Proceedings of Machine Learning Research 304, 2025 ACML 2025

Policy Iteration for Two-Player General-Sum Stochastic Stackelberg Games

Mikoto Kudo MIKOTO@BBO.CS.TSUKUBA.AC.JP
Youhei Akimoto AKIMOTO@CS.TSUKUBA.AC.JP
University of Tsukuba & RIKEN AIP, Tsukuba, Ibaraki 305-8577, Japan

Editors: Hung-yi Lee and Tongliang Liu

Abstract

We address two-player general-sum stochastic Stackelberg games (SSGs), where the leader's policy is optimized considering the best-response follower whose policy is optimal for its reward under the leader. Existing policy gradient and value iteration approaches for SSGs do not guarantee monotone improvement in the leader's policy under the best-response follower. Consequently, their performance is not guaranteed when their limits are not stationary Stackelberg equilibria (SSEs), which do not necessarily exist. In this paper, we derive a policy improvement theorem for SSGs under the best-response follower and propose a novel policy iteration algorithm that guarantees monotone improvement in the leader's performance. Additionally, we introduce Pareto-optimality as an extended optimality of the SSE and prove that our method converges to the Pareto front when the leader is myopic.
Keywords: Stochastic Stackelberg Games; Policy Iteration; Convergence Analysis.

1. Introduction

A Markov decision process (MDP) is a mathematical framework that models the decision-making of agents in dynamic environments, which serves as the foundation of reinforcement learning (RL). The performance of the agent's policy is evaluated by assessing the state-action sequences it produces with the reward function. This performance is quantified by the expected cumulative discounted reward. An optimal policy is one that maximizes this expected return.

One extension of MDPs for multi-agent systems is a stochastic game, also known as a Markov game, where multiple agents with individual reward functions simultaneously attempt to develop policies that maximize their expected cumulative discounted rewards within a single environment. Stochastic games were initially proposed in game theory, and the solution is defined as a tuple of policies that form a specific type of equilibrium, such as a Nash equilibrium. The problem setting is also characterized by the relationship between the reward functions, such as the zero-sum (i.e., competitive) and fully cooperative settings.

In this study, we addressed the problem of determining stationary Stackelberg equilibria (SSEs) in 2-player stochastic games with general relationships in the rewards. An SSE is defined as a tuple of stationary policies of two types of agents, *leader* and *follower*, such that the leader's policy maximizes the leader's payoff when the follower takes the policy that is the best response to the leader's policy with respect to the follower's payoff. We assume that the follower always takes the best response to any leader's policy. We refer to such a follower as a *best-response follower*. Additionally, we assume that the follower's best responses are always computable to the leader and that the environment is known. No other assumptions are made on the follower's algorithm.

This situation arises, for example, in the design of e-commerce platforms aimed at the maximization of the site owner's profit. The site user (follower) repeatedly makes purchasing decisions based on their preferences. In contrast, the site owner (leader) can configure various elements of the platform, such as page transitions, advertisements, and an incentive bonus. Leveraging knowledge of the user's past behavior, the owner seeks to maximize long-term profit by anticipating the user's responses. Assuming the user's responses are always optimal for their preferences and the owner's prediction is accurate enough, the resulting owner's problem can then be formulated as an SSG with a known best-response follower.

As an SSE does not always exist in the general-sum setting (Bucarey et al., 2022), the algorithm should converge a leader's policy to the SSE if it exists; otherwise, it must converge to a policy that achieves reasonably satisfactory performance. However, existing methods based on dynamic programming operators (Bucarey et al., 2022; Zhang et al., 2020a) and policy gradient methods with total derivatives (Zheng et al., 2022; Vu et al., 2022) do not guarantee the convergence to the SSEs in our setting, even if they exist. Moreover, these algorithms may converge leaders' policies to low-quality ones because they do not guarantee monotone improvement of the leaders' performance under the best-response follower. In this work, we focus on policy improvement methods and develop an algorithm that satisfies the above requirements.

1.1. Main Contributions

Our contributions are summarized as follows. First, we explicitly derive the fixed point of the dynamic programming operator used in Bucarey et al. (2022); Zhang et al. (2020a). This result reveals that the fixed point does not necessarily bring a reasonable leader policy when it is not guaranteed to be an SSE (Section 5). Second, a *policy improvement theorem* for general-sum stochastic Stackelberg games with the best-response follower is derived (Section 6.2), based on which a novel policy iteration algorithm is proposed (Section 6.3). Finally, we introduce the concept of *Pareto-optimality* as an extended optimality of the SSE. Policies that realize the Pareto-optimal value functions or their neighborhood with arbitrary precision always exist. Moreover, Pareto-optimality agrees with the SSE it exists (Section 6.1). Then, we prove that the proposed method monotonically improves the state values toward the Pareto front and converges to the front when the leader is myopic (Section 6.3). To the best of our knowledge, this is the first theoretical guarantee in general-sum myopic-leader SSGs.

2. Preliminaries

2.1. Markov Decision Process

An MDP is a stochastic process with rewards in which the state transitions depend on the actions of an agent. An MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, p, \rho, r, \gamma)$: a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a transition function $p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, an initial state distribution

$\rho \in \Delta(\mathcal{S})$, a bounded reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a discount rate $\gamma \in [0, 1)$, where $\Delta(\mathcal{S})$ is a set of probability distributions on $\mathcal{S}$.

The agent determines its action following its Markov policy. Let $\mathcal{W} := \{f : \mathcal{S} \to \Delta(\mathcal{A})\}$ be a set of (stochastic) decision rules. Then, at the time step $t \in \mathbb{N}$, the agent has the decision rule $f_t \in \mathcal{W}$ and selects the action $a_t \in \mathcal{A}$ in the current state $s_t \in \mathcal{S}$ as $a_t \sim f_t(s_t)$. The conditional probability is denoted as $f_t(a_t|s_t)$. If the decision rule is invariant over time ($f_t = f$), the Markov policy is called a *stationary policy* and is simply denoted by $f \in \mathcal{W}$.

The next state s_{t+1} is stochastically determined as $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$ when the agent selects its action a_t in the state s_t. Repeating such transitions with one stationary policy $f \in \mathcal{W}$ generates the state-action sequence $\{(s_t, a_t)\}_{t=0}^{\infty}$ under f. Then, the performance of f can be defined by evaluating the state-action sequence with the reward function, and the optimal policies are defined as policies with the highest performance. Optimal policies are defined as stationary policies that maximize the expectation of the cumulative discounted reward under $f \in \mathcal{W}_A$ conditioned by an initial state $s \in \mathcal{S}$, which is given by

$$V^f(s) := \mathbb{E}^f \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s \right],$$

for all $s \in \mathcal{S}$, where $\mathbb{E}^f [\cdot|s_0 = s]$ is the expectation over $\{(s_t, a_t)\}_{t=0}^{\infty}$ generated under (f, p) given initial state s. $V^f(s)$ is called a state value function of stationary policy f. Hereafter, we consider only stationary policies and refer to them simply as policies.

