta_k \cdot \mathcal{E}\left[n_k^i(t)\right]$, where $n_k^i(t)$ is the number of times agent i pulls the suboptimal arm k up to (and including) round t.

Before presenting our algorithms and regret upper bounds we present some graph terminology.

Definition 7 (Clique covering number)**.** *A clique cover $\mathcal{C}$ of any graph $G = (\mathcal{V}, \mathcal{E})$ is a partition of $\mathcal{V}$ into subgraphs $C \in \mathcal{C}$ such that each subgraph C is fully connected, i.e., a clique. The size of the smallest possible covering $\mathcal{C}^\star$ is known as the clique covering number $\bar{\chi}(G)$.*

Definition 8 (Independence number)**.** *The independence number $\alpha(G)$ of $G = (\mathcal{V}, \mathcal{E})$ is the size of the largest subset of $\mathcal{V}_\alpha \subseteq \mathcal{V}$ such that no two vertices in $\mathcal{V}_\alpha$ are connected.*

Definition 9 (Domination number)**.** *The domination number $\psi(G)$ of $G = (\mathcal{V}, \mathcal{E})$ is the size of the smallest subset $\mathcal{V}_\psi \subseteq \mathcal{V}$ such that each vertex not in $\mathcal{V}_\psi$ is adjacent to at least one agent in $\mathcal{V}_\psi$.*

Organization. In this paper, we study three specific forms of communication errors. Section 10.3 discusses the case when, for both *m*essage-passing and *i*nstantaneous reward-sharing, any message forwarding fails independently with probability p, resulting in stochastic communication failures. Section 10.4 discusses the case when *i*nstantaneous reward-sharing incurs a random (but bounded) delay. Section 10.5 discusses the case when the outgoing reward from any message may be corrupted by an adversarial amount at most ϵ. Finally, in Section 10.6, we discuss an improved algorithm for the case with perfect communication and present minimax lower bounds on the problem. We present all proofs in the Appendix and present proof-sketches highlighting the central ideas in the main paper.

For $t = 1, 2, ...$ **each agent** $i \in \mathcal{V}$

1. Plays arm $A_i(t)$, gets reward $r_i(t)$, computes $\mathbf{m}_i(t) = \langle A_i(t), r_i(t), i, t \rangle$.
2. Adds $\mathbf{m}_i(t)$ to the set of messages $\mathbf{M}_i(t)$, broadcasts all messages in $\mathbf{M}_i(t)$ to its neighbors and receives messages $\mathbf{M}'_i(t)$ from its neighbors.
3. Computes $\mathbf{M}_i(t+1)$ from $\mathbf{M}'_i(t)$ by discarding all messages sent prior to round $t - \gamma$.

This is called *i*nstantaneous reward sharing for $\gamma = 1$ (no delays), and *m*essage-passing for $\gamma > 1$.

Figure 10.1: The cooperative bandit protocol with delay parameter γ.

For $t = 1, 2, ...,$ **each agent** $i \in \mathcal{V}$

1. Calculates, for each arm $k \in [K]$, $Q_k^i(t-1) = \widehat{\mu}_k^i(t-1) + \sigma\sqrt{\frac{2(\xi+1)\log(t-1)}{N_k^i(t-1)}}$, where $N_k^i(t-1)$ is the number of reward samples available for arm k at time t.
2. Plays arm $A_i(t) = \arg\max_k Q_k^i(t-1)$

Figure 10.2: Cooperative `UCB1` which uses additional arm pulls from messages.

10.3 Probabilistic Message Selection for Random Communication Failures

The fundamental advantage of cooperative estimation is the ability to leverage observations about suboptimal arms from neighboring agents to reduce exploration. However, when agents are communicating over an arbitrary graph, the amount of information an agent receives varies according to its connectivity in G. For example, agents with a large number of neighbors receive more information, leading them to begin exploitation earlier than agents with fewer neighbors. This means that well-connected agents exhibit better performance early on, but because they quickly do only exploiting, agents that are poorly connected typically only observe exploitative arm pulls, which requires them to explore for longer in order to obtain similarly good estimates for suboptimal arms, increasing their regret. The disparity between

performance in well-connected versus poorly connected agents is exacerbated in the presence of random *l*ink failures, where any message sent by an agent can fail to reach its recipient with a failure probability $1 - p$ (drawn i.i.d. for each message).

Indeed, it is natural to expect the group regret to decrease with decreasing link failure probability, i.e., increasing communication probability p. However, what we observe experimentally (Section 10.7) is that this holds only for graphs G that are *r*egular (i.e., each agent has the same degree), or close to regular. When G is irregular, as we increase p from 0 to 1, the group performance oscillates. While, in some cases, the improved performance in the well-connected agents can outweigh the degradation observed in the weakly-connected agents (leading to lower *g*roup regret), it is prudent to consider an approach that mitigates this disparity by regulating information flow in the network.

Information Regulation in Cooperative Bandits. Our approach to regulate information is straightforward: we direct each agent i to discard any incoming message with an agent-specific probability $1 - p_i$, while always utilizing its own observations. For specific values of p_i, we can obtain various weighted combinations of internal versus group observations. Our first algorithm `RCL-LF` (Link Failures) is built on this regulation strategy, coupled with `UCB1` exploration using all selected observations for each arm. Essentially, each agent runs `UCB1` using the cumulative set of observations it has received from its network. After pulling an arm, it broadcasts its pulled arm and reward through the network, but incorporates each incoming message *o*nly with a probability p_i. Pseudo code for the algorithm is given in the appendix. We first present a regret bound for `RCL-LF` when run with the *i*nstantaneous reward-sharing protocol.

Theorem 17 (`RCL-LF` Regret with instantaneous reward-sharing)**.** `RCL-LF` *running with the instantaneous reward-sharing protocol (Figure 10.1,* $\gamma = 1$*) obtains cumula-*

tive group regret of

$$\mathsf{Reg}_G(T) \leq g(\xi,\sigma)\left(\sum_{i=1}^{N}(1-p_i\cdot p)+\sum_{\mathcal{C}\in\mathcal{C}}(\max_{i\leq\mathcal{C}}p_i)\cdot p\right)\left(\sum_{k>1}\frac{\log T}{\Delta_k}\right)+f(5N,G)$$

where $\mathcal{C}$ is a non-overlapping clique covering of G.

Proof sketch. We follow an approach similar to the analysis of `UCB1` by [7] with several key modifications. First, we partition the communication graph G into a set of non-overlapping cliques and then analyze the regret of each clique. The group regret can be obtained by taking the summation of the regret over each clique. Two major technical challenges in proving the regret bound for `RCL-LF` are (a) deriving a tail probability bound for probabilistic communication, and (b) bounding the regret accumulated by agents by losing information due to communication failures and message discarding. We overcome the first challenge by noticing that communication is independent of the decision making process thus $\mathbb{E}\left(\exp\left(\lambda\sum_{\tau=1}^{t}X_\tau^i\mathbf{1}\{A_\tau^i=k\}-\mu_kN_k^i(t)-\frac{\lambda^2\sigma_k^2}{2}N_k^i(t)\right)\right)\leq 1$ holds under probabilistic communication. We obtain the tail bound by combining this result with the Markov inequality and optimizing over λ using a peeling type argument. We address the second challenge by proving that the number of times agents do not share information about any suboptimal arm k can be bounded by a term that increases logarithmically with time and scales with number of agents, G, and communication probabilities, as $\sum_{i=1}^{N}(1-p_i\cdot p)+\sum_{\mathcal{C}\in\mathcal{C}}(\max_{i\leq\mathcal{C}}p_i)\cdot p$. □

Remark 13 (Regret bound optimality)**.** *Under perfect communication ($p=1$) and no message discarding, i.e., $p_i=p=1,\forall i\in[N]$ the dominant term in our regret bound scales with $\bar{\chi}(G)$, obtaining identical performance to deterministic communication over G [19]. Alternatively, when $p_i=p=0$, there is no communication, and hence, the regret bound is $\mathcal{O}(N\log T)$. Theorem 17 quantifies the benefit of communication in reducing the group regret under probabilistic link failure and*

when agents incorporate observations with an agent-specific probability. Note that $\sum_{i=1}^{N}(1-p_i\cdot p)+\sum_{\mathcal{C}\in\mathcal{C}}(\max_{i\leq\mathcal{C}}p_i)\cdot p = N - p\cdot\left(\sum_{i=1}^{N}p_i-\sum_{\mathcal{C}\in\mathcal{C}}(\max_{i\leq\mathcal{C}}p_i)\right)$. *Since the clique covering is non-overlapping, the results show that agents obtain improved group performance for any communication probability* $p>0$ *for any nontrivial graph as compared to the case with no communication in which each agent learns on its own.*

Remark 14 (Controlling information disparity)**.** *In order to regulate the information disparity across the network we set* $p_i = \frac{d_{\min}(G)}{d_i(G)}$. *Thus, the agent(s) with minimum degree* $d_{\min}$ *incorporate each message they receive with probability* 1 *and we have that the expected number of messages for each agent is the same, i.e.,* $T\cdot d_{\min}(G)$. *Therefore, every agent receives the same amount of information (in expectation), providing a large performance improvement for irregular graphs (see Section 10.7).*

Message-Passing. Under this communication protocol each agent i communicates with neighbors at distance at most γ, where each hop adds a 1-step delay. Our algorithm `RCL-CF` obtains a similar regret bound in this setting as well, when all agents use the same `UCB1` exploration strategy (Figure 10.2).

Theorem 18 (`RCL-LF` Regret with message-passing)**.** *Let* $\mathcal{C}$ *be a minimal clique covering of* G_γ. *For any* $\mathcal{C}\in\mathcal{C}$ *and* $i,j\in\mathcal{C}$ *let* $\gamma_i = \max_{j\in\mathcal{C}} d(i,j)$ *be the maximum distance (in graph* G*) between agents* i *and* j. `RCL-LF` *running with the message-passing protocol (Figure 10.1) with delay parameter* γ *obtains cumulative group regret of*

$$\mathsf{Reg}_G(T)\leq g(\xi,\sigma)\left(\sum_{i=1}^{N}(1-p_i\cdot p^{\gamma_i})+\bar{\chi}(G_\gamma)\cdot(\max_{i\leq N}\ p_i\cdot p^{\gamma_i})\right)\left(\sum_{k>1}\frac{\log T}{\Delta_k}\right)+f((\gamma+4)N,G_\gamma).$$

Proof sketch. We partition the graph G_γ into non-overlapping cliques, analyze the regret of each clique and take the summation of regrets over cliques to obtain group

regret. In addition to the challenges encountered in Theorem 17 here we are required to account for having different probabilities of failures for messages due to having multiple paths of different length between agents and to account for the delay incurred by each hop when passing messages. We overcome the first challenge by noting that agent i receives each message with at least probability p^{γ_i}. We overcome the second challenge by identifying that regret incurred by delays can be upper bounded using $\left(\sum_{i=1}^{N} \gamma_i - N\right) \sum_{k>1} \Delta_k$. □

Remark 15. *Finding an optimal observation probability* $\{p_i\}_{1=1}^{N}$ *for* `RCL-LF` *with message-passing is difficult due to the delays added by each hop when forwarding messages. If messages are forwarded without a delay, optimal performance can be obtained by using* $p_i = \frac{d_{\min}(G_\gamma)}{d_i(G_\gamma)}$. *For dense* G_γ, *the above choice of observation probability provides near-optimal performance. When* $\gamma = d_\star(G)$ *we have that* G_γ *is a complete graph,* $p_i = \frac{d_{\min}(G_\gamma)}{d_i(G_\gamma)} = 1$, *and agents do not discard any message. However, when* $\gamma < d_\star(G)$, *the graph* G_γ *is not complete. Therefore agents receive different amounts of information which are approximately proportional to the degree distribution of* G_γ. *As explained earlier this information disparity leads to a performance disparity among agents. As a result group performance decreases. In this case we design the algorithm such that each agent* i *discards messages with* $1 - p_i$ *where* $p_i = \frac{d_{\min}(G_\gamma)}{d_i(G_\gamma)}$. *This regulates the information flow mitigating the bias introduced by information disparity. As a result the group obtains near-optimal performance.*

10.4 Instantaneous Reward-sharing Under Stochastic Delays

Next, we consider a communication protocol, where any message is received after an arbitrary (but bounded) stochastic delay. We assume for simplicity that each message is sent only once in the network (and not forwarded multiple times as in

*m*essage-passing), and leave the message-passing setting as future work. We assume, furthermore that the delays are identically and independently drawn from a bounded distribution with expectation $\mathbb{E}[\tau]$ (similar to prior work, e.g., [38, 91]). For this setting, we demonstrate that cooperative `UCB1`, along with incorporating all messages as soon as they are available, provides efficient performance, both empirically and theoretically. We denote this algorithm as `RCL-SD` (Stochastic Delays), and demonstrate that this approach incurs only an extra $\mathcal{O}(\sqrt{N \log T} + \log T)$ overhead compared to perfect communication.

Theorem 19 (`RCL-SD` Regret)**.** *Let* $D_{total} = N \cdot \mathbb{E}[\tau] + 2 \log T + 2\sqrt{N \cdot \mathbb{E}[\tau] \log T}$ *denote an upper bound on the total number of outstanding messages.* `RCL-SD` *obtains, with probability at least* $1 - \frac{1}{T}$*, cumulative group regret of*

$$\mathsf{Reg}_G(T) \leq g(\xi, \sigma) \cdot \bar{\chi}(G) \cdot \left(\sum_{k>1} \frac{\log T}{\Delta_k} \right) + D_{total} \cdot \left(\sum_{k>1} \Delta_k \right) + f(5N, G).$$

Proof sketch. We first demonstrate that the additional group regret due to stochastic delays can be bounded by the maximum number of cumulative outstanding messages over all agents at any given time step. Then we apply a result similar to Lemma 2 of [38] to bound the total number of outstanding messages using the cumulative expected delay $N \cdot \mathbb{E}[\tau]$, giving the result. □

Remark 16. *The* D_{total} *term is a succinct upper bound on the maximum number of cumulative outstanding messages over all agents, and when the expected delay* $\mathbb{E}[\tau] = o(1)$*, we see that the contribution of* D_{total} *is* $\mathcal{O}(\sqrt{N \log T} + \log T)$*. We conjecture that this cannot be improved without restricting communication, as each agent will send* T *messages in total. The result obtained by [38] has a similar dependence for a single agent.*

10.5 Hybrid Arm Elimination for Adversarial Reward Corruptions

In this section, we assume that any reward when transmitted can be corrupted by a maximum value of ϵ, i.e., $\max_{t,n} |r_n(t) - \tilde{r}_n(t)| \leq \epsilon$ where $\tilde{r}_n(t)$ denotes the transmitted reward. Furthermore, we assume that the corruptions can be *a*daptive, i.e., can depend on the prior actions and rewards of each agent. This model includes natural settings, where messages can be corrupted during transmission, as well as *b*yzantine communication [20]. If ϵ were known, we could then extend algorithms for misspecified bandits [27] to create a robust estimator and a subsequent `UCB1`-like algorithm that obtains a regret of $\mathcal{O}(\bar{\chi}(G_\gamma)K(\frac{\log T}{\Delta}) + TNK\epsilon)$. However, this approach has two issues. First, ϵ is typically not known, and the dependence on G_γ can be improved as well. We present an arm-elimination algorithm called `RCL-AC` (Adversarial Corruptions) that provides better guarantees on regret, without knowledge of ϵ in Algorithm 8.

The central motif in `RCL-AC`'s design is to eliminate bad arms by an epoch-based exploration, an idea that has been successful in the past for adversarially-corrupted stochastic bandits [57, 29]. The challenge, however, in a message-passing decentralized setting is two-fold. First, agents have different amounts of information based on their position in the network, and hence badly positioned agents in G may be exploring for much larger periods. Secondly, communication between agents is delayed, and hence any agent naively incorporating stale observations may incur a heavy bias from delays. To ameliorate the first issue, we partition the group of agents into two sets - *e*xploring agents ($\mathcal{I}$) and *i*mitating agents ($\mathcal{V} \setminus \mathcal{I}$). The idea is to only allow well-positioned agents in $\mathcal{I}$ to direct the exploration strategy for their neighboring agents, and the rest simply imitate their exploration strategy. We select $\mathcal{I}$ as a minimal dominating set of G_γ, hence $|\mathcal{I}| = \psi(G_\gamma)$. Furthermore, since $\mathcal{V} \setminus \mathcal{I}$ is a vertex cover, this ensures that each *i*mitating agent is connected (at distance at most γ) to at least

one agent in $\mathcal{I}$. Next, observe that there are two sources of delay: first, any *i*mitating agent must wait at most γ trials to observe the latest action from its corresponding *e*xploring agent, and second, each *e*xploring agent must wait an additional γ trials for the feedback from all of its imitating agents. We propose that each *e*xploring agent run `UCB1` for 2γ rounds after each epoch of arm elimination, using *o*nly local pulls. This prevents a large bias due to these delays, at a small cost of $\mathcal{O}(\log \log T)$ suboptimal pulls.

Theorem 20 (`RCL-RC` Regret)**.** `RCL-RC` *obtains, with probability at least* $1-\delta$*, group regret of*

$$\mathsf{Reg}_G(T) = \mathcal{O}\left(KTN\gamma\epsilon + \psi(G_\gamma)\cdot\sum_{k>1}\frac{\log T}{\Delta_k}\log\left(\frac{K\psi(G_\gamma)\log T}{\delta}\right) + N\sum_{k>1}\Delta_k + \sum_{k>1}\frac{N\log(\gamma\log T)}{\Delta_k}\right).$$

Proof sketch. Since the dominating set covers $\mathcal{V}$, we can decompose the group regret into the cumulative regret of the subgraphs corresponding to each agent in $\psi(G_\gamma)$. For each subgraph, we can consider the cumulative regret incurred when the *e*xploring agent follows `UCB1` versus arm elimination. We have that arm elimination occurs for $\log T$ epochs, and since `UCB1` runs for 2γ rounds between succesive epochs, we have that in any subgraph of size n, the cumulative regret from `UCB1` rounds is of $\mathcal{O}(nK\log(\gamma\log T))$. For arm elimination, we can bound the subgraph regret using a modification of the approach in [29]: the difference in our approach is to construct a multi-agent filtration for arbitrary (reward-dependent) corruptions from message-passing, and then applying Freedman's bound on the resulting martingale sequence. Subsequently, the regret in each epoch is bounded in a manner similar to [29], and finally applying a union bound. □

Remark 17 (Regret Optimality)**.** *Theorem 20 demonstrates a trade-off between communication density and the adversarial error, as seen by the first two terms in the*

regret bound. The first term ($KTN\gamma\epsilon$) is a bound on the cumulative error introduced due to message-passing, which is increasing in γ, whereas the second term denotes the logarithmic regret due to exploration, where $\psi(G_\gamma)$ decreases as γ increases: for $\gamma = d_\star(G), \psi(G_\gamma) = 1$, matching the lower bound in [19]. This too is expected, as fewer exploring agents are needed with a higher communication budget. Furthermore, we conjecture that the first term is optimal (in terms of T, up to graphical constants): a linear lower bound has been demonstrated for the single-agent setting in [57].

Remark 18 (Computational complexity)**.** *While the dominating set problem is known to be NP-complete [40], the problem admits a polynomial-time approximation scheme (PTAS) [16] for certain graphs, for which our bounds hold exactly. However,* `RCL-RC` *can work on any dominating set of size n, and suffer regret of $\widetilde{\mathcal{O}}(KTN\gamma\epsilon + n\sum_{k>1}\frac{\log T}{\Delta_k})$*[1]*.*

10.6 An Algorithm for Perfect Communication and Lower Bounds

For perfect communication, we present Delayed `MP-UCB`, a simple improvement to `UCB1` with message-passing where each agent i only incorporates messages originated prior to $\bar{\gamma} \leq \gamma$ time steps, reducing disparity in information across agents.

Theorem 21 (Delayed `MP-UCB Regret`)**.** `Delayed(MP)-UCB` *obtains cumulative group regret of*

$$\mathsf{Reg}_G(T) \leq g(\xi,\sigma)\bar{\chi}(G_\gamma)\left(\sum_{k>1}\frac{\log T}{\Delta_k}\right) + (N - \bar{\chi}(G_\gamma)\,(\gamma-1)\sum_{k>1}\Delta_k + f(5N, G_\gamma) + h(G_\gamma,\bar{\gamma})$$

where $h(G_\gamma,\bar{\gamma}) = \left((N-\bar{\chi}(G_\gamma)\bar{\gamma} + \sum_{t>\bar{\gamma}}^{T}\frac{\log\left(1-\frac{d_i(G_\gamma)\bar{\gamma}}{(d_i(G_\gamma)+1)t}\right)}{\log 1.3}\frac{1}{t^{(\xi+1)\left(1-\frac{0.09}{16}\right)}}\right)\sum_{k>1}\Delta_k.$

[1] The $\widetilde{\mathcal{O}}$ notation ignores absolute constants and $\log\log(\cdot)$ factors in T.

Proof sketch. Following a similar approach to the proof of Theorem 18 we partition the graph G_γ into a set of non-overlapping cliques, analyze the regret of each clique via a `UCB1` type analysis and take the summation of regret over cliques. However, using less information (due to delayed information usage) in estimates leads to a large confidence bound $C_k^i(t)$ and this reduces the contribution to the regret from tail probabilities. Note that $\log\left(1 - \frac{d_i(G_\gamma)\bar{\gamma}}{(d_i(G_\gamma)+1)t}\right)$ is negative $\forall t > \bar{\gamma}$, and hence lower regret achieved due to low tail probabilities is given by the second term of $h(G_\gamma, \bar{\gamma})$. □

Remark 19. *Incorporating only the messages originated before $\bar{\gamma}$ time steps is similar to communicating over $G_{\bar{\gamma}}$ after a delay of $\bar{\gamma}$ time steps. When G is connected and $\bar{\gamma} = \gamma = d_*$ this is similar to communicating over a complete graph with a delay of d_*. Thus Delayed MP-UCB mitigates the disparity in information used by each agent, leading to improved group performance.*

Lower Bounds. Without strict assumptions, a lower bound of $\mathcal{O}\left(\sum_{k>1} {}^{\log T}/_{\Delta_k}\right)$ has been demonstrated both for $\gamma = 1$ (instantaneous reward-sharing, [42]) and $\gamma > 1$ (message-passing, [19]), which both suggest that a speedup of $\frac{1}{N}$ is potentially achievable. For a more restrictive class of *i*ndividually consistent and *n*on-altruistic policies (i.e., that do not contradict their local feedback), a tighter lower bound of $\mathcal{O}\left(\alpha(G_2)\sum_{k>1} {}^{\log T}/_{\Delta_k}\right)$ can be demonstrated for reward-sharing [42], and consequently $\mathcal{O}\left(\alpha(G_{\gamma+1})\sum_{k>1} {}^{\log T}/_{\Delta_k}\right)$ for message-passing. To supplement these results, we present a lower bound to characterize the minimax optimal rates for the problem. We present first an assumption on multi-agent policies.

Assumption 1 (Agnostic decentralized policies)**.** *A set of N policies $\pi_1, ..., \pi_N$ are termed agnostic decentralized policies, if for every pair (i, j) of agents that communicate in G and each $t \in [T]$, $\pi_i(t)$ is independent of $\{\pi_j(\tau)\}_{\tau=1}^{t-d(i,j)}$ conditioned on the rewards $\{(A_j(\tau), X_j(\tau))\}_{\tau=1}^{t-d(i,j)}$.*

Theorem 22 (Minimax Rate)**.** *For any policy $\mathcal{A}$, there exists a K-armed environment over N agents with $\Delta_k \leq 1$ for any connected graph G and $\gamma \geq 1$ such that, for some absolute constant c,*

$$\mathsf{Reg}_G(\mathcal{A}, T) \geqslant c\sqrt{KN(T + \widetilde{d}(G))}.$$

Furthermore, if $\mathcal{A}$ is an agnostic decentralized policy, there exists a K-armed environment over N agents with $\Delta_k \leq 1$ for any connected graph G and $\gamma \geq 1$ such that, for some absolute constant c',

$$\mathsf{Reg}_G(\mathcal{A}, T) \geqslant c'\sqrt{\alpha^\star(G_\gamma)KNT}.$$

Here $\tilde{d}(G) = \sum_{i=1}^{d^\star(G)} \bar{d}_{=i} \cdot i$ denotes the average delay incurred by message-passing across the network G, and $\alpha^\star(G_\gamma) = \frac{N}{1+\bar{d}_\gamma}$ is Turan's lower bound [90] on $\alpha(G_\gamma)$.

Remark 20 (Tightness of lower bound)**.** *The first minimax bound does not make any assumptions on the policy $\mathcal{A}$, and hence we only see an additive dependence of the average delay incurred by communication over G. This dependence generalizes the minimax rate for delayed multi-armed bandits [72] to graphical feedback. For the latter bound, observe that a variety of cooperative extensions of single-agent bandit algorithms [42, 19, 13] obey this assumption, where the decision-making for any agent is independent of any other agent, conditioned on the observed rewards. In this setting, agents merely treat messages as additional pulls to construct stronger estimators, and do not strategize collectively. This bound is exact (up to constants) for a variety of communication graphs G. For instance, for linear and circular graphs, $\frac{\alpha^\star(G_\gamma)}{\alpha(G_\gamma)} = o(1)$, and for d-regular graphs, $\alpha^\star(G_\gamma) = \alpha(G_\gamma)$ [90].*

10.7 Experimental Results

We consider the 10-armed bandit with rewards drawn from Gaussian distributions with $\sigma_k = 1$ for each arm, such that $\mu_1 = 1$ and $\mu_k = 0.5$ for $k \neq 1$, and the number of agents $N = 50$, where we repeat each experiment 100 times with G selected randomly from different families of random graphs. The bottom row of Figure 10.3 corresponds to Erdos-Renyi graphs with $p = 0.7$. The top row of Figure 10.3 (a), (c) and (d) corresponds to multi-star graphs and (b) and (e) to random tree graphs. We set $\xi = 1.1$ and $\gamma = \max\{3, d_\star(G)/2\}$.

Stochastic Link Failure . Figure 10.3(a) and Figure 10.3(b) summarize performance of `RCL(RS)-LF` and `RCL(MP)-LF`, comparing it with the corresponding reward-sharing and message-passing `UCB`-like algorithms in which $p_i = 1$, $\forall i \in [N]$, for different p values. The group regret is given at $T = 500$. The results validate our claim that probabilistic message discarding improves performance for irregular graphs and provides competitive performance for *near*-regular graphs.

Stochastic Delays. We compare performance of `RCL-SD` with `UCB1`. We draw delays from a bounded distribution with $\mathbb{E}[\tau] = 10$ and $\tau_{\max} = 50$. The results are summarized in Figure 10.3(c).

Adversarial Communication. We compute the (approximate) dominating set using the algorithm provided in `networkx` for each connected component in G_γ. We draw corruptions uniformly from the range $[0, \epsilon]$ for each message, where ϵ is increased from 10^{-3} to 10^{-2}. The group regret at $T = 500$ as a function of ϵ is shown in Figure 10.3(d) and compared against individual `UCB1` and cooperative UCB with message-passing (`MP-UCB`), which incur larger regret increasing linearly with ϵ.

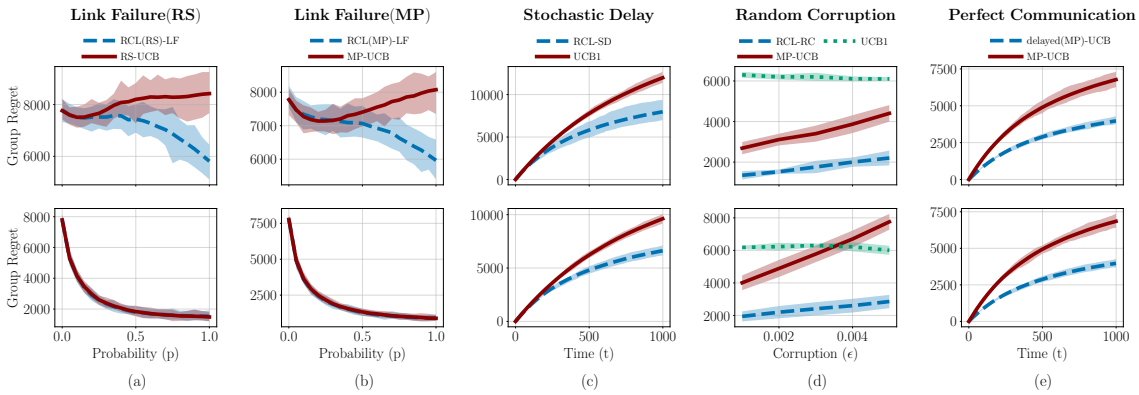


Figure 10.3: Experimental results for various imperfect communication settings.

