Optimal control of robust team stochastic games

Optimal control of robust team stochastic games

Feng Huang, Ming Cao, and Long Wang

***Abstract*—In stochastic dynamic environments, team stochastic games have emerged as a versatile paradigm for studying sequential decision-making problems of fully cooperative multi-agent systems. However, the optimality of the derived policies is usually sensitive to the model parameters, which are typically unknown and required to be estimated from noisy data in practice. To mitigate the sensitivity of the optimal policy to these uncertain parameters, in this paper, we propose a model of "robust" team stochastic games, where players utilize a *robust* optimization approach to make decisions. This model extends team stochastic games to the scenario of incomplete information and meanwhile provides an alternative solution concept of robust team optimality. To seek such a solution, we develop a learning algorithm in the form of a Gauss-Seidel modified policy iteration and prove its convergence. This algorithm, compared with robust dynamic programming, not only possesses a faster convergence rate, but also allows for using approximation calculations to alleviate the curse of dimensionality. Moreover, some numerical simulations are presented to demonstrate the effectiveness of the algorithm by generalizing the game model of social dilemmas to sequential robust scenarios.**

***Index Terms*—Optimal control, team stochastic games, multi-agent reinforcement learning, sequential robust social dilemmas.**

I. INTRODUCTION

STOCHASTIC games [1], also known as Markov games [2], as a general framework of multi-agent reinforcement learning (MARL) have long been an active research topic across the fields of stochastic optimal control, operations research, and artificial intelligence (AI) [3], [4], [5]. Especially, it has attracted much attention in recent years due to some ground-breaking advances made by the MARL in conjunction with deep neural networks in achieving human-level performance in several long-standing sequential decision-making conundrums [6]. Compared with the normal-form matrix games [7], [8] and evolutionary games [9], [10], one of its most distinctive features is the introduction of game-environmental states. In each stage of the play, the game is in an environmental state taken from a given set, and each player then relies on its decision rule as a function of the current game-environmental state to choose actions. The collection of players' actions, together with the current environmental state, subsequently determine the probability distribution of the state that the play will visit at the next stage, and as the consequence of the actions and the game-environmental transitions every player will receive an immediate payoff. Therefore, stochastic games are usually deemed as the generalization of one-shot matrix games to the dynamic multi-stage and multi-state situations where the game-environmental state changes in response to the players' decisions [11]. Also, they extend the Markov decision processes (MDPs) [12] from merely involving a single agent to the setup involving multiple competitive decision-makers.

This work was supported by the National Natural Science Foundation of China (grant no. 62036002). F.H. acknowledges the support from China Scholarship Council (grant no. 201906010075). M.C. was supported in part by the European Research Council (grant no. ERC-CoG-771687) and The Netherlands Organization for Scientific Research (grant no. NWO-vidi-14134). (Corresponding authors: Long Wang and Ming Cao)

F. Huang is with the Center for Systems and Control, College of Engineering, Peking University, Beijing 100871, China and ENTEG, Faculty of Science and Engineering, University of Groningen, Groningen 9747 AG, The Netherlands (e-mail: fenghuang@pku.edu.cn).

M. Cao is with ENTEG, Faculty of Science and Engineering, University of Groningen, The Netherlands (e-mail: m.cao@rug.nl).

L. Wang is with the Center for Systems and Control, College of Engineering, Peking University, Beijing 100871, China (e-mail: longwang@pku.edu.cn).

In contrast to the non-cooperative setting of stochastic games, team stochastic games [13] are a fully cooperative multi-agent system in which a group of players works together, through interactions, coordinations, or information-sharing, to jointly accomplish a task or achieve a common goal [14], [15]. Hence, it is formally defined as the stochastic games where all players have a common payoff function [4], [5]. In particular, such a game model has recently sparked increasing research interest across various disciplines due to its extensive real-world applications, such as multi-robot systems, unmanned aerial vehicles, and communication networks (see [4], [5], [15], [16] for an overview), and its close connection to the theory of team decisions [17], [18], [19]. However, most of the existing work is based on the complete information setting where the structural information of the game, such as players' payoff functions and/or the transition probabilities of the game-environmental states, is the common prior knowledge to all players. Under such an assumption, the notion of the optimal Nash equilibrium [13], i.e. the joint decision rule achieving the maximal expected discounted sum of the gains received by the team, naturally provides a plausible solution to predicting the outcome of the game evolution. This is because at this equilibrium state, no player can improve the long-term expected return of the team by unilaterally deviating from its policy. Although there are some debates about the predictive or prescriptive role of the equilibrium concept in practice (see [20] and its commentaries), such Nash equilibria have played a central role in the corresponding studies [21], [22], [23].

While it has been proven that the Nash equilibrium solution always exists for the stochastic games with complete information [24], in the real world, the problem related to the parameter uncertainty of games is pervasive. For example, in many realistic applications associated with reinforcement learning, the payoff functions of agents are usually required to be designed or learned from interactions, and their properties strongly affect the success of targeted tasks [25]; also, the transition probability distribution of the game-environmental states is generally estimated from historical data, thereby influenced by the statistical errors [26], [27]. In particular, such an issue of data uncertainty has given rise to well-grounded concerns recently, such as in the field of AI safety [28], [29] and uncertain robotic systems [30], and accordingly it prompts the research priorities of robust AI [31]. In game theory, on the other hand, this problem has a long research tradition, and corresponds to the games with incomplete information [8]. The first general analytical framework for analyzing the games with uncertain parameters is from the seminal three-part essay by Harsanyi [32] where a new model named "Bayesian games" is constructed and the notion of Nash equilibrium is extended to the incomplete information scenario, termed "Bayesian equilibrium." By relaxing the common prior knowledge assumptions of Harsanyi's model, i.e. the distribution information of all uncertain parameters is available for players, the *ex post* equilibrium [33] provides a distribution-free alternative as a refinement of the Bayesian equilibrium. Different from these two methods, utilizing the theory of robust optimization [34], Aghassi and Bertsimas open a new avenue to robust games [35] where players make action choices depending on an expected worst-case payoff in response to the uncertainty of payoff parameters in one-shot matrix games. This robust game model not only relaxes the common knowledge assumption of Bayesian games, but also provides a more general equilibrium concept termed *robust-optimization equilibrium* than that of the *ex post* equilibrium. Using a similar technique, these results are later extended to stochastic games [36] and some deep MARL algorithms have been developed to find the equilibria [37]. Parallel to the game-theoretical research, robust optimization [34] is another domain devoting to coping with the decision-making problems with data uncertainty. To mitigate the sensitivity of the optimal policy to the ambiguity of transition probabilities in MDPs, robust dynamic programming (rDP) [26], [27] offers the first novel theoretical paradigm and, meanwhile, two algorithms, the so-called robust value iteration (rVI) and robust policy iteration (rPI), are proposed to find the robust optimal policy. Inspired by these work, Kaufman and Schaefer [38] then develop a robust version of modified policy iteration (rMPI) [12] as the generalization of these two algorithms. However, one common shortcoming of these algorithms is that the convergence rate may be quite slow when the discount factor in the Markov decision problem is close to one. To alleviate the curse of dimensionality in large-scale MDPs, the algorithm of rDP is also scaled up using function approximation [39].

In this work, we aim to deal with the sequential decision-making problems in the team stochastic games with uncertain transition probabilities, and accordingly we propose a model of robust team stochastic games. This game model relaxes the complete information assumption of team stochastic games and meanwhile provides an alternative solution concept of robust team optimal policy. To seek such a solution, we develop a learning algorithm, which we call Gauss-Seidel robust approximate modified policy iteration (GSraMPI). Compared with rDP [26], this algorithm allows for approximation computations and thus can be applied to large-scale decision-making problems. Furthermore, we present the convergence analysis of the algorithm within mild approximation tolerance and calculate its convergence rates. The results manifest that at a faster convergence rate than rDP and rMPI, this algorithm can effectively converge to the robust team near-optimal policy within a finite number of iterations. To demonstrate the effectiveness of the algorithm, we also generalize the canonical game model of social dilemmas to sequential robust scenarios; and using them as benchmarks, numerical simulations show that, as compared with rDP and rMPI, the GSraMPI algorithm admits substantially less iteration time to find the robust team optimal policy.

Notation: Throughout the paper, we use $\mathbb{R}$ and $\mathbb{N}$ to represent the set of real numbers and the set of non-negative integers, respectively. For n sets $\mathcal{A}^i$, $i = 1, 2, \ldots, n$, let $\ltimes_{i=1}^n \mathcal{A}^i$ be their Cartesian product. For a scalar x, we use $|x|$ to denote its absolute value, while for a set $\mathcal{S} = \{s^1, s^2, ..., s^m\}$, we use $|\mathcal{S}|$ to represent its cardinality. The space of the probability distribution on $\mathcal{S}$ is denoted by $\Delta(\mathcal{S})$, and the set of bounded real valued functions on $\mathcal{S}$ is represented by $\mathcal{V}$. For a vector $v = [v(s^1), v(s^2), \ldots, v(s^m))]^T \in \mathcal{V}$, its vector norm is defined by $\|v\| := \sup_{s\in\mathcal{S}} |v(s)|$, where $v(s)$ represents the component of v corresponding to $s \in \mathcal{S}$ and T represents transpose. As such, $(\mathcal{V}, \|\cdot\|)$ forms a normed linear space, and it is also a Banach space. For a matrix $P = [p_{kl}] \in \mathbb{R}^{m\times m}$ with element p_{kl} in row k and column l, its matrix norm and spectral radius are defined by $\|P\| := \sup_{k\in\{1,2,\ldots,m\}} \sum_{l=1}^m |p_{kl}|$ and $\sigma(P) := \limsup_{t\to\infty} \|P^t\|^{1/t}$, respectively. Alternatively, we also use $P(l \mid k)$ to represent the element of P in row k and column l. We use I and $\mathbf{1}$ to denote the identity matrix and the all-ones vector with appropriate dimensions, respectively. When '$>$', '$\geq$', '$<$', and '$\leq$' are applied to vectors and matrices, they refer to componentwise order.

II. Problem formulation and preliminaries

A. Stochastic games

We consider an n-player discounted stochastic game [1], also called Markov game [2], with the infinite decision horizon $\mathcal{T} := \{1, 2, \ldots\}$. At each epoch $\tau \in \mathcal{T}$ of the play, the game enters an environmental state $s_\tau \in \mathcal{S}$, where $\mathcal{S} := \{s^1, s^2, ..., s^m\}$ is the discrete finite set consisting of the m game-environmental states shared by all players. Depending on the current environmental state s_τ, every player $i \in \mathcal{N}$ is allowed to choose an action a^i from the discrete finite set $\mathcal{A}^i$ of its available actions, where $\mathcal{N} := \{1, 2, \ldots, n\}$ is the finite set of the indices of the n players. The collection of all players' actions then forms an action aggregation (i.e. joint action) $a_\tau := (a_\tau^1, a_\tau^2, \ldots, a_\tau^n) \in \mathcal{A}$, where $\mathcal{A} := \ltimes_{i=1}^n \mathcal{A}^i$ is the joint action set of all players. As the consequence of the joint action a_τ, the game environment will transit from its current state s_τ to the state $s_{\tau+1}$ at the next epoch with probability $p(s_{\tau+1}|s_\tau, a_\tau)$, where $p(\cdot|s, a) : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability function of the game-environmental states, which maps from the current state $s \in \mathcal{S}$ to the probability distribution $p(\cdot|s, a) \in \Delta(\mathcal{S})$, given the joint action a. Subsequently, every player $i \in \mathcal{N}$ receives an immediate payoff $r^i(s_\tau, a_\tau, s_{\tau+1})$ as a function of the current state

$s_\tau \in \mathcal{S}$, joint action $a_\tau \in \mathcal{A}$, and the state $s_{\tau+1} \in \mathcal{S}$ at the next epoch. This process is repeated infinitely.