There always exists a deterministic optimal policy (Sutton and Barto, 2018) such that

$$f^*(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \mathbb{E}_{s' \sim p(s'|s,a)} \left[\max_{f \in \mathcal{W}} V^f(s') \right] \right\} \tag{1}$$

for all $s \in \mathcal{S}$. The main algorithms used to determine the optimal policy are value iteration and policy iteration when the transition and reward functions are known, and RL when they are unknown (Sutton and Barto, 2018).

2.2. Stochastic Game and Stackelberg Game

Stochastic games (Shapley, 1953; Solan, 2022), also known as Markov games, are an extension of MDPs for multi-agent settings. When there are two agents, A and B, they have finite action spaces $\mathcal{A}$ and $\mathcal{B}$, bounded reward functions $r_A : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ and $r_B : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$, discount rates $\gamma_A \in [0, 1)$ and $\gamma_B \in [0, 1)$ and sets of policies $\mathcal{W}_A := \{f : \mathcal{S} \to \Delta(\mathcal{A})\}$ and $\mathcal{W}_B := \{g : \mathcal{S} \to \Delta(\mathcal{B})\}$, respectively. The state $s \in \mathcal{S}$ is shared among the agents, and the transition function is defined as $p : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$; that is, the transition probability depends on the actions of all the agents.

The expectation of the cumulative discounted reward under stationary policies $f \in \mathcal{W}_A$ and $g \in \mathcal{W}_B$ is defined for r_A and r_B, respectively. Then, equilibrium policies, such as Nash equilibrium, can be identified as solutions to the problem by considering each cumulative reward as the agents' payoff.

The relationship of the reward functions between agents characterizes the game. In game theory, a game where $r_A = -r_B$ is called a zero-sum game, and the others are called non-zero-sum games. Both are called general-sum games. In the context of a multi-agent RL, situations where $r_A \propto -r_B$ and $r_A \propto r_B$ are called competitive and cooperative, respectively.

The Stackelberg game is a game model in which a leader selects its strategy first, and then a follower selects its strategy that maximizes the follower's payoff against the leader's strategy. This follower's strategy is called the *best response* strategy against the leader's strategy. When the leader's strategy maximizes the leader's payoff under the follower's best response, the pair of strategies is called a Stackelberg equilibrium. In this study, we consider the extension of Stackelberg games to stochastic games, such as the stochastic Stackelberg game (SSG). Unlike the single-agent MDP, the optimal policy, *SSE policy* defined below, is not guaranteed to exist, and even if it does, it is not necessarily deterministic. An example scenario where the SSE policy does not exist is provided in the next section.

3. Problem Setting

A general-sum SSG is represented by $(\mathcal{S}, (\mathcal{A}, \mathcal{B}), \hat{p}, \rho, (\hat{r}_A, \hat{r}_B), (\gamma_A, \gamma_B))$, where the subscripts A and B indicate the *leader* and *follower*, respectively. The sets of policies are defined by

$$\mathcal{W}_A := \{f : \mathcal{S} \to \Delta(\mathcal{A})\}, \quad \mathcal{W}_B := \{g : \mathcal{S} \to \Delta(\mathcal{B})\},$$

and the sets of deterministic policies are defined by

$$\mathcal{W}_A^d := \{f : \mathcal{S} \to \mathcal{A}\}, \quad \mathcal{W}_B^d := \{g : \mathcal{S} \to \mathcal{B}\}.$$

Let $\mathcal{F}_\mathcal{S} := \{v : \mathcal{S} \to \mathbb{R}\}$ be the set of real-valued functions with a state as input.

For conciseness, we define the marginalization of $\hat{r}$ and $\hat{p}$ over the leader's action $a \in \mathcal{A}$ under an action distribution $f_s \in \Delta(\mathcal{A})$ as

$$\begin{aligned} r_i(s, f_s, b) &:= \sum_{a \in \mathcal{A}} f_s(a) \hat{r}_i(s, a, b), \quad (\forall i \in \{A, B\}) \\ p(s'|s, f_s, b) &:= \sum_{a \in \mathcal{A}} f_s(a) \hat{p}(s'|s, a, b), \end{aligned}$$

where $s \in \mathcal{S}$ is a state and $b \in \mathcal{B}$ is a follower's action. Hereafter, the players are indexed by $i \in \{A, B\}$.

The state value functions of the leader and the follower are defined as follows.

Definition 1 (State value function) *Let $s \in \mathcal{S}$ be a state and $f \in \mathcal{W}_A, g \in \mathcal{W}_B$ be stationary policies of the leader and the follower. The state value functions $V_i^{fg} \in \mathcal{F}_\mathcal{S}$ of player $i \in \{A, B\}$ for a pair (f, g) are given by*

$$V_i^{fg}(s) := \mathbb{E}^{fg}\left[\sum_{t=0}^{\infty} \gamma_i^t \hat{r}_i(s_t, a_t, b_t) \middle| s_0 = s\right],$$

where $\mathbb{E}^{fg}[\cdot|s_0 = s]$ is the expectation over the state-action sequence $\{(s_t, a_t, b_t)\}_{t=0}^{\infty}$ generated under $(f, g, \hat{p})$ given initial state s.

The follower's best response is defined by employing the state value functions.

Definition 2 (Follower's best response) *Let $s \in \mathcal{S}$ be a state and $f \in \mathcal{W}_A$ a leader's policy. The set of the follower's best responses $R_B^*(f) \subseteq \mathcal{W}_B^d$ against the leader's policy f is defined as*

$$R_B^*(f) := \{g \in \mathcal{W}_B^d \mid g(s) \in R_B^*(s,f)\ \forall s \in \mathcal{S}\},$$

where the best response action at state s is defined as

$$R_B^*(s,f) := \operatorname*{argmax}_{b \in \mathcal{B}} r_B(s, f(s), b) + \gamma_B \mathbb{E}_{s' \sim p(s'|s,f(s),b)} \left[\max_{g \in \mathcal{W}_B} V_B^{fg}(s') \right].$$

As the optimal policies in single-agent MDPs are represented as in Equation (1), we can observe that the set of best responses $R_B^*(f)$ with fixed leader's stationary policy f is the set of the optimal policies in terms of single-agent MDPs. It is also shown that $R_B^*(f)$ is non-empty for all $f \in \mathcal{W}_A$.

We assume that the follower is a *best-response follower*, meaning that it always adopts the best response to the leader's policy. If the leader adopts a stationary policy $f \in \mathcal{W}_A$, the best-response follower adopts the stationary policy $g \in R_B^*(f) \subseteq \mathcal{W}_B^d$. This setup is essentially equivalent to a situation where we can only control the leader's policy, and the follower independently learns its policy with the ability to find the best responses in a finite amount of time for any leader's policy. No other assumptions are made about the follower's learning process.