Perfect Communication . We compare the regret curve for $T = 1000$ for our `Delayed(MP)-UCB` against regular `MP-UCB` in Figure 10.3(e). We use $\bar{\gamma} = 2$. It is evident that delayed incorporation of messages markedly improves performance across both networks.

10.8 Conclusions

In this paper, we studied the cooperative bandit problem in three different imperfect communication settings. For each setting, we proposed algorithms with competitive empirical performance and provided theoretical guarantees on the incurred regret. Further, we provided an algorithm for perfect communication that comfortably outperforms existing baseline approaches. We additionally provided a tighter network-dependent minimax lower bound for the cooperative bandit problem. We believe that our contributions can be of immediate utility in applications. Moreover, future inquiry can be pursued in several different directions, including multi-agent reinforcement learning and contextual bandit learning.

Ethical Considerations. Our work is primarily theoretical, and we do not foresee any negative societal consequences arising specifically from our contributions in this paper.

10.9 Appendix

10.9.1 Proof of Theorem 17

We consider the case where each message fails with probability $1-p$ and each agent i uses the messages it receives from its neighbors with probability p_i. This is equivalent to each agent i receiving messages from its neighbors with probability $p_i p$. Let $\mathbf{1}\{(i,j) \in E_t\}$ be the indicator random variable that takes value 1 if agent i receives reward value and arm id from agent j at time t and 0 otherwise.

We start by proving some useful lemmas.

Lemma 13. (Restatement of results from [7]) *Let* $\eta_k = \left(\frac{8(\xi+1)\sigma^2}{\Delta_k^2}\right)\log T$. *For any suboptimal arm* k *and* $\forall i, t$ *we have*

$$\mathsf{P}\left(A_i(t+1)=k, N_k^i(t) > \eta_k\right) \leq \mathsf{P}\left(\widehat{\mu}_1^i(t) \leq \mu_1 - C_1^i(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^i(t) \geq \mu_k + C_k^i(t)\right)$$

Proof. Let $Q_k^i(t) = \widehat{\mu}_k^i(t) + C_k^i(t)$. Note that for any $k > 1$ we have

$$\begin{aligned}\{A_i(t+1)=k\} &\subset \left\{Q_k^i(t) \geq Q_1^i(t)\right\} \\ &\subset \left\{\left\{\mu_1 < \mu_k + 2C_k^i(t)\right\} \cup \left\{\widehat{\mu}_1^i(t) \leq \mu_1 - C_1^i(t)\right\} \cup \left\{\widehat{\mu}_k^i(t) \geq \mu_k + C_k^i(t)\right\}\right\}.\end{aligned}$$

Let $\eta_k = \left(\frac{8(\xi+1)\sigma^2}{\Delta_k^2}\right)\log T$. Since $N_k^i(t) > \eta_k$ the event $\{\mu_1 < \mu_k + 2C_k^i(t)\}$ does not occur. Thus we have

$$\mathsf{P}\left(A_i(t+1)=k, N_k^i(t) > \eta_k\right) \leq \mathsf{P}\left(\widehat{\mu}_1^i(t) \leq \mu_1 - C_1^i(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^i(t) \geq \mu_k + C_k^i(t)\right)$$

This concludes the proof of Lemma 13. □

Lemma 14. *Let $\bar{\chi}(G)$ is the clique covering number of graph G. Let $\eta_k = \left(\frac{8(\xi+1)\sigma_k^2}{\Delta_k^2}\right)\log T$. Then we have*

$$\sum_{i=1}^{N}\mathbb{E}[n_k^i(T)] \leq \left(\sum_{i=1}^{N}(1-p_i p)+\bar{\chi}(G)p_{\max}p\right)\eta_k + 2N \tag{10.1}$$

$$+\sum_{i=1}^{N}\sum_{t=1}^{T-1}\left[\mathsf{P}\left(\widehat{\mu}_1^i(t)\leq \mu_1 - C_1^i(t)\right)+\mathsf{P}\left(\widehat{\mu}_k^i(t)\geq \mu_k + C_k^i(t)\right)\right] \tag{10.2}$$

Proof. Let $\mathcal{C}$ be a non overlapping clique covering of G. Note that for each suboptimal arm $k>1$ we have

$$\sum_{i=1}^{N}\mathbb{E}[n_k^i(T)] = \sum_{i=1}^{N}\sum_{t=1}^{T}\mathsf{P}\left(A_i(t)=k\right) = \sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathsf{P}\left(A_i(t)=k\right). \tag{10.3}$$

Let $\tau_{k,\mathcal{C}}$ denote the maximum time step when the total number of times arm k has been played by all the agents in clique $\mathcal{C}$ is at most $\eta_k + |\mathcal{C}|$ times. This can be stated as $\tau_{k,\mathcal{C}} := \max\{t\in[T] : \sum_{i\in\mathcal{C}} n_k^i(t) \leq \eta_k + |\mathcal{C}|\}$. Then, we have that $\eta_k < \sum_{i\in\mathcal{C}} n_k^i(\tau_{k,\mathcal{C}}) \leq \eta_k + |\mathcal{C}|$.

For each agent $i\in\mathcal{C}$ let

$$\bar{N}_k^i(t) := \sum_{j\in\mathcal{C}}\sum_{\tau=1}^{t}\mathbf{1}\{A_j(\tau)=k\}\mathbf{1}\{(i,j)\in E_\tau\},$$

denote the sum of the total number of times agent i pulled arm k and the total number of observations it received from agents in its clique about arm k until time t. Define $\bar{\tau}_{k,\mathcal{C}}^i := \max\{t\in[T] : \bar{N}_k^i(t)\leq\eta_k\}$. Then we have that $\eta_k - |\mathcal{C}| < \bar{N}_k^i(\bar{\tau}_{k,\mathcal{C}}^i) \leq \eta_k$.

Note that $N_k^i(t)\geq\bar{N}_k^i(t),\forall t$, hence for all $i\in\mathcal{C}$ we have $N_k^i(t)>\eta_k,\forall t>\bar{\tau}_{k,\mathcal{C}}^i$. Here we consider that $\bar{\tau}_{k,\mathcal{C}}^{(i)}\geq\tau_{k,\mathcal{C}},\forall i$. From regret results it follows that regret for this case is greater than the regret for the case where $\bar{\tau}_{k,\mathcal{C}}^i<\tau_{k,\mathcal{C}}$ for some (or all) i.

We analyse the expected number of times agents pull suboptimal arm k as follows,

$$\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathbf{1}\{A_i(t)=k\} \tag{10.4}$$

$$=\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\mathbf{1}\{A_i(t)=k\}+\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{\bar{\tau}^i_{k,\mathcal{C}}}\mathbf{1}\{A_i(t)=k\}+\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t>\bar{\tau}^i_{k,\mathcal{C}}}^{T}\mathbf{1}\{A_i(t)=k\} \tag{10.5}$$

$$\leq\sum_{\mathcal{C}\in\mathcal{C}}(\eta_k+|\mathcal{C}|)+\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{\bar{\tau}^i_{k,\mathcal{C}}}\mathbf{1}\{A_i(t)=k\}+|\mathcal{C}| \tag{10.6}$$

$$+\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t>\bar{\tau}^i_{k,\mathcal{C}}}^{T-1}\mathbf{1}\{A_i(t+1)=k\}\mathbf{1}\left\{N^i_k(t)>\eta_k\right\}. \tag{10.7}$$

Taking expectation we have

$$\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathsf{P}\left(A_i(t)=k\right)\leq\sum_{\mathcal{C}\in\mathcal{C}}(\eta_k+2|\mathcal{C}|) \tag{10.8}$$

$$+\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{\bar{\tau}^i_{k,\mathcal{C}}}\mathsf{P}\left(A_i(t)=k\right)+\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t>\bar{\tau}^i_{k,\mathcal{C}}}^{T-1}\mathsf{P}\left(A_i(t+1)=k,N^i_k(t)>\eta_k\right). \tag{10.9}$$

Note that we have

$$\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{\bar{\tau}^i_{k,\mathcal{C}}}\mathbf{1}\{A_i(t)=k\} \tag{10.10}$$

$$=\sum_{i\in\mathcal{C}}\bar{N}^i_k(\bar{\tau}^i_{k,\mathcal{C}})-\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\mathbf{1}\{A_i(t)=k\}-\sum_{i\in\mathcal{C}}\sum_{j\neq i,j\in\mathcal{C}}\sum_{t=1}^{\bar{\tau}^i_{k,\mathcal{C}}}\mathbf{1}\{A_j(t)=k\}\mathbf{1}\{(i,j)\in E_t\} \tag{10.11}$$

$$=\sum_{i\in\mathcal{C}}\bar{N}^i_k(\bar{\tau}^i_{k,\mathcal{C}})-\sum_{i\in\mathcal{C}}n^i_k(\tau_{k,\mathcal{C}})-\sum_{i\in\mathcal{C}}\sum_{j\neq i,j\in\mathcal{C}}\sum_{t=1}^{\bar{\tau}^i_{k,\mathcal{C}}}\mathbf{1}\{A_j(t)=k\}\mathbf{1}\{(i,j)\in E_t\} \tag{10.12}$$

$$\leq |\mathcal{C}|\eta_k - \eta_k - \sum_{i\in\mathcal{C}} \sum_{j\neq i, j\in\mathcal{C}} \sum_{t=1}^{\bar{\tau}^i_{k,\mathcal{C}}} \mathbf{1}\{A_j(t) = k\}\mathbf{1}\{(i,j) \in E_t\} \tag{10.13}$$

$$\leq |\mathcal{C}|\eta_k - \eta_k - \sum_{i\in\mathcal{C}} \sum_{j\neq i, j\in\mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathbf{1}\{A_j(t) = k\}\mathbf{1}\{(i,j) \in E_t\}. \tag{10.14}$$

Taking the expectation

$$\sum_{i\in\mathcal{C}} \sum_{t>\tau_{k,\mathcal{C}}}^{\bar{\tau}^i_{k,\mathcal{C}}} \mathsf{P}\left(A_i(t) = k\right) \leq |\mathcal{C}|\eta_k - \eta_k - \sum_{i\in\mathcal{C}} p_i p \sum_{j\neq i, j\in\mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathsf{P}\left(A_j(t) = k\right) \tag{10.15}$$

$$= |\mathcal{C}|\eta_k - \eta_k - \sum_{i\in\mathcal{C}} p_i p \sum_{j\neq i, j\in\mathcal{C}} \mathbb{E}(n^j_k(\tau_{k,\mathcal{C}})) \tag{10.16}$$

$$= |\mathcal{C}|\eta_k - \eta_k - \left(\sum_{i\in\mathcal{C}} p_i p\right)\left(\sum_{i\in\mathcal{C}} \mathbb{E}(n^i_k(\tau_{k,\mathcal{C}}))\right) + \sum_{i\in\mathcal{C}} p_i p \mathbb{E}(n^i_k(\tau_{k,\mathcal{C}})) \tag{10.17}$$

$$\leq |\mathcal{C}|\eta_k - \eta_k - p\left(\sum_{j\in\mathcal{C}} p_j - p_{\max}\right)\mathbb{E}\left(\sum_{i\in\mathcal{C}} n^i_k(\tau_{k,\mathcal{C}})\right) \tag{10.18}$$

$$\leq |\mathcal{C}|\eta_k - \eta_k - p\left(\sum_{j\in\mathcal{C}} p_j - p_{\max}\right)\eta_k \tag{10.19}$$

$$= \left(|\mathcal{C}| - 1 - p\left(\sum_{j\in\mathcal{C}} p_j - p_{\max}\right)\right)\eta_k. \tag{10.20}$$

Substituting this results to (10.9) we get

$$\sum_{\mathcal{C}\in\mathfrak{C}} \sum_{i\in\mathcal{C}} \sum_{t=1}^{T} \mathsf{P}\left(A_i(t) = k\right) \leq \sum_{\mathcal{C}\in\mathfrak{C}} (\eta_k + 2|\mathcal{C}|) + \sum_{\mathcal{C}\in\mathfrak{C}} \left(|\mathcal{C}| - 1 - p\left(\sum_{j\in\mathcal{C}} p_j - p_{\max}\right)\right)\eta_k \tag{10.21}$$

$$+ \sum_{\mathcal{C}\in\mathfrak{C}} \sum_{i\in\mathcal{C}} \sum_{t>\bar{\tau}^i_{k,\mathcal{C}}}^{T-1} \mathsf{P}\left(A_i(t+1) = k, N^i_k(t) > \eta_k\right). \tag{10.22}$$

Thus from Lemma 13 and (10.22) we have

$$\sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathsf{P}\left(A_i(t)=k\right) \tag{10.23}$$

$$\leq \sum_{\mathcal{C}\in\mathcal{C}}\eta_k + 2N + \sum_{\mathcal{C}\in\mathcal{C}}\left(|\mathcal{C}| - 1 - p\left(\sum_{j\in\mathcal{C}} p_j - p_{\max}\right)\right)\eta_k \tag{10.24}$$

$$+ \sum_{\mathcal{C}\in\mathcal{C}}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{T-1}\left[\mathsf{P}\left(\widehat{\mu}_1^i(t)\leq\mu_1 - C_1^i(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^i(t)\geq\mu_k + C_k^i(t)\right)\right] \tag{10.25}$$

$$\overset{(a)}{=} \bar{\chi}(G)\eta_k + \left(N - \sum_{i=1}^{N} p_i p - \mathcal{X}(G)(1-p_{\max}p)\right)\eta_k + 2N \tag{10.26}$$

$$+ \sum_{i=1}^{N}\sum_{t>\tau_{k,\mathcal{C}}}^{T-1}\left[\mathsf{P}\left(\widehat{\mu}_1^i(t)\leq\mu_1 - C_1^i(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^i(t)\geq\mu_k + C_k^i(t)\right)\right] \tag{10.27}$$

$$\leq \left(\sum_{i=1}^{N}(1-p_i p) + \bar{\chi}(G)p_{\max}p\right)\eta_k + 2N \tag{10.28}$$

$$+ \sum_{i=1}^{N}\sum_{t=1}^{T-1}\left[\mathsf{P}\left(\widehat{\mu}_1^i(t)\leq\mu_1 - C_1^i(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^i(t)\geq\mu_k + C_k^i(t)\right)\right], \tag{10.29}$$

where (a) follows from the fact that clique covering is non overlapping. This concludes the proof of Lemma 14. □

Lemma 15. *Let $d_i(G)$ be the degree of agent i in graph G. For any $\sigma_k > 0$ some constant $\zeta > 1$*

$$\mathsf{P}\left(\left|\widehat{\mu}_k^i(t) - \mu_k\right| > \sigma_k\sqrt{\frac{2(\xi+1)\log t}{N_k^i(t)}}\right) \leq \frac{\log((d_i(G)+1)t)}{\log\zeta}\frac{1}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}. \tag{10.30}$$

Proof. For all k let $X_k^i(t)$ for all i,t be iid copies of X_k. Then we have $X_t^i\mathbf{1}\{A_i(t)=k\} = X_k^i(t)\mathbf{1}\{A_i(t)=k\}$. Recall that reward distribution of arm k has mean μ_k and variance proxy σ_k. Thus $\forall i,t$ we have

$$\mathbb{E}\left(\exp\left(\lambda\left(X_k^i(t)-\mu_k\right)\right)\right) \leq \exp\left(\frac{\lambda^2\sigma_k^2}{2}\right). \tag{10.31}$$

Define local history at every agent i as follows

$$\mathcal{H}_t^i := \sigma\left(X_\tau^i, A_i(\tau), X_\tau^j \mathbf{1}\{(i,j) \in E_\tau\}, A_j(\tau)\mathbf{1}\{(i,j) \in E_\tau\}, \forall\, \tau \in [t], j \in \mathcal{N}_i(G)\right). \tag{10.32}$$

Since $\mathbf{1}\{A_j(\tau) = k\}\mathbf{1}\{(i,j) \in E_\tau\}$ for $j \in \mathcal{N}_i(G)$ is a $\mathcal{H}_{\tau-1}^i$ measurable random variable, we have

$$\mathbb{E}\left(\exp\left(\lambda\left(X_\tau^j - \mu_k\right)\mathbf{1}\{A_j(\tau) = k\}\mathbf{1}\{(i,j) \in E_\tau\}\right)\Big|\, \mathcal{H}_{\tau-1}^i\right) \tag{10.33}$$

$$= \mathbb{E}\left(\exp\left(\lambda\left(X_k^j(\tau) - \mu_k\right)\mathbf{1}\{A_j(\tau) = k\}\mathbf{1}\{(i,j) \in E_\tau\}\right)\Big|\, \mathcal{H}_{\tau-1}^i\right) \tag{10.34}$$

$$\leq \exp\left(\frac{\lambda^2\sigma_k^2}{2}\mathbf{1}\{A_j(\tau) = k\}\mathbf{1}\{(i,j) \in E_\tau\}\right). \tag{10.35}$$

Define a new random variable such that $\forall \tau > 0$.

$$Y_k^i(\tau) = \sum_{j=1}^N \left(X_k^j(\tau)\mathbf{1}\{A_j(\tau) = k\}\mathbf{1}\{(i,j) \in E_\tau\} - \mathbb{E}\left[X_k^j(\tau)\mathbf{1}\{A_j(\tau) = k\}\mathbf{1}\{(i,j) \in E_\tau\}\Big|\mathcal{H}_{\tau-1}^i\right]\right) \tag{10.36}$$

$$= \sum_{j=1}^N \left(X_k^j(\tau) - \mu_k\right)\mathbf{1}\{A_j(\tau) = k\}\mathbf{1}\{(i,j) \in E_\tau\}. \tag{10.37}$$

Note that $\mathbb{E}\left(Y_k^i(\tau)\right) = \mathbb{E}\left(Y_k^i(\tau)|\mathcal{H}_{\tau-1}^i\right) = 0$. Let $Z_k^i(t) = \sum_{\tau=1}^t Y_k^i(\tau)$. For any $\lambda > 0$

$$\mathbb{E}\left(\exp(\lambda Y_k^i(\tau))|\mathcal{H}_{\tau-1}^i\right) \tag{10.38}$$

$$= \mathbb{E}\left(\exp\left(\lambda\sum_{j=1}^N \left(X_k^j(\tau) - \mu_k\right)\mathbf{1}\{A_j(\tau) = k\}\mathbf{1}\{(i,j) \in E_\tau\}\right)\Bigg|\, \mathcal{H}_{\tau-1}^i\right) \tag{10.39}$$

$$= \mathbb{E}\left(\prod_{j=1}^N \exp\left(\lambda\left(X_k^j(\tau) - \mu_k\right)\mathbf{1}\{A_j(\tau) = k\}\mathbf{1}\{(i,j) \in E_\tau\}\right)\Bigg|\, \mathcal{H}_{\tau-1}^i\right) \tag{10.40}$$

$$\overset{(a)}{=} \prod_{j=1}^N \mathbb{E}\left(\exp\left(\lambda\left(X_k^j(\tau) - \mu_k\right)\mathbf{1}\{A_j(\tau) = k\}\mathbf{1}\{(i,j) \in E_\tau\}\right)\Big|\, \mathcal{H}_{\tau-1}^i\right) \tag{10.41}$$

$$\leq \prod_{j=1}^{N} \exp\left(\frac{\lambda^2\sigma_k^2}{2}\mathbf{1}\{A_j(\tau)=k\}\mathbf{1}\{(i,j)\in E_\tau\}\right) \tag{10.42}$$

$$= \exp\left(\frac{\lambda^2\sigma_k^2}{2}\sum_{j=1}^{N}\mathbf{1}\{A_j(\tau)=k\}\mathbf{1}\{(i,j)\in E_\tau\}\right). \tag{10.43}$$

Equality (a) follows from the fact that random variables $\left\{\exp\left(\lambda\left(X_k^j(\tau)-\mu_k\right)\mathbf{1}\{A_j(\tau)=k\}\mathbf{1}\{(i,j)\right.\right.$ are conditionally independent with respect to $\mathcal{H}_{\tau-1}^i$. Since $\mathbf{1}\{A_j(\tau)=k\}, \mathbf{1}\{(i,j)\in E_\tau\}$ are $\mathcal{H}_{\tau-1}^i$ measurable, and so

$$\mathbb{E}\left(\exp\left(\lambda Y_k^i(\tau)-\frac{\lambda^2\sigma_k^2}{2}\sum_{j=1}^{N}\mathbf{1}\{A_j(\tau)=k\}\mathbf{1}\{(i,j)\in E_\tau\}\right)\middle|\,\mathcal{H}_{\tau-1}^i\right)\leq 1. \tag{10.44}$$

Let $N_k^i(t) = \sum_{\tau=1}^{t}\sum_{j=1}^{N}\mathbf{1}\{A_i(\tau)=k\}\mathbf{1}\{(i,j)\in E_\tau\}$. Further, using the tower property of conditional expectation we have

$$\mathbb{E}\left(\exp\left(\lambda Z_k^i(t)-\frac{\lambda^2\sigma_k^2}{2}N_k^i(t)\right)\middle|\,\mathcal{H}_{t-1}^i\right)\leq\exp\left(\lambda Z_k^i(t-1)-\frac{\lambda^2\sigma_k^2}{2}N_k^i(t-1)\right). \tag{10.45}$$

Repeating the above step t times we have

$$\mathbb{E}\left(\exp\left(\lambda Z_k^i(t)-\frac{\lambda^2\sigma_k^2}{2}N_k^i(t)\right)\right)\leq 1. \tag{10.46}$$

Note that we have

$$\mathsf{P}\left(\exp\left(\lambda Z_k^i(t)-\frac{\lambda^2\sigma_i^2}{2}N_k^i(t)\right)\geq\exp\left(2\kappa\vartheta\right)\right) \tag{10.47}$$

$$= \mathsf{P}\left(\lambda Z_k^i(t)-\frac{\lambda^2\sigma_k^2}{2}N_k^i(t)\geq 2\kappa\vartheta\right) \tag{10.48}$$

$$= \mathsf{P}\left(\frac{Z_k^i(t)}{\sqrt{N_k^i(t)}}\geq\frac{2\kappa\vartheta}{\lambda\sqrt{N_k^i(t)}}+\frac{\sigma_k^2}{2}\lambda\sqrt{N_k^i(t)}\right). \tag{10.49}$$

Fix a constant $\zeta > 1$. Then $1 \leq N_k^i(t) \leq \zeta^{D_t}$ where $D_t = \frac{\log((d_i(G)+1)t)}{\log \zeta}$. For $\lambda_l = \frac{2}{\sigma_k}\sqrt{\frac{\kappa\vartheta}{\zeta^{l-1/2}}}$ and $\zeta^{l-1} \leq N_k^i(t) \leq \zeta^l$ we have

$$\frac{2\kappa\vartheta}{\lambda_l}\sqrt{\frac{1}{N_k^i(t)}} + \frac{\sigma_k^2}{2}\lambda_l\sqrt{N_k^i(t)} = \sigma_k\sqrt{\kappa\vartheta}\left(\sqrt{\frac{\zeta^{l-1/2}}{N_k^i(t)}} + \sqrt{\frac{N_k^i(t)}{\zeta^{l-1/2}}}\right) \leq \sqrt{\vartheta}, \tag{10.50}$$

where $\kappa = \frac{1}{\sigma_k^2\left(\zeta^{\frac{1}{4}}+\zeta^{-\frac{1}{4}}\right)^2}$.

Then we have

$$\left\{\frac{Z_k^i(t)}{\sqrt{N_k^i(t)}} \geq \sqrt{\vartheta}\right\} \subset \cup_{l=1}^{D_t}\left\{\frac{Z_k^i(t)}{\sqrt{N_k^i(t)}} \geq \frac{2\kappa\vartheta}{\lambda_l\sqrt{N_k^i(t)}} + \frac{\sigma_k^2}{2}\lambda_l\sqrt{N_k^i(t)}\right\} \tag{10.51}$$

$$= \cup_{l=1}^{D_t}\left\{\lambda_l Z_k^i(t) - \frac{\lambda_l^2\sigma_k^2}{2}N_k^i(t) \geq 2\kappa\vartheta\right\}. \tag{10.52}$$

Recall from the Markov inequality that $\mathsf{P}(Y \geq a) \leq \frac{\mathbb{E}(Y)}{a}$ for any positive random variable Y. Thus from (10.52) and Markov inequality we get,

$$\mathsf{P}\left(\frac{Z_k^i(t)}{\sqrt{N_k^i(t)}} \geq \sqrt{\vartheta}\right) \leq \sum_{l=1}^{D_t}\exp(-2\kappa\vartheta). \tag{10.53}$$

Then we have,

$$\mathsf{P}\left(\frac{Z_k^i(t)}{N_k^i(t)} \geq \sqrt{\frac{\vartheta}{N_k^i(t)}}\right) \leq \sum_{l=1}^{D_t}\exp(-2\kappa\vartheta) \tag{10.54}$$

Substituting $\vartheta = 2\sigma_k^2(\xi+1)\log t$ we get

$$\mathsf{P}\left(\left|\widehat{\mu}_k^i(t) - \mu_k\right| > \sigma_k\sqrt{\frac{2(\xi+1)\log t}{N_k^i(t)}}\right) \leq \frac{\log((d_i(G)+1)t)}{\log\zeta}\exp\left(-\frac{4(\xi+1)\log t}{\left(\zeta^{\frac{1}{4}}+\zeta^{-\frac{1}{4}}\right)^2}\right). \tag{10.55}$$

Note that $\forall \zeta > 1$ we have

$$\frac{4}{\left(\zeta^{\frac{1}{4}} + \zeta^{-\frac{1}{4}}\right)^2} \geq 1 - \frac{(\zeta-1)^2}{16} \tag{10.56}$$

Then we have

$$\mathsf{P}\left(\left|\widehat{\mu}_k^i(t) - \mu_k\right| > \sigma_k \sqrt{\frac{2(\xi+1)\log t}{N_k^i(t)}}\right) \leq \frac{\log((d_i(G)+1)t)}{\log \zeta} \frac{1}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}. \tag{10.57}$$

This concludes the proof of Lemma 15. □

Lemma 16. *Let $\zeta = 1.3, \xi \geq 1.1$, $d_i \geq 0$ and $t \in [T]$. Then we have*

$$\sum_{t=1}^{T-1} \frac{1}{\log \zeta} \frac{\log\left((d_i+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}} \leq 12 \log(3(d_i+1)) + 3\left(\log\left(d_i+1\right)+1\right) \tag{10.58}$$

Proof. For $\zeta = 1.3$ we have $\frac{1}{\log \zeta} < 8.78$. Further $(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right) > 2$ and $\forall t \geq 3$ we see that $\frac{\log((d_i+1)t)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}$ is monotonically decreasing. Thus we have

$$\sum_{t=1}^{T-1} \frac{\log\left((d_i+1)t\right)}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}} \leq 1.362 \log(3(d_i+1)) + \int_3^{T-1} \frac{\log\left((d_i+1)t\right)}{t^2} dt \tag{10.59}$$

Let $z = (d_i+1)t$. Then we have

$$\int_3^{T-1} \frac{\log\left((d_i+1)t\right)}{t^2} dt = (d_i+1) \int_{3(d_i+1)}^{(d_i+1)(T-1)} \frac{\log z}{z^2} dz \tag{10.60}$$

$$= (d_i+1)\left[-\frac{\log z}{z} - \frac{1}{z}\right]_{3((d_i+1)}^{(d_i+1)(T-1)} \tag{10.61}$$

Thus we have

$$\int_3^{T-1} \frac{\log\left((d_i+1)t\right)}{t^2} dt \leq (d_i+1)\left[\frac{\log(d_i+1)}{3(d_i+1)} + \frac{1}{3(d_i+1)}\right] \tag{10.62}$$

$$= \frac{1}{3}\log(d_i+1) + \frac{1}{3} \tag{10.63}$$

Recall that For $\zeta = 1.3$ we have $\frac{1}{\log \zeta} < 8.78$. Thus the proof of Lemma 16 follows from (10.59) and (10.63). □

Now we prove Theorem 17 as follows. Recall that group regret can be given as $\mathsf{Reg}_G(T) = \sum_{i=1}^{N} \sum_{k>1} \Delta_k \cdot \mathbb{E}\left[n_k^i(t)\right]$. Thus using Lemmas 14, 15 and 16 we obtain

$$\mathsf{Reg}_G(T) \leq 8(\xi+1)\sigma_k^2 \left(\sum_{i=1}^{N}(1-p_i p) + \bar{\chi}(G) p_{\max} p\right)\left(\sum_{k>1}\frac{\log T}{\Delta_k}\right) \tag{10.64}$$

$$+5N\sum_{k>1}\Delta_k + 4\sum_{i=1}^{N}\left(3\log(3(d_i(G)+1)) + \left(\log\left(d_i(G)+1\right)\right)\right)\sum_{k>1}\Delta_k \tag{10.65}$$

10.9.2 Proof of Theorem 18

In this section we consider the case where agents pass messages up to γ hop neighbors with each hop adding a delay of 1 time step. Let $\mathcal{C}_\gamma$ be a non overlapping clique covering of G_γ. For any $\mathcal{C} \in \mathcal{C}_\gamma$ and $i, j \in \mathcal{C}$ let $\gamma_i = \max_{j\in\mathcal{C}} d(i,j)$ be the maximum distance (in graph G) between agent i and any other agent j in the same clique in graph G_γ. Let $\mathbf{1}\{(i,j) \in E_{\tau',\tau}\}$ is a random variable that takes value 1 if at time τ agent i receives the message initiated by agent j at time τ'. Recall that each communicated message fails with probability $1-p$ and each agent i incorporates the messages it receives from its neighbors with probability p_i.