Without loss of generality, in this paper we assume that both the payoff functions of players and the transition probability functions of the game-environmental states are stationary. That is, for any state joint-action triple $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, $r^i(s, a, s')$ for $\forall i \in \mathcal{N}$ and $p(s'|s, a)$ do not change with time. Moreover, $r^i(s, a, s')$ is assumed to be bounded, i.e. $|r^i(s, a, s')| \leq R_{\max} < \infty$, for $\forall i \in \mathcal{N}$ and $\forall (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.

At each decision epoch, the action choice of each player is determined by the *decision rule* (akin to the role of control law in control systems), which prescribes a scheme of action selection for each game-environmental state. Generally speaking, one can use a decision rule to rely on the historical data, such as those actions and game-environmental states having been visited. However, note that for any initial game-environmental states, there always exists a Markovian (memoryless) counterpart that can give rise to the same long-term performance as the history-dependent decision rule (see Theorem 5.5.3 in [12]). Therefore, we here restrict our attention exclusively to Markovian decision rules. Formally, for a player $i \in \mathcal{N}$, a Markovian randomized (MR) decision rule is defined by $d^i(\cdot|s) : \mathcal{S} \to \Delta(\mathcal{A}^i)$, which maps the current game-environmental state $s \in \mathcal{S}$ into the probability distribution $d^i(\cdot|s) \in \Delta(\mathcal{A}^i)$. Particularly, if $d^i(a^i|s) = 1$ for some $a^i \in \mathcal{A}^i$, it implies that action a^i will be selected with certainty and accordingly the decision rule degenerates to become deterministic. In this case, we refer to it as the Markovian deterministic (MD) decision rule and denote it by $d^i(s) : \mathcal{S} \to \mathcal{A}^i$, which maps the current game-environmental state $s \in \mathcal{S}$ to a specific action $a^i \in \mathcal{A}^i$. Accordingly, let the set of the MR and MD decision rules available to player i be denoted by $\mathcal{D}^i_{\text{MR}} := \{d^i(\cdot|s) \mid d^i(\cdot|s) : \mathcal{S} \to \Delta(\mathcal{A}^i)\}$ and $\mathcal{D}^i_{\text{MD}} := \{d^i(s) \mid d^i(s) : \mathcal{S} \to \mathcal{A}^i\}$, respectively.

Those decision rules adopted by a player at all decision epochs shape into a time sequence, which prescribes the policy or strategy of the player. More specifically, the *policy* of player $i \in \mathcal{N}$ is defined by $\pi^i := (d^i_1, d^i_2, \ldots, d^i_\tau, \ldots)$, where d^i_τ is the decision rule used by player i at the epoch $\tau \in \mathcal{T}$. By convention [12], [40], we only focus on the stationary policy in this paper, i.e. $d^i_\tau = d^i$ for $\forall \tau \in \mathcal{T}$. In this case, we use $(d^i)^\infty$ to represent the policy $\pi^i = (d^i, d^i, \ldots)$ of player $i \in \mathcal{N}$ for brevity, and accordingly we denote its policy space by $\Pi^i_{\text{MR}} := (\mathcal{D}^i_{\text{MR}})^\infty$ if $d^i \in \mathcal{D}^i_{\text{MR}}$ or $\Pi^i_{\text{MD}} := (\mathcal{D}^i_{\text{MD}})^\infty$ if $d^i \in \mathcal{D}^i_{\text{MD}}$. Similarly, let the joint decision rule and the joint policy of all players be denoted by $d := (d^1, d^2, \ldots, d^n)$ and $\pi := (\pi^1, \pi^2, \ldots, \pi^n) = (d)^\infty$, respectively. Therein, if $d^i \in \mathcal{D}^i_{\text{MR}}$ for $\forall i \in \mathcal{N}$, $d(\cdot|s) : \mathcal{S} \to \Delta(\mathcal{A})$ defines a joint MR decision rule which maps the game-environmental state $s \in \mathcal{S}$ to the probability distribution $d(\cdot|s) \in \Delta(\mathcal{A})$; whereas if $d^i \in \mathcal{D}^i_{\text{MD}}$ for $\forall i \in \mathcal{N}$, $d(s) : \mathcal{S} \to \mathcal{A}$ defines a joint MD decision rule which maps the game-environmental state $s \in \mathcal{S}$ to a specific joint action $a \in \mathcal{A}$. Accordingly, the set of joint decision rules and joint policies is denoted by $\mathcal{D}_{\text{MR}} := \ltimes_{i=1}^n \mathcal{D}^i_{\text{MR}}$ and $\Pi_{\text{MR}} := (\mathcal{D}_{\text{MR}})^\infty$ if $d^i \in \mathcal{D}^i_{\text{MR}}$ for $\forall i \in \mathcal{N}$, or $\mathcal{D}_{\text{MD}} := \ltimes_{i=1}^n \mathcal{D}^i_{\text{MD}}$ and $\Pi_{\text{MD}} := (\mathcal{D}_{\text{MD}})^\infty$ if $d^i \in \mathcal{D}^i_{\text{MD}}$ for $\forall i \in \mathcal{N}$, respectively. To ease the notation, in particular, we adopt the superscript $-i$ to represent the joint quantity in which the index of player i has been removed. As such, the joint decision rule and joint policy can be rewritten as $d = (d^i, d^{-i})$ and $\pi = (\pi^i, \pi^{-i}) = ((d^i)^\infty, (d^{-i})^\infty)$, respectively, where $d^{-i} \in \mathcal{D}^{-i}_{\text{MR}} := \ltimes_{i' \neq i} \mathcal{D}^{i'}_{\text{MR}}$ or $\mathcal{D}^{-i}_{\text{MD}} := \ltimes_{i' \neq i} \mathcal{D}^{i'}_{\text{MD}}$ and $\pi^{-i} \in \Pi^{-i}_{\text{MR}} := (\mathcal{D}^{-i}_{\text{MR}})^\infty$ or $\Pi^{-i}_{\text{MD}} := (\mathcal{D}^{-i}_{\text{MD}})^\infty$.

Given an initial state $s_1 = s \in \mathcal{S}$, when all players adopt a specific joint policy $\pi \in \Pi_{\text{MR}}$ to play the stochastic game, it will induce a probability measure on the trajectory of states and the sequence of joint actions. To evaluate the policy performance of player $i \in \mathcal{N}$ in games, one normally used criterion is to calculate its expected total discounted return:

$$v^i_s(\pi) := \mathbb{E}_\pi \left\{ \sum_{\tau=1}^{\infty} \lambda^{\tau-1} r^i(s_\tau, a_\tau, s_{\tau+1}) \middle| s_1 = s \right\}, \forall s \in \mathcal{S}, \tag{1}$$

where $\mathbb{E}_\pi\{\cdot\}$ represents the expectation over the stochastic process $\{(s_\tau, a_\tau)\}_{\tau \in \mathcal{T}}$ induced by the joint policy π, and $\lambda \in [0, 1)$ is the discount factor, which evaluates the present value of future payoffs. Thus, for player $i \in \mathcal{N}$, its goal is to seek a policy π^i such that this return can be maximized, given the policy π^{-i} of other players. The dominant solution concept of this problem is the (Markov perfect) Nash equilibrium. A joint policy $\pi^* = (\pi^{*^i}, \pi^{*^{-i}}) = ((d^{*^i})^\infty, (d^{*^{-i}})^\infty) \in \Pi_{\text{MR}}$ is said to be a *(Markov perfect) Nash equilibrium* if for $\forall s \in \mathcal{S}$ and $\forall i \in \mathcal{N}$, $v^i_s((d^{*^i})^\infty, (d^{*^{-i}})^\infty) \geq v^i_s((d^i)^\infty, (d^{*^{-i}})^\infty)$ for $\forall d^i \in \mathcal{D}^i_{\text{MR}}$. In particular, it has been proven that this equilibrium solution always exists for multi-player general-sum discounted stochastic games [24].

When all players share an identical payoff function, the aforementioned noncooperative stochastic games reduce to team (or cooperative) stochastic games [4], [13], [15]. It then follows from (1) that every player will attain an identical expected total discounted return, and we denote it by $\bar{v}_s(\pi)$. Since all players have a common objective function in this case, a natural solution concept of the game is the *team-optimal joint policy* $\pi^* = (d^*)^\infty \in \Pi_{\text{MR}}$, which satisfies $\bar{v}_s((d^*)^\infty) \geq \bar{v}_s((d)^\infty)$ for $\forall s \in \mathcal{S}$ and $\forall d \in \mathcal{D}_{\text{MR}}$. Clearly, the team-optimal joint policy is a (Markov perfect) Nash equilibrium, but the converse is not necessarily true.

Besides this identical-payoff based definition, we here consider a slightly more general and appealing alternative based on the team-average payoff. Specifically, we allow different players to have non-identical payoff functions, while their goal is to seek a team-optimal policy such that taking the team-average payoff $r(s, a, s') = \frac{1}{n} \sum_{i=1}^n r^i(s, a, s')$, $\forall (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ as the common payoff function, the expected total discounted return $v_s(\pi)$ of the team is maximal, where $v_s(\pi)$ is calculated by

$$v_s(\pi) = \mathbb{E}_\pi \left\{ \sum_{\tau=1}^{\infty} \lambda^{\tau-1} r(s_\tau, a_\tau, s_{\tau+1}) \middle| s_1 = s \right\}, \forall s \in \mathcal{S}. \tag{2}$$

Compared with the identical-payoff model, this team-average counterpart not only incorporates the former as a special case, but also allows for the heterogeneity of players. Note that for $\forall s \in \mathcal{S}$, one can get an identical optimal $v_s(\pi)$ for both $\pi \in \Pi_{\text{MR}}$ and $\pi \in \Pi_{\text{MD}}$ (see Proposition 6.2.1 in [12]). Thus, we only consider $d \in \mathcal{D}_{\text{MD}}$ in the following. Moreover, to simplify notations, we use $\mathcal{D}$ as the shorthand of $\mathcal{D}_{\text{MD}}$ and Π as the shorthand of Π_{MD} unless otherwise specified.