We assume that the set of the follower's best responses, $R_B^*(s,f)$, is a singleton for each $s \in \mathcal{S}$ and for each $f \in \mathcal{W}_A$. Otherwise, we break ties deterministically. It follows that $R_B^*(f)$ is also a singleton for all $f \in \mathcal{W}_A$; thus, the follower's best response against f is unique. For simplicity, let $R_B^*(f)$ be the best response. This enables us to define a state value function $V_A^{fR_B^*(f)}$ of f under the best-response follower. For simplicity, we denote

$$V_A^{f\dagger} := V_A^{fR_B^*(f)}.$$

We aim to determine an SSE policy, which is defined as the leader's stationary policy that maximizes the leader's state value function for all states under the best-response follower. This policy and the SE value function are defined below.

Definition 3 (SE value function) *Let $s \in \mathcal{S}$ be a state. An SE value function is defined as*

$$V_A^*(s) := \sup_{f \in \mathcal{W}_A} V_A^{f\dagger}(s).$$

Definition 4 (SSE policy) *If a leader's stationary policy $f^* \in \mathcal{W}_A$ satisfies $V_A^{f^*\dagger}(s) = V_A^*(s)$ for all $s \in \mathcal{S}$, then f^* is an SSE policy.*

Game without SSEs We provide an example game where SSEs do not exist. This example is first introduced by Jean-Marie in his seminar talk (Jean-Marie et al., 2022) related to Bucarey et al. (2022). We consider a game with two states $\mathcal{S} = \{s_1, s_2\}$, two leader's actions $\mathcal{A} = \{a_1, a_2\}$, and two follower's actions $\mathcal{B} = \{b_1, b_2\}$. The dynamics and

Table 1: The follower's best response and the leader's value under the best response follower.

Case	$R^*_B(s_1, f), R^*_B(s_2, f)$	$V_A^{f\dagger}(s_1),\, V_A^{f\dagger}(s_2)$
$\gamma_B V^*_B(s_1) > x + \gamma_B V^*_B(s_2)$	b_1, b_2	$\frac{p}{1-\gamma_A}, \frac{p\gamma_A}{1-\gamma_A}$
$\gamma_B V^*_B(s_2) > x + \gamma_B V^*_B(s_1)$	b_2, b_1	$\frac{q\gamma_A}{1-\gamma_A}, \frac{q}{1-\gamma_A}$
$\lvert V^*_B(s_1) - V^*_B(s_2)\rvert < x/\gamma_B$	b_2, b_2	$0, 0$

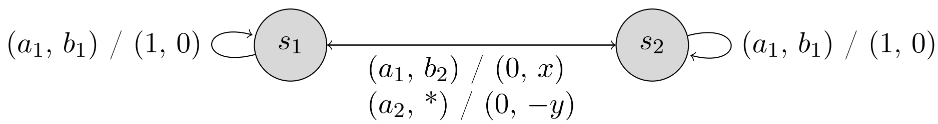

Figure 1: An example of a game without SSEs. Transitions are deterministic and represented by arrows. Each label shows (leader's action, follower's action) / (leader's reward, follower's reward), where $x > 0$ and $y > 0$.

the reward signals are deterministic and represented in Figure 1. Let the leader's policy be denoted as $f_{s_1}(a_1) = p \in [0, 1]$ and $f_{s_2}(a_1) = q \in [0, 1]$.

Given the leader's policy, it reduces to a standard MDP for the follower. The optimal value function for the follower is the solution to

$$V^*_B(s_1) = \max\{\gamma_B V^*_B(s_1), x + \gamma_B V^*_B(s_2)\} \cdot p + (-y + \gamma_B V^*_B(s_2)) \cdot (1 - p);$$
$$V^*_B(s_2) = \max\{\gamma_B V^*_B(s_2), x + \gamma_B V^*_B(s_1)\} \cdot q + (-y + \gamma_B V^*_B(s_1)) \cdot (1 - q).$$

There are three cases depending on the leader's policy f, summarized in Table 1. If $\gamma_B y > x$, we know that $(p, q) = (1, 0)$ leads to the case of $\gamma_B V^*_B(s_1) > x + \gamma_B V^*_B(s_2)$. Then, the maximum value of the leader's value at state s_1 is obtained as $V_A^{f\dagger}(s_1) = \frac{1}{1-\gamma_A}$. Similarly, we can check that $(p, q) = (0, 1)$ leads to the case of $\gamma_B V^*_B(s_2) > x + \gamma_B V^*_B(s_1)$. Then, the maximum value of the leader's value at state s_2 is obtained as $V_A^{f\dagger}(s_2) = \frac{1}{1-\gamma_A}$. We can see that there is a tradeoff between the values at s_1 and s_2 and two Pareto optimal points exist: $(V_A^{f\dagger}(s_1), V_A^{f\dagger}(s_2)) = (\frac{1}{1-\gamma_A}, \frac{\gamma_A}{1-\gamma_A})$ and $(\frac{\gamma_A}{1-\gamma_A}, \frac{1}{1-\gamma_A})$. That is, there does not exist a single optimal policy that maximizes the values at all states simultaneously, implying that an SSE does not exist.

4. Related Work

Bucarey et al. (2022) proposed an algorithm for computing *strong* SSEs in the 2-player general-sum stochastic games, where strong SSEs are introduced to break ties for the follower's best response, which corresponds to a specific tie-breaking mechanism for the follower's best response in our setting. They provide sufficient conditions for their algorithm to converge to the SSEs. However, there is little assurance of the leader's performance of the obtained policy when it does not converge to the SSEs or when there is no SSE. This problem is because the fixed point of the dynamic programming operator used in Bucarey et al. (2022) is different from the SE value function. We discuss this in detail in Section 5.

A primary case where the algorithm of Bucarey et al. (2022) converges and the fixed point coincides with the SE values is when the follower is myopic, i.e., when the follower maximizes its immediate rewards. This setting is adopted in some existing SSG studies. For example, Zhao et al. (2023) derived the upper bound of the regret of the leader's payoff in the cooperative tasks, where the reward function is shared between the leader and the follower, with information asymmetry under the myopic follower. Zhong et al. (2023) proposed an algorithm in the SSG where the group of myopic followers forms a Nash equilibrium among them with analyses on the upper bound of the regret and suboptimality of performance. In contrast to these studies, we focus on the situation where the follower is not myopic.

Zhang et al. (2020a) also considered a similar problem setting, where the reward setting is general-sum and the follower is not myopic. The differences are that the transition function is unknown (i.e., in the RL setting), the follower's policy takes the leader's action as input, and only deterministic stationary policies are considered for the leader's policies. The algorithm proposed by Zhang et al. (2020) shares its principal foundation with that of Bucarey et al. (2022) (see Appendix I). This means that they use the same operator, implying that the limitation of the method of Bucarey et al. (2022) applies to that of Zhang et al. (2020).

