Approximate State Abstraction for Markov Games

Hiroki Ishibashi[1], **Kenshi Abe**[1, 2], **Atsushi Iwasaki**[1]

[1]The University of Electro-Communications
[2]CyberAgent
i2430006@edu.cc.uec.ac.jp, abekenshi1224@gmail.com, atsushi.iwasaki@uec.ac.jp

Abstract

This paper introduces state abstraction for two-player zero-sum Markov games (TZMGs), where the payoffs for the two players are determined by the state representing the environment and their respective actions, with state transitions following Markov decision processes. For example, in games like soccer, the value of actions changes according to the state of play, and thus such games should be described as Markov games. In TZMGs, as the number of states increases, computing equilibria becomes more difficult. Therefore, we consider state abstraction, which reduces the number of states by treating multiple different states as a single state. There is a substantial body of research on finding optimal policies for Markov decision processes using state abstraction. However, in the multi-player setting, the game with state abstraction may yield different equilibrium solutions from those of the ground game. To evaluate the equilibrium solutions of the game with state abstraction, we derived bounds on the duality gap, which represents the distance from the equilibrium solutions of the ground game. Finally, we demonstrate our state abstraction with Markov Soccer, compute equilibrium policies, and examine the results.

1 Introduction

Multi-agent reinforcement learning (MARL) is a framework for sequential decision-making, where multiple agents make decisions in a non-stationary environment to maximize their cumulative rewards. MARL has a wide range of applications, e.g., robotics, distributed control, game AI, and so on (Shalev-Shwartz, Shammah, and Shashua 2016; Silver et al. 2016, 2017; Brown and Sandholm 2018; Perolat et al. 2022). Such an environment is often modeled as two-player zero-sum Markov games (TZMGs) (Littman 1994) and computing the equilibria is said to be empirically tractable. However, it still suffers from the exponential growth of state space size in the number of domain variables.

Markov decision processes (MDPs), which are a single-agent version of Markov games, face the same challenge. Solving MDPs, i.e., computing an optimal policy, is P-Complete in state space size (Littman 1994), while that size often exponentially increases. Several state abstraction techniques have been developed, which aggregate multiple states into one abstract state according to a certain criterion and reduce the state space size. For example, a state is equivalent to another state if choosing an action leads to the same state with same rewards. Such abstraction is effective for agents to solve more complicated MDPs than they would be able to without using abstraction. However, if we abstract the state space by regarding only identical states as equivalent, we are difficult to find all of them within a reasonable time, and even worse no two states might be identical.

In contrast to such *exact state abstraction*, Abel, Hershkowitz, and Littman (2016) proposed *approximate state abstraction*, which characterizes how close a state is to another state in single-agent MDPs. This technique reduces ground MDPs with large state spaces to abstract MDPs with smaller state spaces by aggregating states according to some notion of closeness or similarity. While this relaxation makes the state spaces smaller, the resulting optimal policies in abstract MDPs may become suboptimal. Furthermore, they derive error bounds for the resulting policies based on four different criteria, such as optimal Q-values, according to which states are aggregated.

TZMGs extend MDPs to model decision-making by two interacting agents in a shared, state-dependent environment. Unlike MDPs, the rewards in TZMGs depend on the actions of both agents. For instance, in a soccer game, the rewards vary based on the state of play and the actions chosen by both players. TZMGs provide a framework to formalize such scenarios and have stimulated research in competitive MARL. An agent's optimal policy depends on the policy of its opponent. The goal is to identify the equilibrium policy profile, which consists of mutual best responses, to predict and understand the consequences of their interactions.

Building upon these insights, this paper extends the work by Abel, Hershkowitz, and Littman (2016) from single-agent MDPs to TZMGs. We first describe an approximate state abstraction based on optimal Q-value criteria or minimax values and derive the bound of the duality gap for the resulting equilibrium, which measures the proximity to equilibrium. To establish the bound, we analyze the gains from best responses in the ground game – where the strategy space is unabstracted – against the resulting equilibrium in the abstracted game, where the strategy space has been simplified. Second, we conduct experiments in Markov Soccer and demonstrate how state spaces are reduced and how well

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

the equilibrium strategies in the ground TZMG are approximated. Furthermore, we discuss the extension for the other three criteria.

Related Literature. There is a lot of literature on state abstraction for MDPs initiated by (Dietterich 1998, 1999; Jonsson and Barto 2000). The concept of state equivalence has been developed to reduce state space size by aggregating equivalent states (Givan, Dean, and Greig 2003), and it has been relaxed by allowing some associated actions (Ravindran and Barto 2003, 2004; van der Pol et al. 2020). Ferns, Panangaden, and Precup (2004) proposed a distance between states and aggregates states with zero distance. Similarly, Castro (2020) parameterized the *bisimulation* metric (Ferns, Panangaden, and Precup 2004). Recently, Dadvar, Nayyar, and Srivastava (2023) incorporated state abstraction into policy iteration. Li, Walsh, and Littman (2006) discussed the optimality of policies on five criteria for state abstraction.

In another line of work, state space abstraction has been developed in poker AI (Gilpin 2006; Gilpin and Sandholm 2006, 2007; Gilpin, Sandholm, and Sørensen 2007; Johanson et al. 2013; Waugh 2013; Ganzfried and Sandholm 2013; Burch, Johanson, and Bowling 2014; Kroer and Sandholm 2018). The state spaces in poker can have at most 10×10^{160} states, and this area has been extensively investigated. Initially, states were abstracted manually based on knowledge and experience in poker. Subsequently, automated abstraction techniques were developed. For example, states are aggregated by estimating winning probabilities at information sets (Gilpin and Sandholm 2007; Gilpin, Sandholm, and Sørensen 2007; Johanson et al. 2013; Waugh 2013; Ganzfried and Sandholm 2013). Furthermore, Kroer and Sandholm (2018) presented a unified framework for analyzing abstractions that can express all types of abstractions and solution concepts used in prior work, with performance guarantees, while maintaining comparable bounds on abstraction quality.

2 Preliminaries

2.1 Two-Player Zero-Sum Markov Games

A two-player zero-sum Markov game (TZMG) $\mathcal{M}$ is defined by a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}_1, \mathcal{A}_2, P, R, \gamma \rangle$. Here, $\mathcal{S}$ represents a finite state space, $\mathcal{A}_i$ represents an action space for player $i \in \{1, 2\}$, $P : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \to \Delta(\mathcal{S})$ represents a transition probability function, $R : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \to [0, 1]$ represents a reward function, and $\gamma \in [0, 1)$ represents a discount factor. Let $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$, and let $\boldsymbol{a} = (a_1, a_2) \in \mathcal{A}$ denote the action profile. For a given state $s \in \mathcal{S}$ and an action profile $\boldsymbol{a} \in \mathcal{A}$, the next state is determined according to $P(\cdot|s, \boldsymbol{a})$, and player 1 (resp. player 2) receives a reward of $R(s, \boldsymbol{a})$ (resp. $-R(s, \boldsymbol{a})$).