For ease of the exposition of our main results, here we further introduce some vector and matrix notations, and rewrite (2) in a vector form. Given a joint action $a \in \mathcal{A}$, let $P_a := [p_{kl}]_{|\mathcal{S}|\times|\mathcal{S}|}$ be the transition probability matrix of the game-environmental states, and accordingly let $\mathcal{P} := \{P_a\}_{a\in\mathcal{A}}$ be the collection of those transition probability matrices, where $p_{kl} = p(s^l|s^k, a)$ for $\forall s^k, s^l \in \mathcal{S}$. For a fixed $\mathcal{P}$, moreover, we represent the transition probability matrix induced by the joint decision rule $d \in \mathcal{D}$ by $P_d \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, where the entry of P_d in row k and column l is given by $P_d(l|k) = p(s^l|s^k, d(s^k))$. As such, for the given joint policy $\pi = (d)^\infty$, the probability $Pr(s_\tau = s^l|s_1 = s^k; \pi)$ that starting from the initial state $s^k \in \mathcal{S}$, the game environment transits to state $s^l \in \mathcal{S}$ at the epoch $\tau \in \mathcal{T}$, can be calculated by

$$Pr(s_\tau = s^l|s_1 = s^k; \pi) = (P_d P_d \cdots P_d)(l|k) = (P_d)^{\tau-1}(l|k), \tag{3}$$

where $P_d^0 = I$. Also, given $s \in \mathcal{S}$ and $d \in \mathcal{D}$, one can calculate the expected value $r_{(d,P_d)}(s)$ of the team-average payoff by

$$r_{(d,P_d)}(s) = \sum_{s'\in\mathcal{S}} r(s, d(s), s')p(s'|s, d(s)). \tag{4}$$

From (2)–(4), it follows that $v_{s^k}(\pi) = \sum_{\tau=1}^{\infty}\lambda^{\tau-1}\mathbb{E}_\pi\{r(s_\tau, a_\tau, s_{\tau+1})|s_1 = s^k\} = \sum_{\tau=1}^{\infty}\lambda^{\tau-1}\sum_{l=1}^{m} Pr(s_\tau = s^l|s_1 = s^k; \pi)r_{(d,P_d)}(s^l) = \sum_{\tau=1}^{\infty}\lambda^{\tau-1}\sum_{l=1}^{m}(P_d)^{\tau-1}(l|k)r_{(d,P_d)}(s^l)$ for $\forall s^k \in \mathcal{S}$. Let $v_{(d,P_d)} := [v_{s^1}(\pi), v_{s^2}(\pi), \ldots, v_{s^m}(\pi)]^T$ and $r_{(d,P_d)} := [r_{(d,P_d)}(s^1), r_{(d,P_d)}(s^2), \ldots, r_{(d,P_d)}(s^m)]^T$. Then, the vector form of (2) can be written as $v_{(d,P_d)} = \sum_{\tau=1}^{\infty}\lambda^{\tau-1}{P_d}^{\tau-1}r_{(d,P_d)} = r_{(d,P_d)} + \lambda P_d\left(r_{(d,P_d)} + \lambda P_d r_{(d,P_d)} + \cdots\right) = r_{(d,P_d)} + \lambda P_d v_{(d,P_d)}$. Since $|r^i(s, a, s')| \leq R_{\max} < \infty$ for $\forall i \in \mathcal{N}$ and $\forall (s, a, s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$ according to the assumption and P_d is a row stochastic matrix, $v_s(\pi) \leq R_{\max}/(1-\lambda)$ for $\forall s \in \mathcal{S}$ and $\forall \pi \in \Pi$. It implies that $v_{(d,P_d)} \in \mathcal{V}$ for $\forall d \in \mathcal{D}$ and any P_d. Moreover, note that $\sigma(\lambda P_d) \leq \|\lambda P_d\| = \lambda < 1$. It then follows that $v_{(d,P_d)} = (I - \lambda P_d)^{-1}r_{(d,P_d)} = \sum_{\imath=0}^{\infty}(\lambda P_d)^\imath r_{(d,P_d)}$ is the unique solution to the equation $v = r_{(d,P_d)} + \lambda P_d v$ for $v \in \mathcal{V}$. For the sake of convenience, we call $v_{(d,P_d)}$ the value function afterwards.

B. Robust team stochastic games

The aforementioned model of team stochastic games is based on the assumption of complete information. That is, the structural information of the game, such as payoff functions and transition probabilities, is explicitly known to players. Here, we relax this assumption to the scenario of incomplete information [32], [35], where players do not know the true transition probabilities of the game-environmental states, but rather commonly perceive an uncertainty set of their possible values. More specifically, the transition probability matrix P_a is not predetermined but lies in a discrete finite set $\mathcal{P}_a$ for $\forall a \in \mathcal{A}$. Since the payoff functions of players are dependent on the game-environmental state at the next epoch, the uncertainty of transition probabilities will result in the ambiguity of players' payoffs. Therefore, we refer to a team stochastic game as a *robust team stochastic game* if players' payoffs and/or the transition probabilities of the game-environmental states are uncertain.

By convention, we assume that for $\forall a \in \mathcal{A}$, the uncertainty set $\mathcal{P}_a$ satisfies the so-called "rectangularity" property [26], [27], i.e. $\mathcal{P}_a$ has the form of $\mathcal{P}_a = \ltimes_{k=1}^{m}\mathcal{P}_a(\cdot|k)$ and is independent of historically visited states and actions, where $\mathcal{P}_a(\cdot|k) \subseteq \Delta(\mathcal{S})$ for $\forall k \in \{1, 2, \ldots, m\}$ characterizes the uncertainty of the k-th row of P_a. Then, the admissible set of P_d for a given $d \in \mathcal{D}$ can be represented by $\mathcal{P}_d := \ltimes_{k=1}^{m}\mathcal{P}_{d(s^k)}(\cdot|k)$. In this situation, to systematically mitigate the influence of uncertain transition probabilities on the performance of the optimal policy, players aim to seek a robust (or safe) team-optimal policy $\pi^* = (d^*)^\infty \in \Pi$ such that the value function $v_{(d^*,P^*_{d^*})}$ is maximal with respect to the worst-case $P^*_{d^*}$ in the uncertainty set $\mathcal{P}_{d^*}$.

Definition 1. *A value function $v^* := v_{(d^*,P^*_{d^*})}$ is said to be robust team-optimal if $v_{(d^*,P^*_{d^*})} = \max_{d\in\mathcal{D}}\min_{P_d\in\mathcal{P}_d} v_{(d,P_d)}$, and accordingly $(d^*)^\infty$ is called a robust team-optimal policy.*

It is clear that every robust team-optimal policy is the robust (Markov perfect) Nash equilibrium (see [36] for its definition) of the robust stochastic games, but the converse is not necessarily true. Note that both $\mathcal{D}$ and $\mathcal{P}_d$ for $\forall d \in \mathcal{D}$ are discrete and finite. Therefore, such a robust team-optimal policy always exists [12]. Moreover, in contrast to the robust team-optimal policy, its near-optimal counterpart may be more preferred in practice. A policy $(d^\epsilon)^\infty$ is said to be *ϵ-robust team-optimal* for $\epsilon > 0$ if $\min_{P_{d^\epsilon}\in\mathcal{P}_{d^\epsilon}} v_{(d^\epsilon,P_{d^\epsilon})} \geq v^* - \epsilon\mathbf{1}$.

In the following section, we present our learning algorithm.

III. Learning to seek robust team-optimal policy

In this section, we develop a learning algorithm, named Gauss-Seidel robust approximate modified policy iteration (GSraMPI), to seek the robust team-optimal policy under the assumption of the so-called "classical information pattern" (i.e. all the players fully share information) [41], [42], and meanwhile we prove its convergence to the ϵ-robust team-optimal policy within a finite number of iterations.

A. The GSraMPI Algorithm

The GSraMPI algorithm is shown in Algorithm 1, which primarily encompasses four steps. First, it begins with a value function $v^0 = [v^0_{s^1}, v^0_{s^2}, \ldots, v^0_{s^m}]^T \in \mathcal{V}^0$ at the initial iteration time $t = 0$, where v^0_s is the component of v^0 corresponding to state $s \in \mathcal{S}$ and $\mathcal{V}^0$ is the feasible set of the initial value function. Subsequently, the algorithm calculates an improved policy with respect to an approximate value of $\rho_{(s^k,a)}(v^t)$ by following the process of policy improvement, where $v^t = [v^t_{s^1}, v^t_{s^2}, \ldots, v^t_{s^m}]^T$ is the value function at the

Algorithm 1 Gauss-Seidel robust approximate modified policy iteration (GSraMPI)

Input: $v^0 \in \mathcal{V}^0$, $\epsilon > 0$, $\lambda \in [0,1)$, $\delta : 0 \le \delta < (1-\lambda)^2\epsilon/2\lambda(1+\lambda)$, and $\{M_t\}_{t\in\mathbb{N}}$.

1: Set $t = 0$.

2: (Policy improvement) Set $k = 1$ and go to 2(a).

- **(a)** For $\forall a \in \mathcal{A}$, first compute an approximation $\tilde{\rho}_{(s^k,a)}(v^t)$ of $\rho_{(s^k,a)}(v^t)$ such that
$$|\tilde{\rho}_{(s^k,a)}(v^t) - \rho_{(s^k,a)}(v^t)| \le \lambda\delta; \tag{5}$$
then set
$$u_0^t(s^k) = \max_{a\in\mathcal{A}} \tilde{\rho}_{(s^k,a)}(v^t), \tag{6}$$
and select $d_{t+1}(s^k) \in \arg\max_{a\in\mathcal{A}} \tilde{\rho}_{(s^k,a)}(v^t)$.
- **(b)** If $k = m$, go to 3; otherwise set $k = k+1$ and return to 2(a).

3: (Partial policy evaluation)

- **(a)** If $\|u_0^t - v^t\| < (1-\lambda)\epsilon/2\lambda - \delta$ where $u_0^t := [u_0^t(s^1), u_0^t(s^2), \dots, u_0^t(s^m)]^T$, go to 4; otherwise set $\varsigma = 0$ and go to 3(b).
- **(b)** If $\varsigma = M_t$, go to (f); otherwise set $k = 1$ and go to 3(c).
- **(c)** For $u_\varsigma^t := [u_\varsigma^t(s^1), u_\varsigma^t(s^2), \dots, u_\varsigma^t(s^m)]^T$, compute an approximation $\tilde{\rho}_{(s^k,d_{t+1}(s^k))}(u_\varsigma^t)$ of $\rho_{(s^k,d_{t+1}(s^k))}(u_\varsigma^t)$ such that
$$|\tilde{\rho}_{(s^k,d_{t+1}(s^k))}(u_\varsigma^t) - \rho_{(s^k,d_{t+1}(s^k))}(u_\varsigma^t)| \le \lambda\delta, \tag{7}$$
and set
$$u_{\varsigma+1}^t(s^k) = \tilde{\rho}_{(s^k,d_{t+1}(s^k))}(u_\varsigma^t). \tag{8}$$
- **(d)** If $k = m$, go to 3(e); otherwise set $k = k+1$ and return to 3(c).
- **(e)** Set $\varsigma = \varsigma + 1$ and return to 3(b).
- **(f)** Set $v^{t+1} = u_{M_t}^t$ and $t = t+1$, and return to 2.

4: Set $d^\epsilon(s) = d_{t+1}(s)$ for $\forall s \in \mathcal{S}$ and stop.

Output: ϵ-robust team-optimal policy $(d^\epsilon)^\infty$.

iteration time $t \in \mathbb{N}$ with element v_s^t for state $s \in \mathcal{S}$, and for any given $s^k \in \mathcal{S}$ and $a \in \mathcal{A}$, $\rho_{(s^k,a)}(v^t)$ is calculated by

$$\rho_{(s^k,a)}(v^t) = \min_{p(\cdot|s^k,a)\in\mathcal{P}_a(\cdot|k)} \Bigg\{ \sum_{l=1}^m r(s^k,a,s^l)p(s^l \mid s^k,a) + \lambda \Bigg[\sum_{l<k} p(s^l \mid s^k,a)u_0^t(s^l) + \sum_{l\ge k} p(s^l \mid s^k,a)v_{s^l}^t \Bigg] \Bigg\}. \tag{9}$$

In this process, a parameter $\delta \in [0, (1-\lambda)^2\epsilon/2\lambda(1+\lambda))$ is used to confine the maximal approximation tolerance. Next, the algorithm evaluates an estimation of the value function induced by the improved policy via the process of partial policy evaluation, in which a sequence $\{M_t\}_{t\in\mathbb{N}}$ of non-negative integers is used to limit the iteration times of the evaluation, and $\rho_{(s^k,a')}(u_\varsigma^t)$ for $\forall s^k \in \mathcal{S}$ and $a' = d_{t+1}(s^k)$ is calculated by

$$\rho_{(s^k,a')}(u_\varsigma^t) = \sum_{l=1}^m r(s^k,a',s^l)p^*(s^l \mid s^k,a') + \lambda \Bigg[\sum_{l<k} p^*(s^l \mid s^k,a')u_{\varsigma+1}^t(s^l) + \sum_{l\ge k} p^*(s^l \mid s^k,a')u_\varsigma^t(s^l) \Bigg], \tag{10}$$

where $p^*(\cdot \mid s^k,a')$ is the minimizing transition probability distribution for calculating $\rho_{(s^k,a')}(v^t)$ by (9). Lastly, if the termination condition shown in step 3(a) is satisfied, the algorithm stops and outputs a policy $(d^\epsilon)^\infty$, which will be proven to be the ϵ-robust team-optimal policy. In particular, if $M_t = 0$ for $\forall t \in \mathbb{N}$, $\mathcal{V}^0$ can be set as $\mathcal{V}$; otherwise, $\mathcal{V}^0 = \mathcal{V}_{\mathscr{B}} := \{\nu \in \mathcal{V} \mid \mathscr{B}\nu \ge 0\}$, where $\mathscr{B}$ is an operator defined in (22).