We follow an approach similar to proof of Theorem 17. We star by providing a tail probability bound similar to Lemma 15.

Lemma 17. *Let $d_i(G_\gamma)$ be the degree of agent i in graph G_γ. For any $\sigma_k > 0$ some constant $\zeta > 1$*

$$\mathsf{P}\left(\left|\widehat{\mu}_k^i(t) - \mu_k\right| > \sigma_k\sqrt{\frac{2(\xi+1)\log t}{N_k^i(t)}}\right) \leq \frac{\log((d_i(G_\gamma)+1)t)}{\log\zeta}\frac{1}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}. \tag{10.66}$$

Proof. For all k let $X_k^i(t)$ for all i, t be iid copies of X_k. Then we have $X_t^i\mathbf{1}\{A_i(t)=k\} = X_k^i(t)\mathbf{1}\{A_i(t)=k\}$. Recall that reward distribution of arm k has mean μ_k and variance

proxy σ_k. Thus $\forall i, t$ we have

$$\mathbb{E}\left(\exp\left(\lambda\left(X_k^i(t)-\mu_k\right)\right)\right) \leq \exp\left(\frac{\lambda^2\sigma_k^2}{2}\right). \tag{10.67}$$

Define local history at every agent i as follows

$$\mathcal{H}_t^i := \sigma\left(X_{\tau'}^i, A_i(\tau'), X_{\tau'}^j \mathbf{1}\{(i,j)\in E_{\tau',\tau}\}, A_j(\tau')\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}, \forall\, \tau', \tau\in[t], j\in\mathcal{N}_i(G_\gamma)\right). \tag{10.68}$$

Since $\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}$ for $j\in\mathcal{N}_i(G_\gamma)$ is a $\mathcal{H}_{\tau-1}^i$ measurable random variable, we have $\forall \tau'\leq\tau$

$$\begin{aligned}
&\mathbb{E}\left(\exp\left(\lambda\left(X_{\tau'}^j-\mu_k\right)\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}\right)\Big|\,\mathcal{H}_{\tau-1}^i\right) && (10.69)\\
&=\mathbb{E}\left(\exp\left(\lambda\left(X_k^j(\tau')-\mu_k\right)\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}\right)\Big|\,\mathcal{H}_{\tau-1}^i\right) && (10.70)\\
&\leq \exp\left(\frac{\lambda^2\sigma_k^2}{2}\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}\right). && (10.71)
\end{aligned}$$

Define a new random variable such that $\forall\tau>0$ and $\tau'\leq\tau$

$$\begin{aligned}
Y_k^i(\tau) =& \sum_{j=1}^N\sum_{\tau'=1}^{\tau}\left(X_k^j(\tau')\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}\right. && (10.72)\\
&\left.-\mathbb{E}\left[X_k^j(\tau')\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}\Big|\mathcal{H}_{\tau-1}^i\right]\right) && (10.73)\\
=&\sum_{j=1}^N\sum_{\tau'=1}^{\tau}\left(X_k^j(\tau')-\mu_k\right)\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}. && (10.74)
\end{aligned}$$

Note that $\mathbb{E}\left(Y_k^i(\tau)\right)=\mathbb{E}\left(Y_k^i(\tau)|\mathcal{H}_{\tau-1}^i\right)=0$. Let $Z_k^i(t)=\sum_{\tau=1}^t Y_k^i(\tau)$. For any $\lambda>0$

$$\begin{aligned}
&\mathbb{E}\left(\exp(\lambda Y_k^i(\tau))|\mathcal{H}_{\tau-1}^i\right) && (10.75)\\
&=\mathbb{E}\left(\exp\left(\lambda\sum_{j=1}^N\sum_{\tau'=1}^{\tau}\left(X_k^j(\tau')-\mu_k\right)\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}\right)\Big|\mathcal{H}_{\tau-1}^i\right) && (10.76)
\end{aligned}$$

$$= \mathbb{E}\left(\prod_{j=1}^{N}\prod_{\tau'=1}^{\tau}\exp\left(\lambda\left(X_k^j(\tau')-\mu_k\right)\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}\right)\Big|\mathcal{H}_{\tau-1}^i\right) \quad (10.77)$$

$$\overset{(a)}{=}\prod_{j=1}^{N}\prod_{\tau'=1}^{\tau}\mathbb{E}\left(\exp\left(\lambda\left(X_k^j(\tau')-\mu_k\right)\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}\right)\Big|\mathcal{H}_{\tau-1}^i\right) \quad (10.78)$$

$$\leq\prod_{j=1}^{N}\prod_{\tau'=1}^{\tau}\exp\left(\frac{\lambda^2\sigma_k^2}{2}\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}\right) \quad (10.79)$$

$$=\exp\left(\frac{\lambda^2\sigma_k^2}{2}\sum_{j=1}^{N}\sum_{\tau'=1}^{\tau}\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}\right). \quad (10.80)$$

Equality (a) follows from the fact that $\forall\tau'\leq\tau$ random variables $\left\{\exp\left(\lambda\left(X_k^j(\tau')-\mu_k\right)\mathbf{1}\{A_j(\tau')=k\right.\right.$ are conditionally independent with respect to $\mathcal{H}_{\tau-1}^i$. Since $\mathbf{1}\{A_j(\tau')=k\},\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}$ are $\mathcal{H}_{\tau-1}^i$ measurable, and so

$$\mathbb{E}\left(\exp\left(\lambda Y_k^i(\tau)-\frac{\lambda^2\sigma_k^2}{2}\sum_{j=1}^{N}\sum_{\tau'=1}^{\tau}\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}\right)\Big|\,|\mathcal{H}_{\tau-1}^i\right)\leq 1. \quad (10.81)$$

Let $N_k^i(t)=\sum_{\tau=1}^{t}\sum_{\tau'=1}^{\tau}\sum_{j=1}^{N}\mathbf{1}\{A_i(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\}$. Further, using the tower property of conditional expectation we have

$$\mathbb{E}\left(\exp\left(\lambda Z_k^i(t)-\frac{\lambda^2\sigma_k^2}{2}N_k^i(t)\right)\Big|\,\mathcal{H}_{t-1}^i\right)\leq\exp\left(\lambda Z_k^i(t-1)-\frac{\lambda^2\sigma_k^2}{2}N_k^i(t-1)\right). \quad (10.82)$$

Repeating the above step t times we have

$$\mathbb{E}\left(\exp\left(\lambda Z_k^i(t)-\frac{\lambda^2\sigma_k^2}{2}N_k^i(t)\right)\right)\leq 1. \quad (10.83)$$

Note that we have

$$\mathsf{P}\left(\exp\left(\lambda Z_k^i(t)-\frac{\lambda^2\sigma_i^2}{2}N_k^i(t)\right)\geq\exp\left(2\kappa\vartheta\right)\right) \quad (10.84)$$

$$= \mathsf{P}\left(\lambda Z_k^i(t) - \frac{\lambda^2\sigma_k^2}{2}N_k^i(t) \geq 2\kappa\vartheta\right) \tag{10.85}$$

$$= \mathsf{P}\left(\frac{Z_k^i(t)}{\sqrt{N_k^i(t)}} \geq \frac{2\kappa\vartheta}{\lambda\sqrt{N_k^i(t)}} + \frac{\sigma_k^2}{2}\lambda\sqrt{N_k^i(t)}\right). \tag{10.86}$$

Fix a constant $\zeta > 1$. Then $1 \leq N_k^i(t) \leq \zeta^{D_t}$ where $D_t = \frac{\log((d_i(G_\gamma)+1)t)}{\log\zeta}$. For $\lambda_l = \frac{2}{\sigma_k}\sqrt{\frac{\kappa\vartheta}{\zeta^{l-1/2}}}$ and $\zeta^{l-1} \leq N_k^i(t) \leq \zeta^l$ we have

$$\frac{2\kappa\vartheta}{\lambda_l}\sqrt{\frac{1}{N_k^i(t)}} + \frac{\sigma_k^2}{2}\lambda_l\sqrt{N_k^i(t)} = \sigma_k\sqrt{\kappa\vartheta}\left(\sqrt{\frac{\zeta^{l-1/2}}{N_k^i(t)}} + \sqrt{\frac{N_k^i(t)}{\zeta^{l-1/2}}}\right) \leq \sqrt{\vartheta}, \tag{10.87}$$

where $\kappa = \frac{1}{\sigma_k^2\left(\zeta^{\frac{1}{4}}+\zeta^{-\frac{1}{4}}\right)^2}$.

Then we have

$$\left\{\frac{Z_k^i(t)}{\sqrt{N_k^i(t)}} \geq \sqrt{\vartheta}\right\} \subset \cup_{l=1}^{D_t}\left\{\frac{Z_k^i(t)}{\sqrt{N_k^i(t)}} \geq \frac{2\kappa\vartheta}{\lambda_l\sqrt{N_k^i(t)}} + \frac{\sigma_k^2}{2}\lambda_l\sqrt{N_k^i(t)}\right\} \tag{10.88}$$

$$= \cup_{l=1}^{D_t}\left\{\lambda_l Z_k^i(t) - \frac{\lambda_l^2\sigma_k^2}{2}N_k^i(t) \geq 2\kappa\vartheta\right\}. \tag{10.89}$$

Recall from the Markov inequality that $\mathsf{P}(Y \geq a) \leq \frac{\mathbb{E}(Y)}{a}$ for any positive random variable Y. Thus from (10.89) and Markov inequality we get,

$$\mathsf{P}\left(\frac{Z_k^i(t)}{\sqrt{N_k^i(t)}} \geq \sqrt{\vartheta}\right) \leq \sum_{l=1}^{D_t}\exp(-2\kappa\vartheta). \tag{10.90}$$

Then we have,

$$\mathsf{P}\left(\frac{Z_k^i(t)}{N_k^i(t)} \geq \sqrt{\frac{\vartheta}{N_k^i(t)}}\right) \leq \sum_{l=1}^{D_t}\exp(-2\kappa\vartheta) \tag{10.91}$$

Recall that $\forall\zeta > 1$ we have

$$\frac{4}{\left(\zeta^{\frac{1}{4}}+\zeta^{-\frac{1}{4}}\right)^2} \geq 1 - \frac{(\zeta-1)^2}{16} \tag{10.92}$$

Substituting $\vartheta = 2\sigma_k^2(\xi+1)\log t$ we get

$$\mathsf{P}\left(\left|\widehat{\mu}_k^i(t) - \mu_k\right| > \sigma_k\sqrt{\frac{2(\xi+1)\log t}{N_k^i(t)}}\right) \leq \frac{\log((d_i(G_\gamma)+1)t)}{\log\zeta}\frac{1}{t^{(\xi+1)\left(1-\frac{(\zeta-1)^2}{16}\right)}}. \quad (10.93)$$

This concludes the proof of Lemma 17. □

We prove a Lemma similar to Lemma 14 for message-passing as follows.

Lemma 18. *Let $\bar{\chi}(G_\gamma)$ is the clique number of graph G_γ. Let $\eta_k = \left(\frac{8(\xi+1)\sigma_k^2}{\Delta_k^2}\right)\log T$. Then we have*

$$\sum_{i=1}^{N}\mathbb{E}[n_k^i(T)] \leq \left(\sum_{i=1}^{N}(1-p_i p^{\gamma_i}) + \bar{\chi}(G_\gamma)\max_{i\in[N]} p_i p^{\gamma_i}\right)\eta_k + N(\gamma+1) + \quad (10.94)$$

$$+\sum_{i=1}^{N}\sum_{t=1}^{T-1}\left[\mathsf{P}\left(\widehat{\mu}_1^i(t) \leq \mu_1 - C_1^i(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^i(t) \geq \mu_k + C_k^i(t)\right)\right] \quad (10.95)$$

Proof. Note that for each suboptimal arm $k > 1$ we have

$$\sum_{i=1}^{N}\mathbb{E}[n_k^i(T)] = \sum_{i=1}^{N}\sum_{t=1}^{T}\mathsf{P}\left(A_i(t)=k\right) = \sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathsf{P}\left(A_i(t)=k\right). \quad (10.96)$$

Let $\tau_{k,\mathcal{C}}$ denote the maximum time step when the total number of times arm k has been played by all the agents in clique $\mathcal{C}$ is at most $\eta_k + |\mathcal{C}|$ times. This can be stated as $\tau_{k,\mathcal{C}} := \max\{t \in [T] : \sum_{i\in\mathcal{C}} n_k^i(t) \leq \eta_k + |\mathcal{C}|\}$. Then, we have that $\eta_k < \sum_{i\in\mathcal{C}} n_k^i(\tau_{k,\mathcal{C}}) \leq \eta_k + |\mathcal{C}|$.

For each agent $i \in \mathcal{C}$ let

$$\bar{N}_k^i(t) := \sum_{j\in\mathcal{C}}\sum_{\tau=1}^{t}\sum_{\tau'=1}^{\tau}\mathbf{1}\{A_j(\tau')=k\}\mathbf{1}\{(i,j)\in E_{\tau',\tau}\},$$

denote the sum of the total number of times agent i pulled arm k and the total number of observations it received from agents in its clique about arm k until time

t. Define $\bar{\tau}^i_{k,\mathcal{C}} := \max\{t \in [T] : \bar{N}^i_k(t) \leq \eta_k\}$. For each agent $i \in [N]$ let $\overline{\tau}^i_{k,\mathcal{C}} = \max\{\tau_{k,\mathcal{C}} + \gamma_i - 1, \bar{\tau}^i_{k,\mathcal{C}}\}$.

Note that $N^i_k(t) \geq \bar{N}^i_k(t), \forall t$, hence for all $i \in \mathcal{C}$ we have $N^i_k(t) > \eta_k, \forall t > \overline{\tau}^i_{k,\mathcal{C}}$. Here we consider that $\bar{\tau}^i_{k,\mathcal{C}} \geq \tau_{k,\mathcal{C}}, \forall i$. From regret results it follows that regret for this case is greater than the regret for the case where $\bar{\tau}^i_{k,\mathcal{C}} < \tau_{k,\mathcal{C}}$ for some (or all) i.

We analyse the expected number of times agents pull suboptimal arm k as follows,

$$\sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathbf{1}\{A_i(t)=k\} \tag{10.97}$$

$$=\sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t=1}^{\tau_{k,\mathcal{C}}}\mathbf{1}\{A_i(t)=k\}+\sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{\overline{\tau}^i_{k,\mathcal{C}}}\mathbf{1}\{A_i(t)=k\}+\sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t>\overline{\tau}^i_{k,\mathcal{C}}}^{T}\mathbf{1}\{A_i(t)=k\} \tag{10.98}$$

$$\leq\sum_{\mathcal{C}\in\mathcal{C}_\gamma}(\eta_k+|\mathcal{C}|)+\sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{\overline{\tau}^i_{k,\mathcal{C}}}\mathbf{1}\{A_i(t)=k\} \tag{10.99}$$

$$+\sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t>\overline{\tau}^i_{k,\mathcal{C}}}^{T}\mathbf{1}\{A_i(t)=k\}\mathbf{1}\big\{N^i_k(t-1)>\eta_k\big\}. \tag{10.100}$$

Taking expectation we have

$$\sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T}\mathsf{P}\left(A_i(t)=k\right) \tag{10.101}$$

$$\leq\sum_{\mathcal{C}\in\mathcal{C}_\gamma}(\eta_k+2|\mathcal{C}|)+\sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{\overline{\tau}^i_{k,\mathcal{C}}}\mathsf{P}\left(A_i(t)=k\right) \tag{10.102}$$

$$+\sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t>\overline{\tau}^i_{k,\mathcal{C}}}^{T-1}\mathsf{P}\left(A_i(t+1)=k,N^i_k(t)>\eta_k\right). \tag{10.103}$$

Case 1. For agent i we have that $\tau_{k,\mathcal{C}}+\gamma_i-1 \geq \bar{\tau}^i_{k,\mathcal{C}}$ then we have $\overline{\tau}^i_{k,\mathcal{C}} = \tau_{k,\mathcal{C}}+\gamma_i-1$. Then we have $\sum_{t>\tau_{k,\mathcal{C}}}^{\overline{\tau}^i_{k,\mathcal{C}}}\mathbf{1}\{A_i(t)=k\} \leq \gamma_i - 1$

Case 2. For agent i we have that $\tau_{k,\mathcal{C}} + \gamma_i - 1 < \bar{\tau}^i_{k,\mathcal{C}}$ then we have $\overline{\tau}^i_{k,\mathcal{C}} = \bar{\tau}^i_{k,\mathcal{C}}$.

$$\sum_{t>\tau_{k,\mathcal{C}}}^{\overline{\tau}^i_{k,\mathcal{C}}} \mathbf{1}\{A_i(t) = k\} \tag{10.104}$$

$$= \tilde{N}^i_k(\overline{\tau}^i_{k,\mathcal{C}}) - \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathbf{1}\{A_i(t) = k\} - \sum_{j\neq i, j\in\mathcal{C}} \sum_{t=1}^{\overline{\tau}^i_{k,\mathcal{C}}} \sum_{\tau=1}^{t} \mathbf{1}\{A_j(\tau) = k\}\mathbf{1}\{(i,j) \in E_{\tau,t}\} \tag{10.105}$$

$$\leq \tilde{N}^i_k(\overline{\tau}^i_{k,\mathcal{C}}) - \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathbf{1}\{A_i(t) = k\} - \sum_{j\neq i, j\in\mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}+\gamma_i-1} \sum_{\tau=1}^{t} \mathbf{1}\{A_j(\tau) = k\}\mathbf{1}\{(i,j) \in E_{\tau,t}\}. \tag{10.106}$$

Taking the expectation we have

$$\sum_{i\in\mathcal{C}} \sum_{t>\tau_{k,\mathcal{C}}}^{\overline{\tau}^i_{k,\mathcal{C}}} \mathsf{P}\left(A_i(t) = k\right) \leq |\mathcal{C}|\eta_k - \eta_k + \sum_{i\in\mathcal{C}}(\gamma_i - 1) - \sum_{i\in\mathcal{C}} p_i p^{\gamma_i} \sum_{j\neq i, j\in\mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathsf{P}\left(A_j(t) = k\right) \tag{10.107}$$

$$= |\mathcal{C}|\eta_k - \eta_k + \sum_{i\in\mathcal{C}}(\gamma_i - 1) - \sum_{i\in\mathcal{C}} p_i p^{\gamma_i} \sum_{j\neq i, j\in\mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathbb{E}\left(n^j_k(\tau_{k,\mathcal{C}})\right) \tag{10.108}$$

$$\leq \left(|\mathcal{C}| - 1 - \left(\sum_{j\in\mathcal{C}} p_j p^{\gamma_j} - \max_{i\in[N]} p_i p^{\gamma_i}\right)\right)\eta_k + \sum_{i\in\mathcal{C}}(\gamma_i - 1). \tag{10.109}$$

Substituting these results to (10.103) we get

$$\sum_{\mathcal{C}\in\mathcal{C}_\gamma} \sum_{i\in\mathcal{C}} \sum_{t=1}^{T} \mathsf{P}\left(A_i(t) = k\right) \leq \sum_{\mathcal{C}\in\mathcal{C}_\gamma} \left(|\mathcal{C}| - 1 - \left(\sum_{j\in\mathcal{C}} p_j p^{\gamma_j} - \max_{i\in[N]} p_i p^{\gamma_i}\right)\right)\eta_k + \sum_{i\in[N]}(\gamma_i - 1) \tag{10.110}$$

$$+\sum_{\mathcal{C}\in\mathcal{C}_\gamma}(\eta_k+2|\mathcal{C}|)+\sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t>\overline{\tau}^i_{k,\mathcal{C}}}^{T-1}\mathsf{P}\left(A_i(t+1)=k,N^i_k(t)>\eta_k\right) \tag{10.111}$$

$$\leq\left(\sum_{i=1}^{N}(1-p_ip^{\gamma_i})+\bar{\chi}(G_\gamma)\max_{i\in[N]}p_ip^{\gamma_i}\right)\eta_k+\sum_{i\in[N]}\gamma_i+N \tag{10.112}$$

$$+\sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t>\overline{\tau}^i_{k,\mathcal{C}}}^{T-1}\mathsf{P}\left(A_i(t+1)=k,N^i_k(t)>\eta_k\right) \tag{10.113}$$

This concludes the proof of Lemma 18. □

Now we prove Theorem 18 as follows. Thus using Lemmas 16, 17 and 18 we obtain

$$\mathsf{Reg}_G(T)\leq 8(\xi+1)\sigma_k^2\left(\sum_{i=1}^{N}(1-p_ip^{\gamma_i})+\bar{\chi}(G_\gamma)\max_{i\in[N]}p_ip^{\gamma_i}\right)\left(\sum_{k>1}\frac{\log T}{\Delta_k}\right) \tag{10.114}$$

$$+\left(\sum_{i=1}^{N}\gamma_i+4N\right)\sum_{k>1}\Delta_k+4\sum_{i=1}^{N}\left(3\log(3(d_i(G_\gamma)+1))+(\log\left(d_i(G_\gamma)+1\right))\right)\sum_{k>1}\Delta_k \tag{10.115}$$

10.9.3 Proof of Theorem 19

Agents receive information from their neighbors with a stochastic time delay. Let $\mathcal{N}_D$ be the maximum number of outstanding arm pulls by all the agent. We start by proving a result similar to Lemma 14.

Lemma 19. *Let $\bar{\chi}(G)$ is the clique number of graph G. Let $\eta_k=\left(\frac{8(\xi+1)\sigma_k^2}{\Delta_k^2}\right)\log T$. Then we have*

$$\sum_{i=1}^{N}\mathbb{E}[n^i_k(T)]\leq\bar{\chi}(G)\eta_k+\mathbb{E}[\mathcal{N}_D]+2N+ \tag{10.116}$$

$$+\sum_{i=1}^{N}\sum_{t=1}^{T-1}\left[\mathsf{P}\left(\widehat{\mu}^i_1(t)\leq\mu_1-C^i_1(t)\right)+\mathsf{P}\left(\widehat{\mu}^i_k(t)\geq\mu_k+C^i_k(t)\right)\right] \tag{10.117}$$

Proof. Let $\mathcal{C}$ be a non overlapping clique covering of G. Note that for each suboptimal arm $k > 1$ we have

$$\sum_{i=1}^{N} \mathbb{E}[n_k^i(T)] = \sum_{i=1}^{N} \sum_{t=1}^{T} \mathsf{P}\left(A_i(t) = k\right) = \sum_{\mathcal{C}\in\mathcal{C}} \sum_{i\in\mathcal{C}} \sum_{t=1}^{T} \mathsf{P}\left(A_i(t) = k\right). \tag{10.118}$$

Let $\tau_{k,\mathcal{C}}$ denote the maximum time step such that the total number of arm pulls shared by agents in clique $\mathcal{C}$ from arm k is at most $\eta_k + |\mathcal{C}|$. For each agent $i \in \mathcal{C}$ let $D_i(\tau_{k,\mathcal{C}})$ be the number of outstanding messages by agent i from arm k at time $\tau_{k,\mathcal{C}}$. This can be stated as $\tau_{k,\mathcal{C}} := \max\{t \in [T] : \sum_{i\in\mathcal{C}} n_k^i(t) \leq \eta_k + \sum_{i\in\mathcal{C}} D_i(\tau_{k,\mathcal{C}}) + |\mathcal{C}|\}$. Then, we have that $\eta_k + \sum_{i\in\mathcal{C}} D_i(\tau_{k,\mathcal{C}}) < \sum_{i\in\mathcal{C}} n_k^i(\tau_{k,\mathcal{C}}) \leq \eta_k + \sum_{i\in\mathcal{C}} D_i(\tau_{k,\mathcal{C}}) + |\mathcal{C}|$.

Note that for all $i \in \mathcal{C}$ we have $N_k^i(t) > \eta_k, t > \tau_{k,\mathcal{C}}$.

We analyse the expected number of times agents pull suboptimal arm k as follows,

$$\sum_{\mathcal{C}\in\mathcal{C}} \sum_{i\in\mathcal{C}} \sum_{t=1}^{T} \mathbf{1}\{A_i(t) = k\} \tag{10.119}$$

$$= \sum_{\mathcal{C}\in\mathcal{C}} \sum_{i\in\mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathbf{1}\{A_i(t) = k\} + \sum_{\mathcal{C}\in\mathcal{C}} \sum_{i\in\mathcal{C}} \sum_{t>\tau_{k,\mathcal{C}}}^{T} \mathbf{1}\{A_i(t) = k\} \tag{10.120}$$

$$\leq \sum_{\mathcal{C}\in\mathcal{C}} \left(\eta_k + \sum_{i\in\mathcal{C}} D_i(\tau_{k,\mathcal{C}}) + 2|\mathcal{C}|\right) + \sum_{\mathcal{C}\in\mathcal{C}} \sum_{i\in\mathcal{C}} \sum_{t>\tau_{k,\mathcal{C}}}^{T-1} \mathbf{1}\{A_i(t+1) = k\}\mathbf{1}\{N_k^i(t) > \eta_k\}. \tag{10.121}$$

Taking expectation we have

$$\sum_{\mathcal{C}\in\mathcal{C}_\gamma} \sum_{i\in\mathcal{C}} \sum_{t=1}^{T} \mathsf{P}\left(A_i(t) = k\right) \tag{10.122}$$

$$\leq \bar{\chi}(G_\gamma)\eta_k + \mathbb{E}\left[\max_{t\in[T]} \sum_{i=1}^{N} D_i(t)\right] + 2N + \sum_{i=1}^{N} \sum_{t=1}^{T-1} \mathsf{P}\left(A_i(t+1) = k, N_k^i(t) > \eta_k\right) \tag{10.123}$$

The proof of Lemma 19 follows from Lemma 13 and (10.123). □

We upper bound the expected number of outstanding messages by any agent using results by [38] as follows.

Lemma 20. . *Let D_{total} be the maximum number of outstanding messages by all the agent at any time step $t \in [T]$ and let $\mathbb{E}[\tau]$ be the expected delay of any message. Then with probability at least $1 - \frac{1}{T}$ we have*

$$\mathbb{E}[D_{total}] \leq N\mathbb{E}[\tau] + 2\log T + 2\sqrt{N\mathbb{E}[\tau]\log T}. \tag{10.124}$$

Proof. The proof directly follows from Lemma 2 by [38]. □

From Lemmas 19, 15, 16 and 20 we obtain with probability at least $1 - \frac{1}{T}$

$$\mathsf{Reg}_G(T) \leq 8(\xi+1)\sigma_k^2\bar{\chi}(G)\left(\sum_{k>1}\frac{\log T}{\Delta_k}\right) \tag{10.125}$$

$$+\left(N\mathbb{E}[\tau] + 2\log T + 2\sqrt{N\mathbb{E}[\tau]\log T}\right)\sum_{k>1}\Delta_k \tag{10.126}$$

$$+ 5N\sum_{k>1}\Delta_k + 4\sum_{i=1}^{N}\left(3\log(3(d_i(G)+1)) + (\log(d_i(G)+1))\right)\sum_{k>1}\Delta_k \tag{10.127}$$

10.9.4 Proof of Theorem 20

We first restate the result for clarity.

