Comparative Studies on Sample Complexity Bounds in Multi-Agent Reinforcement Learning

Jiaqi Yang

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-47
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-47.html

May 10, 2022

Acknowledgement

I would like to thank my advisor Jiantao Jiao for his generous and unconditional support during my first year as a graduate student, and Song Mei for serving as the second reader for my report. I would like to thank my collaborators Cong Ma and Banghua Zhu for their help and understanding. Finally, I would like to thank my friends Yunkai Zhang and Yuhang Wu for their help.

Comparative Studies on Sample Complexity Bounds in Multi-Agent Reinforcement Learning

by Jiaqi Yang

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Assistant Professor Jiantao Jiao
Research Advisor

05/03/2022
(Date)

* * * * * * *

Assistant Professor Song Mei
Second Reader

05/05/2022
(Date)

Abstract

Comparative Studies on Sample Complexity Bounds in
Multi-Agent Reinforcement Learning

by

Jiaqi Yang

Master of Science in Computer Science

University of California, Berkeley

In this report, we survey on the existing sample complexity bounds from multi-agent reinforcement learning (MARL) literature and those from game theory literature. Along the way, we give unified notations for game theory and MARL, and summarize different definitions of equilibria in game theory and MARL.

By comparative studies on the existing bounds, we identify several interesting open gaps in MARL, and we take preliminary steps towards answering these open questions. This report can serve as a starting point for future studies in MARL theory.

Contents

Contents **i**

List of Tables **ii**

1 Introduction **1**
1.1 Overview . . . 1
1.2 Organization . . . 2

2 Methodology **3**
2.1 Notations and Definitions . . . 3
2.2 Game Theory . . . 3
2.3 Markov Games . . . 5
2.4 Sample Complexity . . . 8

3 Survey on Existing Results **9**
3.1 General Games . . . 9
3.2 Two-Player Games . . . 12

4 Results **14**
4.1 Pseudo-Polynomial Lower Bound for Finding Approximate Nash Equilibrium in Potential Game . . . 14
4.2 Query Complexity Lower Bound in Two-Player Zero-Sum Games . . . 16

5 Discussion **18**
5.1 Conclusion . . . 18
5.2 Open Problems . . . 18

Bibliography **19**

List of Tables

3.1 Summary on Sample Complexity Bounds for Learning ϵ-Nash Equilibrium in Markov Games and Markov Potential Games. 10
3.2 Summary on Sample Complexity Bounds for Learning ϵ-CCE in Markov Games. 11
3.3 Summary on Sample Complexity Bounds for Learning ϵ-CE in Markov Games. . 11
3.4 Summary on Sample Complexity Bounds for Learning ϵ-Nash Equilibrium in Two-Player Zero-Sum Games. 12

Acknowledgments

I would like to thank my advisor Jiantao Jiao for his generous and unconditional support during my first year as a graduate student, and Song Mei for serving as the second reader for my report. I would like to thank my collaborators Cong Ma and Banghua Zhu for their help and understanding. Finally, I would like to thank my friends Yunkai Zhang and Yuhang Wu for their help.

Chapter 1

Introduction

1.1 Overview

Recently, multi-agent reinforcement learning (MARL) has become a prominent paradigm for solving complex multi-agent systems. MARL has achieved impressive success in many practical applications: it could learn human-level artificial intelligence (AI) agents in video games such as StarCraft II [Vinyals et al., 2019], Dota 2 [Berner et al., 2019], Football [Kurach et al., 2020]. Moreover, MARL has demonstrated the ability of learning complex behaviors that signals intelligence [Bansal et al., 2018; Baker et al., 2020; Open Ended Learning Team et al., 2021], showing the hope of learning artificial general intelligence.

In contrast to the empirical success of MARL, our theoretical understanding towards MARL is rather limited. The theory of MARL is usually studied under the framework of Markov game [Shapley, 1953; Littman, 1994], which is an extension of Markov decision process (MDP) [Sutton and Barto, 2018; Agarwal et al., 2019], the standard formulation in RL theory, and (normal-form) game [Nisan et al., 2007], the standard formulation in game theory. A central problem in MARL theory is the sample complexity, which is the amount of samples needed by an algorithm to find an approximate equilibrium. (Formally definitions are in Chapter 2.) For this problem, we observe a huge gap between our understanding towards (single-agent) reinforcement learning (RL) theory and MARL theory. For single-agent RL, the sample complexity has been well-understood: matching lower and upper bounds have been proved in the tabular case [Jaksch et al., 2010; Li et al., 2021], and preliminary results for various function approximation cases have also been shown [Zhang et al., 2021; Dong et al., 2021; Huang et al., 2021]. However, for MARL, even the simplest tabular case has yet to be fully understood, and the lower bounds have yet to match the upper bounds [Jin et al., 2022; Liu et al., 2021; Song et al., 2022; Ding et al., 2022].

Moreover, we observe a bigger gap between our understanding towards MARL theory and game theory. There are quite a few literature that studies the query complexity in learning game equilibria, showing lower and upper bounds [Fearnley et al., 2015; Fearnley and Savani, 2016; Babichenko and Rubinstein, 2020; Babichenko, 2020; Babichenko and Rubin-

stein, 2021]. Because that game is a special case of Markov game and that sample complexity is a simplified version of query complexity,[1] ideally we should be able to recover the query complexity bounds in game theory from the sample complexity bounds in MARL theory. This is unfortunately not the case. As we will see later in this report, there are gaps between MARL and game that are not well-understood. Furthermore, certain equilibria concepts are closely related to optimization problems, such as no-regret online learning [Vishnoi, 2021; Daskalakis et al., 2021; Anagnostides et al., 2022] and min-max optimization [Ouyang and Xu, 2019]. However, these connections are not well-understood, and there are gaps between MARL theory and optimization theory.

In this report, we take preliminary steps to narrow the gaps between current MARL theory and both game and optimization theory. We perform comparative studies on the results of MARL, game theory, and optimization theory, to reveal several gaps therein. Then we resolve some open questions in MARL with existing bounds in game theory and optimization theory.

1.2 Organization

The remaining parts of this report is organized as follows. In Chapter 2, we systematically introduce notions in game theory and MARL theory. In Chapter 3, we survey on existing results, and come up with some open questions. In Chapter 4, we give preliminary results on the questions we give. In Chapter 5, we discuss about our questions and our results.

[1] Sample complexity counts the number of samples, and query complexity counts the number of queries. Samples are usually assumed to be noisy. When samples are noiseless, the sample complexity should be the same as the query complexity.

Chapter 2

Methodology

2.1 Notations and Definitions

For any integer $n \in \mathbb{N}$, we denote $[\![n]\!] = \{1, \cdots, n\}$. For any set $\mathcal{X}$, we use $\Delta_{\mathcal{X}} = \{x \in \mathbb{R}^{\mathcal{X}} : x \geq 0 \text{ and } \|x\|_1 = 1\}$ to denote the set of all probability distributions over $\mathcal{X}$. We use $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to omit logarithmic factors in complexity bounds.

In this report, query complexity refers to sample complexity for noiseless samples.

2.2 Game Theory

In this section, we formally define the most common concepts in game theory, including game and equilibria. The definitions are from various papers, mainly [Nisan et al., 2007; Lattimore and Szepesvári, 2020].