Policy gradient methods have been proposed for SSGs with non-myopic followers (Zheng et al., 2022; Vu et al., 2022) based on the implicit function theorem (Fiez et al., 2020). They guarantee the convergence to differential SEs (DSEs), which are subsets of local SEs. However, these methods are not applicable in our setting, except when using the surrogate model of the follower, due to their centralized nature of learning. Moreover, the leader's policy obtained by these algorithms does not necessarily form the DSEs under the best-response follower because the follower's strategies in DSEs are not always the best response.

Several studies concentrate on the cooperative setting. Kao et al. (2022) and Zhao et al. (2023) analyzed the cooperative tasks with information asymmetry, where the leader cannot observe the follower's action (Kao et al., 2022) and the reward functions are known only to the follower (Zhao et al., 2023). Kononen (2004) and Zhang et al. (2020a) point out Pareto efficiency and uniqueness as advantages of Stackelberg equilibria to Nash equilibria in cooperative tasks.

We proposed a novel algorithm for general-sum SSGs with the non-myopic best-response follower under the assumptions that the transition function is known and that the follower's best response is computable in a reasonable time. Unlike the previous methods, iterative methods by the operator (Bucarey et al., 2022; Zhang et al., 2020a) and policy gradient methods (Zheng et al., 2022; Vu et al., 2022), our algorithm guarantees monotone improvement of the leader's state values under the best-response follower. Furthermore, we give a convergence guarantee even when no SSE policy exists.

5. Analysis on DP Operators

Dynamic programming operators are the core of the solutions of single-agent MDPs and the foundation of RL algorithms (Sutton and Barto, 2018). For stochastic games, the operator is extended as an operation of solving a *one-step game* by multiple players, and methods for Nash equilibria are proposed with convergence guarantees (Hu and Wellman, 2003).

For SSGs, however, existing methods based on such a one-step game operator (Bucarey et al., 2022; Zhang et al., 2020a) have problems in terms of the leader's performance in our setting. This section describes it in detail. In summary, we analyze the fixed points of the operator and show that, while the equilibrium formed by the fixed point holds under the best-response follower, the obtained leader's policy may perform poorly when the fixed point is not the SSE, since monotone improvement in the leader's performance is not guaranteed.

The *one-step game* operator $T : \mathcal{F}_\mathcal{S} \times \mathcal{F}_\mathcal{S} \to \mathcal{F}_\mathcal{S} \times \mathcal{F}_\mathcal{S}$ is defined for SSGs as

$$(Tv)_i(s) := r_i(s, R_A(s,v), R_B(s, R_A(s,v), v_B)) \\ + \gamma_i \mathbb{E}_{s' \sim p(s'|s, R_A(s,v), R_B(s, R_A(s,v), v_B))} \left[v_i(s') \right] \quad (i \in \{A, B\}), \quad (2)$$

where $v := (v_A, v_B) \in \mathcal{F}_\mathcal{S} \times \mathcal{F}_\mathcal{S}$,

$$R_A(s,v) := \operatorname*{argmax}_{f_s \in \Delta(\mathcal{A})} r_A(s, f_s, R_B(s, f_s, v_B)) + \gamma_A \mathbb{E}_{s' \sim p(s'|s, f_s, R_B(s, f_s, v_B))} \left[v_A(s') \right],$$

which is assumed to exist for all s and v,[1] and

$$R_B(s, f_s, v_B) := \operatorname*{argmax}_{b \in \mathcal{B}} r_B(s, f_s, b) + \gamma_B \mathbb{E}_{s' \sim p(s'|s, f_s, b)} \left[v_B(s') \right],$$

which is assumed to be a singleton for all s, f_s, and v_B.[2] Equation (2) is regarded as a simplified version of the operator defined in Bucarey et al. (2022). The single update of current values v with T as $v'_i(s) = (Tv)_i(s)$ can be seen as solving a normal-form (i.e., one-step) Stackelberg game for player i at given state $s \in \mathcal{S}$ where the payoff functions are given by current Q functions

$$q_i(s, f_s, b) := r_i(s, f_s, b) + \gamma_i \mathbb{E}_{s' \sim p(s'|s, f_s, b)} \left[v_i(s') \right] \quad (i \in \{A, B\})$$

for the leader's mixed strategy f_s and the follower's pure strategy b. In this context, $R_B(s, f_s, v_B)$ is the follower's best response to f_s, $R_A(s,v)$ is the leader's Stackelberg equilibrium strategy, the pair $(R_A(s,v), R_B(s, R_A(s,v), v_B))$ is the Stackelberg equilibrium, and $((Tv)_A(s), (Tv)_B(s))$ is the equilibrium payoffs of the normal-form game for given s and v.

Bucarey et al. (2022) proposed a method to construct a stationary equilibrium policy corresponding to the fixed point of the operator, which is called *fixed point equilibrium* (*FPE*). If there exists a fixed point $V := (V_A, V_B)$ of T, the pair of the value functions of the FPE $(\bar{f}, \bar{g})$ defined as, for any $s \in \mathcal{S}$,

$$\bar{f}(s) \in R_A(s, V), \quad \bar{g}(s) \in R_B(s, \bar{f}(s), V_B)$$

coincides with V (Bucarey et al., 2022). Since T is a contraction mapping for v_A under certain conditions (Bucarey et al., 2022)[3], we can obtain the fixed point asymptotically by repeatedly applying T with arbitrary initial values.

1. This assumption is necessary to derive a policy corresponding to a value function, and Bucarey et al. (2022) implicitly assumes it. However, it is not guaranteed in general because the inside of argmax is not necessarily continuous with respect to f_s, even though $\Delta(\mathcal{A})$ is compact. Conversely, R_B always exists because $\mathcal{B}$ is finite.
2. Bucarey et al. (2022) defines the *strong* version of R_B similarly to the strong SSE. In this paper, we assume that R_B is a singleton to simplify the discussion, in accordance with our goal of finding the SSE policies. As the core of our analysis does not depend on the uniqueness of R_B, the result can be easily extended to the method of Bucarey et al. (2022).
3. One example is a condition that r_A is an affine transformation of r_B (i.e., cooperative or competitive with r_B). In general, T does not always converge, which is empirically demonstrated for the aforementioned example game (Figure 1) in Jean-Marie et al. (2022).

Let us consider the characteristics of the fixed point (V_A, V_B) of the operator T in general situations where the fixed point exists. Theorem 5 reveals a general property of (V_A, V_B). The proof is in Appendix A. (Appendices are provided as supplementary material.)