A Markov policy for player i, denoted as $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$, represents the probability of choosing action $a_i \in \mathcal{A}_i$ at a state $s \in \mathcal{S}$. Letting $\boldsymbol{\pi} = (\pi_1, \pi_2)$ be a policy profile, we further define the state value function, which is the expected discounted sum of rewards at state $s \in \mathcal{S}$ as follows:

$$V^{\boldsymbol{\pi}}(s) := \mathrm{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, \boldsymbol{a}_t) \,\middle|\, s_1 = s, \right.$$
$$\left. \boldsymbol{a}_t \sim \boldsymbol{\pi}(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, \boldsymbol{a}_t), \forall t \geq 0 \right].$$

We similarly define the state-action value function of taking an action profile $\boldsymbol{a} \in \mathcal{A}$ at state $s \in \mathcal{S}$ as follows:

$$Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}) := R(s, \boldsymbol{a}) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \boldsymbol{a}) V^{\boldsymbol{\pi}}(s').$$

From the definition of these functions, the state value function $V^{\boldsymbol{\pi}}$ can be expressed as follows:

$$V^{\boldsymbol{\pi}}(s) = \sum_{\boldsymbol{a} \in \mathcal{A}} \boldsymbol{\pi}(\boldsymbol{a}|s) Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}).$$

2.2 Nash Equilibrium

In TZMGs, a *Nash equilibrium* is defined as the policy profile that satisfies the following condition for all $s \in \mathcal{S}$ simultaneously:

$$\forall (\pi_1, \pi_2),\ V^{\pi_1^*, \pi_2}(s) \geq V^{\boldsymbol{\pi}^*}(s) \geq V^{\pi_1, \pi_2^*}(s).$$

Intuitively, a Nash equilibrium is the policy profile where no player can improve her value by deviating from her own policy. Shapley (1953) has shown that any Nash equilibrium $\boldsymbol{\pi}^*$ in TZMGs satisfies the following condition for all $s \in \mathcal{S}$:

$$\begin{aligned} V^{\boldsymbol{\pi}^*}(s) &= \max_{p \in \Delta(\mathcal{A}_1)} \min_{a_2 \in \mathcal{A}_2} \sum_{a_1 \in \mathcal{A}_1} p(a_1) Q^{\boldsymbol{\pi}^*}(s, \boldsymbol{a}) \\ &= \min_{p \in \Delta(\mathcal{A}_2)} \max_{a_1 \in \mathcal{A}_1} \sum_{a_2 \in \mathcal{A}_2} p(a_2) Q^{\boldsymbol{\pi}^*}(s, \boldsymbol{a}). \end{aligned} \tag{1}$$

It is known that this minimax value is unique for each s (Shapley 1953), thus we can write $V^*(s) := V^{\boldsymbol{\pi}^*}(s)$ and $Q^*(s, \boldsymbol{a}) := Q^{\boldsymbol{\pi}^*}(s, \boldsymbol{a})$.

To measure the proximity to equilibrium for a given policy profile $\boldsymbol{\pi} = (\pi_1, \pi_2)$, we use the *duality gap* defined as follows:

$$\mathrm{GAP}\,(\boldsymbol{\pi}) := \max_{s \in \mathcal{S}, \pi_1', \pi_2'} \left(V^{\pi_1', \pi_2}(s) - V^{\pi_1, \pi_2'}(s) \right).$$

From the definition, we can see that $\mathrm{GAP}\,(\boldsymbol{\pi}) \geq 0$ for any $\boldsymbol{\pi}$, and the equality holds if and only if $\boldsymbol{\pi}$ is a Nash equilibrium.

2.3 Minimax Q-learning

Let us briefly describe Minimax Q-learning (Littman 1994) which is developed to address the limitations of standard Q-learning in adversarial, zero-sum environments. Its robustness and adaptability in competitive settings make it effective for AI in games and security applications. It has been shown that Minimax Q-learning converges to a Nash equilibrium under some appropriate conditions (Szepesvári and Littman 1999). Due to its theoretical guarantee and ease of implementation, Minimax Q-learning is used as a standard algorithm to compute equilibrium policies for TZMGs.

Algorithm 1 illustrates the procedure with finite T iterations.

Algorithm 1: Minimax Q-learning

```
Input: Learning rates $\alpha_t$, exploration parameter $\beta$
$V[s] \leftarrow 0$ for all $s \in \mathcal{S}$
$Q[s, \boldsymbol{a}] \leftarrow 0$ for all $s \in \mathcal{S}$ and $\boldsymbol{a} \in \mathcal{A}$
$\pi_i^0(\cdot|s) \leftarrow \left(\frac{1}{|\mathcal{A}_i|}\right)_{a \in \mathcal{A}_i}$ for all $i \in \{1, 2\}$ and $s \in \mathcal{S}$
Sample initial state $s_0$
for $t = 0, 1, \dots, T-1$ do
    $\pi'_i(\cdot|s_t) \leftarrow (1-\beta)\pi_i^t(\cdot|s_t) + \frac{\beta}{|\mathcal{A}_i|}\mathbf{1}$ for all $i \in \{1, 2\}$
    Sample action profile $\boldsymbol{a}_t \sim \boldsymbol{\pi}'(\cdot|s_t)$
    Next state is sampled $s_{t+1} \sim P(\cdot|s_t, \boldsymbol{a}_t)$
    $Q[s_t, \boldsymbol{a}_t] \leftarrow (1-\alpha_t)Q[s_t, \boldsymbol{a}_t]$
                $+ \alpha_t \left(R(s_t, \boldsymbol{a}_t) + \gamma V^{\boldsymbol{\pi}}[s_{t+1}]\right)$
    $\pi_1^{t+1}(\cdot|s_t) \leftarrow \arg\max_{p \in \Delta(\mathcal{A}_1)} \min_{a_2 \in \mathcal{A}_2} \sum_{a_1 \in \mathcal{A}_1} p(a_1) Q[s_t, \boldsymbol{a}_t]$
    $\pi_2^{t+1}(\cdot|s_t) \leftarrow \arg\min_{p \in \Delta(\mathcal{A}_2)} \max_{a_1 \in \mathcal{A}_1} \sum_{a_2 \in \mathcal{A}_2} p(a_2) Q[s_t, \boldsymbol{a}_t]$
    $V(s_t) \leftarrow \sum_{a_1 \in \mathcal{A}_1} \sum_{a_2 \in \mathcal{A}_2} \pi_1^{t+1}(a_1|s_t)\pi_2^{t+1}(a_2|s_t) Q[s_t, \boldsymbol{a}_t]$
end for
Output: $(\pi_1^T, \pi_2^T)$
```