Definition 1 (Game)**.** An n-player game is defined by a tuple $(n, \{\mathcal{A}_i\}_{i=1}^n, \{R_i : \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \to \mathbb{R}\}_{i=1}^n)$. Here, n is the number of players (or agents).[1] For each agent $i \in [\![n]\!]$, the set $\mathcal{A}_i$ is the set of all actions that agent i can choose, and R_i is its payoff function that it wants to maximize by cleverly choosing its action. We define $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$, $a = (a_1, \cdots, a_n) \in \mathcal{A}$. Usually, we assume $R_i \in [0, 1]$.

For each agent $i \in [\![n]\!]$, its strategy $\pi_i \in \Delta_{\mathcal{A}_i}$ is a distribution over its action set. When π_i is a singleton distribution (i.e., $|\operatorname{supp} \pi_i| = 1$), we say π_i is a pure strategy. In general, we say it is a mixed strategy.

Collectively, we call $\pi = (\pi_1, \cdots, \pi_n)$ as a strategy profile. For convenience, we define $\pi_{-i} = (\pi_1, \cdots, \pi_{i-1}, \pi_{i+1}, \cdots, \pi_n)$ to be the strategy profile excluding agent i. The payoff of a strategy profile π is

$$R_i(\pi) = R_i(\pi_1, \cdots, \pi_n) = \mathop{\mathbb{E}}_{a_i \sim \pi_i} R_i(a) = \mathop{\mathbb{E}}_{a_i \sim \pi_i} R_i(a_1, \cdots, a_n). \tag{2.1}$$

[1] Throughout this report, we use "player" and "agent" interchangeably.

Definition 2 (Nash equilibrium)**.** Let π be a strategy profile in an n-player game. We say π is a Nash equilibrium (NE), if each agent chooses its best response against other agents. Formally, the best response of agent i against π_{-i} is a strategy

$$\pi_i^\dagger \in \arg\max_{\pi_i} R_i(\pi_i, \pi_{-i}). \tag{2.2}$$

We say π is an ϵ-approximate NE (or ϵ-NE in short) if

$$\max_{1\le i\le n} R_i(\pi_i^\dagger, \pi_{-i}) - R_i(\pi) \le \epsilon. \tag{2.3}$$

In particular, when $\epsilon = 0$, we say π is an NE.

Definition 3 (Coarse correlated equilibrium)**.** Let $\sigma \in \Delta_{\mathcal{A}}$ be a distribution over joint actions, which can be seen as a coordinator that suggests actions to each agent according to some distribution. We say σ is a coarse correlated equilibrium (CCE), if for every agent, detouring from the coordination is always no better than following it. Formally, let $R_i(\sigma) = \mathbb{E}_{a\sim\sigma} R_i(a)$. We say σ is an ϵ-approximate CCE (or ϵ-CCE), if

$$\max_{1\le i\le n} \max_{a_i'\in\mathcal{A}_i} \mathop{\mathbb{E}}_{a\sim\sigma} R_i(a_i', a_{-i}) - R_i(\sigma) \le \epsilon. \tag{2.4}$$

We say σ is a CCE when $\epsilon = 0$.

Definition 4 (Correlated equilibrium)**.** We say σ is a correlated equilibrium (CE), if for every agent, exploiting the coordination is always no better than following it, where by exploiting we mean the agent could choose better action based on the action suggested by σ. Formally, we say σ is an ϵ-approximate CE (or ϵ-CE), if

$$\max_{1\le i\le n} \max_{\phi:\mathcal{A}_i\to\mathcal{A}_i} \mathop{\mathbb{E}}_{a\sim\sigma} R_i(\phi(a_i), a_{-i}) - R_i(\sigma) \le \epsilon, \tag{2.5}$$

and we say σ is a CE when $\epsilon = 0$.

Definition 5 (Potential game)**.** An n-player game $(n, \{\mathcal{A}_i\}_{i=1}^n, \{R_i\}_{i=1}^n)$ is a potential game, if there exists a potential function $\Phi : \mathcal{A} \to \mathbb{R}$ that simultaneously captures the incentive of changing actions for all agents, such that for any agent i and any actions $a_i, a_i' \in \mathcal{A}_i, a_{-i} \in \mathcal{A}_{-i}$,

$$R_i(a_i, a_{-i}) - R_i(a_i', a_{-i}) = \Phi(a_i, a_{-i}) - \Phi(a_i', a_{-i}). \tag{2.6}$$

A cooperative game is an n-player game with $R_1 = \cdots = R_n$, which implicitly requires $\mathcal{A}_1 = \cdots = \mathcal{A}_n$. Note that a cooperative game is a potential game with $\Phi = R_i$. We assume $|\Phi| \le \Phi_{\max}$.

Definition 6 (Zero-sum game)**.** A zero-sum game is a 2-player game with $R_1 = -R_2$. In this case, we write $R = R_2$ and denote the game by $(-R, R)$. Usually, we assume $|R| \le 1$, which implies $-1 \le R_1, R_2 \le 1$.

2.3 Markov Games

In this section, we formally define the most common concepts in MARL, including Markov games and equilibria. The definitions are merged from various papers, mainly [Nisan et al., 2007; Lattimore and Szepesvári, 2020; Shapley, 1953; Littman, 1994; Jin et al., 2022; Song et al., 2022; Daskalakis et al., 2022]. We unify the notations from them, which allows us to compare their results much more easily.

Definition 7 (Markov game, finite-horizon)**.** An n-agent finite-horizon Markov game is defined by a tuple $(n, H, \mathcal{S}, \{\mathcal{A}_i\}_{i\in[\![n]\!]}, \{r_{h,i} : \mathcal{S}\times\mathcal{A}\to[0,1]\}_{\substack{h\in[\![H]\!]\\ i\in[\![n]\!]}}, \{P_h : \mathcal{S}\times\mathcal{A}\to\Delta_{\mathcal{S}}\}_{h\in[\![H]\!]}, s_1)$, where $\mathcal{S}$ is the state space, $\mathcal{A}_i$ is agent i's action space, $r_{h,i} : \mathcal{S}\times\mathcal{A}\to[0,1]$ is agent i's (deterministic) reward function at step h, $P_h(\cdot|s,a)\in\Delta_{\mathcal{S}}$ is the transition dynamics, and $s_1\in\mathcal{S}$ is the start state.

There are two types of policies being studied in finite-horizon Markov game: Markovian and non-Markovian.

- A (Markov) policy is a collection of maps from states to distributions over actions: $\pi = \{\pi_h : \mathcal{S}\to\Delta_{\mathcal{A}}\}_{h\in[\![H]\!]}$, where $\pi_h(a|s)$ is the probability of the agents choosing joint action a at step h when they are at state s. It rolls out a trajectory by $a_h\sim\pi_h(\cdot|s_h), s_{h+1}\sim P_h(\cdot|s_h,a_h)$.
- A non-Markov policy is a collection $\pi = \{\pi_h : (\mathcal{S}\times\mathcal{A}\times\mathbb{R})^{h-1}\times\mathcal{S}\to\Delta_{\mathcal{A}}\}$, which means the policy π_h at step h could depend on the history (states, actions, rewards) from step 1 through step $(h-1)$, plus the current state s_h. It rolls out a trajectory by $a_h\sim\pi_h(\cdot|\{(s_{h'},a_{h'},r_{h'})\}_{h'\in[\![h-1]\!]}, s_h), s_{h+1}\sim P_h(\cdot|s_h,a_h)$.