Theorem 23. *Algorithm 8 obtains, with probability at least $1-\delta$, cumulative group regret of*

$$\mathsf{Reg}_G(T) = \mathcal{O}\left(KTN\gamma\epsilon + \psi(G_\gamma)\sum_{k\neq k^\star}\frac{\log T}{\Delta_k}\log\left(\frac{K\psi(G_\gamma)\log T}{\delta}\right) + N\Delta_k + \frac{N\log(N\gamma\log T)}{\Delta_k}\right).$$

Proof. We decompose the regret based on the dominating set and epoch. Let $\mathcal{I} \subseteq \mathcal{V}$ be an dominating set of G_γ and M_i be the number of epochs run for the subgraph

covered by agent i. Observe that the total regret can be written as,

$$\mathsf{Reg}_G(T) = \sum_{i \in \mathcal{I}} \left(\sum_{k=1}^{K} \sum_{t=1}^{T} \Delta_k \cdot \left(\mathsf{P}(A_i(t) = k) + \sum_{j \in \mathcal{N}_i(G_\gamma)} \mathsf{P}(A_j(t) = k) \right) \right). \quad (10.128)$$

First, observe that $A_j(t) = A_i(t - d(i,j))$ for all $j \in \mathcal{N}_i(G_\gamma)$ and all $t \in [d(i,j), T]$. Rearranging the above, we have,

$$\mathsf{Reg}_G(T) \leqslant \sum_{i \in \mathcal{I}} \left(\sum_{k=1}^{K} \Delta_k \cdot \left(\sum_{t=1}^{T} \mathsf{P}(A_i(t) = k) + \sum_{j \in \mathcal{N}_i(G_\gamma)} \left(\sum_{t=1}^{T-d(i,j)} \mathsf{P}(A_i(t) = k) + d(i,j) \right) \right) \right) \quad (10.129)$$

$$\leqslant \sum_{i \in \mathcal{I}} \left(\sum_{k=1}^{K} \Delta_k \cdot |\mathcal{N}_i^+(G_\gamma)| \cdot \left(\sum_{t=1}^{T-\gamma} \mathsf{P}(A_i(t) = k) + \gamma \right) \right) \quad (10.130)$$

$$= \sum_{i \in \mathcal{I}} \left(|\mathcal{N}_i^+(G_\gamma)| \sum_{k=1}^{K} \Delta_k \left(\sum_{t=1}^{T-\gamma} \mathsf{P}(A_i(t) = k) \right) \right) + N\gamma \sum_{k=1}^{K} \Delta_k. \quad (10.131)$$

$$(10.132)$$

Now, observe that we run two algorithms in tandem for each subgraph of G induced by $\mathcal{N}_i^+(G_\gamma)$. Let us split the total number of rounds of the game into epochs that run arm elimination and the intermittent periods of running `UCB1`. We denote the cumulative regret in the i^{th} induced subgraph from rounds γ to T as $\mathsf{Reg}_{\mathcal{N}_i^+(G_\gamma)}(T)$, and analyse it separately.

$$\mathsf{Reg}_{\mathcal{N}_i^+(G_\gamma)}(T) \leqslant |\mathcal{N}_i^+(G_\gamma)| \sum_{k=1}^{K} \left(\Delta_k \left(\sum_{t \leq T-\gamma : t \in \mathcal{M}_i} \mathsf{P}(A_i(t) = k) + \sum_{t \leq T-\gamma : t \notin \mathcal{M}_i} \mathsf{P}(A_i(t) = k) \right) \right). \quad (10.133)$$

Here $\mathcal{M}_i$ denotes the rounds in which arm elimination is played in the agents in the i^{th} induced subgraph. Since each `UCB1` period after each epoch is of length 2γ, we have at most $2\gamma M_i$ rounds of isolated `UCB1`. We analyse the second term in the bound first.

By the standard analysis of the UCB1 algorithm [7], we have that the leader agent, i.e. agent i, incurs $\mathcal{O}(K \log T/\Delta)$ regret. We therefore have,

$$|\mathcal{N}_i^+(G_\gamma)| \sum_{k=1}^{K} \left(\Delta_k \left(\sum_{t \notin \mathcal{M}_i} \mathsf{P}(A_i(t) = k) \right) \right) \leqslant |\mathcal{N}_i^+(G_\gamma)| \cdot \sum_{k=1}^{K} \left(\left(1 + \frac{\pi^2}{3}\right) \Delta_k + \frac{8 \log(2\gamma M_i)}{\Delta_k} \right).$$

Now, we analyse the first term in the regret bound. By Theorem 24, we have that with probability at least $1 - \delta$ simultaneously for each induced subgraph corresponding to agent $i \in \mathcal{I}$,

$$\sum_{k=1}^{K} \left(\Delta_k \left(\sum_{m \in \mathcal{M}_i} \mathbb{E}\left[n_k^i(m)\right] \right) \right) = \mathcal{O}\left(\gamma\epsilon \cdot KT|\mathcal{N}_i^+(G_\gamma)| + \sum_{k>1} \frac{\log T}{\Delta_k} \log\left(\frac{K\psi(G_\gamma)}{\delta} \log T \right) \right).$$

Summing over each leader agent, we have that with probability at least $1 - \delta$,

$$\sum_{i \in \mathcal{I}} \sum_{k=1}^{K} \left(\Delta_k \left(\sum_{m \in \mathcal{M}_i} \mathbb{E}\left[n_k^i(m)\right] \right) \right) = \mathcal{O}\left(\gamma\epsilon \cdot KTN + \sum_{k>1} \frac{\log T}{\Delta_k} \log\left(\frac{K\psi(G_\gamma)}{\delta} \log T \right) \right).$$

Next, observe that for all i, $|\mathcal{M}_i| \leq \log(MT)$ by Lemma 21. Replacing this result in the UCB1 regret for each leader, and summing over all $i \in \mathcal{I}$, we have,

$$\mathsf{Reg}_G(T) = \mathcal{O}\left(\gamma\epsilon \cdot KTN + \sum_{k>1} \psi(G_\gamma) \frac{\log T}{\Delta_k} \log\left(\frac{K\psi(G_\gamma) \log T}{\delta} \right) + N\Delta_k + \frac{N \log(N\gamma \log T)}{\Delta_k} \right).$$

□

Lemma 21. *For any leader i, let $L^i(m)$ denote the length of the m^{th} epoch of arm elimination. Then, we have that $L^i(m)$ satisfies,*

$$2^{2m-2}\lambda \leq L^i(m) \leq K 2^{2m-2}\lambda.$$

Furthermore, the number of arm elimination epochs for agent i satisfies $M_i \leq \log_2(T - 2\gamma)$.

Proof. The proof closely follows the proof of Lemma 2 in [29]. For any leader i, let $\hat{k}$ be the optimal arm under $r^i(m)$, therefore $r^i_\star(m) - r^i_{\hat{k}}(m) \leq 0$ and therefore $\Delta^i_{\hat{k}}(m) = 2^{-m}$, and therefore $L^i(m+1) \geq n^i_{\hat{k}}(m+1) = \lambda(\Delta^i_{\hat{k}}(m))^{-2} \geq 2^{2m}\lambda$. Next, observe that $\Delta^i_k(m) \geq 2^{-m}$ for each arm k, and therefore $n^i_k(m+1) \leq 2^{2m}\lambda$, giving the upper bound.

For the second part, observe that $\sum_{m=1}^{M_i} L^i(m) \leq T - 2\gamma M_i \leq T - 2\gamma$, and that $L^i(m) \geq \frac{2^{2m-2}\lambda}{|\mathcal{N}^+_i(G_\gamma)|}$. Summing over $m \in [M_i]$ and taking the logarithm provides us with the result. □

Lemma 22. *Denote $\mathcal{E}$ to be the event for which,*

$$\left\{ \forall m, i, k, \left| r^i_k(m) - \mu_k \right| \leq 2\gamma\epsilon + \frac{\Delta^i_k(m-1)}{16} \bigwedge \sum_{\substack{t \in \mathcal{M}_i(m) \\ j \in \mathcal{N}^+_i(G_\gamma)}} X^j_k(t + d(i,j)) \leq 2n^i_k(m) \right\}$$

Then, we have that $\mathsf{P}(\mathcal{E}) \geq 1 - \delta$.

Proof. Recall that at each step in the epoch, the leader agent picks an arm k with probability $p^i_k(m) = \frac{n^i_k(m)}{L^i(m)}$, and let $X^j_k(t)$ denote whether agent j picks arm k at time t. Let $C_{j\to i}(t) = \tilde{r}_{j\to i}(t) - r_j(t)$ denote the corruption in the transmitted reward from agent j when it reaches agent i, and $\mathcal{M}_i(m) = [T_i(m-1)+1, \cdots, T_i(m)]$ denote the $L^i(m)$ steps in the m^{th} epoch for the arm elimination algorithm run by the leader i. We then have,

$$r^i_k(m) = \frac{1}{n^i_k(m)} \left(\sum_{\substack{t \in \mathcal{M}_i(m) \\ j \in \mathcal{N}^+_i(G_\gamma)}} X^j_k(t + d(i,j)) \cdot (r_j(t + d(i,j)) + C_{j\to i}(t + d(i,j))) \right)$$

For simplicity, let

$$A_k^i(m) = \sum_{\substack{t\in\mathcal{M}_i(m)\\ j\in\mathcal{N}_i^+(G_\gamma)}} X_k^j(t+d(i,j))\cdot r_j(t+d(i,j)), B_k^i(m) = \sum_{\substack{t\in\mathcal{M}_i(m)\\ j\in\mathcal{N}_i^+(G_\gamma)}} X_k^j(t+d(i,j))\cdot C_{j\to i}(t+d(i,j)).$$

We can bound the first summation by a multiplicative version of the Chernoff-Hoeffding bound [5] as each r_j is bounded within $[0,1]$ and X_k^i is a random variable in $\{0,1\}$ with mean $p_k^i(m)L^i(m)\mu_k \leq n_k^i(m)$. We obtain that with probability at least $1-\beta/2$,

$$\left|\frac{A_k^i(m)}{n_k^i(m)} - \mu_i\right| \leq \sqrt{\frac{3\log(\frac{4}{\beta})}{n_k^i(m)}}.$$

To bound the second term, we must construct a filtration that ensures that the corruption is measurable. For the set $\mathcal{N}_i^+(G_\gamma)$, consider an order σ of the N agents, such that $\sigma[1]=i$, followed by the agents at distance 1 from i, then the agents at distance 2, and so on until distance γ, and next consider the ordering $\{\tilde{r}_\tau\}_{\tau=1}^{|\mathcal{N}_i^+(G_\gamma)|t}$ of the rewards generated by all agents within $\mathcal{M}_i(m)$ where $\tilde{r}_\tau$ is the reward obtained by agent $j=(\sigma(\tau) \mod |\mathcal{N}_i^+(G_\gamma)|)$ during the round $\lfloor\frac{\tau}{|\mathcal{N}_i^+(G_\gamma)|}\rfloor + d(i,j)$, and similarly consider an identical ordering of the pulled arms $\{\widetilde{X}_\tau\}_{\tau=1}^{|\mathcal{N}_i^+(G_\gamma)|t}$. Now consider the filtration $\{\mathcal{F}_t\}_{t=1}^{T|\mathcal{N}_i^+(G_\gamma)|}$ generated by the two stochastic processes of $\tilde{r}$ and $\widetilde{X}$. Clearly, the corruption $C_{\sigma(j)\to i}(t)$ is deterministic conditioned on $\mathcal{F}_{t-1}$. Moreover, we have that the pulled arm satisfies, for all $\tau \in [|\mathcal{N}_i^+(G_\gamma)|t]$ that $\mathbb{E}[\tilde{X}_\tau|\mathcal{F}_{\tau-1}] = p_k^i(m)$. Furthermore, since the corruption in each round is bounded and deterministic, we have that the sequence $Z_\tau = (\widetilde{X}_\tau - p_k^i(m))\cdot\widetilde{C}_\tau$ (where $\widetilde{C}_\tau$ is the corresponding ordering of corruptions) is a martingale difference sequence with respect to $\{\mathcal{F}_\tau\}_{\tau=1}^T$. Now, consider the slice of $[|\mathcal{N}_i^+(G_\gamma)|t]$ that is present within $B_k^i(m)$, and let the corresponding indices be given by the set $\widetilde{\mathcal{M}}_i(m)$. Using the fact that the observed rewards are bounded, we have that,

$$\sum_{\tau\in\widetilde{\mathcal{M}}_i(m)} \mathbb{E}[Z_\tau^2|\mathcal{F}_{\tau-1}] \leq \sum_{\tau\in\widetilde{\mathcal{M}}_i(m)} |\widetilde{C}_\tau|\cdot\mathbb{V}(Z_\tau) \leq p_k^i(m)\cdot\sum_{\tau\in\widetilde{\mathcal{M}}_i(m)}\widetilde{C}_\tau \leq \gamma C L^i(m).$$

We then have by Freedman's inequality that with probability at least $1-\frac{\beta}{4}$,

$$\frac{B_k^i(m)}{n_k^i(m)} \leq \frac{p_k^i(m)}{n_k^i(m)}\left(\sum_{\tau\in\widetilde{\mathcal{M}}_i(m)}\widetilde{C}_\tau + \frac{\gamma C L^i(m)+\log(4/\beta)}{n_k^i(m)}\right) \leq 2\gamma\epsilon + \sqrt{\frac{\log(4/\beta)}{16 n_k^i(m)}}.$$

The last inequality follows from the fact that $n_k^i(m)\geq\lambda\geq 16\ln(4/\beta)$. With the same probability, we can derive a bound for the other tail. Now, observe that since each X_k^i is a random variable with mean p_k^i, we have by the multiplicative Chernoff-Hoeffding bound that the probability that the sum of $L^i(m)$ i.i.d. bernoulli trials with mean $p_k^i(m)$ is greater than $2p_k^i(m)\cdot L^i(m) = 2n_k^i(m)$ is at most $2\exp(-n_k^i(m)/3)\leq 2\exp(-\lambda/3)\leq\beta$.

To conclude the proof, we apply each of the above bounds with $\beta=\frac{\delta}{2K\alpha(G_\gamma)\log T}$ to each epoch and arm. Observe that $\beta\geq 4\exp\left(-\frac{\lambda}{16}\right)$. Now, since $\log(4/\beta)=\lambda/(32)^2$ we have that,

$$\mathbb{P}\left(\left|r_k^i(m)-\mu_k\right| \geq 2\gamma\epsilon + \frac{\Delta_k^i(m-1)}{16}\bigwedge\sum_{\substack{t\in\mathcal{M}_i(m)\\ j\in\mathcal{N}_i^+(G_\gamma)}} X_k^j(t+d(i,j)) \geq 2n_k^i(m)\right) \leq \frac{\delta}{2K\alpha(G_\gamma)\log T}.$$

The proof concludes by a union bound over all epochs, arms and agents in $\mathcal{I}$. □

Lemma 23. *If the event $\mathcal{E}$ (Lemma 22) occurs then for each $i\in\mathcal{I}, m\in\mathcal{M}_i$,*

$$-2\gamma\epsilon - \frac{\Delta_\star^i(m-1)}{8} \leq r_\star^i(m)-\mu_\star \leq 2\gamma\epsilon.$$

Proof. Observe that $r^i_\star(m) \geq r^i_{k^\star}(m) - \frac{1}{16}\Delta^i_{k^\star}(m-1)$. This fact coupled with the fact that $\mathcal{E}$ holds provides the lower bound. The upper bound is obtained by observing that,

$$r^i_\star(m) \leq \max_i \left\{ \mu_i + 2\gamma\epsilon + \frac{\Delta^i_k(m-1)}{16} - \frac{\Delta^i_k(m-1)}{16} \right\} \leq \mu_\star + 2\gamma\epsilon.$$

□

Lemma 24. *If the event $\mathcal{E}$ (Lemma 22) occurs then for each $i \in \mathcal{I}, m \in \mathcal{M}_i$,*

$$\Delta^i_k(m) \geq \frac{\Delta_k}{2} - 6\gamma\epsilon \sum_{n=1}^{m} 8^{n-m} - \frac{3}{4}2^{-m}.$$

Proof. We first bound $\Delta^i_k(m) \leq 2(\Delta_k + 2^{-m} + 2\gamma\epsilon \cdot \sum_{n=1}^{m} 8^{n-m})$ under $\mathcal{E}$ by induction. Observe that when $m = 1$ we have that trivially $\Delta^i_k(1) \leq 1 \leq 2 \cdot 2^{-1}$. Now, if the bound holds for epoch $m-1$ for any agent, we have by Lemma 23,

$$r^i_\star(m) - r^i_k(m) = r^i_\star(m) - \mu_\star + \mu_\star - \mu_k + \mu_k - r^i_k(m) \leq 4\gamma\epsilon + \Delta_k + \frac{\Delta^i_k(m-1)}{16}.$$

Replacing the induction hypothesis in the upper bound, we have,

$$\begin{aligned} r^i_\star(m) - r^i_k(m) &\leq 4\gamma\epsilon + \Delta_k + \frac{1}{8}\left(\Delta_k + 2^{-(m-1)} + 2\gamma\epsilon \cdot \sum_{n=1}^{m-1} 8^{n-m+1}\right) \\ &\leq 2(\Delta_k + 2^{-m} + 2\gamma\epsilon \cdot \sum_{n=1}^{m} 8^{n-m}). \end{aligned}$$

Now, we bound the gaps as,

$$\Delta^i_k(m) \geq r^i_\star(m) - r^i_k(m) \geq \Delta_k - 4\gamma\epsilon - \left(\frac{\Delta^i_{k^\star}(m-1)}{8} - \frac{\Delta^i_k(m-1)}{16}\right).$$

The last inequality follows from Lemma 23 and the event $\mathcal{E}$. Replacing the bound from induction we obtain,

$$\begin{aligned}\Delta_k^i(m) &\geq \Delta_k - 4\gamma\epsilon - \left(\frac{6\gamma\epsilon}{8}\sum_{n=1}^{m} 2^{n-m} + \frac{3}{8}2^{-(m-1)} + \frac{\Delta_k}{8}\right) \\ &\geq \frac{\Delta_k}{2} - 6\gamma\epsilon\sum_{n=1}^{m} 8^{n-m} - \frac{3}{4}2^{-m}.\end{aligned}$$

□

Theorem 24. *The cumulative regret for all agents within each independent set corresponding to leader $i \in \mathcal{I}$ satisfy simultaneously, with probability at least $1-\delta$,*

$$\sum_{m=1}^{\mathcal{M}_i}\sum_{k=1}^{K}\Delta_k\mathbb{E}[n_k^i(m)] = \mathcal{O}\left(\log\left(\frac{K\psi(G_\gamma)}{\delta}\log(T)\right)\log(T)\left(\sum_{k=1}^{K}\frac{1}{\Delta_k}\right) + \gamma\epsilon\cdot KT\cdot|\mathcal{N}_i^+(G_\gamma)|\right).$$

Proof. We bound the regret in each epoch $m \in \mathcal{M}_i$ for each arm $k \neq k^\star$ based on three cases.

Case 1. $0 \leq \Delta_k \leq 4/2^m$: We have that $n_k^i(m) \leq \lambda 2^{2(m-1)}$ since $\Delta_k^i(m-1) \geq 2^{m-1}$, and hence,

$$\Delta_k\mathbb{E}[n_k^i(m)] \leq \frac{4\lambda}{\Delta_k^2}\cdot\Delta_k = 4\lambda\cdot\frac{1}{\Delta_k}.$$

Case 2. $\Delta_k > 4/2^m$ and $\gamma\epsilon\sum_{n=1}^{m} 8^{n-m} \leq \Delta_k/64$: We have by Lemma 24,

$$\Delta_k^i(m) \geq \frac{\Delta_k}{2} - 6\gamma\epsilon\sum_{n=1}^{m} 8^{n-m} - \frac{3}{4}2^{-m} \geq \Delta_k\left(\frac{1}{2} - \frac{3}{32} - \frac{3}{8}\right) = \frac{\Delta_k}{32}.$$

Therefore, we have that $n_k^i(m) \leq \frac{1024\lambda}{\Delta_k^2}$, and hence the regret is,

$$\Delta_k\mathbb{E}[n_k^i(m)] \leq \frac{1024\lambda}{\Delta_k^2}\cdot\Delta_k = 1024\lambda\cdot\frac{1}{\Delta_k}.$$

Case 3. $\Delta_k > 4/2^m$ and $\gamma\epsilon\sum_{n=1}^{m} 8^{n-m} > \Delta_k/64$: This implies that $\Delta_k \leq 64\gamma\epsilon \cdot \sum_{n=1}^{m} 8^{n-m}$. Therefore,

$$\begin{aligned}\Delta_k \mathbb{E}[n_k^i(m)] &\leq 64\lambda\gamma\epsilon \left(\sum_{n=1}^{m} 8^{n-m}\right) \cdot 2^{2(m-1)} \\ &\leq 64\lambda\gamma\epsilon \left(\frac{8^{m+1}}{7}\right) \cdot \frac{2^{2(m-1)}}{2^{3m}} \\ &\leq \frac{512}{7}\gamma\epsilon \cdot L^i(m).\end{aligned}$$

Here the last inequality follows from Lemma 21. Putting it together and summing over all epochs and arms, we have with probability at least $1-\delta$ simultaneously for each $i \in \mathcal{I}$,

$$\sum_{m=1}^{\mathcal{M}_i}\sum_{k=1}^{K} \Delta_k \mathbb{E}[n_k^i(m)] \leq 1024^2 \log\left(\frac{8K\psi(G_\gamma)}{\delta}\log(T)\right)\log(T)\left(\sum_{k=1}^{K}\frac{1}{\Delta_k}\right) + 74\gamma\epsilon \cdot KT \cdot |\mathcal{N}_i^+(G_\gamma)|.$$

□

10.9.5 Proof of Theorem 21

In this section we consider that each agent passes messages upto γ-hop neighbors. Agents do not use the messages received during last $\bar{\gamma}$ number of time steps.

Lemma 25. *Let $\bar{\chi}(G_\gamma)$ is the clique number of graph G_γ. Let $\eta_k = \left(\frac{8(\xi+1)\sigma_k^2}{\Delta_k^2}\right)\log T$. Then we have*

$$\sum_{i=1}^{N}\mathbb{E}[n_k^i(T)] \leq \bar{\chi}(G_\gamma)\eta_k + (N - \bar{\chi}(G_\gamma))\,(\bar{\gamma}+\gamma-1) + 2N + \tag{10.134}$$

$$+ \sum_{i=1}^{N}\sum_{t=1}^{T-1}\left[\mathsf{P}\left(\widehat{\mu}_1^i(t) \leq \mu_1 - C_1^i(t)\right) + \mathsf{P}\left(\widehat{\mu}_k^i(t) \geq \mu_k + C_k^i(t)\right)\right] \tag{10.135}$$

Proof. Let $\mathcal{C}_\gamma$ be a non overlapping clique covering of G_γ. Note that for each suboptimal arm $k > 1$ we have

$$\sum_{i=1}^{N} \mathbb{E}[n_k^i(T)] = \sum_{i=1}^{N} \sum_{t=1}^{T} \mathsf{P}\left(A_i(t) = k\right) = \sum_{\mathcal{C}\in\mathcal{C}_\gamma} \sum_{i\in\mathcal{C}} \sum_{t=1}^{T} \mathsf{P}\left(A_i(t) = k\right). \tag{10.136}$$

Let $\tau_{k,\mathcal{C}}$ denote the maximum time step when the total number of times arm k has been played by all the agents in clique $\mathcal{C}$ is at most $\eta_k + (|\mathcal{C}|-1)(\bar{\gamma}+\gamma-1)+|\mathcal{C}|$ times. This can be stated as $\tau_{k,\mathcal{C}} := \max\{t \in [T] : \sum_{i\in\mathcal{C}} n_k^i(t) \leq \eta_k + (|\mathcal{C}|-1)(\bar{\gamma}+\gamma-1)+|\mathcal{C}|\}$. Then, we have that $\eta_k+(|\mathcal{C}|-1)(\bar{\gamma}+\gamma-1) < \sum_{i\in\mathcal{C}} n_k^i(\tau_{k,\mathcal{C}}) \leq \eta_k+(\mathcal{C}-1)(\bar{\gamma}+\gamma-1)+|\mathcal{C}|$.

For each agent $i \in \mathcal{C}$ let

$$\bar{N}_k^i(t) := \sum_{\tau=1}^{t} \mathbf{1}\{A_i(\tau) = k\} + \sum_{j\neq i, j\in\mathcal{C}} \sum_{\tau=1}^{t-\bar{\gamma}} \sum_{\tau'=1}^{\tau} \mathbf{1}\{A_j(\tau') = k\}\mathbf{1}\{(i,j) \in E_{\tau',\tau}\},$$

denote the sum of the total number of times agent i pulled arm k and the total number of observations it received from agents in its clique about arm k until time t.

Note that for all $i \in \mathcal{C}$ we have $N_k^i(t) > \eta_k, \forall t > \tau_{k,\mathcal{C}}$.

We analyse the expected number of times agents pull suboptimal arm k as follows,

$$\sum_{\mathcal{C}\in\mathcal{C}_\gamma} \sum_{i\in\mathcal{C}} \sum_{t=1}^{T} \mathbf{1}\{A_i(t) = k\} \tag{10.137}$$

$$= \sum_{\mathcal{C}\in\mathcal{C}_\gamma} \sum_{i\in\mathcal{C}} \sum_{t=1}^{\tau_{k,\mathcal{C}}} \mathbf{1}\{A_i(t) = k\} + \sum_{\mathcal{C}\in\mathcal{C}_\gamma} \sum_{i\in\mathcal{C}} \sum_{t>\bar{\tau}_{k,\mathcal{C}}^i}^{T} \mathbf{1}\{A_i(t) = k\} \tag{10.138}$$

$$\leq \sum_{\mathcal{C}\in\mathcal{C}_\gamma} (\eta_k + (|\mathcal{C}|-1)(\bar{\gamma}+\gamma-1) + 2|\mathcal{C}|) + \sum_{\mathcal{C}\in\mathcal{C}_\gamma} \sum_{i\in\mathcal{C}} \sum_{t>\tau_{k,\mathcal{C}}}^{T-1} \mathbf{1}\{A_i(t+1) = k\}\mathbf{1}\left\{N_k^i(t) > \eta_k\right\}. \tag{10.139}$$

Taking expectation we have

$$\sum_{\mathcal{C}\in\mathcal{C}_\gamma} \sum_{i\in\mathcal{C}} \sum_{t=1}^{T} \mathsf{P}\left(A_i(t) = k\right) \tag{10.140}$$

$$\leq \sum_{\mathcal{C}\in\mathcal{C}_\gamma} (\eta_k + (|\mathcal{C}|-1)(\bar{\gamma}+\gamma-1) + 2|\mathcal{C}|) + \sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t>\tau_{k,\mathcal{C}}}^{T-1} \mathsf{P}\left(A_i(t+1)=k, N_k^i(t)>\eta_k\right). \tag{10.141}$$

$$= \bar{\chi}(G_\gamma)\eta_k + (N-\bar{\chi}(G_\gamma))(\bar{\gamma}+\gamma-1) + 2N + \sum_{\mathcal{C}\in\mathcal{C}_\gamma}\sum_{i\in\mathcal{C}}\sum_{t=1}^{T-1} \mathsf{P}\left(A_i(t+1)=k, N_k^i(t)>\eta_k\right) \tag{10.142}$$

The proof of Lemma 25 follows from Lemma 13 and (10.142). □

Now we prove Theorem 21 as follows. Thus using Lemmas 16, 17 and 25 we obtain

$$\mathsf{Reg}_G(T) \leq 8(\xi+1)\sigma_k^2\bar{\chi}(G_\gamma)\left(\sum_{k>1}\frac{\log T}{\Delta_k}\right) + ((N-\bar{\chi}(G_\gamma)(\bar{\gamma}+\gamma-1)+5N)\sum_{k>1}\Delta_k \tag{10.143}$$

$$+ 4\sum_{i=1}^{N}(3\log(3(d_i(G_\gamma)+1)) + (\log(d_i(G_\gamma)+1)))\sum_{k>1}\Delta_k \tag{10.144}$$

10.9.6 Lower Bounds

22

Theorem 25 (Minimax Rate)**.** *For any multi-agent algorithm $\mathcal{A}$, there exists a $K-$armed environment over N agents with $\Delta_k \leq 1$ such that,*

$$\mathsf{Reg}_G(\mathcal{A},T) \geqslant c\sqrt{KN(T+\widetilde{d}(G))}.$$

Furthermore, if $\mathcal{A}$ is an agnostic decentralized policy, there exists a $K-$armed environment over N agents with $\Delta_k \leq 1$ for any connected graph G and $\gamma \geq 1$ such that, for some absolute constant c'

$$\mathsf{Reg}_G(\mathcal{A},T) \geqslant c'\sqrt{\alpha^\star(G_\gamma)KNT}.$$

Where $\tilde{d}(G) = \sum_{i=1}^{d^\star(G)} \bar{d}_{=i} \cdot i$ *denotes the average delay incurred by message-passing across the network* G, $d_{=i} = \frac{1}{N}\sum_{i,j} \mathbf{1}\{d(i,j) = i\}$ *denotes the number of agent pairs that are at distance exactly* i, *and* $\alpha^\star(G_\gamma) = \frac{N}{1+\bar{d}_\gamma}$ *is Turan's lower bound [90] on* $\alpha(G_\gamma)$.