Theorem 5 *For $V := (V_A, V_B) \in \mathcal{F}_\mathcal{S} \times \mathcal{F}_\mathcal{S}$, let $R_A(V) \in \mathcal{W}_A$ be a stationary policy whose action distribution on a state $s \in \mathcal{S}$ is in $R_A(s, V)$ and $R_B(f, V_B)$ be a deterministic stationary policy under $f \in \mathcal{W}_A$ whose action on a state s is in $R_B(s, f(s), V_B)$. Then, if $(TV)_i(s) = V_i(s)$ holds for all $i \in \{A, B\}$ and for all $s \in \mathcal{S}$, it holds that*

$$V_A(s) = \max_{f \in \mathcal{W}_A} V_A^{f R_B(f, V_B)}(s) \ \ \forall s \in \mathcal{S}; \quad V_B(s) = \max_{g \in \mathcal{W}_B} V_B^{R_A(V) g}(s) \ \ \forall s \in \mathcal{S}.$$

If V is the fixed point, because $\bar{f} = R_A(V)$ and $\bar{g} = R_B(\bar{f}, V_B)$, it holds that

$$V_B^{\bar{f}\bar{g}}(s) = V_B(s) = \max_{g \in \mathcal{W}_B} V_B^{\bar{f} g}(s), \tag{3}$$

$$V_A^{\bar{f}\bar{g}}(s) = V_A(s) = \max_{f \in \mathcal{W}_A} V_A^{f R_B(f, V_B)}(s) \tag{4}$$

for all $s \in \mathcal{S}$ in light of Theorem 5. Equation (3) shows that the follower's FPE $\bar{g} = R_B(\bar{f}, V_B)$ is the best response to the leader's FPE $\bar{f}$. However, $R_B(f, V_B)$ is not the best response to arbitrary $f \in \mathcal{W}_A$, except when $V_B(s) = \max_{g \in \mathcal{W}_B} V_B^{fg}(s) \ \forall s \in \mathcal{S}$.

The results of Theorem 5 illuminate that the leader's performance V_A of FPEs is the maximum of the leader's state value on $f \in \mathcal{W}_A$ under $R_B(f, V_B)$, but not the maximum under the follower's best responses $R_B^*(f)$ like SE value functions. Therefore, the leader's performance of FPEs at each state can be less than the SE value, which is empirically demonstrated in Bucarey et al. (2019). In addition, it is difficult to determine any meaning from $R_B(f, V_B)$ for arbitrary $f \in \mathcal{W}_A$, which makes the performance of FPEs unpredictable.

Although Bucarey et al. (2022) derives the sufficient condition for SSEs to exist and for FPEs to coincide with the SSEs, this condition is not compatible with our problem setting or can be a strong assumption. The sufficient condition is equivalent to the union of two conditions (Bucarey et al., 2022). The first is that the follower's discount rate $\gamma_B = 0$, which is a condition on the follower's learning algorithm, but we admit no other assumptions on the follower than the best responsivity. The second is that the transition function does not depend on the follower's actions, which can limit the applications of their approach. In the end, finding FPEs does not guarantee the leader's performance in our setting.

To overcome the problem of SSGs that SSEs do not always exist, in the next section, we propose an algorithm that guarantees monotone improvement in the leader's performance toward alternative equilibria, *Pareto-optimal policies*.

6. Pareto-Optimal Policy Iteration

SSE policies do not always exist in general-sum stochastic games. Therefore, it is desired that an algorithm converges to an SSE policy if it exists and to a policy that is "reasonable" provided an SSE policy does not exist. First, we introduce the notion of *Pareto-optimality* as the reasonable target. A policy that approximates a Pareto-optimal value function with an arbitrary precision always exists, and it admits an SSE policy if an SSE policy exists. Then, we derive a *policy improvement theorem* for SSGs under best-response followers. Based on the policy improvement theorem, we design an algorithm that monotonically improves the policy toward Pareto-optimal policies.

Notation We introduce the notation for the dominance relation. Let $v, v' \in \mathcal{F}_{\mathcal{S}}$. If $v(s) = v'(s)$ for all $s \in \mathcal{S}$, we express $v \stackrel{s}{=} v'$. If $v(s) \geq v'(s)$ for all $s \in \mathcal{S}$, then v weakly dominates v' and $v \succcurlyeq v'$. If $v \succcurlyeq v'$ and $v \stackrel{s}{\neq} v'$, v strictly dominates v' and $v \succ v'$.

6.1. Pareto-Optimal Policy

SSE policies exist if and only if there exists $f \in \mathcal{W}_A$ that maximizes $V_A^{f\dagger}(s)$ for all $s \in \mathcal{S}$. We consider $V_A^{f\dagger}(s)$ as $|\mathcal{S}|$ objective functions with $f \in \mathcal{W}_A$ as a common input. From this multi-objective optimization viewpoint, we define the Pareto-optimal policies. Hereafter, we view a value function v as a vector in $\mathbb{R}^{|\mathcal{S}|}$ and apply topological argument in the standard sense in $\mathbb{R}^{|\mathcal{S}|}$.

Definition 6 (Pareto Optimality) *Let $\mathcal{V} = \{v \in \mathcal{F}_{\mathcal{S}} : v \stackrel{s}{=} V_A^{f\dagger}\ \exists f \in \mathcal{W}_A\}$ be the set of a reachable value function. Let $\partial\mathcal{V}$ be the boundary of $\mathcal{V}$ and $\mathrm{cl}\mathcal{V} = \mathcal{V} \cup \partial\mathcal{V}$ be the closure of $\mathcal{V}$. A Pareto-optimal (PO) value function $v^* \in \mathrm{cl}\mathcal{V}$ is such that there exists no $v \in \mathrm{cl}\mathcal{V}$ satisfying $v \succ v^*$. The set of PO value functions is denoted as $\mathcal{PV}$. A stationary policy $f \in \mathcal{W}_A$ satisfying $V_A^{f\dagger} \in \mathcal{PV}$ is a PO policy.*

There always exists $\mathcal{PV} \neq \emptyset$. Moreover, $\mathcal{PV} \subseteq \partial\mathcal{V}$. A PO policy exists if and only if $\mathcal{PV} \cap \mathcal{V} \neq \emptyset$. For any PO value function $v^* \in \mathcal{PV}$, there always exists a policy whose value function is arbitrarily close to v^*. Meanwhile, there is no stationary policy that realizes a value function better than a PO value function. Moreover, an arbitrary PO policy is an SSE policy if there exist SSE policies, as shown in Theorem 7. Therefore, PO value functions are a reasonable alternative to the SE value function. The proof is presented in Appendix B.

Proposition 7 *The SE value function $V_A^* \in \mathcal{PV}$ if and only if $\mathcal{PV}$ is a singleton. Consequently, the SSE policies exist if and only if $\mathcal{PV}$ is a singleton and $\mathcal{PV} \subseteq \mathcal{V}$.*

6.2. Policy Improvement Theorem

We derive three important lemmas based on which we design the proposed approach.

Theorem 8 is an extension of the policy improvement theorem (Sutton and Barto, 2018) in single-agent MDPs. With this theorem, we can confirm if a policy f is better than a baseline f' without computing the value function for f. The proof is shown in Appendix C. The key idea behind its proof is to express the follower's best response in terms of the leader's policy and then apply reasoning analogous to the policy improvement theorem for single-agent MDPs.