A trajectory is denoted by $\tau = (s_1, a_1, s_2, \cdots, s_H, a_H)$. We say a Markov policy π is a product policy, if $\pi_h(\cdot|s)$ is a product measure, $\pi_h(\cdot|s)\in\Delta_{\mathcal{A}_1}\times\cdots\times\Delta_{\mathcal{A}_n}$ for every $s\in\mathcal{S}, h\in[\![H]\!]$.

For any policy π, we define its value function and quality value function (Q-function) by

$$V_{h,i}^{\pi}(s) = \mathop{\mathbb{E}}_{\tau\sim\pi}\left[\sum_{h'=h}^{H} r_{h,i}(s_{h'},a_{h'}) \mid s_h = s\right], \tag{2.7}$$

$$Q_{h,i}^{\pi}(s,a) = \mathop{\mathbb{E}}_{\tau\sim\pi}\left[\sum_{h'=h}^{H} r_{h,i}(s_{h'},a_{h'}) \mid s_h = s, a_h = a\right], \tag{2.8}$$

where $\mathbb{E}_\tau[\cdot \mid s_h = s]$ means the trajectory τ is rolled out by starting at $s_h = s$, and similarly $\mathbb{E}_\tau[\cdot \mid s_h = s, a_h = a]$ means τ is rolled out by starting with $s_h = s$ and $a_h = a$. To unify notations with Definition 8, we denote $V_i^{\pi} = V_{1,i}^{\pi}$ and $Q_i^{\pi} = Q_{1,i}^{\pi}$.

Definition 8 (Markov game, infinite-horizon)**.** An n-agent infinite-horizon (discounted) Markov game is defined by a tuple $(n, \gamma, \mathcal{S}, \{\mathcal{A}_i\}_{i\in[\![n]\!]}, \{r_i : \mathcal{S}\times\mathcal{A}\to[0,1]\}_{i\in[\![n]\!]}, P : \mathcal{S}\times\mathcal{A}\to\Delta_{\mathcal{S}}, s_1)$, where $\gamma\in(0,1)$ is the discount factor, $\mathcal{S}$ is the state space, $\mathcal{A}_i$ is agent i's action

space, $r_i : \mathcal{S} \times \mathcal{A} \to [0,1]$ is agent i's (deterministic) reward function, $P(\cdot|s,a) \in \Delta_{\mathcal{S}}$ is the transition dynamics, and $s_1 \in \mathcal{S}$ is the start state.

There are three types of policies being studied in infinite-horizon Markov game: stationary Markovian, nonstationary Markovian, non-Markovian.

- A stationary Markov policy is a map from states to distributions over actions, $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$, where $\pi(a|s)$ is the probability of the agents choosing joint action a when they are at state s. It rolls out a trajectory by $a_h \sim \pi(\cdot|s_h), s_{h+1} \sim P(\cdot|s_h, a_h)$.
- A nonstationary Markov policy is a collection of maps, $\pi = \{\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}\}_{h\in\mathbb{N}}$, which rolls out by $a_h \sim \pi_h(\cdot|s_h), s_{h+1} \sim P(\cdot|s_h, a_h)$.
- A non-Markov policy is a collection $\pi = \{\pi_h : (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^{h-1} \times \mathcal{S} \to \Delta_{\mathcal{A}}\}_{h\in\mathbb{N}}$, which rolls out a trajectory by $a_h \sim \pi_h(\cdot|\{(s_{h'}, a_{h'}, r_{h'})\}_{h'\in[\![h-1]\!]}, s_h), s_{h+1} \sim P_h(\cdot|s_h, a_h)$.

A trajectory is denoted by $\tau = (s_1, a_1, s_2, \cdots)$. We say a stationary Markov policy π is a product policy if $\pi(\cdot|s) \in \Delta_{\mathcal{A}_1} \times \cdots \times \Delta_{\mathcal{A}_n}$, and we say a nonstationary Markov policy π is a product policy if $\pi_h(\cdot|s) \in \Delta_{\mathcal{A}_1} \times \cdots \times \Delta_{\mathcal{A}_n}$ for every $s \in \mathcal{S}, h \in \mathbb{N}$.

For any policy π, we define its value function and quality value function (Q-function) by

$$V_i^{\pi}(s) = \mathop{\mathbb{E}}_{\tau\sim\pi}\left[\sum_{h=0}^{\infty} \gamma^h r_i(s_h, a_h) \mid s_1 = s\right], \tag{2.9}$$

$$Q_i^{\pi}(s,a) = \mathop{\mathbb{E}}_{\tau\sim\pi}\left[\sum_{h=0}^{\infty} \gamma^h r_i(s_h, a_h) \mid s_1 = s, a_1 = a\right], \tag{2.10}$$

where $\mathbb{E}_\tau[\cdot \mid s_1 = s]$ means the trajectory τ is rolled out by starting at $s_h = s$, and similarly $\mathbb{E}_\tau[\cdot \mid s_1 = s, a_1 = a]$ means τ is rolled out by starting with $s_1 = s$ and $a_1 = a$.

Definition 9 (Nash equilibrium)**.** In a finite-horizon Markov game, we say a Markov product policy π is an ϵ-approximate NE (ϵ-NE) if

$$\max_{1\le i\le n} \max_{\pi_i\in\Pi_i} V_i^{\pi_i,\pi_{-i}}(s_1) - V_i^{\pi}(s_1) \le \epsilon, \tag{2.11}$$

where $\Pi_i = \{\{\pi_{h,i} : \mathcal{S} \to \mathcal{A}_i\}_{h\in[\![H]\!]}\}$.

In an infinite-horizon Markov game, we say a stationary Markov product policy π is an ϵ-NE if (2.11) holds with $\Pi_i = \{\pi : \mathcal{S} \to \mathcal{A}_i\}$; we say a nonstationary Markov product policy π is an ϵ-NE if (2.11) holds with $\Pi_i = \{\{\pi_{h,i} : \mathcal{S} \to \mathcal{A}_i\}_{h\in\mathbb{N}}\}$.

In all the cases above, when $\epsilon = 0$, we say π is an NE.

There are various definitions of CCE in Markov game.

Definition 10 (Coarse correlated equilibria)**.** In general, in a Markov game, we say a policy π is an ϵ-approximate CCE (ϵ-CCE) if

$$\max_{1\le i\le n} \{\max_{\pi_i\in\Pi_i} V_i^{\pi_i,\pi_{-i}}(s_1) - V_i^{\pi}(s_1)\}, \tag{2.12}$$

with different choices of Π_i's for different types of CCE.