Remark 1. *It is easy for the initialization condition $v^0 \in \mathcal{V}_{\mathscr{B}}$ to be satisfied in practice. For example, selecting $v_s^0 \le (1-\lambda)^{-1}\min_{s^k,s^l\in\mathcal{S},a\in\mathcal{A}} r(s^k,a,s^l)$ for $\forall s \in \mathcal{S}$ will ensure that $v^0 \in \mathcal{V}_{\mathscr{B}}$. Moreover, to implement the approximation computations in Algorithm 1, many state-of-the-art techniques can be used, such as using deep neural networks as function approximation [40] or using empirical samples [43], [44]. In particular, when the number of samples is large enough, it has been proven that a good enough approximate solution can be obtained by sampling-based methods [43].*

B. The convergence analysis

Here, we present the convergence analysis of the GSraMPI algorithm, which includes two parts. The first part shows the convergence of GSraMPI in a degenerated situation, i.e. $M_t = 0$ for $\forall t \in \mathbb{N}$, in which the algorithm reduces to a variant of rVI and we refer to it as Gauss-Seidel robust approximate value iteration (GSraVI). Based on the results in this specific case, the second part establishes the convergence of GSraMPI in the general situation where $\{M_t\}_{t\in\mathbb{N}}$ is an arbitrary sequence of non-negative integers.

1) The degenerated form GSraVI: This section establishes the convergence of GSraMPI when it reduces to GSraVI by setting $M_t = 0$ for $\forall t \in \mathbb{N}$. We first consider the exact case where the tolerance of approximation computations is zero, i.e. $\delta = 0$. In this case, from (5) and (6), the iterative scheme of GSraVI can be given by

$$v_s^{t+1} = \max_{a\in\mathcal{A}} \rho_{(s,a)}(v^t) \text{ for } \forall s \in \mathcal{S}, \tag{11}$$

where $u_0^t(s) = v_s^{t+1}$ is applied to calculate $\rho_{(s,a)}(v^t)$ in view of the setup in step 3(f). Let $\check{d} \in \mathcal{D}$ be the decision rule such that for $\forall s \in \mathcal{S}$, $\check{d}(s) \in \arg\max_{a\in\mathcal{A}} \rho_{(s,a)}(v^t)$, and $P_{\check{d}} \in \mathcal{P}_{\check{d}}$ be the worst-case transition probability matrix with element $P_{\check{d}}(l \mid k)$ in row k and column l when using $\check{d}$ to compute $\rho_{(s,\check{d}(s))}(v^t)$. We then split $P_{\check{d}}$ into $P_{\check{d}} = P_{\check{d}}^L + P_{\check{d}}^U$, where $P_{\check{d}}^L$ is a lower triangular matrix with element $P_{\check{d}}^L(l \mid k) = P_{\check{d}}(l \mid k)$ if $l < k$ and otherwise $P_{\check{d}}^L(l \mid k) = 0$; and $P_{\check{d}}^U$ is an upper triangular matrix with element $P_{\check{d}}^U(l \mid k) = P_{\check{d}}(l \mid k)$ if $l \ge k$ and otherwise $P_{\check{d}}^U(l \mid k) = 0$. As such, from (9), the vector form of the iterative scheme (11) can be given by

$v^{t+1} = r_{(\check{d},P_{\check{d}})} + \lambda\left[P_{\check{d}}^{\ L} v^{t+1} + P_{\check{d}}^{\ U} v^t\right]$. By re-arrangement, it yields

$$\begin{aligned} v^{t+1} &= (\underbrace{I - \lambda P_{\check{d}}^{\ L}}_{Q_{\check{d}}})^{-1} r_{(\check{d},P_{\check{d}})} + (I - \lambda P_{\check{d}}^{\ L})^{-1} \underbrace{\lambda P_{\check{d}}^{\ U}}_{R_{\check{d}}} v^t \\ &= Q_{\check{d}}^{\ -1} r_{(\check{d},P_{\check{d}})} + Q_{\check{d}}^{\ -1} R_{\check{d}} v^t, \end{aligned} \tag{12}$$

where $(I - \lambda P_{\check{d}}^{\ L})^{-1}$ exists because $\sigma(\lambda P_{\check{d}}^{\ L}) \le \|\lambda P_{\check{d}}^{\ L}\| < 1$. It is clear that $(Q_{\check{d}}, R_{\check{d}})$ is a type of the regular splitting of $I - \lambda P_{\check{d}}$ because $Q_{\check{d}} - R_{\check{d}} = I - \lambda P_{\check{d}}$, $Q_{\check{d}}^{\ -1} = \sum_{\imath=0}^{\infty} (\lambda P_{\check{d}}^{\ L})^{\imath} \ge 0$, and $R_{\check{d}} \ge 0$. In the following, we refer to this form of regular splitting as the Gauss-Seidel (GS) regular splitting. From (12), it follows that the vector form of (11) can be given by

$$v^{t+1} = \max_{d\in\mathcal{D}} \min_{P_d\in\mathcal{P}_d} \left\{ Q_d^{-1} r_{(d,P_d)} + Q_d^{-1} R_d v^t \right\}, \tag{13}$$

where (Q_d, R_d) is the GS regular splitting of $I - \lambda P_d$. For any $v \in \mathcal{V}$, let the operator $\mathscr{Y} : \mathcal{V} \to \mathcal{V}$ be defined by

$$\mathscr{Y} v := \max_{d\in\mathcal{D}} \min_{P_d\in\mathcal{P}_d} \left\{ Q_d^{-1} r_{(d,P_d)} + Q_d^{-1} R_d v \right\}. \tag{14}$$

Then, (13) can be rewritten as $v^{t+1} = \mathscr{Y} v^t$. Although the definition of $\mathscr{Y}$ is based on the GS regular splitting, the following lemma will show that $\mathscr{Y}$ is a contraction mapping on $\mathcal{V}$ and the convergence rate of the sequence generated by it is less than 1, as long as (Q_d, R_d) is a regular splitting of $I - \lambda P_d$ under mild conditions.

Lemma 1. *For any given $d \in \mathcal{D}$ and $P_d \in \mathcal{P}_d$, let (Q_d, R_d) be a regular splitting of $I - \lambda P_d$. If (Q_d, R_d) satisfies $\alpha := \sup_{d\in\mathcal{D}, P_d\in\mathcal{P}_d} \|Q_d^{-1} R_d\| < 1$, then (a) the sequence $\{v^t\}$ generated by $v^{t+1} = \mathscr{Y} v^t$ converges in norm to the robust team-optimal value function v^* for $t \to \infty$, and v^* is the unique fixed point of $\mathscr{Y}$; (b) the sequence converges globally at order 1 at a rate no greater than α; its global asymptotic average rate of convergence (AARC) is no greater than α, and it converges globally with $O(\beta^t), \beta \le \alpha$* [1].

Proof: We first prove part (a). For any given $u, v \in \mathcal{V}$, we first consider those states $s \in \mathcal{S}$ satisfying $\mathscr{Y} v(s) - \mathscr{Y} u(s) \ge 0$, where $\mathscr{Y} v(s)$ and $\mathscr{Y} u(s)$ are the components of $\mathscr{Y} v$ and $\mathscr{Y} u$ corresponding to state $s \in \mathcal{S}$, respectively.

Let $d_v \in \arg\max_{d\in\mathcal{D}} \left\{ \min_{P_d\in\mathcal{P}_d} \left[Q_d^{-1} r_{(d,P_d)} + Q_d^{-1} R_d v \right] \right\}$. Then,

$$\begin{aligned} 0 &\le \mathscr{Y} v(s) - \mathscr{Y} u(s) \\ &\le \min_{P_{d_v}\in\mathcal{P}_{d_v}} \left\{ Q_{d_v}^{-1} r_{(d_v,P_{d_v})} + Q_{d_v}^{-1} R_{d_v} v \right\}(s) \\ &\quad - \min_{P_{d_v}\in\mathcal{P}_{d_v}} \left\{ Q_{d_v}^{-1} r_{(d_v,P_{d_v})} + Q_{d_v}^{-1} R_{d_v} u \right\}(s). \\ &\le \left\{ \hat{Q}_{d_v}^{-1} r_{(d_v,\hat{P}_{d_v})} + \hat{Q}_{d_v}^{-1} \hat{R}_{d_v} v \right\}(s) \\ &\quad - \left\{ \hat{Q}_{d_v}^{-1} r_{(d_v,\hat{P}_{d_v})} + \hat{Q}_{d_v}^{-1} \hat{R}_{d_v} u \right\}(s) \\ &= \left\{ \hat{Q}_{d_v}^{-1} \hat{R}_{d_v} (v - u) \right\}(s) \le \|\hat{Q}_{d_v}^{-1} \hat{R}_{d_v} (v - u)\| \\ &\le \|\hat{Q}_{d_v}^{-1} \hat{R}_{d_v}\| \|v - u\| \le \alpha \|v - u\|, \end{aligned} \tag{15}$$

where $\hat{P}_{d_v} \in \arg\min_{P_{d_v}\in\mathcal{P}_{d_v}} \left\{ Q_{d_v}^{-1} r_{(d_v,P_{d_v})} + Q_{d_v}^{-1} R_{d_v} u \right\}$ and $(\hat{Q}_{d_v}, \hat{R}_{d_v})$ is the corresponding regular splitting of $I - \lambda \hat{P}_{d_v}$.

Similarly, applying the same argument for the states $s \in \mathcal{S}$ satisfying $\mathscr{Y} v(s) - \mathscr{Y} u(s) \le 0$, one can obtain $0 \le \mathscr{Y} u(s) - \mathscr{Y} v(s) \le \alpha \|u - v\|$. Consequently, it leads to $\|\mathscr{Y} v - \mathscr{Y} u\| \le \alpha \|v - u\|$. Since $\alpha < 1$, $\mathscr{Y}$ is a contraction mapping on $\mathcal{V}$. From Banach fixed-point theorem (see Theorem 6.2.3 in [12]), it follows that $\{v^t\}$ converges in norm to the unique fixed point $\tilde{v}^*$ of $\mathscr{Y}$.