Proof. Our approach is an extension of the single-agent bandit lower bound [12]. Let $\mathcal{A}$ be a deterministic (multi-agent) algorithm, and let the empirical distribution of arm pulls across all agents be given by $p^i(t) = (p_1^i(t), ..., p_K^i(t))$, where $p_k(t) = \frac{n_i^k(T)}{T}$. Consider the random variable J_t^i drawn according to $p^i(t)$ and P_i denote the law of J_t when drawn from arm k having parameter $\frac{1+\varepsilon}{2}$ (and other arms with parameter $\frac{1-\varepsilon}{2}$). We have,

$$\mathsf{P}_k\left(J_t^i = j\right) = \mathbb{E}_k\left[\frac{n_i^k(T)}{T}\right].$$

Since on pulling any arm $k' \neq k$, we obtain regret ε, we therefore have for the group regret,

$$\begin{aligned}\mathbb{E}_k\left[\sum_{t=1}^{T}\left(N \cdot r_k(t) - \sum_{i\in\mathcal{V}} r_{A_i}(t)\right)\right] &= \varepsilon \cdot T \cdot \sum_{i\in\mathcal{V}} \mathsf{P}_k\left(J_t^i = k'\right) \\ &= \varepsilon \cdot T \cdot \sum_{i\in\mathcal{V}}\left(1 - \sum_{k'\neq k} \mathsf{P}_k\left(J_t^i = k'\right)\right).\end{aligned}$$

By Pinsker's inequality and averaging over all $k \in [K]$, we have for any $i \in \mathcal{V}$,

$$\frac{1}{K}\sum_{k=1}^{K} \mathsf{P}_k\left(J_t^i = k\right) \leqslant \frac{1}{K} + \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{2}\mathsf{KL}(\mathsf{P}_0, \mathsf{P}_k)}.$$

We now bound the R.H.S. using the chain rule for KL-divergence. Since we assume that $\mathcal{A}$ is deterministic, we have that the rewards obtained by the agent i until time t from its neighborhood alone determine uniquely the empirical distribution of plays. Here, the analysis diverges from that of the single-agent bandit as a richer set of

observations is available to each agent. Denote the set of rewards observed by agent i at instant τ be given by $\mathcal{O}_i(\tau)$. First, observe that since each reward is i.i.d., we have for any k,

$$\mathsf{KL}(\mathsf{P}_0(\mathcal{O}_i(\tau)), \mathsf{P}_k(\mathcal{O}_i(\tau))) = |\mathcal{O}_i(\tau)| \cdot \mathsf{KL}\left(\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2}\right)$$

For $k = 0$ the above divergence is 0. When we consider the standard single-agent setting, $|\mathcal{O}_i(\tau)| = 1$, recovering the usual bound. Now, by the chain rule, we have that, at round t for any agent i, and arm $k \in [K]$,

$$\begin{aligned}\mathsf{KL}(\mathsf{P}_0(t), \mathsf{P}_k(t)) &= \mathsf{KL}(\mathsf{P}_0(1), \mathsf{P}_k(1)) + \sum_{\tau=2}^{t} |\mathcal{O}_i(\tau)| \, \mathsf{KL}\left(\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2}\right) \\ &= \mathsf{KL}\left(\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2}\right) \mathbb{E}_0\left[\sum_{j \in \mathcal{V}} n_j^k(t - d(i,j))\right].\end{aligned}$$

Replacing this result in the earlier equation, we have by the concavity of KL divergence:

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K} \mathsf{P}_k\left(J_t^i = k\right) &\leqslant \frac{1}{K} + \frac{1}{K}\sum_{k=1}^{K} \sqrt{\frac{1}{2}\mathsf{KL}(\mathsf{P}_0, \mathsf{P}_k)} \\ &\leqslant \frac{1}{K} + \frac{1}{K}\sum_{k=1}^{K} \sqrt{\mathsf{KL}\left(\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2}\right) \mathbb{E}_0\left[\sum_{j \in \mathcal{V}} n_j^k(T - d(i,j))\right]} \\ &\leqslant \frac{1}{K} + \sqrt{\left(\frac{TN - \sum_{j=1}^{d^\star(G)} d_{=j}(i) \cdot j}{K}\right) \cdot \mathsf{KL}\left(\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2}\right)}.\end{aligned}$$

Now, observe that the KL divergence between Bernoulli bandits can be bounded as

$$\mathsf{KL}(p, q) \leq \frac{(p-q)^2}{q(1-q)}.$$

Substituting we get,

$$\frac{1}{K}\sum_{k=1}^{K}\mathsf{P}_k\left(J_t^i=k\right)\leqslant\frac{1}{K}+\sqrt{\frac{4\varepsilon^2(NT-\sum_{j=1}^{d^\star(G)}d_{=j}(i)\cdot j)}{(1-\varepsilon^2)K}}.$$

Replacing this in the regret and using $\varepsilon\leqslant 1/2$, we get that,

$$\begin{aligned}
&\mathbb{E}_k\left[\sum_{t=1}^{T}\left(N\cdot r_k(t)-\sum_{i\in\mathcal{V}}r_{A_i}(t)\right)\right]\\
&\geqslant\varepsilon\cdot T\cdot\sum_{i\in\mathcal{V}}\left(1-\frac{1}{K}-\sqrt{\frac{4\varepsilon^2(NT-\sum_{j=1}^{d^\star(G)}d_{=j}(i)\cdot j)}{(1-\varepsilon^2)K}}\right)\\
&\geqslant\varepsilon\cdot T\cdot\sum_{i\in\mathcal{V}}\left(\frac{1}{2}-4\varepsilon\sqrt{\frac{(NT-\sum_{j=1}^{d^\star(G)}d_{=j}(i)\cdot j)}{3K}}\right)\\
&=\frac{\varepsilon\cdot NT}{2}-\frac{4\varepsilon^2NT}{\sqrt{K}}\left(\sum_{i,j\in\mathcal{V}}T-d(i,j)\right)^{1/2}
\end{aligned}$$

Setting $\varepsilon=c\cdot\sqrt{\frac{K}{N(T-\sum_{j=1}^{d^\star(G)}\bar{d}_{=j}\cdot j)}}$ where c is a constant to be tuned later, we have,

$$\begin{aligned}
\mathbb{E}_k\left[\sum_{\tau=1}^{T}\left(N\cdot r_{k,t}-\sum_{i\in\mathcal{V}}r_{A_i(t),t}\right)\right]&\geqslant\left(\frac{c}{2}-\frac{4c^2}{\sqrt{3}}\right)\cdot\sqrt{\frac{KN^2T^2}{N(T-\sum_{j=1}^{d^\star(G)}\bar{d}_{=j}\cdot j)}}\\
&\geqslant 0.027\sqrt{KN(T+\sum_{j=1}^{d^\star(G)}\bar{d}_{=j}\cdot j)}.
\end{aligned}$$

This proves the first part of the theorem. Now, when the policies are decentralized and agnostic, the chain rule step can be factored as follows.

$$\begin{aligned}
\mathsf{KL}(\mathsf{P}_0(t),\mathsf{P}_k(t))&=\mathsf{KL}(\mathsf{P}_0(1),\mathsf{P}_k(1))+\sum_{\tau=2}^{t}|\mathcal{O}_i(\tau)|\,\mathsf{KL}\left(\frac{1-\varepsilon}{2},\frac{1+\varepsilon}{2}\right)\\
&=\mathsf{KL}\left(\frac{1-\varepsilon}{2},\frac{1+\varepsilon}{2}\right)\mathbb{E}_0\left[\sum_{j\in\mathcal{N}_\gamma^+(G)}n_j^k(t-d(i,j))\right].
\end{aligned}$$

Note that here instead of taking the cumulative sum over all $\mathcal{V}$ we select only those agents that are within the $\gamma-$neighborhood of i in G, since conditioned on these observations the rewards of the agents are independent of all other rewards (by Assumption), and hence the higher-order KL divergence terms are 0. Replacing this in the analysis gives us the following decomposition (after similar steps as the first part):

$$\begin{aligned}\mathbb{E}_k\left[\sum_{t=1}^{T}\left(Nr_k(t)-\sum_{i\in\mathcal{V}}r_{A_i}(t)\right)\right] &\geqslant \frac{NT\varepsilon}{2}-\frac{4\varepsilon^2 T}{\sqrt{3K}}\cdot\sum_{i\in\mathcal{V}}\left(\sum_{j:\mathcal{N}_\gamma^+(i)}T-d(i,j)\right)^{1/2}\\ &\geqslant \frac{NT\varepsilon}{2}-\frac{4\varepsilon^2 N^{1/2}T}{\sqrt{3K}}\cdot\left(\sum_{i\in\mathcal{V}}\sum_{j:\mathcal{N}_\gamma^+(i)}T-d(i,j)\right)^{1/2}\end{aligned}$$

Setting $\varepsilon = c\cdot\sqrt{\frac{NK}{\sum_{i\in\mathcal{V}}\sum_{j:\mathcal{N}_\gamma^+(i)}T-d(i,j)}}$ where c is a constant to be tuned later, we have,

$$\begin{aligned}\mathbb{E}_k\left[\sum_{t=1}^{T}\left(N\cdot r_k(t)-\sum_{i\in\mathcal{V}}r_{A_i}(t)\right)\right] &\geqslant \left(\frac{c}{2}-\frac{4c^2}{\sqrt{3}}\right)\cdot\sqrt{\frac{N^3T^2}{\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{N}_i^+(G_\gamma)}T-d(i,j)}}\\ &\geqslant \left(\frac{c}{2}-\frac{4c^2}{\sqrt{3}}\right)\cdot\sqrt{\frac{N^3T}{\sum_{i\in\mathcal{V}}1+d_i(G_\gamma)}}\\ &\geqslant \frac{3}{4}\left(\frac{c}{2}-\frac{4c^2}{\sqrt{3}}\right)\sqrt{\alpha^\star(G_\gamma)NT}\\ &\geqslant 0.019\sqrt{\alpha^\star(G_\gamma)NT}.\end{aligned}$$

The constants in both settings are obtained by optimizing c over $\mathbb{R}$. Extending this to random (instead of deterministic) algorithms is straightforward via Fubini's theorem, see Theorem 2.6 of [11]. □

10.9.7 Pseudo code

Algorithm 8: `RCL-RC`: Cooperative Hybrid Arm Elimination

Parameters: Confidence $\delta \in (0,1)$, horizon T, graph G with exploration set $\mathcal{I} \subseteq \mathcal{V}$. Initialize $T_i(0) = K, \forall i \in \mathcal{I} \lambda = 1024 \log\left(\frac{8K\psi(G_\gamma)}{\delta} \log_2 T\right)$ and $\Delta_k^i(0) = 1, \forall\ k \in [K]$ and $i \in \mathcal{I}$.

for each subgraph $\mathcal{N}_i^+(G_\gamma)$ where $i \in \mathcal{I}$ **do**
- **for** $t = 1, ..., K$, each agent $j \in \mathcal{N}_i^+(G_\gamma)$ **do**
  - Play arm K and get reward $r_j(t)$.
- **end**
- **for** epoch $m_i = 1, 2, ...,$ **do**
  - Set $n_k^i(m_i) = \lambda(\Delta_k^i(m_i - 1))^{-2} \forall k \in [K]$.
  - $N_i(m_i) = \sum_k n_k^i(m_i)$ and $T_i(m_i) = T_i(m_i) + N_i(m_i) + 2\gamma$.
  - **for** agent $j \in \mathcal{N}_i^+(G_\gamma)$ **do**
    - **for** $t = T_i(m_i - 1)$ *to* $s = T_i(m_i - 1) + 2\gamma$ **do**
      - **if** $j \neq i$ **then**
        - **if** $t \leq K + d(i,j)$ **then**
          - Pull random arm.
        - **end**
        - **else**
          - Pull $A_j(t) = A_i(t - d(i,j))$ and get reward $r_j(t)$.
        - **end**
      - **end**
      - **else**
        - Pull $A_j(t) =$ `UCB1`(t)
      - **end**
    - **end**
    - **for** $t = T_i(m_i - 1) + 2\gamma$ *to* $T_i(m_i)$ **do**
      - **if** $j \neq i$ **then**
        - Pull $A_j(t) = A_i(t - d(i,j))$ and get reward $r_j(t)$.
      - **end**
      - **else**
        - Pull an arm $A_i(t) = k \in [K]$ with probability $n_k^i(m_i)/N_k(m_i)$.
      - **end**
    - **end**
  - **end**
- **end**

end

Algorithm 9: RCL-LF

Input: Arms $k \in [K]$, variance proxy upper bound σ^2, parameter ξ
Initialize: $N_k^i(0) = \widehat{\mu}_k^i(0) = C_k^i(0) = 0, \forall k, i$
for each iteration $t \in [T]$ **do**
  for each agent $i \in [N]$ **do**
    /* Sampling phase */
    if $t = 1$ **then**
      $A_t^i \leftarrow \textsc{RandomArm}([K])$
    end
    else
      $A_t^i \leftarrow \arg\max_k \widehat{\mu}_k^i(t-1) + C_k^i(t-1)$
    end
    /* Send messages */
    $\textsc{Create}\left(\mathbf{m}_t^i := \left\langle A_t^i, r_t^i, i, t \right\rangle\right)$
    $\textsc{Send}\left(\mathbf{M}_t^i \leftarrow \mathbf{M}_{t-1}^i \cup \mathbf{m}_t^i\right)$
  end
  for each agent $i \in [N]$ **do**
    /* Receive messages */
    for each neighbor $j \in \mathcal{N}_i(G_\gamma)$ **do**
      /* Discard messages with probability $1 - p_i$ */
      for each message $\mathbf{m} \in \mathbf{M}_t^j$ **do**
        with probability p_i, $\mathbf{M}_t^i \leftarrow \mathbf{M}_t^i \cup \mathbf{m}$
        with probability $1 - p_i$, $\mathbf{M}_t^i \leftarrow \mathbf{M}_t^i$
      end
    end
    /* Update estimates */
    for each arm $k \in [K]$ **do**
      $\textsc{Calculate}\left(N_k^i(t), \widehat{\mu}_k^i(t), C_k^i(t)\right)$
    end
  end
end

Chapter 11

Heterogeneous Explore-Exploit Strategies on Multi-Star Networks

Udari Madhushani and Naomi Ehrich Leonard

We investigate the benefits of heterogeneity in multi-agent explore-exploit decision making where the goal of the agents is to maximize cumulative group reward. To do so we study a class of distributed stochastic bandit problems in which agents communicate over a multi-star network and make sequential choices among options in the same uncertain environment. Typically, in multi-agent bandit problems, agents use homogeneous decision-making strategies. However, group performance can be improved by incorporating heterogeneity into the choices agents make, especially when the network graph is irregular, i.e. when agents have different numbers of neighbors. We design and analyze new heterogeneous explore-exploit strategies, using the multi-star as the model irregular network graph. The key idea is to enable center agents to do more exploring than they would do using the homogeneous strategy, as a means of providing more useful data to the peripheral agents. In the case all agents broadcast their reward values and choices to their neighbors with the same probability, we provide theoretical guarantees that group performance improves under

the proposed heterogeneous strategies as compared to under homogeneous strategies. We use numerical simulations to illustrate our results and to validate our theoretical bounds.

11.1 Introduction

The influence of agent heterogeneity on cooperation in social learning has been a recent focus of research in many fields, including ecology, sociology, and decision theory [**?**]. Studies on evolutionary human behavior provide evidence that individual differences can be leveraged to enhance collective prosperity [**?**]. Motivated by applications such as social foraging and multi-robot coordination tasks, we study and design cooperative strategies for a group of agents making sequential explore-exploit decisions in an uncertain environment. The strategies we design incorporate agent heterogeneity to optimize the performance of the group through collective learning.

Consider a group of agents, each making a sequence of choices among options in an uncertain environment in order to maximize collective payoff. At each time step in the sequence, each agent chooses an option depending on the knowledge it has acquired about the environment up to that time step. Maximizing payoff necessitates striking a balance between making choices that yield high immediate payoff, i.e., exploiting, and making choices that yield high information content and possibly high future payoffs, i.e., exploring. When an agent fails to acquire sufficient information about the environment to make optimal decisions, it must sacrifice exploitation potential in order to explore. However, in the group setting, agents can recover exploitation potential by gaining information through cooperation i.e., through collective learning.

Sequential decision making in uncertain environments that requires trading off exploitation and exploration is modeled mathematically by the bandit framework [79]. In the multi-armed bandit (MAB) problem, an agent is repeatedly faced with

the task of choosing an option from a given set of options. At each time step the agent receives a stochastic reward drawn from a fixed probability distribution associated with the chosen option. The agent's goal is to maximize the cumulative reward by the end of the decision-making process. This requires choosing frequently enough the optimal option i.e., the option with highest expected reward. In order to meet this requirement, the agent must simultaneously choose options that are known to provide high rewards (exploit) and choose lesser known options (explore) that might potentially provide even higher rewards [44, 7].

Maximizing cumulative reward is equivalent to minimizing cumulative regret, defined as the loss incurred by an agent choosing a sub-optimal option instead of the optimal option. Since the probability distribution associated with each option is fixed, cumulative regret can be minimized by reducing the number of times sub-optimal options are chosen. Performance of the proposed algorithms for this problem is measured using expected cumulative regret. The paper [44] establishes that any efficient policy chooses suboptimal options asymptotically logarithmically in time. The paper [7] proposes an Upper Confidence Bound (UCB) based sampling rule that achieves a logarithmic expected cumulative regret uniformly in time.

The papers [46, **?**, 66, 84, 63, 62, 14, 61, 48, 94, 42] extend to the multi-agent setting and capture different aspects of collective learning. In [46, **?**, 66, 84], agents share their estimates of the expected reward of options with neighbors according to fixed communication structures. The papers [46, **?**] use a running consensus algorithm to update estimates and provide graph-structure-dependent performance measures that predict the relative performance of agents and networks. The paper [**?**] also addresses the case of a constrained reward model in which agents that choose the same option at the same time step receive no reward. The paper [66] proposes an accelerated consensus procedure assuming agents know the spectral gap of the communication graph and designs a decentralized UCB algorithm based on delayed rewards. The paper [84]

considers a P2P communication where an agent is only allowed to communicate with two other agents at each time step.

The papers [63, 62, 14, 61, 48, 94, 42] consider the case in which agents share reward values and choices with neighbors. In [63, 62, 14], agents use stochastic communication structures that depend on the decision-making process. In [63], each agent observes rewards and actions of its neighbors when it is exploring. In [62], each agent instead broadcasts its rewards and actions to its neighbors when it is exploring. In [14], at each time step, agents decide either to sample an option or to broadcast the last obtained reward to the entire group.

The setup in our earlier paper [61] is closest to that in the present paper: agents observe reward values and actions of their neighbors defined by a network graph that changes in time according to probabilistic edge weights. An underlying fixed network graph is given, and each agent k observes its neighbors with probability p_k. The communication structure is independent of the decision-making process.

The papers [46, ?, 66, 84, 63, 62, 14, 61] consider homogeneous protocols, whereas the papers [48, 94, 42] consider protocols where some agents (followers) copy actions of others (leaders). In [48], followers observe rewards and choices of their neighbors. In [94] one leader explores and estimates the mean reward of options, while all other agents choose the option with highest estimated mean per the leader. The paper [42] proposes the FYL algorithm, which uses a deterministic communication protocol and exploits degree heterogeneity of the communication network graph. FYL outperforms our algorithm when $p_k = p = 1$; however, our algorithm provides a method to exploit agent heterogeneity when agents share information with probability $0 < p < 1$.

When communication among agents is defined by an irregular network graph, e.g., some agents serve as information hubs, group performance can be improved by using heterogeneous explore-exploit strategies. To understand this, consider an environment with unconstrained resources. Then, agents can only influence the decisions

of one another through the information they share, and the structure of interactions that defines neighbors, i.e., who is sharing information with whom, strongly affects the quality and quantity of information received by each individual.

We consider the case that all agents broadcast their instantaneous rewards and actions to their neighbors with probability p. This communication protocol is motivated by real-world applications in which estimates of mean rewards or the sum of collected rewards, which rely on the history of choices and rewards, are deliberately not disclosed to protect privacy [23]. For example, in user targeted recommender systems [96] (or clinical trials [89]), sharing user (patient) history of choices can reveal sensitive information about users (patients). Even when an agent is broadcasting only its current rewards and actions to neighbors, an adversarial agent can listen to the broadcasts and access the history of choices made by the agent. To reduce such privacy leakage we consider agents that broadcast instantaneous rewards and actions probabilistically. Further, if communication failures are possible, then having agents broadcast only current rewards and actions avoids problems associated with agents losing track of what information has and has not been received by neighbors. In this context, $1 - p$ represents the probability of communication failure.

In irregular and centralized networks like the multi-star, center agents have more neighbors and thus receive more information than peripheral agents. This leads to an imbalanced exploitation potential across the group [**?**, **?**], and group performance degrades with increasing number of peripheral agents. We investigate improving group performance by leveraging heterogeneity in the exploitation potential of agents. To do so we propose heterogeneous explore-exploit strategies that require center agents to explore more and thus increase the exploitation potential of peripheral agents.

The multi-star network models recommender systems, where there are many small servers, assigned to different regions, that each make sequential recommendations based on user feedback and communicate only with a large central server. Perfor-

mance can be improved by using the central server to suggest more exploratory recommendations which allows the system to gather more information about user preferences. Probabilistic communication accounts for random communication failures between servers.

The paper is organized as follows. In Section **??** we provide the problem formulation and notation. Section **??** presents the proposed algorithm and intuition. We analyze performance of the proposed algorithm in Section **??** and provide improved theoretical bounds for the expected cumulative group regret. In Section 9.10.11 we show numerical simulations to illustrate and validate the theoretical results. We conclude in Section **??**.

11.2 Problem Formulation

In this section we present the problem formulation and relevant mathematical notations. Consider a group of K agents, each faced with the same N-armed bandit problem for T time steps. At each time step $t \in \{1, \dots, T\}$, each agent chooses an option and receives a stochastic reward associated with the chosen option. Let X_i be a sub-Gaussian random variable that denotes the *reward associated with option* $i \in \{1, \dots, N\}$. Sub-Gaussian rewards include widely used distributions such as Bernoulli, Gaussian, and bounded rewards. Define $\mu_i = \mathbb{E}(X_i)$ and σ_i^2 as the *expected reward* and *variance proxy* associated with option i, respectively. Let $i^* = \arg_i \max\{\mu_1, \dots, \mu_N\}$ be the *optimal option* with highest expected reward. Define $\Delta_i = \mu_{i^*} - \mu_i$ as the *expected reward gap* between option i^* and option i.

Let $G(\mathcal{V}, \mathcal{E})$ be a fixed undirected network graph that defines the structure of the interactions between agents. This captures the inherent hard communication constraints of the system. Here $\mathcal{V}$ is a set of K vertices such that each vertex corresponds to an agent. Each edge $e(k, j) \in \mathcal{E}$ in the graph denotes that agent k and

agent j are *neighbors*. At each time step, each agent broadcasts its reward value and action to its neighbors with *broadcasting probability* p. Let $\mathbb{I}^t_{\{k,j\}}$ be the indicator random variable that takes value 1 if agent k receives information from agent j at time t and 0 otherwise. Then, for every time t, $\mathbb{E}(\mathbb{I}^t_{\{k,j\}}) = p, \forall k, j$ such that $e(k,j) \in \mathcal{E}$, and $\mathbb{E}(\mathbb{I}^t_{\{k,j\}}) = 0$ otherwise. We define $\mathbb{I}^t_{\{k,k\}} = 1, \forall k, t$.

Let d_k be the *degree* (number of neighbors) of agent k and $d_{avg} = \frac{1}{K}\sum_{k=1}^{K} d_k$ be the *average degree of the network*. Let d_k^{avg} be the *average degree of neighbors of agent* k: $d_k^{avg} = \frac{1}{d_k}\sum_{e(k,j)\in\mathcal{E}}^{K} d_j$.

We focus on *multi-star* graphs defined as follows. Let there be m center agents and $K-m$ peripheral agents. Without loss of generality let each agent k, $k \leq m$, be a *center agent*. All center agents are neighbors of one another, i.e., $e(k,j) \in \mathcal{E}, \forall k, j \leq m$, and a center agent's degree d_k is at least $m-1$. Each *peripheral agent* k, $k > m$, has exactly one neighbor ($d_k = 1$), and the neighbor is a center agent. To reduce complexity, we assume the graph is symmetric, which implies that all center agents have the same number of neighbors. Thus $K-m$ is an integer multiple of m. If $K > 2$ and $m < K$, the multi-star graph is *irregular*, i.e., the degree of center agents differs from the degree of peripheral agents. Let d_{cen} be the degree of each center agent. Then, $d_{cen} = \frac{K-m}{m} + m - 1$. When $m = 1$ the graph is a star, the most irregular multi-star graph. When $m = K$, there are no peripheral agents and the graph is all-to-all and thus regular.

Let φ_t^k be a random variable that denotes the *option chosen by agent* $k \in \{1, \ldots, K\}$ *at time* $t \in \{1, \ldots, T\}$. Let $\mathbb{I}_{\{\varphi_t^k = i\}}$ be an indicator random variable that takes value 1 if agent k chooses option i at time t and 0 otherwise. Let $n_i^k(t)$ be the *total number of times agent* k *chooses option* i *until time* t and let $N_i^k(t)$ be the *total number of times agent* k *observes option* i *until time* t. The total number of observations is the sum of the number of samples taken from option i by agent k and

the number of broadcasts on option i by its neighbors:

$$n_i^k(t) = \sum_{\tau=1}^{t} \mathbb{I}_{\{\varphi_\tau^k = i\}}, \quad N_i^k(t) = \sum_{\tau=1}^{t} \sum_{j=1}^{K} \mathbb{I}_{\{\varphi_\tau^j = i\}} \mathbb{I}_{\{k,j\}}^{\tau}. \tag{11.1}$$

Let $\widehat{\mu}_i^k(t)$ denote the estimate of expected reward of agent k for option i at time t. Then, $\widehat{\mu}_i^k(t) = \frac{S_i^k(t)}{N_i^k(t)}$, where $S_i^k(t) = \sum_{\tau=1}^{t} \sum_{j=1}^{K} X_i \mathbb{I}_{\{\varphi_\tau^j = i\}} \mathbb{I}_{\{k,j\}}^{\tau}$.

Expected regret is defined as the expected loss suffered by agents by sampling sub-optimal options. Let $R(t)$ be the *cumulative group regret* at time t. Then *expected cumulative group regret* can be computed as

$$\mathbb{E}\left(R(t)\right) = \sum_{i=1}^{N} \sum_{k=1}^{K} \Delta_i \mathbb{E}\left(n_i^k(t)\right). \tag{11.2}$$

11.3 Algorithm

To realize the goal of maximizing cumulative group reward, agents should minimize the number of times they sample sub-optimal options. Each agent employs an agent-based strategy that captures the trade-off between exploring and exploiting by constructing an objective function that strikes a balance between the estimation of the expected reward and the uncertainty associated with the estimate [7].

Since center agents have more neighbors they are more likely to obtain a high number of observations. This reduces the uncertainty associated with their estimate of the expected reward of options. Thus, identifying the optimal option requires less exploring, which increases their exploitation potential. Since peripheral agents only have one neighbor they are more likely to obtain a low number of observations. Thus, identifying the optimal option requires more exploring, which decreases their exploitation potential. Further, since center agents do less exploring, the usefulness of the information they broadcast is reduced, also decreasing the peripheral agents' exploitation potential. Accordingly, homogeneous sampling rules in irregular, multi-

star networks lead to imbalanced exploitation potential across the group and thus degraded group performance.

To improve group performance, we propose heterogeneous explore-exploit strategies that regulate exploitation potential across the network. When center agents are more exploratory their performance degrades, but the usefulness of the information they broadcast increases and so the performance of peripheral agents improves. When there are more peripheral agents than center agents, and broadcasting probability p is sufficiently high, the performance improvement obtained by peripheral agents outweighs the performance degradation incurred by center agents, and group performance increases. If p is too small, for example, when broadcasting is costly or risky, center agents do not broadcast enough information to benefit peripheral agents. Thus it doesn't pay for center agents to increase their exploration. Indeed, when $p = 0$ all agents have the same exploitation potential.