Daskalakis et al. [2022] defines two types of CCE in an infinite-horizon Markov game:[2]

- a stationary Markov policy π is an ϵ-approximate stationary Markov CCE (ϵ-stationary Markov CCE) if it satisfies (2.12) with $\Pi_i = \{\pi : \mathcal{S} \to \mathcal{A}_i\}$;
- a nonstationary Markov policy π is an ϵ-approximate nonstationary Markov CCE (ϵ-nonstationary Markov CCE) if it satisfies (2.12) with $\Pi_i = \{\{\pi_{h,i} : \mathcal{S} \to \mathcal{A}_i\}_{h\in\mathbb{N}}\}$.

Song et al. [2022] defines a type of CCE in a finite-horizon Markov game:

- A non-Markov policy π is an ϵ-CCE if it satisfies (2.12) with $\Pi_i = \{\{\pi_h : (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^{h-1} \times \mathcal{S} \to \Delta_{\mathcal{A}_i}\}_{h\in[\![H]\!]}\}$.

Liu et al. [2021] defines a type of CCE in a finite-horizon Markov game:

- A Markov policy π is an ϵ-CCE if it satisfies (2.12) with $\Pi_i = \{\{\pi_{h,i} : \mathcal{S} \to \mathcal{A}_i\}_{h\in[\![H]\!]}\}$.

In all cases above, when $\epsilon = 0$, we omit the prefix "ϵ-" and "ϵ-approximate" when describing the CCE.

Definition 11 (Correlated equilibria)**.** In a finite-horizon Markov game, a Markov policy [Liu et al., 2021], or non-Markov policy [Song et al., 2022], π, is an ϵ-approximate CE (ϵ-CE) if

$$\max_{1\le i\le n} \max_{\psi_i\in\Psi_i} V_i^{\psi\diamond\pi}(s_1) - V_i^{\pi}(s_1) \le \epsilon, \tag{2.13}$$

where $\Psi_i = \{\{\phi_{h,s} : \mathcal{A}_i \to \mathcal{A}_i\}_{\substack{h\in[\![H]\!]\\ s\in\mathcal{S}}}\}$ and $\psi\diamond\pi(\cdot|\cdot)$ is a distribution induced by first generating $a \sim \pi(\cdot|\cdot)$, then choosing $(\psi(a_i), a_{-i})$.

Definition 12 (Markov potential game)**.** In [Ding et al., 2022], an infinite-horizon Markov game is a Markov potential game if there exists a potential function $\Phi^{\pi}(s) : \Pi \times \mathcal{S} \to \mathbb{R}$ such that

$$V_i^{\pi_i,\pi_{-i}}(s_1) - V_i^{\pi_i',\pi_{-i}}(s_1) = \Phi^{\pi_i,\pi_{-i}}(s) - \Phi^{\pi_i',\pi_{-i}}(s) \tag{2.14}$$

for any $\pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$, where $\Pi_i = \{\pi_i : \mathcal{S} \to \mathcal{A}_i\}$ and $\Pi_{-i} = \{\pi_{-i} : \mathcal{S} \to \mathcal{A}_{-i}\}$.

In [Song et al., 2022], a finite-horizon Markov game is a Markov potential game if there exists a potential function $\Phi(\pi) : \Pi \to \mathbb{R}$ such that

$$V_i^{\pi_i,\pi_{-i}}(s_1) - V_i^{\pi_i',\pi_{-i}}(s_1) = \Phi(\pi_i, \pi_{-i}) - \Phi(\pi_i', \pi_{-i}) \tag{2.15}$$

for any $\pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$, where $\Pi_i = \{\{\pi_{h,i} : \mathcal{S} \to \mathcal{A}_i\}_{h\in[\![H]\!]}\}$ and $\Pi_{-i} = \{\{\pi_{-i} : \mathcal{S} \to \mathcal{A}_{-i}\}_{h\in[\![H]\!]}\}$. We assume $|\Phi| \le \Phi_{\max}$.

A (finite- or infinite-horizon) Markov cooperative game is an n-agent Markov game with $R_1 = \cdots = R_n$. Similar to cooperative game, this implicitly requires $\mathcal{A}_1 = \cdots = \mathcal{A}_n$, and that a Markov cooperative game is a Markov potential potential game with $\Phi = V_i$.

[2] Actually, they defined three types, but we only present two here.

Definition 13 (Zero-sum Markov game)**.** A two-agent infinite-horizon Markov game is called zero-sum if $r_1 = -r_2$. A two-agent finite-horizon Markov game is called zero-sum if $r_{h,1} = -r_{h,2}$ for every $h \in [\![H]\!]$.

2.4 Sample Complexity

Definition 14 (Bandit, Semi-bandit)**.** We study the minimum number of queries required to learn an approximate equilibrium (can be ϵ-Nash, ϵ-CCE, ϵ-CE) in game. Formally, we define two interactive protocols, bandit learning and semi-bandit learning, with two query constraints, pure and mixed query, and two noise levels, noisy and noiseless. In particular, for the noisy cases, we assume the noise has variance σ^2. We begin with the definition in the noisy mixed bandit learning case.

For each round $t = 1, \cdots, T$, the algorithm queries a strategy profile $\pi^{(t)} = (\pi_1^{(t)}, \cdots, \pi_n^{(t)})$ and sees the outcome $\mathbb{E}_{a\sim\pi^{(t)}} R_i(a) + \mathcal{N}(0, \sigma^2)$. We study two goals. The first is to minimize the cumulative regret, defined as the cumulative suboptimality gap, summing over $\pi^{(1)}, \cdots, \pi^{(t)}$. The definition of suboptimality gap is the left-hand side of (2.3), (2.4), or (4) if we are studying Nash, CCE, or CE, respectively. The second goal is to minimize the number of samples required for a given ϵ. Formally, the goal is to find the minimum T such that there exists an algorithm that finds an ϵ-approximate equilibrium with constant probability.

Next, we define the noiseless, pure, semi-bandit cases. For noiseless cases, we assume $\sigma = 0$. For pure query cases, we require that every $\pi_i^{(t)}$ is a pure strategy. For semi-bandit cases, the algorithm can observe $\mathbb{E}_{a\sim\pi^{(t)}} R_i(a_i, a_{-i}) + \mathcal{N}(0, \sigma^2)$ for every $a_i \in \mathcal{A}_i$.

Chapter 3

Survey on Existing Results

3.1 General Games

In this section, we survey on existing sample complexity bounds for Markov games. By "general", we refers to that we consider n agents in general. In the next section, we will specifically narrow down to the $n = 2$ agents case.

To save space, we only present the state-of-the-art bounds.

Learning Approximate Nash Equilibrium

We begin with a survey on learning ϵ-Nash equilibrium using MARL. The main existing results are presented in Table 3.1.

We find that learning approximate Nash equilibrium in Markov games can be very hard, as Song et al. [2022] shows an exponential lower bound on sample complexity. This indeed matches the hardness result in game theory [Rubinstein, 2016]. Albeit this, polynomial sample complexity bounds do exist for the two-agent case [Liu et al., 2021; Daskalakis et al., 2020]. We emphasize that there is no contradiction between their results and the exponential lower bound, because their bound scales with the product of the action spaces of two agents, which is at best exponential in the number of agents.