We now show that $\tilde{v}^* = v^*$. Since $\tilde{v}^*$ is the unique fixed point of $\mathscr{Y}$, for any given $d \in \mathcal{D}$, we have $\tilde{v}^* = \mathscr{Y} \tilde{v}^* \ge \min_{P_d\in\mathcal{P}_d} \left\{ Q_d^{-1} r_{(d,P_d)} + Q_d^{-1} R_d \tilde{v}^* \right\}$. Also, for any $\epsilon > 0$, there exists a $P_d^\epsilon \in \mathcal{P}_d$ such that $\min_{P_d\in\mathcal{P}_d} \left\{ Q_d^{-1} r_{(d,P_d)} + Q_d^{-1} R_d \tilde{v}^* \right\} \ge Q_d^{\epsilon\ -1} r_{(d,P_d^\epsilon)} + Q_d^{\epsilon\ -1} R_d^\epsilon \tilde{v}^* - \epsilon \mathbf{1}$, where $(Q_d^\epsilon, R_d^\epsilon)$ is the corresponding regular splitting of $I - \lambda P_d^\epsilon$. As a result, $\tilde{v}^* \ge Q_d^{\epsilon\ -1} r_{(d,P_d^\epsilon)} + Q_d^{\epsilon\ -1} R_d^\epsilon \tilde{v}^* - \epsilon \mathbf{1}$. By re-arrangement, it leads to

$$\begin{aligned} \tilde{v}^* &\ge (I - Q_d^{\epsilon\ -1} R_d^\epsilon)^{-1} Q_d^{\epsilon\ -1} r_{(d,P_d^\epsilon)} - \epsilon (I - Q_d^{\epsilon\ -1} R_d^\epsilon)^{-1} \mathbf{1} \\ &= (Q_d^\epsilon - R_d^\epsilon)^{-1} r_{(d,P_d^\epsilon)} - \epsilon (I - Q_d^{\epsilon\ -1} R_d^\epsilon)^{-1} \mathbf{1} \\ &= (I - \lambda P_d^\epsilon)^{-1} r_{(d,P_d^\epsilon)} - \epsilon \sum_{\imath=0}^{\infty} (Q_d^{\epsilon\ -1} R_d^\epsilon)^{\imath} \mathbf{1} \\ &= \sum_{\imath=0}^{\infty} (\lambda P_d^\epsilon)^{\imath} r_{(d,P_d^\epsilon)} - \epsilon \sum_{\imath=0}^{\infty} (Q_d^{\epsilon\ -1} R_d^\epsilon)^{\imath} \mathbf{1} \\ &= v_{(d,P_d^\epsilon)} - \epsilon \sum_{\imath=0}^{\infty} (Q_d^{\epsilon\ -1} R_d^\epsilon)^{\imath} \mathbf{1}, \end{aligned} \tag{16}$$

where $(I - \lambda P_d^\epsilon)^{-1} = \sum_{\imath=0}^{\infty} (\lambda P_d^\epsilon)^{\imath}$ and $(I - Q_d^{\epsilon\ -1} R_d^\epsilon)^{-1} = \sum_{\imath=0}^{\infty} (Q_d^{\epsilon\ -1} R_d^\epsilon)^{\imath}$ hold because $\sigma(\lambda P_d^\epsilon) \le \|\lambda P_d^\epsilon\| = \lambda < 1$ and $\sigma(Q_d^{\epsilon\ -1} R_d^\epsilon) \le \|Q_d^{\epsilon\ -1} R_d^\epsilon\| \le \alpha < 1$. Since $(Q_d^{\epsilon\ -1} R_d^\epsilon)^{\imath} \mathbf{1} \le \mathbf{1} \sup_{k\in\{1,2,\ldots,m\}} \sum_{l=1}^{m} |(Q_d^{\epsilon\ -1} R_d^\epsilon)^{\imath} (l|k)| = \mathbf{1} \|(Q_d^{\epsilon\ -1} R_d^\epsilon)^{\imath}\| \le \mathbf{1} \|Q_d^{\epsilon\ -1} R_d^\epsilon\|^{\imath} \le \alpha^{\imath} \mathbf{1}$ for any nonnegative integer $\imath$, $\tilde{v}^* \ge v_{(d,P_d^\epsilon)} - \epsilon \sum_{\imath=0}^{\infty} (Q_d^{\epsilon\ -1} R_d^\epsilon)^{\imath} \mathbf{1} \ge v_{(d,P_d^\epsilon)} - \epsilon \sum_{\imath=0}^{\infty} \alpha^{\imath} \mathbf{1} = v_{(d,P_d^\epsilon)} - \epsilon \mathbf{1}/(1 - \alpha)$. As $\epsilon > 0$ is arbitrary, $\tilde{v}^* \ge v_{(d,P_d^\epsilon)} \ge \min_{P_d\in\mathcal{P}_d} v_{(d,P_d)}$. In addition, note that $d \in \mathcal{D}$ is arbitrary. Thus, $\tilde{v}^* \ge v^* = \max_{d\in\mathcal{D}} \min_{P_d\in\mathcal{P}_d} v_{(d,P_d)}$.

On the other hand, since $\tilde{v}^* = \mathscr{Y} \tilde{v}^* = \max_{d\in\mathcal{D}} \min_{P_d\in\mathcal{P}_d} \left\{ Q_d^{-1} r_{(d,P_d)} + Q_d^{-1} R_d \tilde{v}^* \right\}$, it follows that for any $\epsilon > 0$, there is a $\tilde{d} \in \mathcal{D}$ such that $\tilde{v}^* \le \min_{P_{\tilde{d}}\in\mathcal{P}_{\tilde{d}}} \left\{ Q_{\tilde{d}}^{-1} r_{(\tilde{d},P_{\tilde{d}})} + Q_{\tilde{d}}^{-1} R_{\tilde{d}} \tilde{v}^* \right\} + \epsilon \mathbf{1}$. Thus, for any $\tilde{P}_{\tilde{d}} \in \mathcal{P}_{\tilde{d}}$, $\tilde{v}^* \le \tilde{Q}_{\tilde{d}}^{-1} r_{(\tilde{d},\tilde{P}_{\tilde{d}})} + \tilde{Q}_{\tilde{d}}^{-1} \tilde{R}_{\tilde{d}} \tilde{v}^* + \epsilon \mathbf{1}$, where $(\tilde{Q}_{\tilde{d}}, \tilde{R}_{\tilde{d}})$ is the corresponding regular splitting of $I - \lambda \tilde{P}_{\tilde{d}}$. Then, using a similar calculation to (16) leads to $\tilde{v}^* \le v_{(\tilde{d},\tilde{P}_{\tilde{d}})} + \epsilon \mathbf{1}/(1 - \alpha)$. Since both $\epsilon > 0$ and $\tilde{P}_{\tilde{d}} \in \mathcal{P}_{\tilde{d}}$ for the given $\tilde{d} \in \mathcal{D}$ are arbitrary, $\tilde{v}^* \le \min_{P_{\tilde{d}}\in\mathcal{P}_{\tilde{d}}} v_{(\tilde{d},P_{\tilde{d}})}$. Consequently, $\tilde{v}^* \le \max_{d\in\mathcal{D}} \min_{P_d\in\mathcal{P}_d} v_{(d,P_d)} = v^*$. Therefore, we have $\tilde{v}^* = v^*$.

To complete, we establish part (b). Since $\mathscr{Y}$ is a contraction mapping on $\mathcal{V}$ and its unique fixed point is v^*, $\|v^{t+1} - v^*\| = \|\mathscr{Y} v^t - \mathscr{Y} v^*\| \le \alpha \|v^t - v^*\|$ for $\forall t \in \mathbb{N}$. By recursion, using this inequality repeatedly yields $\|v^t - v^*\| \le \alpha^t \|v^0 - v^*\|$. Dividing both sides by $\|v^0 - v^*\|$ and taking the t-th root show that $\beta := \limsup_{t\to\infty} \left[\|v^t - v^*\| / \|v^0 - v^*\| \right]^{1/t} \le$

[1] See section 6.3.1 in [12] for the definition of convergence order and rates.

α. From this inequality, it is immediate to get that $\limsup_{t\to\infty}\|v^t - v^*\|/\beta^t \le \|v^0 - v^*\| < \infty$. Thus, we have the results shown in part (b). ■

From this lemma, we now show that the degenerated form GSraVI will converge to an ϵ-robust team-optimal policy within a finite number of iterations when the tolerance of approximation computations is zero.

***Theorem* 1.** *If $M_t = 0$ for $\forall t \in \mathbb{N}$ and the approximation calculations are exact, i.e. $\delta = 0$, in Algorithm 1, then for any $v^0 \in \mathcal{V}$, (a) the sequence $\{v^t\}$ generated by Algorithm 1 converges in norm to the robust team-optimal value function v^* for $t \to \infty$; it converges globally at order 1 at a rate no greater than λ; its global AARC is no greater than λ, and it converges globally with $O(\beta^t), \beta \le \lambda$; (b) the algorithm terminates within a finite number of iterations with an ϵ-robust team-optimal policy $(d^\epsilon)^\infty$ and its corresponding value function v^ϵ satisfies $\|v^\epsilon - v^*\| < \epsilon$.*

Proof: Note first that in the iterative scheme (13) of the GSraVI algorithm, (Q_d, R_d) is the GS regular splitting of $I - \lambda P_d$ for any given $d \in \mathcal{D}$ and $P_d \in \mathcal{P}_d$. Moreover, since $Q'_d = I$ and $R'_d = \lambda P_d$ are a trivial regular splitting of $I - \lambda P_d$, and meanwhile it satisfies $R_d \le R'_d$, it follows from Proposition 6.3.5 in [12] that $\|Q_d^{-1}R_d\| \le \|{Q'_d}^{-1}R'_d\| = \lambda < 1$. Consequently, from Lemma 1, part (a) is obtained.