Using this intuition, we propose the following heterogeneous sampling rules. Assume that variance proxy σ_i^2 for each option i is known to all agents.

Definition 10. (Heterogeneous Sampling Rules) *The sampling rule* $\{\varphi_t^k\}_1^T$ *of agent* k *at time* $t \in \{1, \ldots, T\}$ *is*

$$\mathbb{I}_{\{\varphi_{t+1}^k = i\}} = \begin{cases} 1 & , \quad i = \arg\max\{Q_1^k(t), \cdots, Q_N^k(t)\} \\ 0 & , \quad \text{o.w.} \end{cases}$$

with

$$Q_i^k(t) = \widehat{\mu}_i^k(t) + C_i^k(t) \tag{11.3}$$

$$C_i^k(t) = \sigma_i \sqrt{\frac{2(1+\alpha_k)(\xi+1)\log t}{N_i^k(t)}} \tag{11.4}$$

where $\xi > 1$ *and*

$$\alpha_k = \begin{cases} \frac{p^{1-p}(d_k - d_k^{avg})}{d_k} & , \quad k \leq m \\ 0 & , \quad k > m. \end{cases} \tag{11.5}$$

$C_i^k(t)$ in (11.4) represents *agent k's uncertainty in its estimated mean of option* i, and Definition 10 implies that for any agent k, when $C_i^k(t)$ is high, agent k will more likely explore. By (11.4), $C_i^k(t)$ can be high when N_i^k, the number of agent k's observations of option i, is low, i.e., when option i is under-sampled. $C_i^k(t)$ can also be high when *agent k's exploration bias* $\alpha_k > 0$ is high.

By (11.5), $\alpha_k \neq 0$ only for center agents. Since peripheral agents have one center agent neighbor, $d_k^{avg} \leq d_k$ and thus $\alpha_k \geq 0$ for every center agent $k \leq m$. In fact, $\alpha_k \geq 0$ is designed to grow with increasing irregularity: in the regular case (all-to-all) when $m = K$, $\alpha_k = 0$, and in the most irregular case (star) when $m = 1$, $d_1 = K - 1$ and $d_1^{avg} = 1$ so $(d_1 - d_1^{avg})/d_1 = (K-2)/(K-1)$. Further, α_k grows with p according to the factor p^{1-p}, which grows rapidly for intermediate values of p and is large (i.e., saturates to 1) only when center agents are broadcasting their reward values and actions with sufficiently high probability p.

Definition 11. *To get the* corresponding homogeneous sampling rules *let* $\alpha_k = 0, \forall k$, *in Definition 10. Heterogeneous and homogeneous rules for peripheral agents are the same.*

By design, the heterogeneous rules of Definition 10 drive center agents to explore more than the corresponding homogeneous rules and only when it benefits group performance.

11.4 Performance Analysis

In this section we analyze the performance of the heterogeneous sampling rules of Definition 10. Using an approach similar to [7] with a few key modifications, we upper bound the expected cumulative group regret $\mathbb{E}(R(T))$. We show that the bound is lower than the upper bound in the case of the corresponding homogeneous sampling rules, and so we can conclude that the designed heterogeneous strategies provide better group performance than the homogeneous strategies.

By (11.2), we upper bound $\mathbb{E}(R(T))$ if we upper bound $\sum_{k=1}^{K}\mathbb{E}(n_i^k(T))$, where $n_i^k(T)$ is the number of times agent k samples sub-optimal option i until time T. By Definition 10, agent k chooses sub-optimal option i at time t if $Q_i^k(t) \geq Q_{i^*}^k(t)$. Then, $n_i^k(t) = \sum_{\tau=1}^{t}\mathbb{I}_{\{\varphi_\tau^k=i\}} \leq \sum_{\tau=1}^{t}\mathbb{I}_{\{Q_i^k(\tau)\geq Q_{i^*}^k(\tau)\}}$. For each option i and agent k let $\{\eta_i^k(t)\}_1^T$ be a sequence of nonnegative nondecreasing functions. Then,

$$\sum_{k=1}^{K}\mathbb{E}\left(n_i^k(T)\right) \leq \sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}\left(\mathbb{I}_{\{\varphi_t^k=i\}}, N_i^k(t)\leq \eta_i^k(t)\right) + \sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{P}\left(Q_i^k(t)\geq Q_{i^*}^k(t), N_i^k(t) > \eta_i^k(t)\right). \tag{11.6}$$

It remains to upper bound the right hand side of (11.6) and we do so in two steps. First, we upper bound the second summation term of (11.6) as follows. From (11.3) we have

$$\left\{Q_i^k(t)\geq Q_{i^*}^k(t)\right\} \subseteq \left\{\mu_{i^*} < \mu_i + 2C_i^k(t)\right\} \cup \left\{\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\right\} \cup \left\{\widehat{\mu}_i^k(t)\geq \mu_i + C_i^k(t)\right\}. \tag{11.7}$$

For all k let

$$\eta_i^k(t) = (1+\alpha_k)\eta_i(t), \quad \eta_i(t) = \frac{8\sigma_i^2(\xi+1)\log t}{\Delta_i^2}. \tag{11.8}$$

Then, by (11.4), $\{\mu_{i^*} < \mu_i + 2C_i^k(t)\} \cap \{N_i^k(t) > \eta_i^k(t)\} = \emptyset$ where $\emptyset$ is the empty set. Using (11.7) we obtain

$$\begin{aligned}
&\mathbb{P}\left(Q_i^k(t) \geq Q_{i^*}^k(t), N_i^k(t) > \eta_i^k(t)\right) \leq \\
&\mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\right) + \mathbb{P}\left(\widehat{\mu}_i^k(t) \geq \mu_i + C_i^k(t)\right). \qquad (11.9)
\end{aligned}$$

To upper bound the right hand side of (11.9) we use the tail probability bound provided in the following lemma.

Lemma 26. *For any $\xi > 1$, some $\zeta > 1$ and for $\sigma_i > 0$ in the uncertainty $C_i^k(t)$ given by (11.4), we get*

$$\mathbb{P}\left(\left|\widehat{\mu}_i^k(t) - \mu_i\right| > C_i^k(t)\right) \leq \frac{1}{\log \zeta} \frac{\log\left((1+d_k)t\right)}{t^{(\xi+1)(1+\alpha_k)}}.$$

Proof. From Theorem 1 in the paper [61] we have for some $\zeta > 1$ and for $\sigma_i > 0$ there exists a $\vartheta_k > 0$ such that

$$\mathbb{P}\left(\widehat{\mu}_i^k(T) - \mu_i > \sqrt{\frac{\vartheta_k}{N_i^k(T)}}\right) \leq \frac{\nu \log((d_k+1)T)}{\exp(2\kappa\vartheta_k)}$$

where, $\nu = \frac{1}{\log \zeta}$, $\kappa = \frac{1}{\sigma_i^2\left(\zeta^{\frac{1}{4}} + \zeta^{-\frac{1}{4}}\right)^2}$. Since $\alpha_k \geq 0, \forall k$, we can use $\vartheta_k = 2\sigma_i^2(1+\alpha_k)(\xi+1)\log t$ to get the statement of the lemma. □

Using the statement of Lemma 26 in (11.9),

$$\begin{aligned}
&\mathbb{P}\left(Q_i^k(t) \geq Q_{i^*}^k(t), N_i^k(t) > \eta_i^k(t)\right) \\
&\qquad \leq \frac{2}{\log \zeta} \frac{\log\left((1+d_k)t\right)}{t^{(\xi+1)(1+\alpha_k)}}. \qquad (11.10)
\end{aligned}$$

Summing the right hand side of (11.10) over t we get

$$\sum_{t=1}^{T} \frac{\log\left((1+d_k)t\right)}{t^{(\xi+1)(1+\alpha_k)}} \leq \log(1+d_k) + \frac{\log(1+d_k)(\xi\alpha_k+\xi+\alpha_k)+1}{(\xi\alpha_k+\xi+\alpha_k)^2}. \tag{11.11}$$

Since log is concave, substituting (11.11) into (11.10) we get

$$\begin{aligned}\sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(Q_i^k(t) \geq Q_{i^*}^k(t), N_i^k(t) > \eta_i^k(t)\right) &\leq \frac{2K}{\log\zeta}\log(1+d_{avg}) \\ +\frac{2}{\log\zeta}\sum_{k=1}^{K}\frac{\log(1+d_k)(\xi\alpha_k+\xi+\alpha_k)+1}{(\xi\alpha_k+\xi+\alpha_k)^2},\end{aligned} \tag{11.12}$$

which upper bounds the second summation of (11.6).

Next, we upper bound the first summation term of (11.6) as follows. Since we restrict to symmetric graphs where all center agents have the same number and type of neighbors, $\alpha_k = \alpha, \forall k \leq m$. Then, by (11.8) we have $\eta_i^k(t) = (1+\alpha)\eta_i(t), \forall k \leq m$, and $\eta_i^k(t) = \eta_i(t), \forall k > m$. Let $[x]^+ = \max\{x, 0\}$.

Lemma 27. *Let G be a symmetric multi-star graph with m center agents and $K-m$ peripheral agents. Let $\{\eta_i^k(t)\}_1^T$ be the sequence of nonnegative nondecreasing functions given by (11.8). Then with some high probability $1-\delta(p)$*

$$\begin{aligned}\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{P}\left(\mathbb{I}_{\{\varphi_t^k=i\}}=1, N_i^k(t) \leq \eta_i^k(t)\right) &\leq (K-m)\eta_i(T) \\ +\frac{m}{1+p(m-1)}\left[1-p\frac{K-m}{m}\right]^+(1+\alpha)\eta_i(T).\end{aligned}$$

Proof. Recall the definitions of $n_i^k(t)$ and $N_i^k(t)$ in (**??**). Since the communication structure is independent of the decision-making process $\forall k$,

$$\mathbb{E}\left(n_i^k(t)\right) + p \sum_{e(k,j)\in\mathcal{E}}^{K} \mathbb{E}\left(n_i^j(t)\right) = \mathbb{E}\left(N_i^k(t)\right). \tag{11.13}$$

Since $N_i^k(t)$ is a nonnegative random variable, $N_i^k(t) \leq \eta_i^k(t) \implies \mathbb{E}\left(N_i^k(t)\right) \leq \eta_i^k(t)$. Thus, from (11.13), for all k,

$$\begin{aligned} &\mathbb{E}\left(n_i^k(t), N_i^k(t) \leq \eta_i^k(t)\right) \\ &\quad + p \sum_{e(k,j)\in\mathcal{E}}^{K} \mathbb{E}\left(n_i^j(t), N_i^k(t) \leq \eta_i^k(t)\right) \leq \eta_i^k(t). \end{aligned} \tag{11.14}$$

To upper bound $\sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\mathbb{I}_{\{\varphi_t^k=i\}}= 1, N_i^k(t) \leq \eta_i^k(t)\right)$ we maximize $\sum_{k=1}^{K} \mathbb{E}(n_i^k(t))$ subject to the constraint given by (11.14). This is the linear programming optimization problem: maximize $\sum_{k=1}^{K} \mathbb{E}\left(n_i^k(t), N_i^k(t) \leq \eta_i^k(t)\right)$ subject to (11.14) and $\mathbb{E}\left(n_i^k(t), N_i^k(t) \leq \eta_i^k(t)\right) \geq 0$ for all k. For $p = 1$ the solution is the sum of $\eta_i^k(t)$ over the maximal independent set of G, which for a multi-star graph is the set of peripheral agents $k \geq m+1$. Thus, for general p we have

$$\begin{aligned} \sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\mathbb{I}_{\{\varphi_t^k=i\}}= 1, N_i^k(t) \leq \eta_i^k(t)\right) &\leq \sum_{k=m+1}^{K} \eta_i^k(T) \\ &+ \sum_{k=1}^{m} \frac{1}{1+p(m-1)} \left[1 - p\frac{K-m}{m}\right]^+ \eta_i^k(T), \end{aligned}$$

and the statement of the lemma follows. □

This concludes upper bounding the first summation of (11.6).

Theorem 26. *Consider a distributed stochastic bandit problem with N options, K agents, and T time steps. Let communication graph G be a symmetric multi-star graph with m center agents and $K-m$ peripheral agents. If all agents sample according to*

the heterogeneous sampling rules defined in Definition 10, with some high probability $1-\delta(p)$ *the expected cumulative group regret satisfies*

$$\begin{aligned}\mathbb{E}\left(R(T)\right) &\leq c_1(K,m,\alpha,p)\sum_{i=1}^{N}\frac{8\sigma_i^2(\xi+1)\log T}{\Delta_i}\\ &+\frac{2}{\log\zeta}\sum_{i=1}^{N}\Delta_i\Bigg(K\log(1+d_{avg})+(K-m)\frac{\xi\log 2+1}{\xi^2}\\ &\qquad+m\frac{\log(1+d_{cen})(\xi\alpha+\xi+\alpha)+1}{(\xi\alpha+\xi+\alpha)^2}\Bigg),\\ c_1(K,m,\alpha,p) &= K-m+\frac{m(1+\alpha)}{1+p(m-1)}\left[1-p\frac{K-m}{m}\right]^+,\end{aligned}$$

where Δ_i *is the expected reward gap between the options* i^* *and* i*,* σ_i^2 *is the variance proxy, and* $\xi,\zeta>1$*.*

Proof. Result follows from (11.2), (11.6), (11.12) and Lemma 27. □

Remark 21. *Recall that under the corresponding homogeneous sampling rules we have* $\alpha_k=0,\forall k$*. Thus, we can recover the expected cumulative group regret bound for the homogeneous sampling rules as follows:*

$$\begin{aligned}\mathbb{E}\left(R(T)\right) &\leq c_2(K,m,p)\sum_{i=1}^{N}\frac{8\sigma_i^2(\xi+1)\log T}{\Delta_i}\\ &+\frac{2}{\log\zeta}\sum_{i=1}^{N}\Delta_i\Bigg(K\log(1+d_{avg})+(K-m)\frac{\xi\log 2+1}{\xi^2}\\ &\qquad+m\frac{\log(1+d_{cen})\xi+1}{\xi^2}\Bigg),\end{aligned}$$

$c_2(K,m,p)=K-m+\frac{m}{1+p(m-1)}\left[1-p\frac{K-m}{m}\right]^+$.

When the network graph has a large enough ratio of peripheral agents to center agents and a sufficiently high broadcasting probability p, i.e. $p(K-m)/m>1$, we

have $\left[1 - p\frac{K-m}{m}\right]^+ = 0$, which implies $c_1 = c_2 = K - m$. And since $\alpha > 0$ we have

$$\frac{\log(1 + d_{cen})(\xi\alpha + \xi + \alpha) + 1}{(\xi\alpha + \xi + \alpha)^2} < \frac{\log(1 + d_{cen})\xi + 1}{\xi^2}.$$

Plugging these results into the bounds of Theorem 26 and Remark 21, we see that the heterogeneous sampling rules provide a lower theoretical regret bound than the corresponding homogeneous sampling rules, which implies that the heterogeneous sampling rules provide better group performance than the homogeneous sampling rules.

Remark 22. *Our bounds hold for sub-exponential reward distributions, where X_i is a sub-exponential random variable with mean μ_i and parameters (σ_i^2, b) with $b \leq \frac{\sigma_i}{2\sqrt{2(\xi+1)\log T}}$.*

11.5 Simulation Results

In this section we provide numerical simulations to illustrate results and validate theoretical bounds. For all simulations, we consider 10 options ($N = 10$) with Gaussian reward distributions. Expected reward for the optimal option is $\mu_{i^*} = 11$ and for all sub-optimal options $i \neq i^*$ is $\mu_i = 10$. We let variance associated with all options i be $\sigma_i^2 = 1$. Because the expected reward gaps $\Delta_i = 1$, $i \neq i^*$, are equal to the variances $\sigma_i^2 = 1$, it is a challenging problem to distinguish the optimal option from the sub-optimal options. For all simulations, we consider 1000 time steps ($T = 1000$) and use 1000 Monte Carlo simulations with $\xi = 1.01$.

We show simulation results for performance of a group of $K = 36$ agents that communicate over two different symmetric multi-star graphs and use the heterogeneous sampling rules of Definition 10. We compare to the case when agents use the corresponding homogeneous sampling rules of Definition 11. The first multi-star graph has $m = 2$ center agents and $K - m = 34$ peripheral agents, with each cen-

ter agent communicating with 17 peripheral agents and the other center agent. The second multi-star graph has $m = 3$ center agents and $K - m = 33$ peripheral agents, with each center agent communicating with 11 peripheral agents and the other center agents. In each case, center agents are interchangeable and peripheral agents are interchangeable, so the average performance of a center (peripheral) agent is the same as the individual performance of a center (peripheral) agent.

Figure 11.1 shows how average expected cumulative group regret varies with broadcasting probability p for agents using the heterogeneous rules (dotted) and homogeneous rules (solid). Regret is inversely related to performance: lower group regret implies higher group performance. Results are plotted on the left for the graph with 2 center agents and on the right for the graph with 3 center agents. When $p = 0$ there is no communication at all. So when p becomes even just a little positive and agents learn about options from their neighbors, regret falls, i.e., group performance rises.

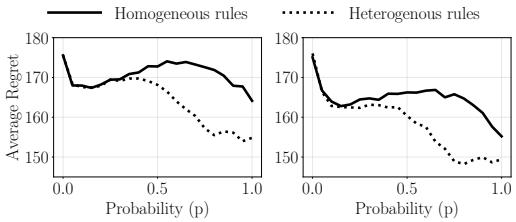


Figure 11.1: Average expected cumulative group regret for $K = 36$ agents at time $t = 1000$ as a function of broadcasting probability p with communication over a symmetric multi-star graph. Left: 2 center and 34 peripheral agents. Right: 3 center and 33 peripheral agents. Dotted lines and solid line shows average regret when agents use heterogeneous and homogeneous sampling rules, respectively.

In the case of the homogeneous rules, as p increases through intermediate values, center agents do less and less exploring and the usefulness of the information received by peripheral agents decreases. This leads to increased regret for peripheral agents, and the group overall, and thus degraded group performance. When p approaches 1, center agents receive sufficient information from their peripheral neighbors such that

their improved performance outweighs the degraded performance of peripheral agents. This leads to a final decrease in group regret and increase in group performance.

The improvement in performance provided by the heterogeneous rules relative to the homogeneous rules, as predicted by Theorem 26 and Remark 21, can be clearly seen in Figure 11.1 by observing how much lower the dotted regret curve is than the solid regret curve. The growth in regret in the homogeneous case, as p increases through intermediate values, is reduced in the heterogeneous case. This is because, by design, center agents are biased toward more exploring, which improves the information that peripheral agents receive. The group performance increase that comes, as p increases further, occurs in the heterogeneous case well before p approaches 1.

The influence of irregularity of the graph can be observed in Figure 11.1 by comparing the left plot (2 center agents and more irregular) to the right plot (3 center agents and less irregular). The results suggest that performance is higher with more center agents, i.e., with greater regularity in the graph.

Figure 11.2 shows expected cumulative regret as a function of time t for center (blue), peripheral (pink), and average (black) agents, when $p = 0.8$ and agents use the heterogeneous rules (dotted) and homogeneous rules (solid). Results are plotted on the left for the graph with 2 center agents and on the right for the graph with 3 center agents. It can be observed that, as predicted for the heterogeneous rules, the peripheral agent performance increases and the center agent performance decreases, such that group performance (as represented by the average agent) improves. Further, a comparison of left and right plots suggests that group performance improves with more center agents (more regularity).

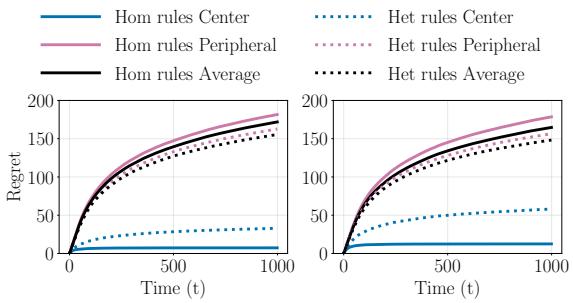


Figure 11.2: Expected cumulative regret of center agent, peripheral agent, and average agent for $K = 36$ agents as a function of time t for $p = 0.8$ and the same two symmetric multi-star graphs as in Figure 11.1: 2 center agents (left) and 3 center agents (right) where agents use heterogeneous (dotted) and homogeneous (solid) sampling rules.

11.6 Conclusion

We have designed and analyzed new heterogeneous rules for how a group of agents that share information over a network should sample an uncertain environment to maximize group reward. We consider communication networks defined by symmetric multi-star graphs, since these exemplify realistic settings. Using the multi-armed bandit problem as the explore-exploit framework, we show how sampling rules for center agents that favor exploring over exploiting make the information that center agents broadcast to their neighbors more useful, thereby increasing the total reward accumulated by the group.

Our analysis and design advance understanding of the role that heterogeneity does and can play in collective decision making. And our demonstration that heterogeneity can be leveraged to improve the performance of a cooperative multi-agent system suggests that further investigation is warranted.

Chapter 12

Heterogeneous Social Value Orientation Improves Meaningful Diversity in Various Incentive Structures

Udari Madhushani, Kevin R. McKee, John P. Agapiou, Joel Z. Leibo, Richard Everett, Thomas Anthony, Edward Hughes, Karl Tuyls and Edgar A. Duéñez-Guzmán

In reinforcement learning, Social Value Orientation (SVO) is an intrinsic motivation that remaps agent rewards based on particular target distributions of group reward. Prior studies show that groups of agents endowed with heterogeneous SVO learn diverse policies, particularly in settings that resemble the Prisoner's dilemma. Our work extends this body of results and demonstrate that (1) heterogeneous SVO leads to meaningfully diverse policies across a range of incentive structures in sequential social dilemmas, as measured by task-specific diversity metrics; and (2) learning a best response to these diverse policies leads to better zero-shot generalization in se-

quential social dilemmas with multiple equilibria. We show that these best-response agents learning a conditional policy, which we posit is the reason for improved zero-shot generalization results.

12.1 Introduction

In psychology research, Social Value Orientation (SVO) is a cognitive construct reflecting a person's preference for resource allocation between themselves and others [28, 54, 71]. While some individuals may solipsistically focus on maximizing their personal success, others demonstrate different motivations, including maximizing the difference between their own and others' outcomes (a competitive orientation), maximizing collective welfare (a prosocial orientation), or maximizing other peoples' benefit (an altruistic orientation). In artificial intelligence research, various algorithms draw inspiration from these insights in their design or implementation [68, 80]. In reinforcement learning, SVO is an intrinsic motivation that transforms an agent's reward based on its particular target distribution between its reward and the reward of others. Recently, there's been research investigating the role of SVO in situations where a group of agents or players interact in ways that involve trade-offs between their self-interest and the collective interest of the group. This research has generated valuable insights into the impact of SVO on the emergence of diverse behaviors and cooperation [68], generalization [69] wherein agents interact with novel scenarios during test time, and partner choices [67] in sequential social dilemmas. SVO research has focused primarily on social dilemmas with underlying incentive structures resembling the *prisoner's dilemma* [77], wherein each player has an incentive to defect, even though both would be better off if they both cooperated.

Sequential social dilemmas [51] are a class of social dilemma in which the decision-making process of the interacting agents is temporally and spatially extended. Per-

forming well in a sequential social dilemmas tends to require the consideration of long-term consequences, interdependence, and cooperation among group members. Research on sequential social dilemmas has been widely studied in the context of emergence and maintenance of cooperation [52, 75], inequity aversion [37], partner choices [21, 67] wherein agents have a choice with whom to interact, and generalization [69, 1] wherein agents interact with novel scenarios during test time.

In sequential social dilemmas, it is useful to think of players as having an *intrinsic reward* in addition to the environment-provided *extrinsic reward*. Intrinsic reward can be used to capture the social preferences of players, and are typically functions of the vector of all players' reward. In most research in sequential social dilemmas, all players either have no *intrinsic reward*, or they all have the same function (i.e. they have homogeneous social preferences) [52, 93]. However, it has been observed that having a population of agents who differ in their intrinsic reward function (i.e. they have heterogeneous social preferences) can lead to higher levels of cooperation [37]. In [68, 69, 67], the authors showed that heterogeneity can produce behavioral diversity.

Diversity in policies has been demonstrated to improve various aspects of agent performance, such as exploration [100], adaptation to environmental changes [18], positive group outcomes [68, 85], generalization to novel co-players [56], and collaboration with humans [82]. One way to quantify diversity is through state-action variation, which measures the distribution of state-action pairs that an agent explores during training. State-action diversity can be assessed by measuring differences in the state visitation frequency [100], action selection frequency in a given state [69], or differences between state-action trajectories starting from a specific state [56]. To complement these methods, diversity can also be quantified by examining the reward an agent obtains when interacting with different co-players (often called *strategic diversity*) [9, 26], which can provide a complementary measure of diversity in behavior.

However, defining a universal diversity metric from trajectories can be challenging, and so it is possible instead to use environment-specific measures of diversity.

Zero-shot generalization [36, 35, 82, 50, 69] seeks to develop general agents that are capable of successfully interacting with novel agents during test time (i.e., agents they have not seen during training). In such situations, the policies of the novel agents encountered at test time can be out-of-distribution for the agents, leading to poor coordination in purely cooperative settings [36, 56], and getting exploited in competitive settings [74]. In mixed-motive games, failure to generalize to novel agents can lead to deadweight loss by missing an opportunity to cooperate [50]. Learning a best response to partners/opponents with meaningfully diverse policies has emerged as a promising approach to zero-shot generalization [82]. The intuition behind this approach is that training with a set of diverse policies decreases the likelihood of encountering out-of-distribution policies at test time. Despite this promise these best response techniques have not yet been applied in a wide range of incentive structures.

In this work, we assess heterogeneous SVO in a range of incentive structures in sequential social dilemmas. We include temporally and spatially extended environments with an underlying structure that is like: *Prisoner's dilemma*; *Chicken*, where both players have an incentive to choose a risky behavior, but where the worst outcome is if both choose the high risk; and *Stag hunt* wherein players have a safe choice, and an incentive to coordinate on a high-reward strategy that carries a risk of costly miscoordination. Chicken and Stag hunt are equilibrium selection social dilemmas.

We show that heterogeneous SVO leads to diverse policies, as measured with task-specific diversity metrics. We also show that this diversity can be leveraged via best response results in better zero-shot generalization in equilibrium selection sequential social dilemmas. We found that best-response agents adapted to partners/opponents with diverse behaviors by learning a conditional policy during training. However, when the test scenario contained conditional policies and the sequen-

tial social dilemma was not an equilibrium-selection problem, we found that best responses to diverse behaviors collapsed to one unconditional policy, leading to poor zero-shot generalization. Our results extend to multi-player games with more than 2 players.

The paper is organized as follows. Section 2 outlines the methodology employed in the paper. In Subsection 2.1, we present the formulation of the N-agent partially observable Markov process used in the paper. Subsection 2.2 describes the Social Value Orientation (SVO) framework and its implementation. In Subsection 2.3, we discuss the various environments used in the study and their characteristics. Subsection 2.4 details the procedure for generating diverse policies in sequential social dilemmas. In Subsection 2.5, we present the process for training a best response agent with a population of agents and evaluating zero-shot generalization performance. Furthermore, we provide a description of the agent's architecture in Subsection 2.5. Section 3 presents the results of the work. In Subsection 3.1 and 3.2, we present the results obtained from generating diverse policies in environments with different incentive structures. In Subsection 3.3, we present the results of zero-shot generalization performance evaluation. Finally, in Section 4, we provide additional discussions and conclusions. The section summarizes the main contributions of the work and discuss potential societal impacts.