Theorem 8 *Let $f \in \mathcal{W}_A, f' \in \mathcal{W}_A$ be stationary policies, and a Q-function $Q_A^{f'\dagger}(s, f) : \mathcal{S} \times \mathcal{W}_A \to \mathbb{R}$ is expressed as*

$$Q_A^{f'\dagger}(s, f) := r_A(s, f(s), R_B^*(s, f)) + \gamma_A \mathbb{E}_{s'} \left[V_A^{f'\dagger}(s') \right],$$

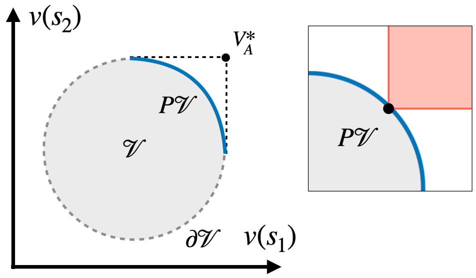

Figure 2: Pareto optimal value functions in the two-state value function space. Thick blue line: Pareto optimal value functions. Red shaded area (the upper right of the dot in the box): the area of value functions dominating the value function represented by the black dot, which is empty for PO value functions.

where $\mathbb{E}_{s'}$ *is considered under* $p(s'|s, f(s), R_B^*(s, f))$. *Then, the following conditions hold:*

$$
\begin{aligned}
&(a) \quad Q_A^{f'\dagger}(\cdot, f) \succcurlyeq V_A^{f'\dagger} \implies V_A^{f\dagger} \succcurlyeq Q_A^{f'\dagger}(\cdot, f) \succcurlyeq V_A^{f'\dagger};\\
&(b) \quad Q_A^{f'\dagger}(\cdot, f) \overset{s}{=} V_A^{f'\dagger} \iff V_A^{f\dagger} \overset{s}{=} V_A^{f'\dagger};\\
&(c) \quad Q_A^{f'\dagger}(\cdot, f) \succ V_A^{f'\dagger} \iff V_A^{f\dagger} - V_A^{f'\dagger} \succ \gamma_A \mathbb{E}_{s'}\left[V_A^{f\dagger}(s') - V_A^{f'\dagger}(s')\right].
\end{aligned}
$$

A necessary condition for PO policies is given below, which shows that no further performance improvement can be made by Theorem 8 (a) from PO policies. This proof is in Appendix D.

Theorem 9 (Necessary condition for PO policies) *If* $f' \in \mathcal{W}_A$ *is a PO policy, it holds that for any* $f \in \mathcal{W}_A$

$$Q_A^{f'\dagger}(\cdot, f) \succcurlyeq V_A^{f'\dagger} \implies V_A^{f\dagger} \overset{s}{=} V_A^{f'\dagger}. \tag{5}$$

The following result shows a sufficient condition for PO policies. This proof is presented in Appendix E.

Theorem 10 (Sufficient condition for PO policies) $f' \in \mathcal{W}_A$ *is a PO policy if, for any* $f \in \mathcal{W}_A$*, either of the following two conditions hold:*

$$
\begin{aligned}
&\text{(i)} \quad Q_A^{f'\dagger}(\cdot, f) \overset{s}{=} V_A^{f'\dagger};\\
&\text{(ii)} \quad \exists s \in \mathcal{S}, Q_A^{f'\dagger}(s, f) < V_A^{f'\dagger}(s) - \frac{\gamma_A}{1-\gamma_A}\delta(f, f'),
\end{aligned}
$$

where $\delta(f, f') := \max_{s \in \mathcal{S}} \mathbb{E}_{s' \sim p(\cdot|s, f(s), R_B^*(s, f))}\left[Q_A^{f'\dagger}(s', f) - V_A^{f'\dagger}(s')\right]$.

When $\gamma_A = 0$, Theorem 10 implies that no further policy update by Theorem 8 (a) with strict improvement is a sufficient condition for PO policies. Combined with Theorem 9, f' is a PO policy if and only if there is no room for policy improvement when $\gamma_A = 0$.

6.3. Policy Iteration for PO Policies

We propose an approach that monotonically improves policy toward PO policies and is guaranteed to converge.

PO value functions are not always unique. To guide the algorithm, we introduce scalarization. Given a stationary policy $f \in \mathcal{W}_A$, we maximize a Pareto-compliant scalarization $\mathcal{L}$ of the state value on f, namely, $\mathcal{L}[V_A^{f\dagger}]$, where the Pareto-compliant scalarization[4] $\mathcal{L} : \mathrm{cl}\mathcal{V} \to \mathbb{R}$ is a continuous function such that $v \succ v' \implies \mathcal{L}[v] > \mathcal{L}[v']$. Then, the maximum of the scalarized value in $\mathrm{cl}\mathcal{V}$ is obtained by a PO value function, namely, $\operatorname{argmax}_{v\in\mathrm{cl}\mathcal{V}} \mathcal{L}[v] \subseteq \mathcal{PV}$.

Directly maximizing $\mathcal{L}[V_A^{f\dagger}]$ is, however, intractable. This is because the computation of $V_A^{f\dagger}$ for each $f \in \mathcal{W}_A$ requires repeatedly applying the Bellman expectation operator under f and $R_B^*(f)$,

$$(T_A^{fR_B^*(f)} V)(s) := r_A(s, f(s), R_B^*(s, f)) + \gamma_A \mathbb{E}_{s'\sim p(s'|s,f(s),R_B^*(s,f))}\left[V(s')\right],$$

to an initial function $V_0 \in \mathcal{F}_\mathcal{S}$ until it converges, where the convergence is guaranteed as in a single-agent MDP.

To alleviate this difficulty and obtain a sequence $\{f_t\}_{t\geq 0}^\infty$ satisfying $\mathcal{L}[V_A^{f_{t+1}\dagger}] \geq \mathcal{L}[V_A^{f_t\dagger}]$, we employ the policy improvement theorem, as in the policy iteration in single-agent MDPs. Rather than computing $V_A^{f\dagger}$ for each f, we calculate $Q_A^{f_t\dagger}(\cdot, f)$, requiring only $V_A^{f_t\dagger}$. Then, we construct the set $\mathcal{W}_{\succcurlyeq}(f_t)$ of policies satisfying the policy improvement condition from f_t, namely,

$$\mathcal{W}_{\succcurlyeq}(f_t) := \left\{ f \in \mathcal{W}_A \mid Q_A^{f_t\dagger}(\cdot, f) \succcurlyeq V_A^{f_t\dagger} \right\}. \tag{6}$$

Selecting the next policy f_{t+1} from $\mathcal{W}_{\succcurlyeq}(f_t)$ guarantees $\mathcal{L}[V_A^{f_{t+1}\dagger}] \geq \mathcal{L}[V_A^{f_t\dagger}]$ in light of Theorem 8 (a). Ideally,[5] the proposed algorithm selects f_{t+1} from the set of policies that maximizes the scalarized value, namely,

$$f_{t+1} \in \mathcal{W}(f_t) := \operatorname*{argmax}_{f\in\mathcal{W}_{\succcurlyeq}(f_t)} \mathcal{L}\left[Q_A^{f_t\dagger}(\cdot, f)\right]. \tag{8}$$

Theorem 11 guarantees that the proposed algorithm always converges and that the limit satisfies the necessary condition for PO policies (i.e., $v_\infty \in \partial\mathcal{V}$). The proof is shown in Appendix F.