We now prove part (b). Since $\{v^t\}$ converges to v^*, it is a Cauchy sequence. Hence, the termination condition in step 3(a) will be satisfied for any $\epsilon > 0$ after a finite number of iterations. Without loss of generality, suppose that the algorithm terminates at $t = N$ with $\|v^{N+1} - v^N\| < (1-\lambda)\epsilon/2\lambda$, and meanwhile it outputs a decision rule d^ϵ, which satisfies $d^\epsilon \in \arg\max_{d\in\mathcal{D}}\left\{\min_{P_d\in\mathcal{P}_d}\left[Q_d^{-1}r_{(d,P_d)} + Q_d^{-1}R_d v^N\right]\right\}$ in view of the setup in step 2(a). In this case, let $P^\epsilon_{d^\epsilon} \in \arg\min_{P_{d^\epsilon}\in\mathcal{P}_{d^\epsilon}}\left[Q_{d^\epsilon}^{-1}r_{(d^\epsilon,P_{d^\epsilon})} + Q_{d^\epsilon}^{-1}R_{d^\epsilon}v^N\right]$, where $(Q_{d^\epsilon}, R_{d^\epsilon})$ is the GS regular splitting of $I - \lambda P_{d^\epsilon}$. For $v \in \mathcal{V}$, define the operator $\mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})} : \mathcal{V} \to \mathcal{V}$ by $\mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})}v = {Q^\epsilon_{d^\epsilon}}^{-1}r_{(d^\epsilon,P^\epsilon_{d^\epsilon})} + {Q^\epsilon_{d^\epsilon}}^{-1}R^\epsilon_{d^\epsilon}v$, where $(Q^\epsilon_{d^\epsilon}, R^\epsilon_{d^\epsilon})$ is the GS regular splitting of $I - \lambda P^\epsilon_{d^\epsilon}$. It is easy to see that $\mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})}$ is a contraction mapping on $\mathcal{V}$ because $\|\mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})}u - \mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})}v\| = \|{Q^\epsilon_{d^\epsilon}}^{-1}R^\epsilon_{d^\epsilon}(u-v)\| \le \|{Q^\epsilon_{d^\epsilon}}^{-1}R^\epsilon_{d^\epsilon}\|\|u - v\| \le \|I^{-1}(\lambda P^\epsilon_{d^\epsilon})\|\|u-v\| = \lambda\|u-v\|$ holds for any $u, v \in \mathcal{V}$, where $\|{Q^\epsilon_{d^\epsilon}}^{-1}R^\epsilon_{d^\epsilon}\| \le \|I^{-1}(\lambda P^\epsilon_{d^\epsilon})\|$ follows from Proposition 6.3.5 in [12]. On the other hand, by solving the fixed point equation $v = \mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})}v = {Q^\epsilon_{d^\epsilon}}^{-1}r_{(d^\epsilon,P^\epsilon_{d^\epsilon})} + {Q^\epsilon_{d^\epsilon}}^{-1}R^\epsilon_{d^\epsilon}v$, one can get that $v_{(d^\epsilon,P^\epsilon_{d^\epsilon})}$ is the unique fixed point of $\mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})}$. Also, from the definition of $\mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})}$, one can see that $\mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})}v^N = \mathscr{Y}v^N = v^{N+1}$. It follows that $\|v_{(d^\epsilon,P^\epsilon_{d^\epsilon})} - v^{N+1}\| = \|\mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})}v_{(d^\epsilon,P^\epsilon_{d^\epsilon})} - v^{N+1}\| \le \|\mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})}v_{(d^\epsilon,P^\epsilon_{d^\epsilon})} - \mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})}v^{N+1}\| + \|\mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})}v^{N+1} - \mathscr{L}_{(d^\epsilon,P^\epsilon_{d^\epsilon})}v^N\| \le \lambda\|v_{(d^\epsilon,P^\epsilon_{d^\epsilon})} - v^{N+1}\| + \lambda\|v^{N+1} - v^N\|$. By re-arrangement, one can get $\|v_{(d^\epsilon,P^\epsilon_{d^\epsilon})} - v^{N+1}\| \le \frac{\lambda}{1-\lambda}\|v^{N+1} - v^N\| < \epsilon/2$. Moreover, since $\|v^{N+1} - v^*\| \le \|v^{N+1} - \mathscr{Y}v^{N+1}\| + \|\mathscr{Y}v^{N+1} - v^*\| \le \lambda\|v^{N+1} - v^N\| + \lambda\|v^{N+1} - v^*\|$ by applying the contraction property of $\mathscr{Y}$, $\|v^{N+1} - v^*\| \le \frac{\lambda}{1-\lambda}\|v^{N+1} - v^N\| < \epsilon/2$ by re-arrangement. Consequently, $\|v_{(d^\epsilon,P^\epsilon_{d^\epsilon})} - v^*\| \le \|v_{(d^\epsilon,P^\epsilon_{d^\epsilon})} - v^{N+1}\| + \|v^{N+1} - v^*\| < \epsilon$. ■

Theorem 1 establishes the convergence of the degenerated form GSraVI of GSraMPI in the exact case, i.e. $\delta = 0$. In the following, we extend this result to the inexact case where the tolerance of approximation calculations is relaxed to $\lambda\delta$ for $0 \le \delta < (1-\lambda)^2\epsilon/2\lambda(1+\lambda)$. To distinguish these two cases, let $\{\tilde{v}^t\}$ be the sequence generated by the GSraVI algorithm in the inexact case. Then, the following inequality will hold.

Lemma 2. *If $v^0 = \tilde{v}^0$, then $\{v^t\}$ and $\{\tilde{v}^t\}$ satisfy*

$$v^t - \theta\mathbf{1} \le \tilde{v}^t \le v^t + \theta\mathbf{1}, \text{ for } \forall t \in \mathbb{N}, \tag{17}$$

where $\theta := \lambda\delta/(1-\lambda) \in [0, \epsilon/2)$.

Proof: We prove this result by induction. For $t = 0$, (17) holds according to the assumption $v^0 = \tilde{v}^0$. Suppose now that (17) is satisfied for $t = \kappa \ge 0$, and we consider the case $t = \kappa + 1$. First, from (9), substituting each term of (17) yields $\rho_{(s^1,a)}(v^\kappa - \theta\mathbf{1}) = \rho_{(s^1,a)}(v^\kappa) - \lambda\theta \le \rho_{(s^1,a)}(\tilde{v}^\kappa) \le \rho_{(s^1,a)}(v^\kappa) + \lambda\theta = \rho_{(s^1,a)}(v^\kappa + \theta\mathbf{1})$ for $\forall a \in \mathcal{A}$. Taking the maximum of each term of this inequality subject to $a \in \mathcal{A}$, we have

$$v^{\kappa+1}_{s^1} - \lambda\theta \le \max_{a\in\mathcal{A}}\rho_{(s^1,a)}(\tilde{v}^\kappa) \le v^{\kappa+1}_{s^1} + \lambda\theta. \tag{18}$$

Since $\tilde{\rho}_{(s^1,a)}(\tilde{v}^\kappa) - \lambda\delta \le \rho_{(s^1,a)}(\tilde{v}^\kappa) \le \tilde{\rho}_{(s^1,a)}(\tilde{v}^\kappa) + \lambda\delta$ from (5), taking the maximum of each term subject to $a \in \mathcal{A}$ yields

$$\tilde{v}^{\kappa+1}_{s^1} - \lambda\delta \le \max_{a\in\mathcal{A}}\rho_{(s^1,a)}(\tilde{v}^\kappa) \le \tilde{v}^{\kappa+1}_{s^1} + \lambda\delta. \tag{19}$$

Then, according to (18) and (19), it follows that

$$v^{\kappa+1}_{s^1} - \lambda(\delta+\theta) \le \tilde{v}^{\kappa+1}_{s^1} \le v^{\kappa+1}_{s^1} + \lambda(\delta+\theta). \tag{20}$$

Since $\lambda(\delta+\theta) = \theta$, (20) is changed to $v^{\kappa+1}_{s^1} - \theta \le \tilde{v}^{\kappa+1}_{s^1} \le v^{\kappa+1}_{s^1} + \theta$. Combing this inequality with $v^\kappa_{s^j} - \theta \le \tilde{v}^\kappa_{s^j} \le v^\kappa_{s^j} + \theta$, $j = 2, 3, \ldots, m$, and from (9), one can obtain that for $\forall a \in \mathcal{A}$, $\rho_{(s^2,a)}(v^\kappa) - \lambda\theta \le \rho_{(s^2,a)}(\tilde{v}^\kappa) \le \rho_{(s^2,a)}(v^\kappa) + \lambda\theta$. Taking the maximum of each term subject to $a \in \mathcal{A}$ then leads to $v^{\kappa+1}_{s^2} - \lambda\theta \le \max_{a\in\mathcal{A}}\rho_{(s^2,a)}(\tilde{v}^\kappa) \le v^{\kappa+1}_{s^2} + \lambda\theta$. Likewise, for $s^2 \in \mathcal{S}$, since $\tilde{\rho}_{(s^2,a)}(\tilde{v}^\kappa) - \lambda\delta \le \rho_{(s^2,a)}(\tilde{v}^\kappa) \le \tilde{\rho}_{(s^2,a)}(\tilde{v}^\kappa) + \lambda\delta$ from (5), we have $\tilde{v}^{\kappa+1}_{s^2} - \lambda\delta \le \max_{a\in\mathcal{A}}\rho_{(s^2,a)}(\tilde{v}^\kappa) \le \tilde{v}^{\kappa+1}_{s^2} + \lambda\delta$ by taking the maximum of each term. Using these two inequalities, one can similarly obtain $v^{\kappa+1}_{s^2} - \theta \le \tilde{v}^{\kappa+1}_{s^2} \le v^{\kappa+1}_{s^2} + \theta$. Then, by recursion, applying the same argument for s^j, $j = 3, 4, \ldots, m$, one can get $v^{\kappa+1}_{s^j} - \theta \le \tilde{v}^{\kappa+1}_{s^j} \le v^{\kappa+1}_{s^j} + \theta$. Therefore, $v^{\kappa+1} - \theta\mathbf{1} \le \tilde{v}^{\kappa+1} \le v^{\kappa+1} + \theta\mathbf{1}$. ■

Leveraging this inequality, we now proceed to show the convergence of GSraVI and calculate its convergence rates in the inexact case.

***Theorem* 2.** *If $M_t = 0$ for $\forall t \in \mathbb{N}$ and the tolerance of approximation calculations is within $\lambda\delta$ for $\delta \in [0, (1-\lambda)^2\epsilon/2\lambda(1+\lambda))$ in Algorithm 1, then for any $v^0 \in \mathcal{V}$, (a) the sequence $\{\tilde{v}^t\}$ generated by Algorithm 1 with the initialization $\tilde{v}^0 = v^0$ converges in norm to the robust team-optimal value function v^* for $t \to \infty$; it converges globally at order 1 at a rate no greater than λ; its global AARC is no greater than λ, and it converges globally with $O(\beta^t), \beta \le \lambda$; (b) the algorithm terminates within a finite number of iterations with* *an ϵ-robust team-optimal policy $(d^\epsilon)^\infty$ and its corresponding value function v^ϵ satisfies $\|v^\epsilon - v^*\| < \epsilon$.*

Proof: Since it has been proven in Theorem 1 that the sequence $\{v^t\}$ generated by Algorithm 1, when $M_t = 0$ for $\forall t \in \mathbb{N}$ and $\delta = 0$, converges to v^*, there is a positive integer N for any $\epsilon > 0$ such that $\|v^t - v^*\| < \epsilon/2$ holds for $\forall t \geq N$. Moreover, from Lemma 2, one can obtain that for $\forall t \in \mathbb{N}$,

$$\|\tilde{v}^t - v^*\| \leq \|\tilde{v}^t - v^t\| + \|v^t - v^*\| \leq \|v^t - v^*\| + \theta. \quad (21)$$

Consequently, it leads to $\|\tilde{v}^t - v^*\| \leq \epsilon/2 + \theta < \epsilon$ for $\forall t \geq N$, which implies that $\tilde{v}^t$ converges to v^*.

We now calculate the convergence rates of $\{\tilde{v}^t\}$. From Theorem 1, we first note that $\{v^t\}$ satisfies $\|v^t - v^*\| \leq \lambda\|v^{t-1} - v^*\|$ for $\forall t \in \mathbb{N}$. Moreover, from Lemma 2, $\|v^{t-1} - v^*\| \leq \|v^{t-1} - \tilde{v}^{t-1}\| + \|\tilde{v}^{t-1} - v^*\| \leq \|\tilde{v}^{t-1} - v^*\| + \theta$ holds for all t. In view of these two inequalities, it follows from (21) that $\|\tilde{v}^t - v^*\| \leq \|v^t - v^*\| + \theta \leq \lambda\|v^{t-1} - v^*\| + \theta \leq \lambda\|\tilde{v}^{t-1} - v^*\| + (\lambda+1)\theta$. Since $\theta \in [0, \epsilon/2)$ can be arbitrarily close to zero in view of the arbitrariness of ϵ, $\|\tilde{v}^t - v^*\| \leq \lambda\|\tilde{v}^{t-1} - v^*\|$ for all t. Thus, it suggests that $\{\tilde{v}^t\}$ converges globally at order 1 and the convergence rate is no greater than λ. Also, from (21), we have $\|\tilde{v}^t - v^*\| \leq \|v^t - v^*\|$. It follows that $\|\tilde{v}^t - v^*\|/\|v^0 - v^*\| \leq \|v^t - v^*\|/\|v^0 - v^*\|$ and $\|\tilde{v}^t - v^*\|/\beta^t \leq \|v^t - v^*\|/\beta^t$ for $\forall t \in \mathbb{N}$. From Theorem 1, it implies that the global AARC of $\{\tilde{v}^t\}$ is no greater than λ and it converges globally with $O(\beta^t), \beta \leq \lambda$.