1. At each iteration $t \geq 0$, each player $i \in \{1, 2\}$ chooses an action $a_{i,t} \in \mathcal{A}_i$ randomly with probability β, otherwise based on her policy $\pi_i^t(\cdot|s_t) \in \Delta(\mathcal{A}_i)$.
2. State s_t transits to s_{t+1} according to the transition probability function $P(\cdot \mid s_t, a_{1,t}, a_{2,t})$.
3. State-action value function Q_t is updated with the obtained reward $R(s_t, \boldsymbol{a}_t)$ and the learning rate α_t:
$$\begin{aligned} Q_{t+1}(s_t, \boldsymbol{a}_t) :=& (1-\alpha_t)Q_t(s_t, \boldsymbol{a}_t) \\ &+ \alpha_t(R(s_t, \boldsymbol{a}_t) + V_t(s_{t+1})). \end{aligned}$$
4. Players update their policies using linear programming:
$$\pi_1(\cdot|s_t) := \arg\max_{p \in \Delta(\mathcal{A}_1)} \min_{a_2 \in \mathcal{A}_2} \sum_{a_1 \in A_1} p(a_1) Q_{t+1}(s_t, \boldsymbol{a}),$$
$$\pi_2(\cdot|s_t) := \arg\min_{p \in \Delta(\mathcal{A}_2)} \max_{a_1 \in \mathcal{A}_1} \sum_{a_2 \in A_2} p(a_2) Q_{t+1}(s_t, \boldsymbol{a}).$$
5. They update their state value function with current policy:
$$V_{t+1}(s_t) := \sum_{a_1 \in \mathcal{A}_1} \sum_{a_2 \in \mathcal{A}_2} \pi_1(a_1|s_t)\pi_2(a_2|s_t) Q_{t+1}(s_t, \boldsymbol{a}).$$

We implement this procedure when we compute the duality gap in Section 5. Note that our proposed state abstraction does not depend on Minimax Q-learning as well as in (Abel, Hershkowitz, and Littman 2016). In fact, we are going to discuss the extensions to different criteria of aggregating states.

3 State Abstraction

In this section, we extend state abstraction for an MDP to a TZMG. State abstraction is a method for reducing the state space by aggregating similar states to decrease the time of calculating the equilibria. In the previous research of Abel et al. (Abel, Hershkowitz, and Littman 2016), they propose four different state abstraction approaches for an MDP and theoretically analyze how well the optimal policy in the abstract MDP achieves performance in the ground MDP using various metrics. In this section, we extend their approach based on state-action value function similarity. The other approaches, including model similarity, Boltzmann distribution similarity, and multinomial distribution similarity, are introduced in the Discussion section.

3.1 Abstract Two-Player Zero-Sum Markov Games

This section defines an abstract TZMG $\mathcal{M}_A = \langle \mathcal{S}_A, \mathcal{A}_1, \mathcal{A}_2, P_A, R_A, \gamma \rangle$ by using the notation introduced by Abel, Hershkowitz, and Littman (2016); Li, Walsh, and Littman (2006). Here, $\mathcal{S}_A$ is an abstract state space, and $P_A : \mathcal{S}_A \times \mathcal{A}_1 \times \mathcal{A}_2 \to \Delta(\mathcal{S}_A)$ is an abstract transition probability function, and $R_A : \mathcal{S}_A \times \mathcal{A}_1 \times \mathcal{A}_2 \to [0, 1]$ is an abstract reward function.

Let $\phi : \mathcal{S} \to \mathcal{S}_A$ be a *state aggregation function* that maps from the ground state space $\mathcal{S}$ to an abstract state space $\mathcal{S}_A$. Given ϕ, we can define a set of states $G_A(s_A)$ that are aggregated into the same abstract state $s_A \in \mathcal{S}_A$:

$$G_A(s_A) := \{g \in \mathcal{S} \mid \phi(g) = s_A\}.$$

For a given ground state $s \in \mathcal{S}$, we further define a set of states $G(s)$ that are aggregated into the same abstract state as s:

$$G(s) := \{g \in \mathcal{S} \mid \phi(g) = \phi(s)\}.$$

In order to construct an abstract transition probability function P_A and an abstract reward function R_A, we introduce a *weight function* $w : \mathcal{S} \to [0, 1]$, which satisfies the following condition:

$$\forall s_A \in \mathcal{S}_A, \quad \sum_{g \in G_A(s_A)} w(g) = 1.$$

For example, $w(s) = 1/|G(s)|$ is a representative weight function. Using this function, we define the abstract transition probability function P_A as follows:

$$P_A(s'_A|s_A, \boldsymbol{a}) := \sum_{s \in G_A(s_A)} \sum_{s' \in G_A(s'_A)} P(s'|s, \boldsymbol{a}) w(s).$$

Similarly, the abstract reward function $R_A : \mathcal{S}_A \times \mathcal{A} \to [0, 1]$ is defined as follows:

$$R_A(s_A, \boldsymbol{a}) := \sum_{s \in G_A(s_A)} R(s, \boldsymbol{a}) w(s).$$

The policy profile in the abstract TZMG $\mathcal{M}_A$ is defined similarly to the ground TZMG $\mathcal{M}$. Let $\pi_{A,i} : \mathcal{S}_A \to \Delta(\mathcal{A}_i)$ be a policy for player i in $\mathcal{M}_A$. Similarly, $V_A^{\boldsymbol{\pi}_A}(s_A)$ denotes a state value for a given policy profile $\boldsymbol{\pi}_A$ at state $s_A \in \mathcal{S}_A$ in the abstract TZMG, and $Q_A^{\boldsymbol{\pi}_A}(s_A, \boldsymbol{a})$ is defined as a state-action value for $s_A \in \mathcal{S}_A$ and $\boldsymbol{a} \in \mathcal{A}$. Finally, letting $\boldsymbol{\pi}^*_A$ be a Nash equilibrium in the abstract TZMG $\mathcal{M}_A$, $\boldsymbol{\pi}^*_A$ must satisfy the following condition for all $s_A \in \mathcal{S}_A$ and $(\pi_{A,1}, \pi_{A,2})$:

$$V_A^{\pi^*_{A,1}, \pi_{A,2}}(s_A) \geq V_A^{\boldsymbol{\pi}^*_A}(s_A) \geq V_A^{\pi_{A,1}, \pi^*_{A,2}}(s_A).$$

4 Abstraction Based on Minimax Values

In this section, we analyze the performance of a Nash equilibrium $\boldsymbol{\pi}_A^*$ in an abstract TZMG $\mathcal{M}_A$ under a specific aggregation function ϕ when applied to the ground TZMG $\mathcal{M}$. To this end, we define the policy profile $\boldsymbol{\pi}_{GA}^*(s)$ in the ground TZMG, which is induced by $\boldsymbol{\pi}_A^*$. Formally, $\boldsymbol{\pi}_{GA}^*(s)$ for all s in $\mathcal{S}$ is given by:

$$\boldsymbol{\pi}_{GA}^*(s) := \boldsymbol{\pi}_A^*(\phi(s)).$$

In a later section, we derive the upper bound on the duality gap of $\boldsymbol{\pi}_{GA}^*$ under various aggregation functions ϕ.

4.1 Suboptimality of Nash Equilibria in Abstract TZMGs

We examine the aggregation function ϕ^{Q^*}, which is constructed based on the state-action value function Q^*. Specifically, in the abstraction ϕ^{Q^*}, states are aggregated to the same abstract state when their minimax state-action values are close within ϵ.

Assumption 1. *The aggregation function ϕ^{Q^*} satisfies the following property for some non-negative constant $\epsilon \geq 0$:*

$$\begin{aligned} &\phi^{Q^*}(s_1) = \phi^{Q^*}(s_2) \\ &\Rightarrow \forall \boldsymbol{a} \in \mathcal{A},\ |Q^*(s_1, \boldsymbol{a}) - Q^*(s_2, \boldsymbol{a})| \leq \epsilon. \end{aligned} \tag{2}$$

Under Assumption 1, it can be shown that the suboptimality of $\boldsymbol{\pi}_{GA}^*$ is no more than $\mathcal{O}(\epsilon)$.