We highlight that polynomial sample complexity bounds do exist for an important special class of Markov games, namely the Markov potential games, as we see in the table [Song et al., 2022; Ding et al., 2022]. However, there is a huge unexplained gap between these polynomial bounds and the exponential lower bounds in Markov game: these polynomial bounds has the dependency on $\Phi_{\max}$, which, at a first glance, seems unrelated to the statistical complexity

[1] κ_ρ, D denotes the distribution shift (or distribution mismatch) coefficients. Please see the references for the exact definitions. These constants are typical for the theoretical analysis for policy gradient algorithms [Agarwal et al., 2021; Xiao, 2022].

[2] Exact bounds are not presented in their paper.

[3] This bound only holds for cooperative Markov game, which is a special case of Markov potential game with same rewards for all agents.

Table 3.1: Summary on Sample Complexity Bounds for Learning ϵ-Nash Equilibrium in Markov Games and Markov Potential Games.

	Value-Based MARL	Policy Gradient[1]
Markov Game	$\Omega(e^n)$ [Song et al., 2022]	
	$\widetilde{O}(\frac{H^3}{\epsilon^2}\lvert\mathcal{S}\rvert\prod_{i=1}^{2}\lvert\mathcal{A}_i\rvert)$ $\widetilde{O}(\frac{H^4}{\epsilon^2}\lvert\mathcal{S}\rvert^2\prod_{i=1}^{n}\lvert\mathcal{A}_i\rvert)$ [Liu et al., 2021]	Two-player case: Polynomial sample complexity[2] [Daskalakis et al., 2020]
Markov Potential Game	$\widetilde{O}(H^3\Phi_{\max}(\frac{\lvert\mathcal{S}\rvert}{\epsilon^3}+\frac{\lvert\mathcal{S}\rvert^2}{\epsilon^2})\sum_{i=1}^{n}\lvert\mathcal{A}_i\rvert)$ [Song et al., 2022]	$O(\frac{\kappa_\rho^2 n^2\Phi_{\max}}{(1-\gamma)^9\epsilon^4}\max_{1\le i\le n}\lvert\mathcal{A}_i\rvert^2)$ $O(\frac{\kappa_\rho^4 n\Phi_{\max}}{(1-\gamma)^6\epsilon^2}\max_{1\le i\le n}\lvert\mathcal{A}_i\rvert)$ $O(\frac{\kappa_\rho n}{(1-\gamma)^4\epsilon^2}\max_{1\le i\le n}\lvert\mathcal{A}_i\rvert)$[3] [Ding et al., 2022] $O(\frac{\gamma D^2 n\Phi_{\max}\lvert\mathcal{S}\rvert}{(1-\gamma)^5\epsilon^2}\max_{1\le i\le n}\lvert\mathcal{A}_i\rvert)$ [Leonardos et al., 2022]

of the problem, because the rewards are assumed to be 1-subgaussian. Meanwhile, sample complexity bounds independent of $\Phi_{\max}$ would instead scale exponentially with the number of agents [Liu et al., 2021].

We will partly explain this gap in Section 4.1, where we will show the surprising result that these polynomial bounds are indeed pseudo-polynomial bounds,[4] and that the best of Liu et al. [2021] and Song et al. [2022] is what we can hope. Formally, we will show that the sample complexity would inevitably scale with $2^{\Omega(\min\{n,\log\frac{\Phi_{\max}}{\epsilon}\})}$. Note that for cooperative Markov games, we have $\Phi_{\max}\le O(\frac{1}{1-\gamma})$, so this novel lower bound would not contradict [Ding et al., 2022]. We leave a full explanation for this gap as an open question.

Learning Approximate CCE

Next, we survey on learning ϵ-CCE. Table 3.2 presents main existing results.

We find that there are many acceptable definitions of CCE for Markov games, and different definitions lead to different results. Specifically, Daskalakis et al. [2022] shows that finding ϵ-stationary Markov CCE is PPAD-hard, which implies that any algorithm requires at least $2^{\Omega(\text{poly}(n))}$ time to find it [Fearnley et al., 2020]. This is however, not a statistical

[4]Pseudo-polynomial means that the bounds scale with the magnitude of the input. Here, we mean that $\Phi_{\max}$ dependency is inevitable for bounds that are polynomial in the number of agents.

Table 3.2: Summary on Sample Complexity Bounds for Learning ϵ-CCE in Markov Games.

	ϵ-stationary Markov	ϵ-nonstationary Markov	ϵ-non-Markov
Value-Based MARL	PPAD-hard [Daskalakis et al., 2022]	$\widetilde{O}(\frac{H^4}{\epsilon^2}\lvert\mathcal{S}\rvert^2\prod_{i=1}^{n}\lvert\mathcal{A}_i\rvert)$ [Liu et al., 2021] $\widetilde{O}(\frac{H^{10}}{\epsilon^3}\lvert\mathcal{S}\rvert^3\max_{1\le i\le n}\lvert\mathcal{A}_i\rvert)$ [Daskalakis et al., 2022]	$\widetilde{O}(\frac{H^5}{\epsilon^2}\lvert\mathcal{S}\rvert\max_{1\le i\le n}\lvert\mathcal{A}_i\rvert)$ [Song et al., 2022]
Policy Gradient		N/A	

Table 3.3: Summary on Sample Complexity Bounds for Learning ϵ-CE in Markov Games.

	ϵ-nonstationary Markov	ϵ-non-Markov
Value-Based MARL	$\widetilde{O}(\frac{H^4}{\epsilon^2}\lvert\mathcal{S}\rvert^2\prod_{i=1}^{n}\lvert\mathcal{A}_i\rvert)$ [Liu et al., 2021]	$\widetilde{O}(\frac{H^6}{\epsilon^2}\lvert\mathcal{S}\rvert\max_{1\le i\le n}\lvert\mathcal{A}_i\rvert^2)$ [Song et al., 2022]
Policy Gradient	N/A	

lower bound, and there could still exist algorithms that finds the ϵ-stationary Markov CCE with polynomial sample complexity.

For finding ϵ-nonstationary Markov CCE, there are two algorithms [Liu et al., 2021; Daskalakis et al., 2022]. Here, similar to the case of ϵ-Nash, the sample complexity of the first algorithm [Liu et al., 2021] scales with ϵ^{-2} but is exponential in the number of actions. That of the second algorithm [Daskalakis et al., 2022] is linear in n but scales with ϵ^{-3}. It is therefore interesting to ask if we could show if it is inevitable to have either ϵ^{-3} or 2^n. We leave this as an open question.

For finding ϵ-non-Markov CCE, we are only aware of the result in [Song et al., 2022]. We note that for this CCE definition, we could get both ϵ^{-2} and polynomial dependency on n. A second open question would be to see if there is a separation between it and nonstationary Markovian CCE.

Learning Approximate CE

Finally, we survey on learning ϵ-CE in Table 3.2. We do not find many results in MARL for learning ϵ-CE. Song et al. [2022] shows an $\widetilde{O}(\frac{H^6}{\epsilon^2}\lvert\mathcal{S}\rvert\max_{1\le i\le n}\lvert\mathcal{A}_i\rvert^2)$ sample complexity bound, which they claims to improve upon the $\widetilde{O}(\frac{H^4}{\epsilon^2}\lvert\mathcal{S}\rvert^2\prod_{i=1}^{n}\lvert\mathcal{A}_i\rvert)$ obtained by Liu et al. [2021]. However, we note, that the result in [Song et al., 2022] is for ϵ-non-Markov CE, while the one

Table 3.4: Summary on Sample Complexity Bounds for Learning ϵ-Nash Equilibrium in Two-Player Zero-Sum Games.