We next prove part (b). Since $\{\tilde{v}^t\}$ converges to v^* as shown in part (a), the termination condition in step 3(a) will be satisfied after a finite number of iterations. Moreover, since when $\delta = 0$, Theorem 1(b) shows that the algorithm will terminate within a finite number of iterations, there exists a positive integer N for any $\epsilon_1 > 0$ such that $\|v^{N+1} - v^N\| < (1-\lambda)\epsilon_1/2\lambda$. Then, from Lemma 2, we have $\|\tilde{v}^{N+1} - \tilde{v}^N\| \leq \|\tilde{v}^{N+1} - v^{N+1}\| + \|v^{N+1} - v^N\| + \|v^N - \tilde{v}^N\| \leq \|v^{N+1} - v^N\| + 2\theta < (1-\lambda)\epsilon_1/2\lambda + 2\lambda\delta/(1-\lambda)$. Given that ϵ_1 is arbitrary, we therefore select a specific ϵ_1 such that $\epsilon_1 \leq \epsilon - 2\lambda(1+\lambda)\delta/(1-\lambda)^2$ for any given $\epsilon > 0$. As such, $\|\tilde{v}^{N+1} - \tilde{v}^N\| \leq (1-\lambda)\epsilon/2\lambda - \delta$ holds. It implies that the GSraVI algorithm in the inexact case can also terminate at $t = N$. Moreover, since for any given $s \in \mathcal{S}$, one can find that there exists a constant gap between $\rho_{(s,a)}(v^t)$ and $\rho_{(s,a)}(\tilde{v}^t)$ from the proof of Lemma 2, and an approximation error between $\rho_{(s,a)}(\tilde{v}^t)$ and $\tilde{\rho}_{(s,a)}(\tilde{v}^t)$ from (5) for $\forall a \in \mathcal{A}$ and $\forall t \in \mathbb{N}$, which do not affect the selection of the maximizing actions, $\arg\max_{a\in\mathcal{A}} \rho_{(s,a)}(v^N) = \arg\max_{a\in\mathcal{A}} \tilde{\rho}_{(s,a)}(\tilde{v}^N)$. Without loss of generality, we choose the same decision rule in the exact and inexact cases at $t = N$. Then, from Theorem 1, we have $\|v^\epsilon - v^*\| < \epsilon$. ■

2) The general form: We now analyze the convergence of the general form of GSraMPI. That is, M_t is a non-negative integer for $\forall t \in \mathbb{N}$. Using a similar analytical process to the GSraVI, we first derive the convergence results of GSraMPI when the approximation calculations are exact, i.e. $\delta = 0$, and then extend them to the inexact case where the tolerance of approximation calculations is within $\lambda\delta$ for $0 \leq \delta < (1-\lambda)^2\epsilon/2\lambda(1+\lambda)$.

We begin with considering the exact case, i.e. $\delta = 0$. Let the operator $\mathscr{B} : \mathcal{V} \to \mathcal{V}$ for $v \in \mathcal{V}$ be defined by

$$\mathscr{B}v := \mathscr{Y}v - v = \max_{d\in\mathcal{D}} \min_{P_d\in\mathcal{P}_d} \left\{ Q_d^{-1} r_{(d,P_d)} + Q_d^{-1} R_d v - v \right\}, \quad (22)$$

where (Q_d, R_d) is the GS regular splitting of $I - \lambda P_d$. From this definition, it is easy to see that the fixed-point of $\mathscr{Y}$ is the same as the zero-point of $\mathscr{B}$. Moreover, for any given $v \in \mathcal{V}$, we denote the set of v-improving decision rules by $\mathcal{D}_v := \arg\max_{d\in\mathcal{D}} \left\{ \min_{P_d\in\mathcal{P}_d} \left[Q_d^{-1} r_{(d,P_d)} + Q_d^{-1} R_d v \right] \right\}$, and accordingly denote the set of d_v-decreasing transition probability matrix for the given $d_v \in \mathcal{D}_v$ by $\mathcal{P}^*_{d_v} := \arg\min_{P_{d_v}\in\mathcal{P}_{d_v}} \left[{Q_{d_v}}^{-1} r_{(d_v,P_{d_v})} + {Q_{d_v}}^{-1} R_{d_v} v \right]$. After having these concepts at hand the following proposition can be obtained immediately.

Proposition 1. *For any given $u, v \in \mathcal{V}$ and $d_v \in \mathcal{D}_v$, there exists a $P'_{d_v} \in \mathcal{P}_{d_v}$ such that $\mathscr{B}u \geq \mathscr{B}v + ({Q'_{d_v}}^{-1} R'_{d_v} - I)(u - v)$, where (Q'_{d_v}, R'_{d_v}) is the GS regular splitting of $I - \lambda P'_{d_v}$.*

Proof: From the definition of $\mathscr{B}$, it is immediate to get that $\mathscr{B}u \geq \min_{P_{d_v}\in\mathcal{P}_{d_v}} \left\{ Q_{d_v}^{-1} r_{(d_v,P_{d_v})} + Q_{d_v}^{-1} R_{d_v} u - u \right\}$ and $\mathscr{B}v = \min_{P_{d_v}\in\mathcal{P}_{d_v}} \left\{ Q_{d_v}^{-1} r_{(d_v,P_{d_v})} + Q_{d_v}^{-1} R_{d_v} v - v \right\}$. Moreover, since there exists a $P'_{d_v} \in \mathcal{P}_{d_v}$ for any $\epsilon > 0$ such that $\min_{P_{d_v}\in\mathcal{P}_{d_v}} \left\{ Q_{d_v}^{-1} r_{(d_v,P_{d_v})} + Q_{d_v}^{-1} R_{d_v} u - u \right\} \geq {Q'_{d_v}}^{-1} r_{(d_v,P'_{d_v})} + {Q'_{d_v}}^{-1} R'_{d_v} u - u - \epsilon\mathbf{1}$, and meanwhile $\min_{P_{d_v}\in\mathcal{P}_{d_v}} \left\{ Q_{d_v}^{-1} r_{(d_v,P_{d_v})} + Q_{d_v}^{-1} R_{d_v} v - v \right\} \leq {Q'_{d_v}}^{-1} r_{(d_v,P'_{d_v})} + {Q'_{d_v}}^{-1} R'_{d_v} v - v$, subtracting $\mathscr{B}v$ from $\mathscr{B}u$ leads to $\mathscr{B}u - \mathscr{B}v \geq ({Q'_{d_v}}^{-1} R'_{d_v} - I)(u - v) - \epsilon\mathbf{1}$, where (Q'_{d_v}, R'_{d_v}) is the GS regular splitting of $I - \lambda P'_{d_v}$. Note that $\epsilon > 0$ is arbitrary. Therefore, $\mathscr{B}u - \mathscr{B}v \geq ({Q'_{d_v}}^{-1} R'_{d_v} - I)(u - v)$. ■

We now proceed to derive the vector expression of the iterative scheme of GSraMPI. Let d_{v^t} be the decision rule satisfying $d_{v^t}(s) \in \arg\max_{a\in\mathcal{A}} \rho_{(s,a)}(v^t)$ for $\forall s \in \mathcal{S}$, and $P^*_{d_{v^t}}$ be the transition probability matrix with the element $p^*(s^l \mid s^k, d_{v^t}(s^k))$ in row k and column l for $\forall s^k, s^l \in \mathcal{S}$, where $p^*(\cdot \mid s^k, d_{v^t}(s^k))$ is the minimizing transition probability distribution for calculating $\rho_{(s^k, d_{v^t}(s^k))}(v^t)$ by (9). Then, from (5) and (6) and applying an analogous argument to the derivation of (12), the vector form of (6) when $\delta = 0$ can be given by

$$\begin{aligned} u_0^t &= (\underbrace{I - \lambda {P^*_{d_{v^t}}}^L}_{Q^*_{d_{v^t}}})^{-1} r_{(d_{v^t}, P^*_{d_{v^t}})} + (I - \lambda {P^*_{d_{v^t}}}^L)^{-1} \underbrace{\lambda {P^*_{d_{v^t}}}^U}_{R^*_{d_{v^t}}} v^t \\ &= {Q^*_{d_{v^t}}}^{-1} r_{(d_{v^t}, P^*_{d_{v^t}})} + {Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}} v^t, \end{aligned} \quad (23)$$

where ${P^*_{d_{v^t}}}^L$ and ${P^*_{d_{v^t}}}^U$ have the same formula as ${P_{\check{d}}}^L$ and ${P_{\check{d}}}^U$ by replacing $P_{\check{d}}(l \mid k)$ with $p^*(s^l \mid s^k, d_{v^t}(s^k)), \forall s^k, s^l \in \mathcal{S}$, respectively; and $(Q^*_{d_{v^t}}, R^*_{d_{v^t}})$ is the GS regular splitting of $I - \lambda P^*_{d_{v^t}}$. In view of (23), for any given $d \in \mathcal{D}$ and $P_d \in \mathcal{P}_d$, we define the operator $\mathscr{T}_{(d,P_d)} : \mathcal{V} \to \mathcal{V}$ for $v \in \mathcal{V}$ by $\mathscr{T}_{(d,P_d)} v := {Q_d}^{-1} r_{(d,P_d)} + {Q_d}^{-1} R_d v$, where (Q_d, R_d) is the GS regular splitting of $I - \lambda P_d$. Similar to the $\mathscr{L}_{(d^\epsilon, P^\epsilon_{d^\epsilon})}$ defined in the proof of Theorem 1, one can show that $\mathscr{T}_{(d,P_d)}$ is a contraction mapping on $\mathcal{V}$ with constant λ. In this case,

$u_0^t = \mathscr{T}_{(d_{v^t}, P^*_{d_{v^t}})} v^t$. On the other hand, from (7), (8), and (10), one can analogously obtain the vector form of (8) when $\delta = 0$ by $u^t_{\varsigma+1} = \mathscr{T}_{(d_{v^t}, P^*_{d_{v^t}})} u^t_\varsigma$. Thus, the vector representation of $v_s^{t+1} = u^t_{M_t}(s)$ for $\forall s \in \mathcal{S}$ can be given by

$$\begin{aligned} v^{t+1} &= \left(\mathscr{T}_{(d_{v^t}, P^*_{d_{v^t}})}\right)^{M_t+1} v^t \\ &= \sum_{\varsigma=0}^{M_t} \left({Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}}\right)^\varsigma {Q^*_{d_{v^t}}}^{-1} r_{(d_{v^t}, P^*_{d_{v^t}})} \\ &\quad + \left({Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}}\right)^{M_t+1} v^t \\ &= v^t + \sum_{\varsigma=0}^{M_t} \left({Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}}\right)^\varsigma \left({Q^*_{d_{v^t}}}^{-1} r_{(d_{v^t}, P^*_{d_{v^t}})} \right. \\ &\quad \left. + {Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}} v^t - v^t\right) \\ &= v^t + \sum_{\varsigma=0}^{M_t} \left({Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}}\right)^\varsigma (\mathscr{B} v^t) . \end{aligned} \tag{24}$$

From this equation, it is easy to see that the GSraMPI algorithm incorporates the GSraVI as a special case, because (24) degenerates to the iterative scheme of GSraVI, $v^{t+1} = \mathscr{Y} v^t$, when $M_t = 0$. On the other hand, when $M_t \to \infty$, it defines a variant of rPI,

$$\begin{aligned} v^{t+1} &= v^t + \sum_{\varsigma=0}^{\infty} \left({Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}}\right)^\varsigma (\mathscr{B} v^t) \\ &= v^t + \left(I - {Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}}\right)^{-1} \left[{Q^*_{d_{v^t}}}^{-1} r_{(d_{v^t}, P^*_{d_{v^t}})} \right. \\ &\quad \left. - \left(I - {Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}}\right) v^t\right] \\ &= \left(I - {Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}}\right)^{-1} {Q^*_{d_{v^t}}}^{-1} r_{(d_{v^t}, P^*_{d_{v^t}})} \\ &= \left(I - \lambda P^*_{d_{v^t}}\right)^{-1} r_{(d_{v^t}, P^*_{d_{v^t}})} = v_{(d_{v^t}, P^*_{d_{v^t}})}. \end{aligned} \tag{25}$$

Namely, after each iteration, the value function at the next time step evaluates the performance of the improved policy derived from the current value function.