Theorem 1. *When the ground states are aggregated by the aggregation function ϕ^{Q^*} satisfying Assumption 1 with $\epsilon \geq 0$, then $\boldsymbol{\pi}_{GA}^*$ satisfies:*

$$\mathrm{GAP}\left(\boldsymbol{\pi}_{GA}^*\right) \leq \frac{12\epsilon}{(1-\gamma)^3}.$$

To perform an initial abstraction, minimax Q-learning is used to calculate Q-values for the ground game. If the results do not meet a sufficient solution criterion, the process iteratively updates the Q-values and refines the abstraction (Li, Walsh, and Littman 2006). Developing efficient algorithms for discovering abstractions remains open. Also, such abstractions have potential utility beyond a single game, enabling transferable knowledge across related games (Jong and Stone 2005).

4.2 Proofs for Theorem 1

This section provides the proofs for Theorem 1.

Proof of Theorem 1. By the definition of the duality gap:

$$\begin{aligned} \mathrm{GAP}\left(\boldsymbol{\pi}_{GA}^*\right) &= \max_{s\in\mathcal{S},\pi_1,\pi_2}\left(V^{\pi_1,\pi_{GA,2}^*}(s) - V^{\pi_{GA,1}^*,\pi_2}(s)\right) \\ &\leq \max_{s\in\mathcal{S},\pi_1}\left(V^{\pi_1,\pi_{GA,2}^*}(s) - V^{\boldsymbol{\pi}_{GA}^*}(s)\right) \\ &\quad + \max_{s\in\mathcal{S},\pi_2}\left(V^{\boldsymbol{\pi}_{GA}^*}(s) - V^{\pi_{GA,1}^*,\pi_2}(s)\right). \end{aligned} \tag{3}$$

Hence, it is sufficient to derive the upper bound on $\max_{\pi_i} V^{\pi_1,\pi_{GA,2}^*}(s) - V^{\boldsymbol{\pi}_{GA}^*}(s)$ and $V^{\boldsymbol{\pi}_{GA}^*}(s) - \max_{\pi_i} V^{\pi_{GA,1}^*,\pi_2}(s)$ for each $s \in \mathcal{S}$. Here, letting $\pi_1^\dagger$ be an optimal policy against $\pi_{GA,2}^*$ in the ground TZMG, we obtain the following result on the performance difference between $\pi_1^\dagger$ and $\pi_{GA,1}^*$ against $\pi_{GA,2}^*$. The proof for Lemma 1 is provided in Appendix A.1.

Lemma 1. *Assume that the aggregation function ϕ satisfies the following condition for some non-negative constant $\delta \geq 0$:*

$$\forall s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A},\ \left|Q_A^{\boldsymbol{\pi}_A^*}(\phi(s), \boldsymbol{a}) - Q^{\pi_1^\dagger,\pi_{GA,2}^*}(s, \boldsymbol{a})\right| \leq \delta.$$

Then, we have for any $s \in \mathcal{S}$:

$$V^{\pi_1^\dagger,\pi_{GA,2}^*}(s) - V^{\boldsymbol{\pi}_{GA}^*}(s) \leq \frac{2\delta}{1-\gamma}.$$

Next, we show that the assumption in Lemma 1 holds with some constant δ. By the triangle inequality, the difference between $Q_A^{\boldsymbol{\pi}_A^*}(\phi^{Q^*}(s), \boldsymbol{a})$ and $Q^{\pi_1^\dagger,\pi_{GA,2}^*}(s, \boldsymbol{a})$ can be bound as follows:

$$\begin{aligned} &\left|Q_A^{\boldsymbol{\pi}_A^*}(\phi^{Q^*}(s), \boldsymbol{a}) - Q^{\pi_1^\dagger,\pi_{GA,2}^*}(s, \boldsymbol{a})\right| \\ &\leq \left|Q_A^{\boldsymbol{\pi}_A^*}(\phi^{Q^*}(s), \boldsymbol{a}) - Q^*(s, \boldsymbol{a})\right| \\ &\qquad + \left|Q^*(s, \boldsymbol{a}) - Q^{\pi_1^\dagger,\pi_{GA,2}^*}(s, \boldsymbol{a})\right|. \end{aligned} \tag{4}$$

The first term in (4) means the difference between the minimax state-action values in the abstract TZMG $\mathcal{M}_A$ and ground TZMG $\mathcal{M}$. The second term in (4) represents the performance gap between $\boldsymbol{\pi}^*$ and $(\pi_1^\dagger, \pi_{GA,2}^*)$ in $\mathcal{M}$. Each term can be bounded as follows:

Lemma 2. *In the same setup of Theorem 1, we have for any $s \in \mathcal{S}$ and $\boldsymbol{a} \in \mathcal{A}$:*

$$\left|Q_A^{\boldsymbol{\pi}_A^*}(\phi^{Q^*}(s), \boldsymbol{a}) - Q^*(s, \boldsymbol{a})\right| \leq \frac{\epsilon}{1-\gamma}.$$

Lemma 3. *In the same setup of Theorem 1, we have for any $s \in \mathcal{S}$ and $\boldsymbol{a} \in \mathcal{A}$:*

$$\left|Q^*(s, \boldsymbol{a}) - Q^{\pi_1^\dagger,\pi_{GA,2}^*}(s, \boldsymbol{a})\right| \leq \frac{2\epsilon}{(1-\gamma)^2}.$$

By combining (4) and Lemmas 2 and 3, we get for any $s \in \mathcal{S}$ and $\boldsymbol{a} \in \mathcal{A}$:

$$\left|Q_A^{\boldsymbol{\pi}_A^*}(\phi^{Q^*}(s), \boldsymbol{a}) - Q^{\pi_1^\dagger,\pi_{GA,2}^*}(s, \boldsymbol{a})\right| \leq \frac{3\epsilon}{(1-\gamma)^2}.$$

This inequality implies that, under the aggregation function ϕ^{Q^*}, the assumption in Lemma 1 holds with $\delta = \frac{3\epsilon}{(1-\gamma)^2}$. Thus, we can apply Lemma 1, and then obtain the following inequality:

$$V^{\pi_1^\dagger,\pi_{GA,2}^*}(s) - V^{\boldsymbol{\pi}_{GA}^*}(s) \leq \frac{6\epsilon}{(1-\gamma)^3}. \tag{5}$$

By a similar procedure, we can show that:

$$V^{\boldsymbol{\pi}_{GA}^*}(s) - V^{\pi_{GA,1}^*,\pi_2^\dagger}(s) \leq \frac{6\epsilon}{(1-\gamma)^3}. \tag{6}$$

By combining (3), (5), and (6), we can upper bound the duality gap of $\boldsymbol{\pi}_{GA}^*$ as follows:

$$\mathrm{GAP}\left(\boldsymbol{\pi}_{GA}^*\right) \leq \frac{6\epsilon}{(1-\gamma)^3} \times 2 = \frac{12\epsilon}{(1-\gamma)^3}. \quad \square$$