4. For example, a weighted sum scalarization $\mathcal{L}[v] = \sum_{s\in\mathcal{S}} \alpha_s v(s)$ is a Pareto-compliant scalarization for $\boldsymbol{\alpha} := \{\alpha_s \in \mathbb{R}_{>0}\}_{s\in\mathcal{S}}$ such that $\sum_{s\in\mathcal{S}} \alpha_s = 1$. If we choose α such that $\alpha_s = \rho(s)$ for all $s \in \mathcal{S}$, then $\mathcal{L}[v]$ coincides with the expected return under the initial state distribution ρ.
5. In practice, argmax is not necessary, and sometimes it does not exist or is not computable. An alternative approach is to select a policy f_{t+1} from a set $\mathcal{W}_{1-\epsilon}(f_t)$ of ϵ-optimal policies in terms of $\mathcal{L}$, such as

$$\begin{aligned}\mathcal{W}_{1-\epsilon}(f_t) := \Bigg\{ f \in \mathcal{W}_{\succcurlyeq}(f_t) \mid \mathcal{L}\left[Q_A^{f_t\dagger}(\cdot, f)\right] - \mathcal{L}\left[V_A^{f_t\dagger}\right] \\ \geq (1-\epsilon) \sup_{f\in\mathcal{W}_{\succcurlyeq}(f_t)} \left(\mathcal{L}\left[Q_A^{f_t\dagger}(\cdot, f)\right] - \mathcal{L}\left[V_A^{f_t\dagger}\right]\right)\Bigg\}\end{aligned} \tag{7}$$

for some $\epsilon \in [0, 1)$. Notably, $\mathcal{W}(f_t) = \mathcal{W}_1(f_t)$. Therefore, the update in Equation (8) is a special case of Equation (7). If it is still intractable, selecting $f_{t+1} \in \mathcal{W}_{\succcurlyeq}(f_t)$ s.t. $Q_A^{f_t\dagger}(\cdot, f_{t+1}) \succ V_A^{f_t\dagger}$ ensures that (a) and (b) in Theorem 11 hold.

Algorithm 1 Pareto-Optimal Policy Iteration

Input: Maximum number of iterations M.

Randomly initialize $f_0 \in \mathcal{W}_A$.
for $t = 0$ to $M - 1$ **do**
  Compute $V_A^{f_t\dagger}$ by repeatedly applying $T_A^{f_t R_B^*(f_t)}$
  Compute $\mathcal{W}_{\succcurlyeq}(f_t)$ in Equation (6)
  if $Q_A^{f_t\dagger}(s, f) \stackrel{s}{=} V_A^{f_t\dagger}$ for all $f \in \mathcal{W}_{\succcurlyeq}(f_t)$ **then**
    return $f^* \leftarrow f_t$
  else
    Compute $\mathcal{W}(f_t)$ in Equation (8) or Equation (7)
    Randomly sample f_{t+1} from $\mathcal{W}(f_t)$
  end if
end for
return $f^* \leftarrow f_M$

Output: A stationary policy f^*

Theorem 11 *Let $f_0 \in \mathcal{W}_A$ be an arbitrary initial policy and a policy sequence $\{f_t \in \mathcal{W}(f_{t-1})\}_{t=1}^{\infty}$ be obtained by Equation (7). Let $\{v_t = V_A^{f_t\dagger}\}_{t=0}^{\infty}$ be the corresponding value functions. Then, the following statements hold:*

(a) $\{v_t\}_{t=0}^{\infty}$ monotonically increases in the sense that $v_{t+1} \succcurlyeq v_t$ and converges to $v_\infty \stackrel{s}{=} \lim_{t\to\infty} v_t$;

(b) $v_{t+1} \stackrel{s}{=} v_t$ if and only if $v_t \stackrel{s}{=} v_\infty$;

(c) $v_\infty \in \partial\mathcal{V}$ and $\min_{s\in\mathcal{S}} v(s) - v_\infty(s) \leq \gamma_A \left(\max_{s\in\mathcal{S}} v(s) - v_\infty(s)\right)$ for all $v \in \mathcal{V}$;

(d) $v_\infty \in \mathcal{PV}$ if $\gamma_A = 0$.

As shown in the proof (Appendix F), the algorithm ceases to improve if and only if $Q_A^{f_t\dagger}(\cdot, f) \stackrel{s}{=} V_A^{f_t\dagger}$ for all $f \in \mathcal{W}_{\succcurlyeq}(f_t)$. The algorithm terminates when it holds and returns f_t. The entire proposed algorithm is shown in Algorithm 1.

The convergence guarantee and the monotone improvement of Algorithm 1 (Statements (a) and (b) in Theorem 11) show the advantages of the proposed approach. However, compared with the policy iteration approach in single-agent MDPs, where the convergence to the optimal policy is guaranteed, the fixed point of Algorithm 1 is not assured to be PO value functions. Rather, Theorem 11 (c) guarantees that the limit is at the boundary $\partial\mathcal{V}$, including all the PO value functions. In a special case of $\gamma_A = 0$, the limit is guaranteed to be a PO value function (Theorem 11 (d)). The visualization of (c) and (d) is in Figure 3.

This condition, the leader's discount rate $\gamma_A = 0$, is a condition on the leader's learning process, meaning Algorithm 1 does not need any assumptions on the follower's learning process to guarantee the convergence and the leader's "reasonable" performance. This property allows us to apply this algorithm to various applications modeled by decentralized SSGs in which we are the leader. We give an example of such applications in Appendix J.

There is another algorithmic difference from the policy iteration for the single-agent MDP. In single-agent MDPs, a state-wise maximization of the action value function of f_t, $a_s \in \operatorname{argmax}_{a\in\mathcal{A}} Q^{f_t}(s, a)$ is performed to construct the next (deterministic) policy. This is

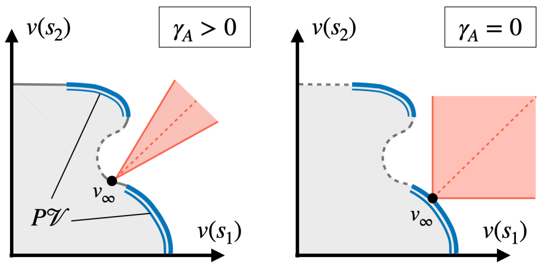

Figure 3: Visualization of Theorem 11 (c) and (d). For v_∞, there must not exist a value function in the red-shaded cone representing the area of v violating the inequality in (c). The upper and lower slopes of the cone are $1/\gamma_A$ and γ_A, respectively. v_∞ can be on the non-Pareto boundary (grey solid line) when $\gamma_A > 0$ (left), but v_∞ is on $\mathcal{PV}$ when $\gamma_A = 0$ (right) because the cone equals the area of dominating value functions.