12.2 Method

12.2.1 N-agent POMDP

We consider a multi-agent partially observable Markov decision process defined by the tuple $\left\langle N, \mathcal{S}, \mathcal{A}, \mathcal{R}, P, \gamma \right\rangle$, where N is the number of agents, $\mathcal{S}$ is the joint state space, $\mathcal{A} = \times_{i=1}^{N} \mathcal{A}^i$ is the joint action space, P is the state transition probability distribution, $\mathcal{R}$ is the reward function and γ is the discount factor. At each time

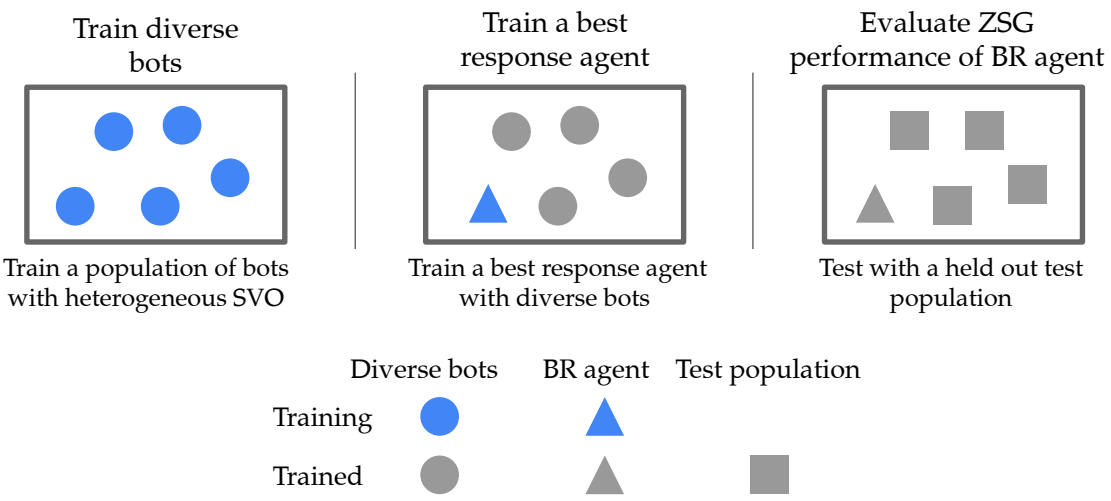


Figure 12.1: Overview of the methodology.

step t, each agent $i \in 1, \ldots, N$ observes a private (local) observation o_t^i and takes an action a_t^i from a set of actions $\mathcal{A}^i$. The joint action of all agents at time step t is denoted as $a_t = (a_t^1, \ldots, a_t^N)$. The state s_t is unobservable, and the partial observation o_t^i depends on the current state of the environment s_t and the agent's observation function. The observation function for agent i is denoted as $O^i(o_t^i|s_t)$. Each agent i receives a reward r_t^i which is a function of the joint action a_t and the state s_t of the environment. The state of the environment transitions according to a probability distribution $P(s_{t+1}|s_t, a_t)$.

The objective of each agent i is to maximize their cumulative expected discounted reward, over a given finite time horizon, defined as $J^i = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t^i\right]$, where $\gamma \in [0, 1]$ balances the importance of immediate and future rewards. The agents' policies are defined as the mapping from the agent's observation history to an action, i.e., $\pi^i(a_t^i|o_1^i, \cdots, o_t^i)$. The policies are updated using a multi-agent reinforcement learning algorithm that maximizes the agents' objective functions.

12.2.2 Social Value Orientation

Omitting the dependence on t, let r^i be the reward of agent i. Let $\bar{r}^{-i}$ be the average reward of all the agent except agent i. Then we have

$$\bar{r}^{-i} = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} r^j.$$

Let svo^i denote the SVO value of agent i. Following the definition given in [67], we define the effective reward $\hat{r}^i$ of agent i as

$$\hat{r}^i = r^i \cos(svo^i) + \bar{r}^{-i} \sin(svo^i).$$

Then agent i optimizes the objective function

$$\hat{J}^i = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \hat{r}_t^i\right].$$

12.2.3 Environments

We provide a brief description of the environments. For all experiments in this paper, we use environments from Meltingpot 2.0 [1].

Intertemporal "in the matrix" repeated games: The "in the matrix" repeated games are a family of sequential social dilemmas in Melting Pot 2.0 where two-players interact. In the beginning of each episode the environment is initialized according to a given resource layout, and a set of fixed points where players can spawn. The map consists of two types of resources which can be distinguished by their colour; red corresponds to defection and blue corresponds to cooperation (see Figure 12.3). Players can pick up resources by walking over them, and these resources go into a player inventory. Players spawn with one of each resource type in their inventory.

After spawning, each player can move around the map, collect resources, and interact with the co-player by firing an interaction beam. When players interact (by one player hitting the other using their interaction beam), each player gets a reward equal to the expected payoff calculated from the inventory counts and environment-specific payoff matrix. The agent who zaps the other agent is considered as the row player. The inventory count of each player defines a mixed strategy where the probability of playing each pure strategy is equivalent to the percentage of the corresponding resource. Let N_r^i and N_g^i denote the inventory count, number of red resources and green resources respectively, for agent $i \in 1, 2$. For each agent i their mixed strategy is given as

$$p = \Big[\frac{N_r^i}{N_r^i + N_g^i}, \frac{N_g^i}{N_r^i + N_g^i}\Big]$$

Let A be the payoff matrix for both row player and column player. Let r_{row} and r_{col} be the reward of row player and column player respectively. Let p_{row} and p_{col} be the mixed strategy probability vector of row player and column player respectively. Then the rewards can be defined as

$$r_{row} = p_{row}^T A p_{col}, \quad r_{col} = p_{col}^T A^T p_{row}$$

These reward calculations correspond to those used in game theory for matrix games and iterated social dilemmas [97].

The payoff matrices A used are given in Figure 12.2. After interacting, players receive their reward from interaction, have their inventory counts reset (to one of each), and get re-spawned after a delay. Players can have multiple interactions within an episode. Once a resource is picked up, it begins to regenerate after a delay of 10 steps, with a 20% chance of regenerating on each subsequent step.

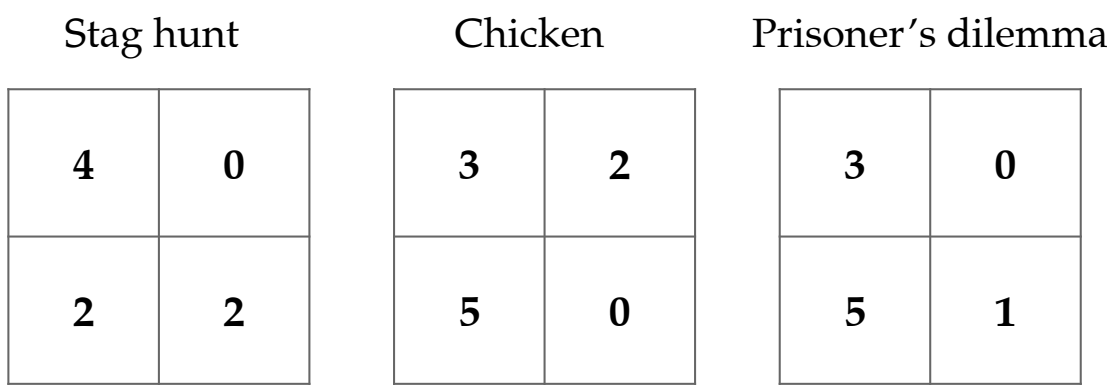


Figure 12.2: Payoff matrices

Externality mushrooms: Externality Mushrooms is sequenctial social dilemma where players immediately get affected from pro(anti)social behaviors of their co-players. This is a 5-player game where players eat mushrooms in order to receive rewards. Four types of mushrooms grow (in different amounts) on the map: red, green, blue, and orange. Eating a red (fize: full internality zero externality) mushroom gives a reward of 1 to the player who consumed the mushroom. Eating a green (hihe: half internality half externality) mushroom gives a total reward of 2/5 *to all players*. Eating a blue (zife: zero internality full externality) mushroom gives a total reward of 3/4 divided equally among all players *excluding the player who consumed it*. Eating an orange (nize: negative internality zero externality) mushroom causes red mushrooms to be destroyed, each with probability 0.25, and gives a reward of -0.1 to the player who consumed it. After eating a mushroom, the player who consumed it freezes for the mushroom's digestion time: 0 (red), 10 (green), 15 (blue), and 15 steps (orange). After spawning, a mushroom is removed from the map after its perishing time: 200 (red), 100 (green), and 75 steps (blue). Orange mushrooms never perish. Mushrooms respawn from spores depending on consumption of other mushrooms. Eating a red, green, or blue mushroom releases 3 spores for red mushrooms, each spore will spawn a mushroom with probability 0.25. Eating a green or blue mushrooms also releases 3 spores for green mushrooms which spawn with probability 0.4. Eating a

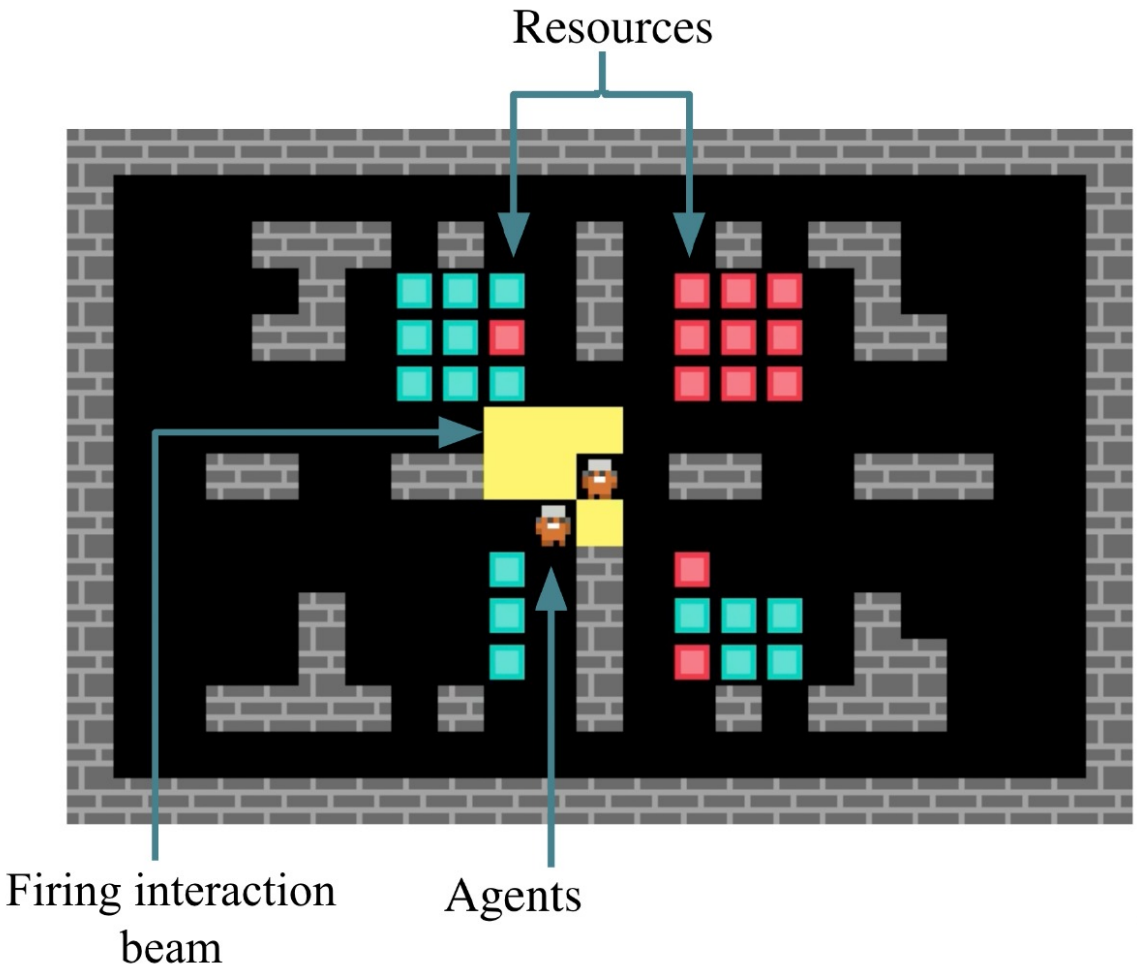


Figure 12.3: "in the matrix" repeated games. This is a 2-player game where agents can gather 2 types of resources. When agents interact (using an interaction beam) they get rewards according to their inventory counts and a game specific payoff matrix. The payoff matrix can be Stag hunt, Chicken or Prisoner's dilemma type payoff matrix

blue mushroom also releases a blue spore which spawn with probability 0.6. Eating an orange mushroom releases a spore for a new orange mushroom which spawns with probability 1.

Externality mushrooms has an incentive structure similar to Chicken, where reward is maximized selfishly by consuming red mushrooms while the others are consuming blue or green mushrooms. But if everyone else is eating red mushrooms, the selfish strategy is to eat green mushrooms, as otherwise all mushrooms would be eventually depleted.

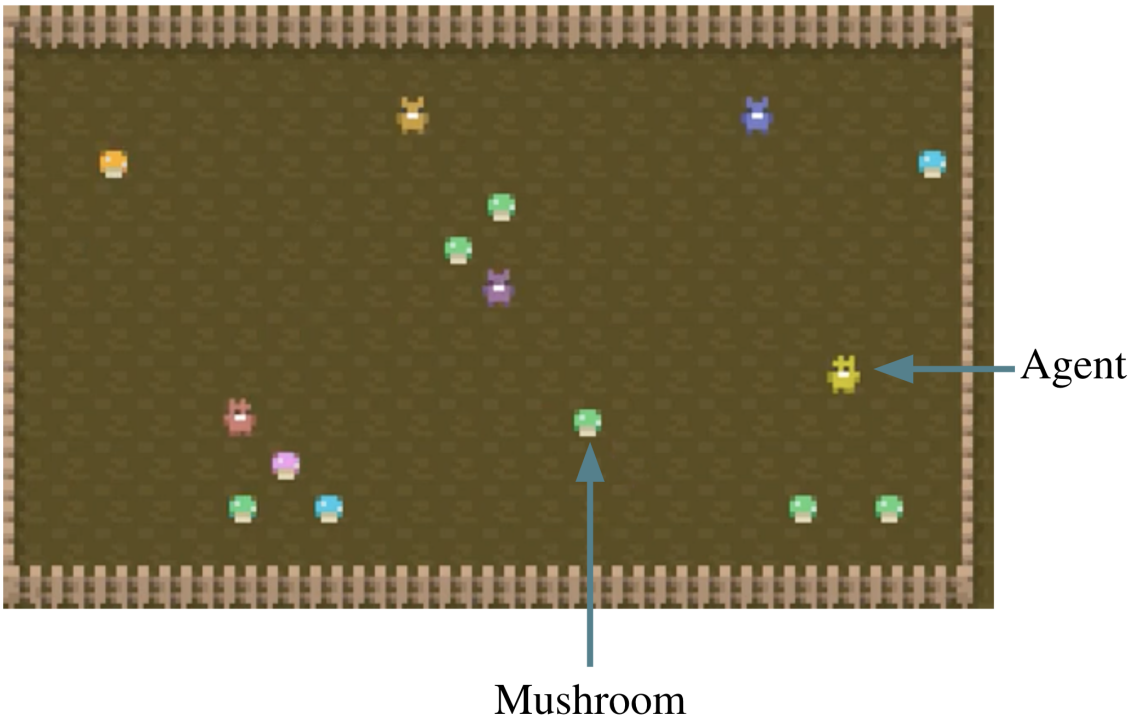


Figure 12.4: Externality Mushrooms. This is a 5-player sequential social dilemma game with immediate feedback. Agents instantaneously share rewards with others depending on the mushroom they are picking.

12.2.4 Generating diverse policies in sequential social dilemmas

In the beginning of the training process we define distinct SVO angles for each agent. Each environment has a fixed number of players. We train the agents in a distributed asynchronous manner by initializing 'arenas' to train a population of agents. Arenas run in parallel and each arena is a copy of the environment with the number of players specified for that environment. This is a multi-agent version of A3C [70] that is commonly used for multi-agent reinforcement learning [1]. Players in each arena plays the game for one episode either in self-play or in population-play (with equal probability). During population-play we sample agents without replacement. We train each agent for 10^9 learner steps.

12.2.5 Training a best-response agent and zero-shot generalization performance evaluation

We train a selfish naive learner without intrinsic reward, to best respond against the policies generated using heterogeneous SVO. In order to avoid confusion we use the term *best-response agent* for the training agent, and *SVO bots* for the pre-trained diverse agents trained with heterogeneous SVO values. In each episode the best-response agent plays with a set of SVO bots sampled without replacement. We train the best-response agent for 10^9 learner steps.

We use environments from Melting Pot 2.0 [1], which provides a protocol for evaluating generalization to novel social partners, which are packaged with the suite as a held-out set of co-players in a suite of test scenarios. We measure the performance of the best-response agent using the Melting Pot test protocol. blueWe provide an overview of the end to end methodological pipeline in Figure 12.1.

12.2.6 Agent architecture

The neural network of the agent consists of two convolutional layers, a two-layer perceptron, and an LSTM—all separated by ReLU activation functions. The convolutional layers have 16 and 32 output channels, kernel shapes of 8 and 4, and strides of 8 and 1. The perceptron layers are 64 neurons each, and the LSTM layer has 128 units. The policy and baseline for the critic are created by multilayer perceptrons (256 hidden units with ReLU activations) connected to the output of the LSTM.

Representation shaping is achieved through the use of an auxiliary loss and contrastive predictive coding [73], which is used to differentiate between nearby time points via LSTM state representations. PopArt [33] is used to adjust for the different reward scales of the different substrates. The optimization method used is RMSProp with a learning rate of 4×10^{-4}, epsilon of 10^{-5}, zero momentum, decay of 0.99, and

batch size of 256. The baseline cost for the critic is 0.5, and the entropy regularization cost for the policy is 0.003.

12.3 Experimental results

12.3.1 Experiment 1: Generating diverse policies in "in the matrix" repeated games

Experimental setup: We consider Stag hunt, Chicken and Prisoners' dilemma "in the matrix" repeated games. In each of the 3 games, there is a 10% chance that the episode will end after every 100 steps, with a minimum of 1000 steps per episode. For each game we average the results over 3 random seeds. For each seed, we train four agents with SVO values of $-15°, 0°, 60°$, and $75°$, respectively. The "in the matrix" repeated games are 2-player games. At the beginning of every episode, we randomly select between two options with equal probability: training one agent in self-play, or training two agents sampled without replacement in population-play. In addition to SVO bots we also train a set of selfish-baseline bots, i.e., no intrinsic reward, using the same procedure for comparison.

Finding 1: Heterogeneous SVO bots learn meaningfully diverse policies

We use the inventory count of the bots at the time of interaction as an environment-specific diversity measure. Since the inventory counts define the mixed strategy probability vectors, sufficiently distinct ratios of inventory counts indicate distinct mixed strategies. During evaluation agents play in either self-play or population-play with equal chance.

Figure 12.6 shows the inventory counts for the 4 bots averaged over the last 500 interactions during evaluation after the completion of training. Top and bottom rows

(a) Stag hunt (b) Chicken (c) Prisoner's dilemma

Figure 12.5: *Comparing meaningful diversity of policies of selfish-baseline bots and SVO bots.* Each subfigure shows average inventory counts during evaluation for 4 agents, trained with 50% self-play and 50% population play, evaluated in self-play and population-play in repeated "in the matrix" games. Dotted and solid bars correspond to self-play and population respectively. The bottom row corresponds to SVO bots with $svo^i \in \{-15°, 0°, 60°, 75°\}$ and the top row corresponds to selfish-baseline bots. Green and red represents cooperative and defective resource counts respectively. Error bars show the standard deviation of results over 3 random seeds.

correspond to resource counts of selfish-baseline bots and SVO bots respectively. Figures 12.6(a), 12.6(b) and 12.6(c) correspond to Stag hunt, Chicken and Prisoners' dilemma respectively. The error bars presented in the figure correspond to the average results of 3 independent runs. The results demonstrate that in each game, all 4 selfish-baseline bots have comparable inventory count ratios, suggesting that their policies lack diversity. Conversely, the 4 SVO bots exhibit varied inventory count ratios, indicating diverse behaviors. For each "in the matrix" repeated game, resource counts correspond to SVO bots with $svo = [-15°, 0°, 60°, 75°]$, where $svo^i = svo[i]$, $i \in$

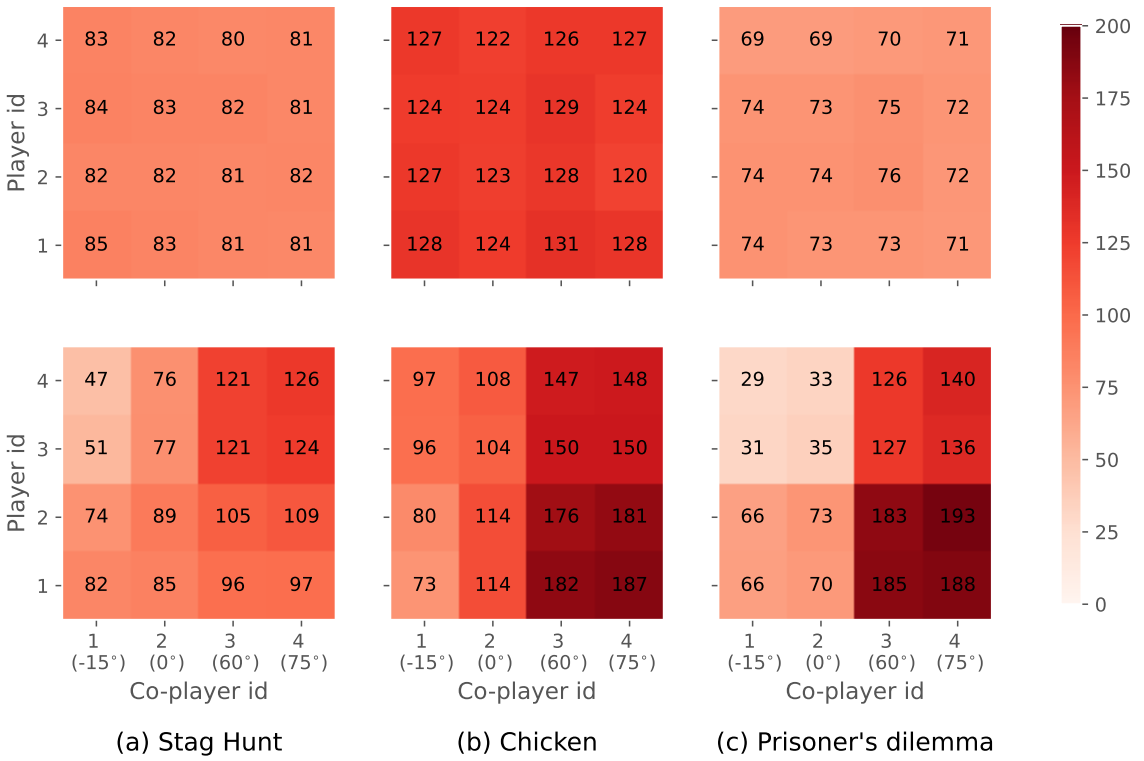


Figure 12.6: *Comparing meaningful diversity of policies of selfish-baseline bots and SVO bots.* Each subfigure shows average return during evaluation for for different pair of agents, trained with 50% self-play and 50% population play, evaluated in self-play and population-play in repeated "in the matrix" games. The return value represents the return obtained by agent when playing with the corresponding co-player. The bottom row corresponds to SVO bots with $svo^i \in \{-15°, 0°, 60°, 75°\}$ and the top row corresponds to selfish-baseline bots. Results are averaged over 3 independent runs

$\{1, 2, 3, 4\}$. We denote the cooperative resource counts and defective resource counts using green and red respectively. As the SVO angles increase from $-15°$ to $75°$, the ratio between the red and green resource counts increases, indicating a more cooperative, prosocial or altruistic behavior.

Similarly we present results for the return agents obtain when they play with different co-players. For selfish-baseline agents obtain similar returns suggesting that agents are learning similar policies. For SVO agents obtain different returns suggesting that agents learn diverse policies.

12.3.2 Experiment 2: Generating diverse policies in Externality Mushrooms

Experimental setup: Similar to "in the matrix" repeated games, in Externality Mushrooms each episode runs for at least 1000 steps. Following that the episode terminates with probability 0.2 at every 100 steps. Similar to the training process in "in the matrix" repeated game we average the results from 3 random seeds. For each seed we train 5 agents with SVO values of $-15°, 0°, 60°, 75°$, and $90°$, respectively. In addition to SVO bots we also train a set of selfish-baseline bots, using the same procedure for comparison. Similar to the "in the matrix" repeated games in Externality mushrooms game also as the SVO angles increase from $-15°$ to $90°$, the ratios between the red and green mushroom fractions, the ratios between the green and blue mushroom fractions, increases, indicating a more cooperative, prosocial or altruistic behavior.

Finding 2: The results extends to multi-player games with more than 2 players

We show that our method scales to games with more than 2 players. Figure 12.7 shows that in Externality Mushrooms, agents trained using heterogeneous SVO learn diverse policies. We use the count of mushrooms consumed of each type as the environment-specific diversity metric. The selfish-baseline bots tend to consume mushrooms at similar ratios across different types, whereas the SVO bots consume varying ratios of different mushroom types exhibiting meaningfully diverse behaviors. Agents with low (or negative) SVO consume the selfish mushroom (red), and even the spiteful mushroom (orange), whereas those with high SVO, tend to consume more of the prosocial mushrooms (green and blue).

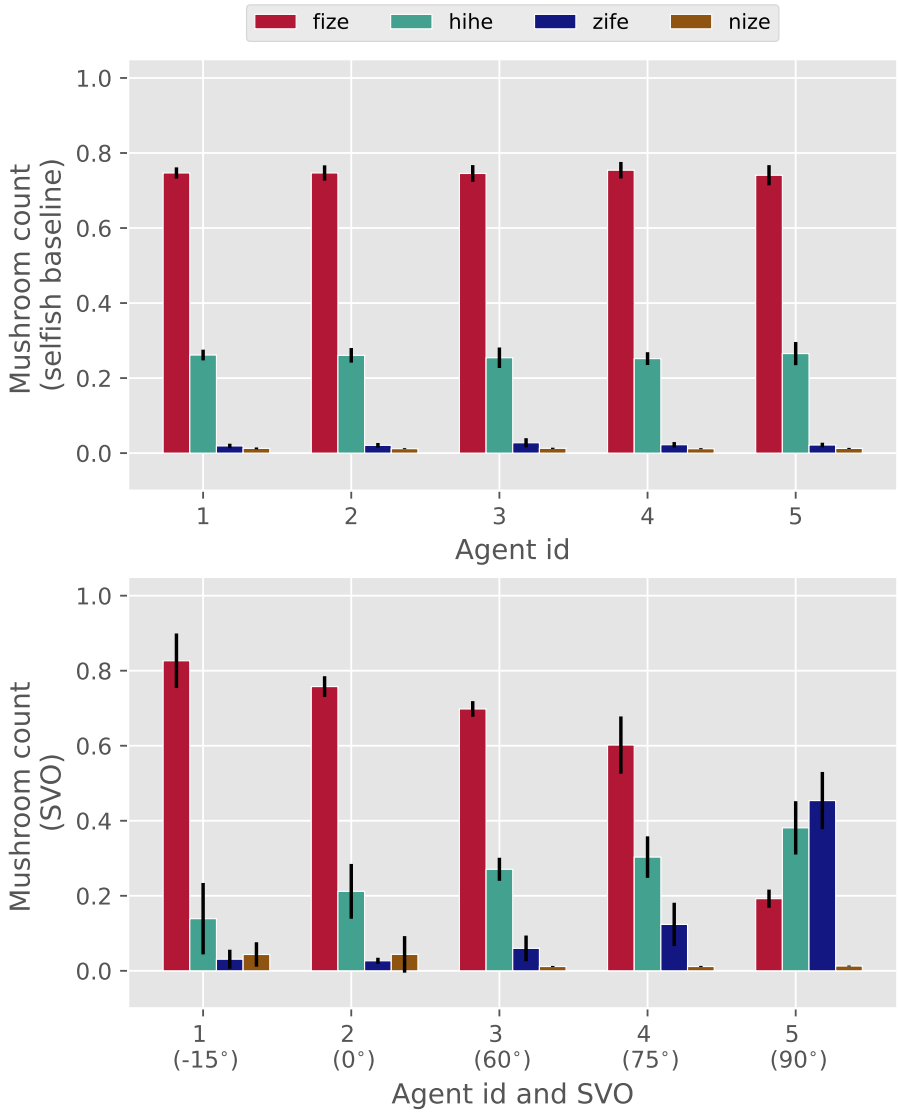


Figure 12.7: *Comparing meaningful diversity of policies of selfish-baseline bots and SVO bots.* Each plot shows average fraction of mushrooms consumed by 5 agents during evaluation, trained with 50% self-play and 50% population play, evaluated in self-play and population-play in Externality mushrooms dense game. Dotted and solid bars correspond to self-play and population respectively. The bottom row corresponds to SVO agents with $svo^i \in \{-15°, 0°, 60°, 75°, 90°\}$ and the top row corresponds to selfish-baseline agents. Error bars show the standard deviation of results over 3 random seeds.