Proof of Lemma 2. From the definition of the state-action value function in the abstract TZMG $\mathcal{M}_A$:

$$\begin{aligned}
&Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s), \boldsymbol{a}) \\
&= R_A(\phi^{Q^*}(s), \boldsymbol{a}) + \gamma \sum_{s'_A \in \mathcal{S}_A} P_A(s'_A | \phi^{Q^*}(s), \boldsymbol{a}) V_A^{\boldsymbol{\pi}^*_A}(s'_A) \\
&= \sum_{g \in G(s)} w(g) \left(R(g, \boldsymbol{a}) + \gamma \sum_{s'_A \in \mathcal{S}_A} \sum_{s' \in G_A(s'_A)} P(s' | g, \boldsymbol{a}) V_A^{\boldsymbol{\pi}^*_A}(s'_A) \right) \\
&= \sum_{g \in G(s)} w(g) \left(R(g, \boldsymbol{a}) + \gamma \sum_{s' \in \mathcal{S}} P(s' | g, \boldsymbol{a}) V_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s')) \right).
\end{aligned}$$

Hence, the difference between $\sum_{g \in G(s)} w(g) Q^*(g, \boldsymbol{a})$ and $Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s), \boldsymbol{a})$ can be bounded as follows:

$$\begin{aligned}
&\sum_{g \in G(s)} w(g) Q^*(g, \boldsymbol{a}) - Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s), \boldsymbol{a}) \\
&= \gamma \sum_{g \in G(s)} w(g) \sum_{s' \in \mathcal{S}} P(s' | g, \boldsymbol{a}) \left(V^*(s') - V_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s')) \right) \\
&= \gamma \sum_{g \in G(s)} w(g) \sum_{s' \in \mathcal{S}} P(s' | g, \boldsymbol{a}) \\
&\quad \cdot \left(\max_{p \in \Delta(A_1)} \min_{a'_2 \in \mathcal{A}_2} \sum_{a'_1 \in \mathcal{A}_1} p(a'_1) Q^*(s', \boldsymbol{a}') \right. \\
&\quad \left. - \max_{p \in \Delta(A_1)} \min_{a'_2 \in \mathcal{A}_2} \sum_{a'_1 \in \mathcal{A}_1} p(a'_1) Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s'), \boldsymbol{a}') \right) \\
&\leq \gamma \max_{(s', \boldsymbol{a}') \in \mathcal{S} \times \mathcal{A}} \left(Q^*(s', \boldsymbol{a}') - Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s'), \boldsymbol{a}') \right), \qquad (7)
\end{aligned}$$

where the second equality follows from the fact that V^* and $V_A^{\boldsymbol{\pi}^*_A}$ are the minimax values in the ground and abstract TZMG, respectively.

On the other hand, under Assumption 1, we have the following lower bound on the state-action value difference:

$$\begin{aligned}
&\sum_{g \in G(s)} w(g) Q^*(g, \boldsymbol{a}) - Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s), \boldsymbol{a}) \\
&\geq \min_{g \in G(s)} Q^*(g, \boldsymbol{a}) - Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s), \boldsymbol{a}) \\
&\geq -\epsilon + Q^*(s, \boldsymbol{a}) - Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s), \boldsymbol{a}). \qquad (8)
\end{aligned}$$

By combining (7) and (8), and then taking the maximum value of both sides, we obtain:

$$\begin{aligned}
&\max_{(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}} \left(Q^*(s, \boldsymbol{a}) - Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s), \boldsymbol{a}) \right) \\
&\leq \epsilon + \gamma \max_{(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}} \left(Q^*(s, \boldsymbol{a}) - Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s), \boldsymbol{a}) \right).
\end{aligned}$$

Rearranging this inequality, we have for any $s \in \mathcal{S}$ and $\boldsymbol{a} \in \mathcal{A}$:

$$\begin{aligned}
&Q^{\pi^*}(s, \boldsymbol{a}) - Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s), \boldsymbol{a}) \\
&\leq \max_{(s', \boldsymbol{a}') \in \mathcal{S} \times \mathcal{A}} \left(Q^{\pi^*}(s', \boldsymbol{a}') - Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s'), \boldsymbol{a}') \right) \leq \frac{\epsilon}{1 - \gamma}.
\end{aligned}$$

By a similar procedure, we can show that:

$$\begin{aligned}
&Q^*(s, \boldsymbol{a}) - Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s), \boldsymbol{a}) \\
&\geq \min_{(s', \boldsymbol{a}') \in \mathcal{S} \times \mathcal{A}} \left(Q^*(s', \boldsymbol{a}') - Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s'), \boldsymbol{a}') \right) \geq -\frac{\epsilon}{1 - \gamma}.
\end{aligned}$$

In summary, we have for any $s \in \mathcal{S}$ and $\boldsymbol{a} \in \mathcal{A}$:

$$\left| Q^*(s, \boldsymbol{a}) - Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s), \boldsymbol{a}) \right| \leq \frac{\epsilon}{1 - \gamma}. \quad \square$$

Proof of Lemma 3. By the definition of the state value function $V_A^{\boldsymbol{\pi}^*_A}$ in the abstract TZMG $\mathcal{M}_A$, we have for any $s \in \mathcal{S}$:

$$V_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s)) = \max_{p \in \Delta(A_1)} \min_{a_2 \in \mathcal{A}_2} \sum_{a_1 \in \mathcal{A}_1} p(a_1) Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s), \boldsymbol{a}).$$

Applying Lemma 2 to this equation, we obtain the following upper bound on $V_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s))$:

$$\begin{aligned}
V_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s)) &\leq \max_{p \in \Delta(A_1)} \min_{a_2 \in \mathcal{A}_2} \sum_{a_1 \in \mathcal{A}_1} p(a_1) Q^*(s, \boldsymbol{a}) + \frac{\epsilon}{1 - \gamma} \\
&= V^*(s) + \frac{\epsilon}{1 - \gamma}.
\end{aligned}$$

Similarly, we can derive the lower bound on $V_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s))$:

$$\begin{aligned}
V_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s)) &\geq \max_{p \in \Delta(A_1)} \min_{a_2 \in \mathcal{A}_2} \sum_{a_1 \in \mathcal{A}_1} p(a_1) Q^*(s, \boldsymbol{a}) - \frac{\epsilon}{1 - \gamma} \\
&= V^*(s) - \frac{\epsilon}{1 - \gamma}.
\end{aligned}$$

Hence, we have for any $s \in \mathcal{S}$:

$$\left| V^*(s) - V_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s)) \right| \leq \frac{\epsilon}{1 - \gamma}.$$