	Noiseless	Noisy
Pure Query	$O(N^2)$[6] $\Omega(N^2)$ when $\epsilon < \frac{1}{4N}$ [Fearnley and Savani, 2016]	$\widetilde{\Theta}(\frac{N}{\epsilon^2})$[7]
Mixed Query	$O(N^2)$ $\widetilde{O}(\frac{N}{\epsilon})$ [Daskalakis et al., 2011]	

in [Liu et al., 2021] is for ϵ-Markov CE, so it is hard to say if this is indeed an improvement. We note that all results are studying the finite-horizon case, so an open question would be showing similar results for CE.

3.2 Two-Player Games

In previous section, we study the sample complexity for Markov game in the most general case, where our focus is to see if it is possible to simultaneously obtain ϵ^{-2} and poly(n) for the sample complexity. In this section, our main issue is to see how the sample complexity would scale with the action set size. This requires a fine-grained examination on the sample complexity. In this section, we will fix $n = 2$, and we denote $N = \max\{|\mathcal{A}_1|, |\mathcal{A}_2|\}$. Recall that $|\mathcal{A}| = |\mathcal{A}_1||\mathcal{A}_2|$. We will study the model in Definition 14.

General-Sum Games

The sample complexity bounds for general sum games has a rather complete picture. Göös and Rubinstein [0] shows the $\Omega(N^{2-o(1)})$ lower bound on sample complexity for noiseless query, which almost matches the $O(N^2)$ upper bound obtained by querying every pair of pure strategies.

For noisy query, the MARL result [Liu et al., 2021] sugests an $\widetilde{O}(\frac{N^2}{\epsilon^2})$ upper bound when we reduce from Markov game to game. It is open whether there could be a matching lower bound, and the noiseless query lower bound strongly suggests that the answer is yes.[5]

Zero-sum Games

The results in zero-sum games are summarized in Table 3.4.

[5]Our guess is based on that the statistical hardness is decoupled from the game theoretical hardness.

We found that the sample complexity has been settled in the noisy query model [Lattimore and Szepesvári, 2020; Jin et al., 2022], but it is not well-understood in the noiseless query model. Moreover, the bounds in Table 3.4 are proved using very different methods. The $O(N^2)$ upper bound is simple. Fearnley and Savani [2016] studied the noiseless pure query model, and proved an $\Omega(N^2)$ lower bound when ϵ is small using combinatorial arguments. For the noiseless mixed query model, Daskalakis et al. [2011] gave an algorithm based on the excessive gap technique [Nesterov, 2005], which is from the optimization theory. In contrast, the lower bounds in the noisy query models are proved using statistical arguments based on Fano's inequality [e.g., Cover and Thomas, 2005; Wainwright, 2019; Lattimore and Szepesvári, 2020]. It is then interesting to study how the Nesterov gap technique would change as the noise level transits from $\sigma = 0$ to $\sigma = 1$. Formally, we ask the following question.

Problem 1 (Noise transition)**.** For $\sigma = [0, 1)$, what are the lower and upper bounds for the bandit query models in Definition 14 for zero-sum games?

In this report, we take an initial step by answering this question for $\sigma = 0$.

[6]This is obtained by querying all pairs of pure strategies.

[7]This is easily proved from standard lower bounds for bandits [Lattimore and Szepesvári, 2020], because finding ϵ-Nash is at least as hard as finding the best arm for N-armed bandits. This is also recovered by sample complexity bounds in MARL [Jin et al., 2022].

Chapter 4

Results

4.1 Pseudo-Polynomial Lower Bound for Finding Approximate Nash Equilibrium in Potential Game

We present a novel lower bound for potential games, showing that finding an ϵ-Nash equilibrium in n-player potential games is at least $2^{\Omega(\min\{n,\log\frac{\Phi_{\max}}{\epsilon}\})}$. Our result suggests that finding an ϵ-Nash is at best pseudo-polynomial or exponential, which negatively answers an open problem raised in [Song et al., 2022].

Problem 2. Find an ϵ-Nash equilibrium (Definition 2) in potential games (Definition 5) using noiseless pure query (Definition 14).

Proposition 1 (Babichenko and Rubinstein [2020], Corollary 2)**.** *For any algorithm $\mathcal{O}$, there exists an instance $\mathcal{I}$ of Problem 2 with $2(n+1)$ players and $\epsilon = 0$, such that $\mathcal{O}$ needs at least $2^{\Omega(n)}$ queries to solve $\mathcal{I}$ with constant probability.*

Remark 1. Although Babichenko and Rubinstein [2020] did not point out that the algorithm can be randomized, they actually allow it, because their result was based on a reduction to the lower bound in [Hubáček and Yogev, 2020], which was about the randomized query complexity.

Remark 2. We have the following observations on the instance $\mathcal{I}$ constructed in the proof [Babichenko and Rubinstein, 2020, Section 4.2] of Proposition 1. Let Φ be the potential function of $\mathcal{I}$. We have the following two claims.

- $\operatorname{Im}\Phi \subseteq \mathbb{Z}$;
- $|\Phi| \leq \Phi_{\max} = Cn30^{2n}$ for some numeric constant $C > 0$.

Proof. In their construction, the instance $\mathcal{I}$ has $2(n+1)$ players and $\mathcal{A}_i = \{0, \cdots, 29\}$ for each player $i \in [\![n]\!]$. In their Section 4.2, the potential function is defined by

$$\Phi(a) = -\sum_{i=1}^{n+1}(a_i - a_{i+n+1})^{2n} - 2\phi(a_{1:(n+1)}) - 2\phi(a_{(n+2):(2n+2)}), \tag{4.1}$$

where $\phi : \{0, \cdots, 29\}^{n+1} \to \mathbb{N}$ is defined their Section 3.2. We note that $|\phi| \leq 88 \cdot (2T+1) + |d_1| + 58$, where $T = 2^{n/2}$ and $d_1 \leq n \cdot 30$, so we conclude that $|\phi| \leq O(T)$ and thus by (4.1), we conclude that $\Phi_{\max} \leq C \cdot n \cdot 30^{2n}$ for some numeric constant $C > 0$. □

Corollary 2. *For any algorithm $\mathcal{O}$, there exists an instance $\mathcal{I}$ of Problem 2 with $2(n+1)$ players and $\epsilon < \frac{1}{30}$, such that $\mathcal{O}$ needs at least $2^{\Omega(n)}$ queries to solve $\mathcal{I}$ with constant probability.*

Proof. Let $\mathcal{I}$ be the instance constructed in Proposition 1. Let $\pi = (\pi_1, \cdots, \pi_n)$ be the strategy profile given by $\mathcal{O}$ when taking $\mathcal{I}$ as input. Note that by Proposition 3 in [Babichenko and Rubinstein, 2020, Section 4.2], the instance $\mathcal{I}$ has a unique (among pure and mixed) Nash equilibrium. Let $a^\star$ be the Nash equilibrium, $a_i = \arg\max_{a_i \in \mathcal{A}_i} \pi_i(a_i)$. We claim that $a_i^\star = a_i$. Then the lower bound in Proposition 1 implies our Corollary 2.