12.3.3 Experiment 3: Zero-shot generalization evaluation

We evaluate the zero-shot generalization performance of a learned best response to the SVO bots trained using heterogeneous SVO. We use the Melting Pot test

scenarios for evaluation in Stag hunt, Chicken, Prisoners' dilemma "in the matrix " repeated games and Externality mushrooms. Test scenario details are provided below.

Test scenarios for "in the matrix" repeated:

S0: Best-response agent encounters either a cooperator or a defector with 0.5 probability

S1: Best-response agent encounters a cooperator

S2: Best-response agent encounters a defector

S3: Best-response agent encounters a player who starts by cooperating and defect for the rest the episode when best-response agent defects once

S4: Best-response agent encounters a player who starts by cooperating and defect for the rest the episode when best-response agent defects twice

S5: Best-response agent encounters a player who plays tit-for-tat

S6: Best-response agent encounters a player who a player who plays tit-for-tat and occasionally unconditionally defect. (noisy tit-for-tat)

S7: Best-response agent encounters a player who cooperate during the first few interactions and defect for the rest of the episode

S8: Best-response agent encounters a player who starts with defection and switch to tit-for-tat strategy when best-response agent defects

S9: Best-response agent encounters a player who starts with defection and switch to noisy tit-for-tat strategy when best-response agent defects

Test scenarios for Externality mushrooms:

S0: Best response agent encounters 4 cooperators

S1: Best response agent encounters 4 defectors

S2: 2 copies of the best-response agent encounter 3 cooperators

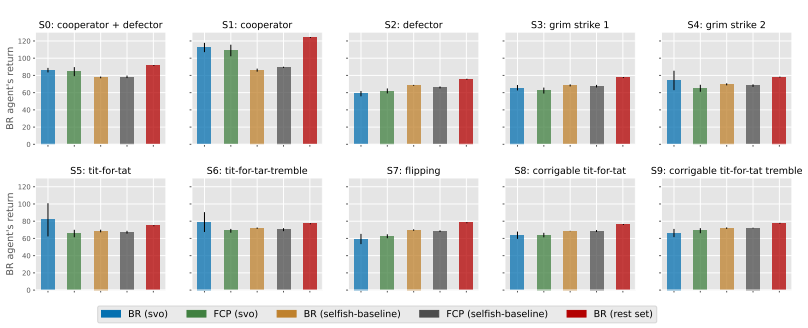
S3: 2 copies of the best-response agent encounter 3 defectors

Baselines: We compare the performance of a learned best response policy for SVO bots with a best response to selfish-baseline bots, Fictitious co-play (FCP, a type of best response that includes also earlier checkpoints of the agents to best respond to) [82] and exploiters (i.e., a best response agent trained on the test scenario directly) [1]. We train one exploiter for each test scenario. To train FCP agents we train a naive learning agent with 3 checkpoints for each bot from a bot population. Here we use the first checkpoint, mid checkpoint and last checkpoint. The mid checkpoint is the time during training where the agent first obtains half of its final reward, of the policies of the bots. We report results for FCP applied to the heterogeneous SVO bots FCP(SVO), as well as to selfish baselines FCP(selfish-baseline).

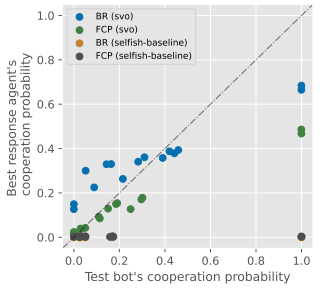
Experimental setup: We train best-response agents for the selfish-baseline bots and SVO bots. Recall that we trained each type of bots, i.e., selfish-baseline or SVO, for 3 random seeds in this setup. We train a best-response agent for bots from each seed. For each type of bots we show the average performance evaluation runs correspond to these 3 training runs.

Finding 3: Our method outperforms the baselines

Figure 12.8 and Figure 12.9 show zero-shot generalization performance results for best-response agents trained with SVO bots and selfish-baseline bots, FCP best response agents and exploiters for Stag hunt and Chicken in "in the matrix" games respectively. Figure12.8(a) illustrates that in most of the scenarios of Stag hunt best-response agents trained with SVO bots outperform or perform comparably to



(a) Zero-shot generalization - Stag hunt



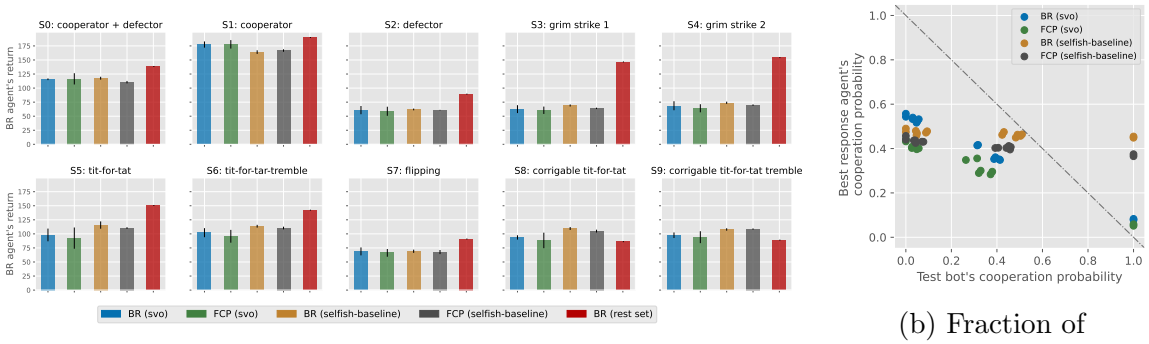
(b) Fraction of cooperation

Figure 12.8: *(a) Comparing zero-shot generalization performance of best-response agents. (b) Comparing how well best-response agents learn conditional policies.* Results for zero-shot generalization performance of best-response agents in Stag hunt. The best-response agents play in Melting Pot scenarios. We show results for best response to SVO bots and best response to selfish-baseline bots. Also we show results for FCP for SVO bots and selfish-baseline bots. Figure (a) shows average reward obtained by best-response agents during evaluation. Figure (b) shows the fraction of interactions wherein the best-response agent cooperated Vs the fraction of interactions wherein the agents in the test population cooperated.

best-response agents trained with selfish-baseline bots. Figure 12.9(a) shows that in scenarios of Chicken best-response agents trained with SVO bots mostly perform comparably to best-response agents trained with selfish-baseline bots. In the next section we show that the best-response agents trained with SVO bots in Chicken learn better conditional policies.

Finding 4: Best-response agents learn a conditional behaviour

In order to get a better understanding about the learned policies of the best-response agents we further analyze the behaviour of the best-response agents during test time. For each test bot, Figures 12.8(b) and 12.9(b) show the fraction of interactions where the best-response agent cooperated with a bot with respect to the fraction of interactions where the bot cooperated with the best-response agent. Figure 12.8(b) corresponds to Stag hunt "in the matrix" repeated and 12.9(b) corresponds to Chicken "in the matrix" repeated. In this analysis we define the

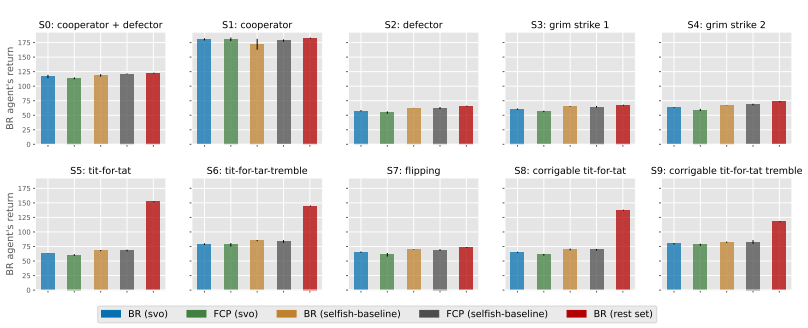


(a) Zero-shot generalization - Chicken

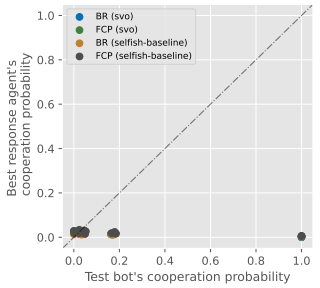
(b) Fraction of cooperation

Figure 12.9: *(a) Comparing zero-shot generalization performance of best-response agents. (b) Comparing how well best-response agents learn conditional policies.* Results for zero-shot generalization performance of best-response agents in Chicken. The best-response agents play in Melting Pot scenarios. We show results for best response to SVO bots and best response to selfish-baseline bots. Also we show results for FCP for SVO bots and selfish-baseline bots. Figure (a) shows average reward obtained by best-response agents during evaluation. Figure (b) shows the fraction of interactions wherein the best-response agent cooperated Vs the fraction of interactions wherein the agents in the test population cooperated.

best-response agent's interaction as a cooperation when they have higher number of cooperative resources than defective resources in their inventory at the time of interaction. In Stag hunt both agents cooperating, i.e., both agents playing Stag, yields a higher reward, but it is a riskier strategy. Defecting, yields a secure payoff. Both agents cooperating or both defecting are Nash equilibria, that is, there is no incentive to unilaterally deviate from that strategy. An agent who cooperates with a defector gets 0 reward. When trained in Stag hunt selfish-baseline bots learn to defect. The best response to unconditional defectors is defecting. Hence the best-response agents trained with selfish-baseline bots learn to unconditionally defect. In contrast the heterogeneous SVO bot population consists of both defectors and cooperators with different levels of cooperation and defection. Best-response agents training with SVO bots encounter both cooperators and defectors and subsequently learn a conditional policy that tends to cooperate with cooperators and defect with defectors. In Chicken the two Nash equilibria are for one agent to cooperate (swerve)



(a) Zero-shot generalization - Prisoners' dilemma



(b) Fraction of cooperation

Figure 12.10: *(a) Comparing zero-shot generalization performance of best-response agents. (b) Comparing how well best-response agents learn conditional policies.* Results for zero-shot generalization performance of best-response agents in Prisoners' Dilemma. The best-response agents play in Melting Pot scenarios. We show results for best response to SVO bots and best response to selfish-baseline bots. Also we show results for FCP for SVO bots and selfish-baseline bots. Figure (a) shows average reward obtained by best-response agents during evaluation. Figure (b) shows the fraction of interactions wherein the best-response agent cooperated Vs the fraction of interactions wherein the agents in the test population cooperated.

and the other agent to defect (straight). In this case selfish-baseline agents learn to do both defection and cooperation. Hence the best-response agents trained with selfish-baseline bots also learn to defect and cooperate. However in Figure 12.9(b) we see that this behaviour is not conditional. In contrast best-response agents training with SVO bots encounter mostly cooperative and mostly defective bots, leading to best-response agents learning a conditional behavior where they tend to cooperate with defectors and defect against cooperators.

Finding 5: Failure case with Prisoners' dilemma

Figure 12.6(c) shows that SVO bots learn diverse policies when trained with heterogeneous SVO values. This indicates that SVO bots are learning policies with different levels of cooperation. However, we see in Figure 4(a) that best-response agents trained with SVO bots perform similarly to the best-response agents trained with selfish-baseline bots. Further, from Figure 4(b) we can see that all the best-

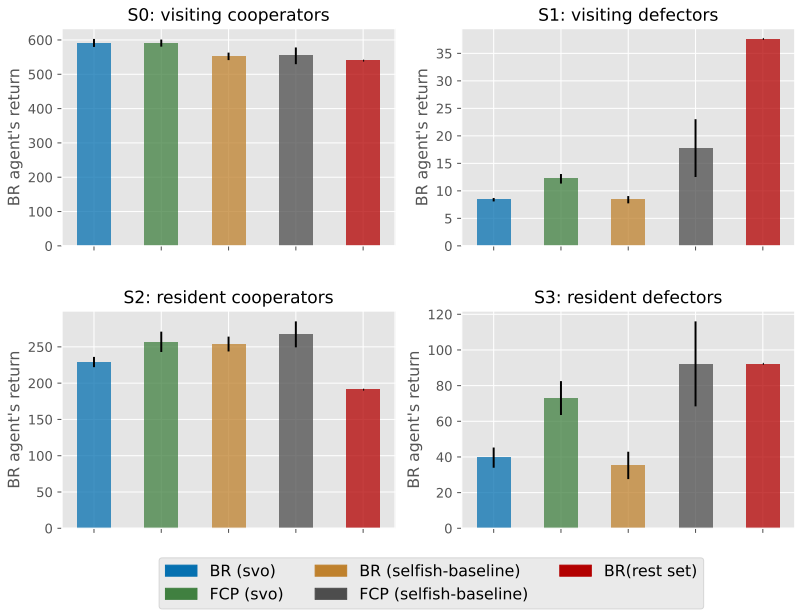


Figure 12.11: **Comparing zero-shot generalization performance of best-response agents.** Results for zero-shot generalization performance, average reward during evaluation, of best-response agents in Externality mushrooms. The best-response agents play in Melting Pot scenarios. We show results for best response to SVO bots and best response to selfish-baseline bots. Also we show results for FCP for SVO bots and selfish-baseline bots.

response agents are learning to defect regardless of the level of cooperation of their partners. This is due to the fact that the best response to an unconditional cooperator and an unconditional defector is unconditionally defecting.

Finding 6: Zero-shot generalization in Externality mushrooms

Figure 12.11 shows results for zero-shot generalization performance of best-response agents for SVO bots and selfish bots. The figure also shows results for FCP best-response agents and exploiters. We see that when the best-response agent is visiting a group of cooperators and defectors, the best response to SVO, and FCP to SVO outperform or perform comparably to their counter-parts trained with selfish-baseline bots. In scenario two, best-response agents trained with SVO bots

perform worse than baseline methods. In scenario 3 best-response agents perform comparably.

12.4 Discussion

In this paper we investigated the impact of heterogeneous social value orientation on different incentive structures in sequential social dilemmas. We tested whether the presence of heterogeneous SVO leads to diverse policies and if learning a best response to these policies improves zero-shot generalization. The study found that the presence of heterogeneous SVO does indeed lead to measurable diversity in policies, and this diversity often results in better zero-shot generalization for agents that best respond to them.

The best-response agents achieve better performance by learning a conditional policy that adapts to novel agents during test time. The study also revealed that when the sequential social dilemma is not an equilibrium-selection problem, this method still generates meaningful diversity in policies, but it fails to achieve better zero-shot generalization performance. This occurs because the best response to a diverse set of policies collapses to one unconditional policy that performs poorly when encountering conditional policies during test time.

Additionally, the study demonstrated that the results extend to multi-player games with more than two players. Our findings have implications for understanding how heterogeneous SVO impacts incentive structures and policy diversity, and how agents can learn to adapt to diverse policies during test time to achieve better zero-shot generalization performance. Our findings provide new insights into the behavior of agents in sequential social dilemmas and highlights the importance of considering the role of heterogeneity in SVO in the design of incentive structures.

In the method proposed in this paper the agent learning a best response to a population with diverse policies, is optimizing for its own reward. This may not align with the well-being of the other agents in the population, leading to negative consequences for them. In order to prevent these negative externalities, it is essential to ensure that the policies of best response agents align with human values.

One method of achieving this is by incorporating ethical considerations, such as fairness and safety, into the agent's reward function or constraints. This can help to ensure that the policies of agents align with human values and that it does not harm others in its pursuit of its own reward. By implementing these mechanisms, it is possible to mitigate the negative impacts of using a best-response agent.

We observed that SVO agents were able to learn cooperative policies in all of the environments we tested. This hints at the potential value of using SVO to capture at least some of the aspects necessary to align agents with human values.

Bibliography

[1] John P Agapiou, Alexander Sasha Vezhnevets, Edgar A Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, et al. Melting pot 2.0. *arXiv preprint arXiv:2211.13746*, 2022.

[2] Rajeev Agrawal. Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

[3] Jason M Altschuler, Victor-Emmanuel Brunel, and Alan Malek. Best arm identification for contaminated bandits. *J. Mach. Learn. Res.*, 20(91):1–39, 2019.

[4] Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, 1987.

[5] Dana Angluin and Leslie G Valiant. Fast probabilistic algorithms for hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 18(2):155–193, 1979.

[6] Lucy M Aplin, Damien R Farine, Julie Morand-Ferron, Andrew Cockburn, Alex Thornton, and Ben C Sheldon. Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature*, 518(7540):538–541, 2015.

[7] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[8] Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

[9] David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech Czarnecki, Julien Perolat, Max Jaderberg, and Thore Graepel. Open-ended learning in symmetric zero-sum games. In *International Conference on Machine Learning*, pages 434–443. PMLR, 2019.

[10] Yogev Bar-On and Yishay Mansour. Individual regret in cooperative nonstochastic multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3116–3126, 2019.

[11] Sébastien Bubeck. *Bandits games and clustering foundations*. PhD thesis, Université des Sciences et Technologie de Lille-Lille I, 2010.

[12] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[13] Nicol'o Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622. PMLR, 2016.

[14] Mithun Chakraborty, Kai Yee Phoebe Chua, Sanmay Das, and Brendan Juba. Coordinated versus decentralized exploration in multi-agent multi-armed bandits. In *IJCAI*, pages 164–170, 2017.

[15] Ronshee Chawla, Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. The gossiping insert-eliminate algorithm for multi-agent bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3471–3481. PMLR, 2020.

[16] Pierluigi Crescenzi, Viggo Kann, and M Halldórsson. A compendium of np optimization problems, 1995.

[17] Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121, 1974.

[18] Kenneth Derek and Phillip Isola. Adaptable agent populations via a generative model of policies. *Advances in Neural Information Processing Systems*, 34:3902–3913, 2021.

[19] Abhimanyu Dubey et al. Cooperative multi-agent bandits with heavy tails. In *International conference on machine learning*, pages 2730–2739. PMLR, 2020.

[20] Abhimanyu Dubey and Alex Pentland. Private and byzantine-proof cooperative decision-making. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 357–365, 2020.

[21] Edgar A Duéñez-Guzmán, Kevin R McKee, Yiran Mao, Ben Coppin, Silvia Chiappa, Alexander Sasha Vezhnevets, Michiel A Bakker, Yoram Bachrach, Suzanne Sadedin, William Isaac, et al. Statistical discrimination in learning agents. *arXiv preprint arXiv:2110.11404*, 2021.

[22] Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D. Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *MLHC*, 2018.

[23] Raphaël Féraud, Réda Alami, and Romain Laroche. Decentralized exploration in multi-armed bandits. *arXiv preprint arXiv:1811.07763*, 2018.

[24] James Fisher. The opening of milkbottles by birds. *Brit. Birds*, 42:347–357, 1949.

[25] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

[26] Marta Garnelo, Wojciech Marian Czarnecki, Siqi Liu, Dhruva Tirumala, Junhyuk Oh, Gauthier Gidel, Hado van Hasselt, and David Balduzzi. Pick your battles: Interaction graphs as population-level objectives for strategic diversity. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1501–1503, 2021.

[27] Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Misspecified linear bandits. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[28] Donald W Griesinger and James W Livingston Jr. Toward a model of interpersonal motivation in experimental games. *Behavioral science*, 18(3):173–188, 1973.

[29] Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR, 2019.

[30] Samarth Gupta, Shreyas Chaudhari, Gauri Joshi, and Osman Yağan. Multi-armed bandits with correlated arms. *IEEE Transactions on Information Theory*, 2021.

[31] Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259, 1944.

[32] Joseph Henrich. *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. princeton University press, 2016.

[33] Matteo Hessel, Hubert Soyer, Lasse Espeholt, Wojciech Czarnecki, Simon Schmitt, and Hado van Hasselt. Multi-task deep reinforcement learning with popart. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3796–3803, 2019.

[34] Saghar Hosseini, Airlie Chapman, and Mehran Mesbahi. Online distributed convex optimization on dynamic networks. *IEEE Transactions on Automatic Control*, 61(11):3545–3550, 2016.

[35] Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, and Jakob Foerster. Off-belief learning. In *International Conference on Machine Learning*, pages 4369–4379. PMLR, 2021.

[36] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pages 4399–4410. PMLR, 2020.

[37] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems*, 31, 2018.

[38] Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461. PMLR, 2013.

[39] Dileep Kalathil, Naumaan Nayyar, and Rahul Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

[40] Richard M Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103. Springer, 1972.

[41] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600. PMLR, 2012.

[42] Ravi Kumar Kolla, Krishna Jagannathan, and Aditya Gopalan. Collaborative learning of stochastic bandits over a social network. *IEEE/ACM Transactions on Networking*, 26(4):1782–1795, 2018.

[43] Lifeng Lai, Hai Jiang, and H Vincent Poor. Medium access in cognitive radio networks: A competitive multi-armed bandit framework. In *42nd Asilomar Conference on Signals, Systems and Computers*, pages 98–102. IEEE, 2008.

[44] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[45] Anusha Lalitha and Andrea Goldsmith. Bayesian algorithms for decentralized stochastic bandits. *IEEE Journal on Selected Areas in Information Theory*, 2(2):564–583, 2021.

[46] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 167–172. IEEE, 2016.

[47] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. On distributed cooperative decision-making in multiarmed bandits. In *European Control Conference (ECC)*, pages 243–248. IEEE, 2016.

[48] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Social imitation in cooperative multi-armed bandits: partition-based algorithms with strictly local information. In *Conference on Decision and Control*, pages 5239–5244, 2018.

[49] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision making in multi-agent multi-armed bandits. *Automatica*, 125:109445, 2021.

[50] Joel Z Leibo, Edgar A Dueñez-Guzman, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *International Conference on Machine Learning*, pages 6187–6199. PMLR, 2021.

[51] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473, 2017.

[52] Adam Lerer and Alexander Peysakhovich. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*, 2017.

[53] Jurij Leskovec. *Dynamics of large networks*. PhD thesis, Carnegie Mellon University, School of Computer Science, Machine Learning, 2008.

[54] Wim BG Liebrand and Charles G McClintock. The ring measure of social values: A computerized procedure for assessing individual differences in information processing and social value orientation. *European journal of personality*, 2(3):217–230, 1988.

[55] Nathan Linial. Locality in distributed graph algorithms. *SIAM Journal on computing*, 21(1):193–201, 1992.

[56] Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot coordination. In *International Conference on Machine Learning*, pages 7204–7213. PMLR, 2021.

[57] Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

[58] Udari Madhushani, Abhimanyu Dubey, Naomi Leonard, and Alex Pentland. One more step towards reality: Cooperative bandits with imperfect communication. *Advances in Neural Information Processing Systems*, 34:7813–7824, 2021.

[59] Udari Madhushani and Naomi Leonard. It doesn't get better and here's why: A fundamental drawback in natural extensions of ucb to multi-agent bandits. In *"I Can't Believe It's Not Better!"NeurIPS 2020 workshop*, 2020.

[60] Udari Madhushani and Naomi Leonard. When to call your neighbor? strategic communication in cooperative stochastic bandits. *arXiv preprint arXiv:2110.04396*, 2021.

[61] Udari Madhushani and Naomi Ehrich Leonard. Heterogeneous stochastic interactions for multiple agents in a multi-armed bandit problem. In *European Control Conference*, pages 3502–3507, 2019.

[62] Udari Madhushani and Naomi Ehrich Leonard. Distributed learning: Sequential decision making in resource-constrained environments. In *"Practicle Machine Learning for Developing Countries"ICLR 2020 workshop*, 2020.

[63] Udari Madhushani and Naomi Ehrich Leonard. A dynamic observation strategy for multi-agent multi-armed bandit problem. In *European Control Conference*, pages 1677–1682, 2020.

[64] Udari Madhushani and Naomi Ehrich Leonard. Distributed bandits: Probabilistic communication on d-regular graphs. In *European Control Conference*, 2021.

[65] Udari Madhushani and Naomi Ehrich Leonard. Heterogeneous explore-exploit strategies on multi-star networks. *IEEE Control Systems Letters*, 5(5):1603–1608, 2021.

[66] David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 4531–4542, 2019.

[67] Kevin R McKee, Xuechunzi Bai, and Susan T Fiske. Warmth and competence in human-agent cooperation. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 898–907, 2022.

[68] Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duèñez-Guzmán, Edward Hughes, and Joel Z Leibo. Social diversity and social preferences in mixed-motive reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 869–877, 2020.

[69] Kevin R McKee, Joel Z Leibo, Charlie Beattie, and Richard Everett. Quantifying the effects of environment and population diversity in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 36(1):1–16, 2022.

[70] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

[71] Ryan O Murphy, Kurt A Ackermann, and Michel JJ Handgraaf. Measuring social value orientation. *Judgment and Decision making*, 6(8):771–781, 2011.

[72] Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. *Advances in Neural Information Processing Systems*, 23, 2010.

[73] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[74] Nicolas Perez-Nieves, Yaodong Yang, Oliver Slumbers, David H Mguni, Ying Wen, and Jun Wang. Modelling behavioural diversity for learning in open-ended games. In *International Conference on Machine Learning*, pages 8514–8524. PMLR, 2021.

[75] Alexander Peysakhovich and Adam Lerer. Consequentialist conditional cooperation in social dilemmas with imperfect information (short workshop version). In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[76] Wenquan Qin, Shucong Lin, Xuan Chen, Jian Chen, Lei Wang, Hongpeng Xiong, Qinxi Xie, Zhaohui Sun, Xiujun Wen, and Cai Wang. Food transport of red imported fire ants (hymenoptera: Formicidae) on vertical surfaces. *Scientific Reports*, 9(1):1–12, 2019.

[77] Anatol Rapoport. Prisoner's dilemma—recollections and observations. In *Game Theory as a Theory of a Conflict Resolution*, pages 17–34. Springer, 1974.

[78] Paul B Reverdy, Vaibhav Srivastava, and Naomi Ehrich Leonard. Modeling human decision making in generalized gaussian multiarmed bandits. *Proceedings of the IEEE*, 102(4):544–571, 2014.

[79] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[80] Wilko Schwarting, Alyssa Pierson, Javier Alonso-Mora, Sertac Karaman, and Daniela Rus. Social behavior for autonomous vehicles. *Proceedings of the National Academy of Sciences*, 116(50):24972–24978, 2019.

[81] Vaibhav Srivastava, Paul Reverdy, and Naomi Ehrich Leonard. Surveillance in an abruptly changing world via multi-armed bandits. In *Conference on Decision and Control*, pages 692–697. IEEE, 2014.

[82] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, 34:14502–14515, 2021.

[83] Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

[84] Balázs Szörényi, Róbert Busa-Fekete, István Hegedűs, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *Journal of Machine Learning Research Workshop and Conference Proceedings*, volume 2, pages 1056–1064. International Machine Learning Societ, 2013.

[85] Zhenggang Tang, Chao Yu, Boyuan Chen, Huazhe Xu, Xiaolong Wang, Fei Fang, Simon Shaolei Du, Yu Wang, and Yi Wu. Discovering diverse multi-agent strategic behavior via reward randomization. In *International Conference on Learning Representations*, 2021.

[86] Chao Tao, Qin Zhang, and Yuan Zhou. Collaborative learning with limited interaction: Tight bounds for distributed exploration in multi-armed bandits. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 126–146. IEEE, 2019.

[87] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[88] Michael Tomasello. *Why we cooperate*. MIT press, 2009.

[89] Aristide CY Tossou and Christos Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[90] Paul Turán. On an external problem in graph theory. *Mat. Fiz. Lapok*, 48:436–452, 1941.

[91] Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

[92] Timothy Verstraeten, Eugenio Bargiacchi, Pieter JK Libin, Jan Helsen, Diederik M Roijers, and Ann Nowé. Multi-agent thompson sampling for bandit applications with sparse neighbourhood structures. *Scientific reports*, 10(1):1–13, 2020.

[93] Jane X Wang, Edward Hughes, Chrisantha Fernando, Wojciech M Czarnecki, Edgar A Duéñez-Guzmán, and Joel Z Leibo. Evolving intrinsic motivations for altruistic behavior. *arXiv preprint arXiv:1811.05931*, 2018.

[94] Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4120–4129, 2020.

[95] Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. In *International Conference on Learning Representations*, 2020.

[96] Romain Warlop, Alessandro Lazaric, and Jérémie Mary. Fighting boredom in recommender systems with linear reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1757–1768, 2018.

[97] Jörgen W Weibull. *Evolutionary game theory*. MIT press, 1997.

[98] Marcelo J Weinberger and Erik Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, 2002.

[99] Tao Yang, Xinlei Yi, Junfeng Wu, Ye Yuan, Di Wu, Ziyang Meng, Yiguang Hong, Hong Wang, Zongli Lin, and Karl H Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019.

[100] Tom Zahavy, Yannick Schroecker, Feryal Behbahani, Kate Baumli, Sebastian Flennerhag, Shaobo Hou, and Satinder Singh. Discovering policies with domino: Diversity optimization maintaining near optimality. *arXiv preprint arXiv:2205.13521*, 2022.