By using this inequality, we obtain for any $s \in \mathcal{S}$ and $\boldsymbol{a} \in \mathcal{A}$,

$$\begin{aligned}
&Q^{\pi_1^\dagger, \pi^*_{GA,2}}(s, \boldsymbol{a}) - Q^*(s, \boldsymbol{a}) \\
&= \gamma \sum_{s' \in \mathcal{S}} P(s' | s, \boldsymbol{a}) \left(V^{\pi_1^\dagger, \pi^*_{GA,2}}(s') - V^*(s') \right) \\
&= \gamma \sum_{s' \in \mathcal{S}} P(s' | s, \boldsymbol{a}) \left(V_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s')) - V^*(s') \right) \\
&\quad + \gamma \sum_{s' \in \mathcal{S}} P(s' | s, \boldsymbol{a}) \left(V^{\pi_1^\dagger, \pi^*_{GA,2}}(s') - V_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s')) \right) \\
&\leq \frac{\gamma \epsilon}{1 - \gamma} \\
&\quad + \gamma \sum_{s' \in \mathcal{S}} P(s' | s, \boldsymbol{a}) \left(V^{\pi_1^\dagger, \pi^*_{GA,2}}(s') - V_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s')) \right), \qquad (9)
\end{aligned}$$

where the first equality follows from the definition of the state-action value function. Here, we can upper bound the term of $V^{\pi_1^\dagger, \pi^*_{GA,2}}(s') - V_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s'))$ as follows:

$$\begin{aligned}
&V^{\pi_1^\dagger, \pi^*_{GA,2}}(s') - V_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s')) \\
&= \max_{a'_1 \in \mathcal{A}_1} \sum_{a'_2 \in \mathcal{A}_2} \pi^*_{GA,2}(a'_2 | s') Q^{\pi_1^\dagger, \pi^*_{GA,2}}(s', \boldsymbol{a}') \\
&\quad - \max_{a'_1 \in \mathcal{A}_1} \sum_{a'_2 \in \mathcal{A}_2} \pi^*_{GA,2}(a'_2 | s') Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s'), \boldsymbol{a}') \\
&\leq \max_{\boldsymbol{a}' \in \mathcal{A}} \left(Q^{\pi_1^\dagger, \pi^*_{GA,2}}(s', \boldsymbol{a}') - Q_A^{\boldsymbol{\pi}^*_A}(\phi^{Q^*}(s'), \boldsymbol{a}') \right). \qquad (10)
\end{aligned}$$

By combining (9), (10), and Lemma 2, we have for any $s \in \mathcal{S}$ and $\boldsymbol{a} \in \mathcal{A}$:

$$Q^{\pi_1^{\dagger},\pi_{GA,2}^{*}}(s,\boldsymbol{a}) - Q^*(s,\boldsymbol{a}) \leq \frac{2\gamma\epsilon}{1-\gamma} + \gamma \max_{(s',\boldsymbol{a}')\in\mathcal{S}\times\mathcal{A}} \left(Q^{\pi_1^{\dagger},\pi_{GA,2}^{*}}(s',\boldsymbol{a}') - Q^*(s',\boldsymbol{a}')\right).$$

Taking the maximum value of both sides:

$$\max_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}} \left(Q^{\pi_1^{\dagger},\pi_{GA,2}^{*}}(s,\boldsymbol{a}) - Q^{\pi^*}(s,\boldsymbol{a})\right) \leq \frac{2\gamma\epsilon}{1-\gamma} + \gamma \max_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}} \left(Q^{\pi_1^{\dagger},\pi_{GA,2}^{*}}(s,\boldsymbol{a}) - Q^{\pi^*}(s,\boldsymbol{a})\right).$$

Rearranging this inequality, we have for any $s \in \mathcal{S}$ and $\boldsymbol{a} \in \mathcal{A}$:

$$Q^{\pi_1^{\dagger},\pi_{GA,2}^{*}}(s,\boldsymbol{a}) - Q^*(s,\boldsymbol{a}) \leq \max_{(s',\boldsymbol{a}')\in\mathcal{S}\times\mathcal{A}} \left(Q^{\pi_1^{\dagger},\pi_{GA,2}^{*}}(s',\boldsymbol{a}') - Q^*(s',\boldsymbol{a}')\right) \leq \frac{2\epsilon}{(1-\gamma)^2}. \tag{11}$$

On the other hand, from the property of minimax state-action value in (1), we have:

$$Q^*(s,\boldsymbol{a}) - Q^{\pi_1^{\dagger},\pi_{GA,2}^{*}}(s,\boldsymbol{a}) \leq 0. \tag{12}$$

By combining (11) with (12), we get for any $s \in \mathcal{S}$ and $\boldsymbol{a} \in \mathcal{A}$:

$$\left|Q^{\pi_1^{\dagger},\pi_{GA,2}^{*}}(s,\boldsymbol{a}) - Q^*(s,\boldsymbol{a})\right| \leq \frac{2\epsilon}{(1-\gamma)^2}. \quad \square$$

5 Experiments

This section demonstrates our state abstraction developed so far in Markov Soccer (Littman 1994; Abe and Kaneko 2021). We here focus only on the minimax values, since the theoretical results on the other criteria in Section 6 rely on the minimax values.

5.1 Markov Soccer

We describe the Markov soccer experiment as 1 vs 1 game on 4 × 5 as shown in Figure 1. Two players "1" and "2" occupy distinct squares of the grid, respectively and the circled player, "1" here, has a "ball," which specifies the states of the game. Figure 1 shows the initial positions of the players. Which player has the ball at the initial turn is determined at uniformly random. In each turn, each player can move to one of the neighboring cells or stand at the place, i.e., their set of actions includes "Up", "Left", "Down", "Right", and "Stand."

After both select their actions, these two moves are executed in random order. When a player tries to move to the cell occupied by the other player, the ball's possession goes to the stationary player, and the positions of both players remain unchanged. Also, if a player's choice lets him or her be out of the pitch, the position of the player remains unchanged. When a player keeping the ball steps into his or her goal (the right side for player 1 and the left side for player 2), the game is over. At the same time, that player scores 1 point and the opponent scores -1 point. The positions of the players and the ball's possession are initialized as shown in Figure 1.

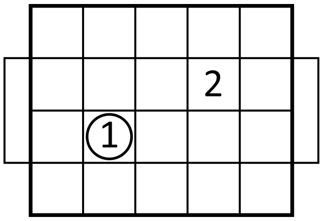

Figure 1: An initial state of the Markov soccer game in which player 1 has the ball.

5.2 Training and Evaluating

In Markov soccer, we build state abstraction, by first solving the game, then greedily aggregating ground states into abstract states that satisfy the Q^* criterion. We calculate the number of states in the abstract Markov game $|\mathcal{S}_A|$, varying ϵ from 0.0 to 2.0 in increments of 0.1. Then, for each ϵ, we compute the equilibrium policies $\boldsymbol{\pi}^*$ and $\boldsymbol{\pi}_A^*$ for the ground and the abstract Markov soccer games, respectively, via the minimax Q-learning. We here assume that the total number of learning iterations T is 1,000,000, the discount factor γ is 0.9, and the learning rate α_t is set to $10^{-\frac{2}{T}t}$ for learning iterations $t \geq 0$. We further evaluate the duality gaps for those equilibrium policies. However, it is demanding to calculate the true one, so we use the approximation obtained from Q-learning.

5.3 Results

We first compute the number of states in the abstract Markov soccer game for different of ϵ in Figure 2. The x- and y-axes represent ϵ and the number of states $|\mathcal{S}_A|$, respectively. We observed that as ϵ increases, the number of states decreases almost linearly. Note that the number of states of the ground Markov soccer is 760, as illustrated at $\epsilon = 0.0$. We observe the value of ϵ up to two, since the difference between the state-action value functions is bound up to 2 from the definition of the game. The number of states of the abstract Markov soccer is 1 at $\epsilon = 2.0$.