invalid in SSGs because such a policy does not necessarily satisfy the condition of Theorem 8 (a) and may not improve the values. Moreover, the optimal policy is not necessarily a deterministic one. Because of these differences, Equation (8) cannot be simplified to the state-wise maximization. This can be a limitation from a practical viewpoint when Equation (8) (or Equation (7)) is intractable due to the large search space of $\mathcal{W}_A$. We propose a practical strategy of splitting the policy space to find the next policy efficiently, avoiding exhaustive search over the entire space $\mathcal{W}_A$. Its details are provided in Appendix G

7. Limitation and Future Work

This study proposed a novel algorithm with a convergence guarantee in two-player general-sum SSGs under best-response followers. We introduced the notion of the Pareto-optimal value function to target it even if there are no SSEs, and developed an algorithm that monotonically improves the policy toward the Pareto front. While existing methods have little guarantee under best-response followers, especially in games where the SSEs do not exist, our proposed approach can monotonically improve the leader's performance and guarantees its limit to be Pareto optimal in state values when the leader is myopic. To the best of our knowledge, this is the first theoretical guarantee in general-sum myopic-leader SSGs.

However, there is room for improvement both from theoretical and practical viewpoints. Further research must focus on addressing the limitations of the current work listed below. First, the Pareto-optimality of the proposed algorithm is not guaranteed. A necessary and sufficient condition for the Pareto-optimality of the algorithm must be derived. The development of a restart strategy to satisfy such a sufficient condition is desired. Second, a computationally efficient update rule to select the next policy (cf. Equation (8) or Equation (7)) is required. Practically, a state-wise update similar to the policy iteration in single-agent MDPs is desired. Finally, our approach requires knowledge of the follower's best response and transition probability, as in the previous studies. A sample approximation of the algorithm by reinforcement learning is desired to widen the applicability of the proposed approach.

References

Vahid Behzadan and Arslan Munir. Vulnerability of deep reinforcement learning to policy induction attacks. In *Machine Learning and Data Mining in Pattern Recognition*, pages 262–275. Springer International Publishing, 2017.

Víctor Bucarey, Eugenio Della Vecchia, Alain Jean-Marie, and Fernando Ordóñez. Stationary strong stackelberg equilibrium in discounted stochastic games. Technical Report RR-9271, INRIA, May 2019.

Víctor Bucarey, Eugenio Della Vecchia, Alain Jean-Marie, and Fernando Ordoñez. Stationary strong stackelberg equilibrium in discounted stochastic games. *IEEE Trans. Automat. Contr.*, 68(9):5271–5286, November 2022.

Tanner Fiez, Benjamin Chasnov, and Lillian Ratliff. Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In Hal Daumé Iii and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, pages 3133–3144. 2020.

Ken Harada, Jun Sakuma, and Shigenobu Kobayashi. Local search for multiobjective function optimization: pareto descent method. In *Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation*, GECCO '06, page 659–666, New York, NY, USA, 2006. Association for Computing Machinery.

Junling Hu and Michael P. Wellman. Nash q-learning for general-sum stochastic games. *J. Mach. Learn. Res.*, 4:1039–1069, dec 2003. ISSN 1532-4435.

Yunhan Huang and Quanyan Zhu. Deceptive reinforcement learning under adversarial manipulations on cost signals. In *Decision and Game Theory for Security*, pages 217–237. Springer International Publishing, 2019.

Alain Jean-Marie, Víctor Bucarey López, Eugenio Della Vecchia, and Fernando Ordoñez. Stationary strong stackelberg equilibrium in discounted stochastic games. Dynamic Games and Applications Seminar, October 2022. URL https://www-sop.inria.fr/members/Alain.Jean-Marie/Talks/13Oct2022/Presentation_GERAD_2022.pdf.

Hsu Kao, Chen-Yu Wei, and Vijay Subramanian. Decentralized cooperative reinforcement learning with hierarchical information structure. In Sanjoy Dasgupta and Nika Haghtalab, editors, *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, pages 573–605. 2022.

V Kononen. Asymmetric multiagent reinforcement learning. In *IEEE/WIC International Conference on Intelligent Agent Technology, 2003. IAT 2003*, volume 2, pages 105–121. IEEE Comput. Soc, 2004.

Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7974–7984. 2020.

L S Shapley. Stochastic games. *Proc. Natl. Acad. Sci. U. S. A.*, 39(10):1095–1100, October 1953.

Eilon Solan. *A Course in Stochastic Game Theory*. Cambridge University Press, May 2022.

Yanchao Sun, Da Huo, and Furong Huang. Vulnerability-Aware poisoning mechanism for online RL with unknown dynamics. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, September 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning, second edition: An Introduction*. MIT Press, November 2018.

Quoc-Liem Vu, Zane Alumbaugh, Ryan Ching, Quanchen Ding, Arnav Mahajan, Benjamin Chasnov, Sam Burden, and Lillian J Ratliff. Stackelberg policy gradient: Evaluating the performance of leaders and followers. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*, 2022.

Haifeng Zhang, Weizhe Chen, Zeren Huang, Minne Li, Yaodong Yang, Weinan Zhang, and Jun Wang. Bi-level actor-critic for multi-agent coordination. *Proc. Conf. AAAI Artif. Intell.*, 34(05):7325–7332, April 2020a.

Haoqi Zhang and David Parkes. Value-based policy teaching with active indirect elicitation. https://dash.harvard.edu/bitstream/handle/1/4039771/Zhang_Value.pdf?sequence=2&isAllowed=y, 2008. Accessed: 2023-2-6.

Haoqi Zhang, David C Parkes, and Yiling Chen. Policy teaching through reward function learning. In *Proceedings of the 10th ACM conference on Electronic commerce*, EC '09, pages 295–304, New York, NY, USA, July 2009. Association for Computing Machinery.

Xuezhou Zhang, Yuzhe Ma, Adish Singla, and Xiaojin Zhu. Adaptive Reward-Poisoning attacks against reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11225–11234. 2020b.

Geng Zhao, Banghua Zhu, Jiantao Jiao, and Michael Jordan. Online learning in stackelberg games with an omniscient follower. In *Proceedings of the 40th International Conference on Machine Learning*, pages 42304–42316. 2023.

Liyuan Zheng, Tanner Fiez, Zane Alumbaugh, Benjamin Chasnov, and Lillian J Ratliff. Stackelberg actor-critic: Game-theoretic reinforcement learning algorithms. *Proc. Conf. AAAI Artif. Intell.*, 36(8):9217–9224, June 2022.

Han Zhong, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Can reinforcement learning find Stackelberg-Nash equilibria in General-Sum markov games with myopically rational followers? *J. Mach. Learn. Res.*, 24(35):1–52, 2023.