We proceed to prove $a_i^\star = a_i$ by contradiction. By the proof in their Lemma 2 and Proposition 3 in Section 4.2, for any $i \in [\![n]\!]$ such that $a_i \neq a_i^\star$, there exists $a_i' \in \mathcal{A}_i$, such that

$$R_i(a_i', \pi_{-i}) \geq R_i(a_i, \pi_{-i}^\star) + 1, \tag{4.2}$$

so if we define

$$\pi_i'(\alpha_i) = \begin{cases} \pi_i(a_i') + \pi_i(a_i), & \alpha_i = a_i', \\ 0, & \alpha_i = a_i, \\ \pi_i(\alpha_i), & \text{otherwise}, \end{cases} \tag{4.3}$$

then

$$R_i(\pi_i', \pi_{-i}) \geq R_i(\pi_i, \pi_{-i}) + \pi_i(a_i) \geq R_i(\pi_i, \pi_{-i}) + \frac{1}{|\mathcal{A}_i|} > R_i(\pi_i, \pi_{-i}) + \epsilon, \tag{4.4}$$

which contradicts to the supposition that π is an ϵ-Nash equilibrium. □

By taking $\Phi_{\max}$ into consideration, we prove the following lower bound.

Theorem 3. *Let $\Phi_{\max}$ be the upper bound of the potential function in the potential game. We have the following lower bounds on the sample complexity of Problem 2. Let $C > 0$ be some numeric constant.*

- *If $\frac{\Phi_{\max}}{\epsilon} > 30Cn30^{2n}$, then the sample complexity is at least $2^{\Omega(n)}$.*

- *If* $\frac{\Phi_{\max}}{\epsilon} \leq 30Cn30^{2n}$*, then it is at least* $2^{\Omega(\log \frac{\Phi_{\max}}{\epsilon})}$.

Proof. We note that the approximate Nash equilibrium is homogeneous. Formally, for any factor $c > 0$, an ϵ-Nash equilibrium in a game with reward R and potential Φ is an $c\epsilon$-Nash equilibrium in the game with reward cR and potential $c\Phi$.

Next, let C be the numeric constant given by Remark 2. If $\frac{\Phi_{\max}}{\epsilon} > 30Cn30^{2n}$, then we rescale the problem by $c = \frac{1}{30\epsilon}$. The new problem would have $\epsilon = \frac{1}{30}$ and $\Phi_{\max} \geq Cn30^{2n}$. Therefore, we conclude the sample complexity lower bound by Corollary 2.

If $\frac{\Phi_{\max}}{\epsilon} \leq 30Cn30^{2n}$, we can shrink the problem size by reducing the number of agents to $n' = \Omega(\log \frac{\Phi_{\max}}{\epsilon})$, so that $\frac{\Phi_{\max}}{\epsilon} > 30Cn'30^{2n'}$. Then we conclude by the previous proof. □

Theorem 3 can be interpreted as an $2^{\Omega(\min\{n, \log \frac{\Phi_{\max}}{\epsilon}\})}$ lower bound on the sample complexity for learning ϵ-Nash equilibrium in Markov potential games. This partly answers the open question raised in [Song et al., 2022], and most importantly, it suggests that one could not significantly improve upon [Liu et al., 2021] and [Song et al., 2022].

There are still two gaps between Theorem 3 and the open question in [Song et al., 2022]. The first is ϵ dependency: as pointed out in Section 3.1, the upper bound in [Song et al., 2022] scales with ϵ^{-3} while Theorem 3 scales with $\text{poly}(\epsilon^{-1})$. The second is action-set dependency: Theorem 3 does not show $\Omega(|\mathcal{A}|)$ but just $2^{\Omega(n)}$. However, we shall note that the second gap is not yet resolved even in games, let alone in Markov games; and the first gap is not as important as we might thought. This is because $\frac{\Phi_{\max}}{\epsilon^3} = O(2^{\log \frac{\Phi_{\max}}{\epsilon^3}}) = 2^{O(\log \frac{\Phi_{\max}}{\epsilon})}$, and the difference between the two bounds is significant only when $\epsilon^{-1} \geq 2^{\Omega(n)}$, which is not the typical regime of interest for the parameters. Moreover, an $\Omega(2^n)$ lower bound is not proved even in games. Would it be proved in the future, our method could be used to improve it to take $\Phi_{\max}$ into consideration.

4.2 Query Complexity Lower Bound in Two-Player Zero-Sum Games

In this section, we initialize the study of Problem 1 in by proving an $\Omega(N)$ lower bound.

Theorem 4. *For the* $\sigma = 0$ *case in Problem 1, the sample complexity of any algorithm is at least* $\Omega(N)$*, for any* $\epsilon \in [0, \frac{1}{2})$.

Proof. Let $N_i = |\mathcal{A}_i|$ for $i \in [\![2]\!]$. Let e_i be the unit vector in $\mathbb{R}^{N_1}$ or $\mathbb{R}^{N_2}$ (which should be clear from context) such that $e_i = (\cdots, 0, 1, 0, \cdots)$ is the identity vector in the i-th dimension. Without loss of generality, we assume that $2 \leq N_1 \leq N_2 = N$. Let $\mathcal{O}$ be any algorithm. We also assume that $\mathcal{O}$ returns the last strategy profile it queried as the ϵ-Nash equilibrium it claims to have found. The proof idea is to note that finding an ϵ-Nash equilibrium is at least as hard as computing the best response.

Formally, we construct the hard instance class with payoff matrix in the form of $uv^\intercal$, where $u = (1, \cdots, 1) \in \mathbb{R}^{N_1}$ and $v \in \mathbb{R}^{N_2}$. Let (x_i, y_i) be the strategy profile queried by $\mathcal{O}$ for

$i \in [\![N]\!]$. We consider the adversary that always outputs $x_i^\intercal R y_i = 0$ for $i \leq N-2$. Because $x_i \in \Delta_{N_1}$, we have $x_i^\intercal u = 1$ and thus

$$x_i^\intercal R y_i = x_i^\intercal u v^\intercal y_i = v^\intercal y_i. \tag{4.5}$$

Define $Y = \{y_1, \cdots, y_{N-1}\}$. Note that

$$Y^\perp = \{v \in \mathbb{R}^{N_2} : v^\intercal y_i = 0 \text{ for } 1 \leq i \leq N-1\} \neq \emptyset, \tag{4.6}$$

so we can take $\gamma \in Y^\perp$ such that $\gamma \neq 0$. Let $k = \arg\max_{i \in [\![N_2]\!]} \gamma^\intercal e_i$. We assume $\|\gamma\|_\infty = \gamma^\intercal e_k = 1$. Otherwise we consider $\pm \frac{\gamma}{\|\gamma\|_\infty}$, respectively. Let $R = u\gamma^\intercal$. Then $x_i^\intercal R y_i = x_i^\intercal u \alpha^\intercal y_i = 0$ for every $i \in [\![N-1]\!]$, but

$$\max_{y \in \Delta_N} x_i^\intercal R y - x_i^\intercal R y_i \geq x_i^\intercal u \alpha^\intercal e_k - 0 = 1, \tag{4.7}$$

so (x_i, y_i) cannot be an ϵ-Nash. Note that in (4.7) we get $\epsilon = 1$, but we stated the theorem for $\epsilon < \frac{1}{2}$. This is because for zero-sum people usually take $R_1 = -R$ and $R_2 = R$, and assume $|R| \leq 1$. However, Definition 1 requires that $0 \leq R_i \leq 1$, so we need to offset by $R_1 = \frac{1-R}{2}, R_2 = \frac{1+R}{2}$ and multiply ϵ by $\frac{1}{2}$. □

Chapter 5

Discussion

5.1 Conclusion

In this report, we summarize and unify the definitions in MARL from various paper, and we survey on existing results in game theory and MARL theory. By comparing the results, we find various gaps inside the study of MARL theory, and between the game theory and MARL theory. We believe this report could serve as a good start point for future studies in MARL theory.