Figure 3 illustrates the duality gap in the number of learning iterations where x- and y-axes represent learning iterations and the gap, respectively, varying ϵ. We label "Ground" the gap for the ground Markov soccer game ($\epsilon = 0.0$) and draw the trajectories varying $\epsilon \in \{0.2, 0.6, 1.0, 1.4, 1.8\}$. We observe that the minimax Q-learning approximately solves the ground game, i.e., the gap converges to zero. If ϵ is sufficiently small, i.e., less than 0.6, the duality gap in the abstract Markov soccer game approximates the ground one. Otherwise, the gaps become significantly worse. In these cases, agents often repeat previous actions without progress. Such deadlocks increase the duality gap, as agents can more easily identify best responses by fully exploring the ground game's strategy space.

6 Extentions

We have extended the approximate function that preserves near-optimal behavior by aggregating states on similar optimal Q-values to two-player zero-sum Markov games. This

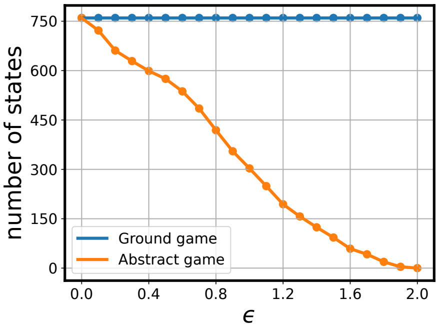

Figure 2: Number of states in the abstract Markov soccer games with respect to ϵ.

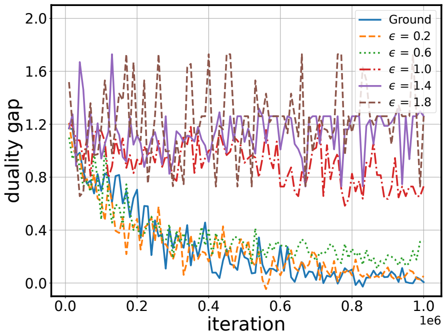

Figure 3: Duality gap at each iteration in minimax Q-learning. Note that the policies are trained in the abstract game, and their duality gap values are computed in the ground game.

section further extends three other types of criteria: *Model Similarity*; *Boltzmann Distribution Similarity*; *Multinomial Distribution similarity*. We then derive the bounds of the gap function for each criterion. Since the theoretical results on these three criteria rely on the one on the minimax Q-values, the proofs are placed in Appendix B-D.

Let us now consider Model Similarity where states are aggregated together when their rewards and transitions are within ϵ (Li, Walsh, and Littman 2006; Abel, Hershkowitz, and Littman 2016). Specifically, when a aggregation function ϕ aggregates states $s_1, s_2 \in \mathcal{S}$, the difference between the probability functions $P(s_1, \boldsymbol{a})$ and $P(s_2, \boldsymbol{a})$ is bounded within $\epsilon \in [0, \infty)$. As well, the difference between the reward functions $R(s_1, \boldsymbol{a})$ and $R(s_2, \boldsymbol{a})$ is bounded within the same error amount.

Assumption 2. *The aggregation function ϕ^{model} satisfies the following property for some non-negative constant $\epsilon \geq 0$: $\phi^{\mathrm{model}}(s_1) = \phi^{\mathrm{model}}(s_2) \Rightarrow$*

$$\max_{\boldsymbol{a} \in \mathcal{A}} |R(s_1, \boldsymbol{a}) - R(s_2, \boldsymbol{a})| \leq \epsilon, \text{ and}$$

$$\max_{s'_A \in \mathcal{S}_A} \left| \sum_{s' \in G_A(s'_A)} \left(P(s'|s_1, \boldsymbol{a}) - P(s'|s_2, \boldsymbol{a})\right) \right| \leq \epsilon.$$

We derive the bound of the duality gap for the Model Similarity criteria as follows.

Theorem 2. *When the ground states are aggregated by the aggregation function ϕ^{model} satisfying Assumption 2 with $\epsilon \geq 0$, then $\boldsymbol{\pi}^*_{GA}$ satisfies:*

$$\mathrm{GAP}\left(\boldsymbol{\pi}^*_{GA}\right) \leq \frac{4(1 + \gamma\left(|\mathcal{S}| - 1\right))\epsilon}{(1-\gamma)^3}.$$

Let us next examine Boltzmann Distribution Similarity from the classic textbook (Sutton and Barto 1998), which balances exploration and exploitation and allows for aggregating states when the ratios of Q-values are similar, but the magnitudes are different.

Assumption 3. *The aggregation function ϕ^{bolt} satisfies the following property for some non-negative constants $\epsilon \geq 0$ and $k \geq 0$: $\phi^{\mathrm{bolt}}(s_1) = \phi^{\mathrm{bolt}}(s_2) \Rightarrow$*

$$\max_{\boldsymbol{a} \in \mathcal{A}} \left| \frac{e^{Q^*(s_1, \boldsymbol{a})}}{\sum_{\boldsymbol{b} \in \mathcal{A}} e^{Q^*(s_1, \boldsymbol{b})}} - \frac{e^{Q^*(s_2, \boldsymbol{a})}}{\sum_{\boldsymbol{b} \in \mathcal{A}} e^{Q^*(s_2, \boldsymbol{b})}} \right| \leq \epsilon, \text{ and}$$

$$\left| \sum_{\boldsymbol{b} \in \mathcal{A}} e^{Q^*(s_1, \boldsymbol{b})} - \sum_{\boldsymbol{b} \in \mathcal{A}} e^{Q^*(s_2, \boldsymbol{b})} \right| \leq k\epsilon.$$

Theorem 3. *When the ground states are aggregated by the aggregation function ϕ^{bolt} satisfying Assumption 3 with $\epsilon \geq 0$ and $k \geq 0$, then $\boldsymbol{\pi}^*_{GA}$ satisfies:*

$$\mathrm{GAP}\left(\boldsymbol{\pi}^*_{GA}\right) \leq \frac{12 e^{\frac{2}{1-\gamma}} \left(|\mathcal{A}_1||\mathcal{A}_2| + k \frac{e^{\frac{1}{1-\gamma}}}{|\mathcal{A}_1||\mathcal{A}_2|} \right) \epsilon}{(1-\gamma)^3}.$$

We finally consider Multinomial Distribution similarity, which is a bit simpler and has close properties to Boltzmann Distribution Similarity.

Assumption 4. *The aggregation function ϕ^{mult} satisfies the following property for some non-negative constants $\epsilon \geq 0$ and $k \geq 0$: $\phi^{\mathrm{mult}}(s_1) = \phi^{\mathrm{mult}}(s_2) \Rightarrow$*

$$\max_{\boldsymbol{a} \in \mathcal{A}} \left| \frac{Q^*(s_1, \boldsymbol{a})}{\sum_{\boldsymbol{b} \in \mathcal{A}} Q^*(s_1, \boldsymbol{b})} - \frac{Q^*(s_2, \boldsymbol{a})}{\sum_{\boldsymbol{b} \in \mathcal{A}} Q^*(s_2, \boldsymbol{b})} \right| \leq \epsilon, \text{ and}$$

$$\left| \sum_{\boldsymbol{b} \in \mathcal{A}} Q^*(s_1, \boldsymbol{b}) - \sum_{\boldsymbol{b} \in \mathcal{A}} Q^*(s_2, \boldsymbol{b}) \right| \leq k\epsilon.$$

Theorem 4. *Suppose that the ground states are aggregated by the aggregation function ϕ^{mult} satisfying Assumption 4 with $\epsilon \geq 0$ and $k \geq 0$. If there exists some positive constant $\delta > 0$ such that $\left|\sum_{\boldsymbol{b} \in \mathcal{A}} Q^*(s, \boldsymbol{b})\right| \geq \delta$ for any states $s \in \mathcal{S}$,*

$$\mathrm{GAP}\left(\boldsymbol{\pi}^*_{GA}\right) \leq \frac{12\left(|\mathcal{A}_1||\mathcal{A}_2| + \frac{k}{\delta}\right)\epsilon}{(1-\gamma)^4}.$$

7 Conclusion

This paper investigates approximate state abstraction, which was originally developed for single-agent MDPs, and extends it for TZMGs, which potentially have many real-world applications. Future works include conducting experiments on games with larger state spaces, such as "The Chasing Game on Gridworld (Wang and Klabjan 2018)" and "Snake Games (Guibas et al. 2022)."

Acknowledgments

We would like to thank Yuki Shimano for useful discussions. Atsushi Iwasaki is supported by JSPS KAKENHI Grant Numbers 21H04890 and 23K17547.

References

Abe, K.; and Kaneko, Y. 2021. Off-Policy Exploitability-Evaluation in Two-Player Zero-Sum Markov Games. In *AAMAS*, 78–87.

Abel, D.; Hershkowitz, D. E.; and Littman, M. L. 2016. Near optimal behavior via approximate state abstraction. In *ICML*, 2915–2923.

Brown, N.; and Sandholm, T. 2018. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374): 418–424.

Burch, N.; Johanson, M.; and Bowling, M. 2014. Solving imperfect information games using decomposition. In *AAAI*, 602–608.

Castro, P. S. 2020. Scalable Methods for Computing State Similarity in Deterministic Markov Decision Processes. In *AAAI*, 10069–10076.

Dadvar, M.; Nayyar, R. K.; and Srivastava, S. 2023. Conditional abstraction trees for sample-efficient reinforcement learning. In *UAI*, 485–495.

Dietterich, T. 1998. The MAXQ Method for Hierarchical Reinforcement Learning. In *ICML*, 118–126.

Dietterich, T. 1999. State abstraction in MAXQ hierarchical reinforcement learning. In *NeurIPS*, 994–1000.

Ferns, N.; Panangaden, P.; and Precup, D. 2004. Metrics for finite Markov decision processes. In *UAI*, 162–169.

Ganzfried, S.; and Sandholm, T. 2013. Action translation in extensive-form games with large action spaces: axioms, paradoxes, and the pseudo-harmonic mapping. In *IJCAI*, 120–128.

Gilpin, A. 2006. A Competitive Texas Hold'em Poker Player via Automated Abstraction and Real-Time Equilibrium Computation. In *AAAI*, 1007–1013.

Gilpin, A.; and Sandholm, T. 2006. Finding equilibria in large sequential games of imperfect information. In *EC*, 160–169.

Gilpin, A.; and Sandholm, T. 2007. Better automated abstraction techniques for imperfect information games, with application to Texas Hold'em poker. In *AAMAS*, 1–8.

Gilpin, A.; Sandholm, T.; and Sørensen, T. B. 2007. Potential-aware automated abstraction of sequential games, and holistic equilibrium analysis of Texas Hold'em poker. In *AAAI*, 50–57.

Givan, R.; Dean, T.; and Greig, M. 2003. Equivalence notions and model minimization in Markov decision processes. *Artificial intelligence*, 147(1–2): 163–223.

Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; and Catanzaro, B. 2022. Efficient Token Mixing for Transformers via Adaptive Fourier Neural Operators. In *ICLR*.

Johanson, M.; Burch, N.; Valenzano, R.; and Bowling, M. 2013. Evaluating state-space abstractions in extensive-form games. In *AAMAS*, 271–278.

Jong, N. K.; and Stone, P. 2005. State abstraction discovery from irrelevant state variables. In *IJCAI*, 752—757.

Jonsson, A.; and Barto, A. G. 2000. Automated state abstraction for options using the U-Tree algorithm. In *NeurIPS*, 1010–1016.

Kroer, C.; and Sandholm, T. 2018. A unified framework for extensive-form game abstraction with bounds. In *NeurIPS*, 613–624.

Li, L.; Walsh, T. J.; and Littman, M. L. 2006. Towards a Unified Theory of State Abstraction for MDPs. In *ISAIM*.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *ICML*, 157–163.

Perolat, J.; Vylder, B. D.; Hennes, D.; Tarassov, E.; Strub, F.; de Boer, V.; Muller, P.; Connor, J. T.; Burch, N.; Anthony, T.; McAleer, S.; Elie, R.; Cen, S. H.; Wang, Z.; Gruslys, A.; Malysheva, A.; Khan, M.; Ozair, S.; Timbers, F.; Pohlen, T.; Eccles, T.; Rowland, M.; Lanctot, M.; Lespiau, J.-B.; Piot, B.; Omidshafiei, S.; Lockhart, E.; Sifre, L.; Beauguerlange, N.; Munos, R.; Silver, D.; Singh, S.; Hassabis, D.; and Tuyls, K. 2022. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623): 990–996.

Ravindran, B.; and Barto, A. G. 2003. SMDP homomorphisms: an algebraic approach to abstraction in semi-Markov decision processes. In *IJCAI*, 1011–1016.

Ravindran, B.; and Barto, A. G. 2004. Approximate homomorphisms: A framework for non-exact minimization in Markov decision processes. In *International Conference on Knowledge Based Computer Systems*, 19–22.

Shalev-Shwartz, S.; Shammah, S.; and Shashua, A. 2016. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*.

Shapley, L. S. 1953. Stochastic Games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T. P.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; and Hassabis, D. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587): 484–489.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T. P.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; and Hassabis, D. 2017. Mastering the game of Go without human knowledge. *Nature*, 550(7676): 354–359.

Sutton, R. S.; and Barto, A. G. 1998. *Reinforcement Learning: An Introduction*. MIT Press.

Szepesvári, C.; and Littman, M. L. 1999. A Unified Analysis of Value-Function-Based Reinforcement-Learning Algorithms. *Neural Computation*, 11(8): 2017–2060.

van der Pol, E.; Kipf, T.; Oliehoek, F. A.; and Welling, M. 2020. Plannable Approximations to MDP Homomorphisms: Equivariance under Actions. In *AAMAS*, 1431–1439.

Wang, X.; and Klabjan, D. 2018. Competitive Multi-agent Inverse Reinforcement Learning with Sub-optimal Demonstrations. In *ICML*, 5143–5151.

Waugh, K. 2013. A fast and optimal hand isomorphism algorithm. In *AAAI Workshop on Computer Poker and Imperfect Information*.