5.2 Open Problems

For better indexing, we list all open problems raised in this report. We came up with four open question in Section 3.1: one for ϵ-Nash, two for ϵ-CCE, and one for ϵ-CE. We came up with one open question in Section 3.2, which is Problem 1.

Bibliography

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021. URL http://jmlr.org/papers/v22/19-736.html.

Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Tuomas Sandholm. Uncoupled learning dynamics with $o(\log t)$ swap regret in multiplayer games, 2022. URL https://arxiv.org/abs/2204.11417.

Yakov Babichenko. Informational bounds on equilibria (a survey). *SIGecom Exch.*, 17 (2):25–45, jan 2020. doi: 10.1145/3381329.3381333. URL https://doi.org/10.1145/3381329.3381333.

Yakov Babichenko and Aviad Rubinstein. Communication complexity of nash equilibrium in potential games (extended abstract). In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1439–1445, 2020. doi: 10.1109/FOCS46700.2020.00137.

Yakov Babichenko and Aviad Rubinstein. *Settling the Complexity of Nash Equilibrium in Congestion Games*, page 1426–1437. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380539. URL https://doi.org/10.1145/3406325.3451039.

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkxpxJBKwS.

Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. Emergent complexity via multi-agent competition. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sy0GnUxCb.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d. O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning, 2019. URL https://arxiv.org/abs/1912.06680.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 04 2005. doi: 10.1002/047174882x. URL https://doi.org/10.1002/047174882x.

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '11, page 235–254, USA, 2011. Society for Industrial and Applied Mathematics.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5527–5540. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/3b2acfe2e38102074656ed938abf4ac3-Paper.pdf.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27604–27616. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/e85cc63b4f0f312f11e073fc68ccffd5-Paper.pdf.

Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of markov equilibrium in stochastic games, 2022. URL https://arxiv.org/abs/2204.03991.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo R. Jovanović. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence, 2022. URL https://arxiv.org/abs/2202.04129.

Kefan Dong, Jiaqi Yang, and Tengyu Ma. Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26168–26182. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/dc5d637ed5e62c36ecb73b654b05ba2a-Paper.pdf.

John Fearnley and Rahul Savani. Finding approximate nash equilibria of bimatrix games via payoff queries. *ACM Transactions on Economics and Computation*, 4(4), 08 2016. ISSN 2167-8375. doi: 10.1145/2956579. URL https://doi.org/10.1145/2956579.

John Fearnley, Martin Gairing, Paul W. Goldberg, and Rahul Savani. Learning equilibria of games via payoff queries. *Journal of Machine Learning Research*, 16(39):1305–1344, 2015. URL http://jmlr.org/papers/v16/fearnley15a.html.

John Fearnley, Spencer Gordon, Ruta Mehta, and Rahul Savani. Unique end of potential line. *Journal of Computer and System Sciences*, 114:1–35, 2020. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2020.05.007. URL https://www.sciencedirect.com/science/article/pii/S0022000020300520.

Mika Göös and Aviad Rubinstein. Near-optimal communication lower bounds for approximate nash equilibria. *SIAM Journal on Computing*, 0(0):FOCS18–316–FOCS18–348, 0. doi: 10.1137/19M1242069. URL https://doi.org/10.1137/19M1242069.

Baihe Huang, Kaixuan Huang, Sham Kakade, Jason D Lee, Qi Lei, Runzhe Wang, and Jiaqi Yang. Going beyond linear rl: Sample efficient neural function approximation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8968–8983. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/4b4edc2630fe75800ddc29a7b4070add-Paper.pdf.

Pavel Hubáček and Eylon Yogev. Hardness of continuous local search: Query complexity and cryptographic lower bounds. *SIAM Journal on Computing*, 49(6):1128–1172, 2020. doi: 10.1137/17M1118014. URL https://doi.org/10.1137/17M1118014.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010. URL http://jmlr.org/papers/v11/jaksch10a.html.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning – a simple, efficient, decentralized algorithm for multiagent RL. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*, 2022. URL https://openreview.net/forum?id=Bx-evj5k6x9.

Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. Google research football: A novel reinforcement learning environment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4501–4510, Apr. 2020. doi: 10.1609/aaai.v34i04.5878. URL https://ojs.aaai.org/index.php/AAAI/article/view/5878.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gfwON7rAm4.

Gen Li, Laixi Shi, Yuxin Chen, Yuantao Gu, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17762–17776. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/94739e5a5164b4d2396e253a11d57044-Paper.pdf.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning*, ICML'94, page 157–163, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1558603352.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7001–7010. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liu21z.html.

Yu. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, 16(1):235–249, 2005. doi: 10.1137/S1052623403422285. URL https://doi.org/10.1137/S1052623403422285.

Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007. doi: 10.1017/CBO9780511800481.

Open Ended Learning Team, Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, Nat McAleese, Nathalie Bradley-Schmieg, Nathaniel Wong, Nicolas Porcel, Roberta Raileanu, Steph Hughes-Fitt, Valentin Dalibard, and Wojciech Marian Czarnecki. Open-ended learning leads to generally capable agents, 2021. URL https://arxiv.org/abs/2107.12808.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1-2):1–35, August 2019. doi: 10.1007/s10107-019-01420-0. URL https://doi.org/10.1007/s10107-019-01420-0.

Aviad Rubinstein. Settling the complexity of computing approximate two-player nash equilibria. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 258–265, 2016. doi: 10.1109/FOCS.2016.35.

L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953. doi: 10.1073/pnas.39.10.1095. URL https://www.pnas.org/doi/abs/10.1073/pnas.39.10.1095.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=6MmiS0HUJHR.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, October 2019. doi: 10.1038/s41586-019-1724-z. URL https://doi.org/10.1038/s41586-019-1724-z.

Nisheeth K. Vishnoi. *Algorithms for Convex Optimization*. Cambridge University Press, 2021. doi: 10.1017/9781108699211.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Lin Xiao. On the convergence rates of policy gradient methods, 2022. URL https://arxiv.org/abs/2201.07443.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Improved variance-aware confidence sets for linear bandits and linear mixture mdp. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4342–4355. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/228bbc2f87caeb21bb7f6949fddcb91d-Paper.pdf.