In view of (24), for any given $M \in \mathbb{N}$ and $v \in \mathcal{V}$, we define the operator $\mathscr{W}^M : \mathcal{V} \to \mathcal{V}$ by $\mathscr{W}^M v := (\mathscr{T}_{(d_v, P^*_{d_v})})^{M+1} v$ where $d_v \in \mathcal{D}_v$ and $P^*_{d_v} \in \mathcal{P}^*_{d_v}$, and accordingly the operator $\mathscr{U}^M : \mathcal{V} \to \mathcal{V}$ by $\mathscr{U}^M v := \max_{d\in\mathcal{D}} \max_{P_d \in \mathcal{P}_d} \Phi(d, P_d, v)$ where $\Phi(d, P_d, v) = \sum_{\varsigma=0}^{M} \left(Q_d^{-1} R_d\right)^\varsigma Q_d^{-1} r_{(d,P_d)} + \left(Q_d^{-1} R_d\right)^{M+1} v$ and (Q_d, R_d) is the GS regular splitting of $I - \lambda P_d$. For these two operators, some useful properties are shown in the following lemmas, which provide the basis for proving the convergence of GSraMPI.

Lemma 3. *For the operator $\mathscr{U}^M$, the following properties hold: (a) it is a contraction mapping on $\mathcal{V}$ with constant λ^{M+1}; (b) for any $\omega^0 \in \mathcal{V}$, the sequence $\{\omega^t\}$ generated by $\omega^{t+1} = \mathscr{U}^M \omega^t$ converges in norm to the robust team-optimal value function v^*, which is the unique fixed-point of $\mathscr{U}^M$; and (c) $\|\omega^{t+1} - v^*\| \le \lambda^{M+1} \|\omega^t - v^*\|$.*

Proof: Since in the definition of $\mathscr{U}^M$, (Q_d, R_d) is the GS regular splitting of $I - \lambda P_d$ for any given $d \in \mathcal{D}$ and $P_d \in \mathcal{P}_d$, applying Proposition 6.3.5 in [12] leads to $\|{Q_d}^{-1} R_d\| \le \|I^{-1}(\lambda P_d)\| = \lambda < 1$. It implies that $\sup_{d\in\mathcal{D}, P_d\in\mathcal{P}_d} \|Q_d^{-1} R_d\| \le \lambda$. Hence, one can use a similar argument to Lemma 1(a) to prove part (a). For any given $u, v \in \mathcal{V}$, we start with considering those states $s \in \mathcal{S}$ satisfying $\mathscr{U}^M v(s) - \mathscr{U}^M u(s) \ge 0$. Let $d'_v \in \arg\max_{d\in\mathcal{D}} \{\max_{P_d\in\mathcal{P}_d} \Phi(d, P_d, v)\}$. Then, we have $\mathscr{U}^M v(s) = \left[\max_{P_{d'_v} \in \mathcal{P}_{d'_v}} \Phi(d'_v, P_{d'_v}, v)\right](s)$ and $\mathscr{U}^M u(s) \ge \left[\max_{P_{d'_v} \in \mathcal{P}_{d'_v}} \Phi(d'_v, P_{d'_v}, u)\right](s)$. Since there exists a $P'_{d'_v} \in \mathcal{P}_{d'_v}$ for any $\epsilon > 0$ such that $\max_{P_{d'_v} \in \mathcal{P}_{d'_v}} \Phi(d'_v, P_{d'_v}, v) \le \Phi(d'_v, P'_{d'_v}, v) + \epsilon\mathbf{1}$ and meanwhile $\max_{P_{d'_v} \in \mathcal{P}_{d'_v}} \Phi(d'_v, P_{d'_v}, u) \ge \Phi(d'_v, P'_{d'_v}, u)$, subtracting $\mathscr{U}^M u(s)$ from $\mathscr{U}^M v(s)$ leads to $\mathscr{U}^M v(s) - \mathscr{U}^M u(s) \le \left[\Phi(d'_v, P'_{d'_v}, v) + \epsilon\mathbf{1} - \Phi(d'_v, P'_{d'_v}, u)\right](s)$. Note that $\epsilon > 0$ is arbitrary. Therefore, $\mathscr{U}^M v(s) - \mathscr{U}^M u(s) \le \left[\Phi(d'_v, P'_{d'_v}, v) - \Phi(d'_v, P'_{d'_v}, u)\right](s) = \left[({Q'_{d'_v}}^{-1} R'_{d'_v})^{M+1}(v - u)\right](s)$, where $(Q'_{d'_v}, R'_{d'_v})$ is the GS regular splitting of $I - \lambda P'_{d'_v}$. Moreover, note that $\left[({Q'_{d'_v}}^{-1} R'_{d'_v})^{M+1}(v - u)\right](s) \le \|({Q'_{d'_v}}^{-1} R'_{d'_v})^{M+1}(v - u)\| \le \|{Q'_{d'_v}}^{-1} R'_{d'_v}\|^{M+1} \|v - u\| \le \lambda^{M+1} \|v - u\|$. Therefore, $0 \le \mathscr{U}^M v(s) - \mathscr{U}^M u(s) \le \lambda^{M+1} \|v - u\|$.

Similarly, applying the same argument for the states $s \in \mathcal{S}$ satisfying $\mathscr{U}^M v(s) - \mathscr{U}^M u(s) \le 0$, one can obtain $0 \le \mathscr{U}^M u(s) - \mathscr{U}^M v(s) \le \lambda^{M+1} \|u - v\|$. Consequently, it follows that $\|\mathscr{U}^M v - \mathscr{U}^M u\| \le \lambda^{M+1} \|v - u\|$. Since $\lambda^{M+1} < 1$, $\mathscr{U}^M$ is a contraction mapping on $\mathcal{V}$. From Banach fixed-point theorem, then, for any $\omega^0 \in \mathcal{V}$, the sequence $\{\omega^t\}$ generated by $\omega^{t+1} = \mathscr{U}^M \omega^t$ will converge in norm to the unique fixed point ω^* of $\mathscr{U}^M$.

Now, we show $\omega^* = v^*$. First, since v^* is the unique fixed point of $\mathscr{Y}$ in view of Lemma 1, i.e. $v^* = \mathscr{Y} v^* = \max_{d\in\mathcal{D}} \min_{P_d\in\mathcal{P}_d} \{Q_d^{-1} r_{(d,P_d)} + Q_d^{-1} R_d v^*\}$, there exists a $d_{v^*} \in \mathcal{D}_{v^*}$ and a $P^*_{d_{v^*}} \in \mathcal{P}^*_{d_{v^*}}$ such that $v^* = {Q^*_{d_{v^*}}}^{-1} r_{(d_{v^*}, P^*_{d_{v^*}})} + {Q^*_{d_{v^*}}}^{-1} R^*_{d_{v^*}} v^*$, where $(Q^*_{d_{v^*}}, R^*_{d_{v^*}})$ is the GS regular splitting of $I - \lambda P^*_{d_{v^*}}$. Leveraging $\mathscr{B} v^* = \mathscr{Y} v^* - v^* = 0$ and from the definition of $\mathscr{W}^M$ and $\mathscr{U}^M$, we then have $v^* = \mathscr{W}^M v^* = \sum_{\varsigma=0}^{M} ({Q^*_{d_{v^*}}}^{-1} R^*_{d_{v^*}})^\varsigma {Q^*_{d_{v^*}}}^{-1} r_{(d_{v^*}, P^*_{d_{v^*}})} + ({Q^*_{d_{v^*}}}^{-1} R^*_{d_{v^*}})^{M+1} v^* \le \mathscr{U}^M v^*$. Applying this inequality recursively for $\imath \in \mathbb{N}$ times leads to $v^* \le (\mathscr{U}^M)^\imath v^*$. In particular, this inequality holds for $\imath \to \infty$. Thus, $v^* \le \omega^*$.

Moreover, for $v \in \mathcal{V}$, we define the operator $\mathscr{M} : \mathcal{V} \to \mathcal{V}$ by $\mathscr{M} v := \max_{d\in\mathcal{D}} \max_{P_d\in\mathcal{P}_d} \{Q_d^{-1} r_{(d,P_d)} + Q_d^{-1} R_d v\}$. Using a similar argument to Lemma 1(a), one can show that $\mathscr{M}$ is a contraction mapping on $\mathcal{V}$, and the sequence generated by it will converge in norm to v^*. Since $\omega^* = \mathscr{U}^M \omega^* \le \mathscr{M}^M \omega^*$ for $\forall M \in \mathbb{N}$ and especially it holds for $M \to \infty$, it yields $\omega^* \le v^*$. Consequently, it follows that $\omega^* = v^*$.

Finally, part (c) can be established by noticing $\|\omega^{t+1} - v^*\| = \|\mathscr{U}^M \omega^t - \mathscr{U}^M v^*\| \le \lambda^{M+1} \|\omega^t - v^*\|$. ■

Lemma 4. *For any $u, v \in \mathcal{V}$ satisfying $u \ge v$, $\mathscr{U}^M u \ge \mathscr{W}^M v$ holds for any $M \in \mathbb{N}$. Moreover, if $u \in \mathcal{V}_{\mathscr{B}} = \{\nu \in \mathcal{V} \mid \mathscr{B}\nu \ge 0\}$, then $\mathscr{W}^M u \ge \mathscr{Y} v$.*

Proof: Let $d_v \in \mathcal{D}_v \subseteq \mathcal{D}$ be the v-improving decision rule and $P^*_{d_v} \in \mathcal{P}^*_{d_v} \subseteq \mathcal{P}_{d_v}$ be the d_v-decreasing transition probability matrix. Then, for any $M \in \mathbb{N}$, we have $\mathscr{W}^M v = \Phi(d_v, P^*_{d_v}, v)$ from the definition of $\mathscr{W}^M$ and $\Phi(\cdot)$, and meanwhile $\mathscr{U}^M u \geq \Phi(d_v, P^*_{d_v}, u)$ from the definition of $\mathscr{U}^M$. Subtracting $\mathscr{W}^M v$ from $\mathscr{U}^M u$ and utilizing $u \geq v$ lead to $\mathscr{U}^M u - \mathscr{W}^M v \geq \Phi(d_v, P^*_{d_v}, u) - \Phi(d_v, P^*_{d_v}, v) = ({Q^*_{d_v}}^{-1} R^*_{d_v})^{M+1}(u - v) \geq 0$, where $(Q^*_{d_v}, R^*_{d_v})$ is the GS regular splitting of $I - \lambda P^*_{d_v}$. Moreover, if $\mathscr{B}u \geq 0$, then from the definition of $\mathscr{W}^M$, we have $\mathscr{W}^M u = u + \sum_{\varsigma=0}^{M} \left({Q^*_{d_u}}^{-1} R^*_{d_u}\right)^{\varsigma} (\mathscr{B}u) \geq u + \mathscr{B}u = \mathscr{Y}u$. Since for any $d \in \mathcal{D}$ and $P_d \in \mathcal{P}_d$, $Q_d^{-1} r_{(d,P_d)} + Q_d^{-1} R_d u \geq Q_d^{-1} r_{(d,P_d)} + Q_d^{-1} R_d v$ holds for any regular splitting (Q_d, R_d) of $I - \lambda P_d$, $\mathscr{Y}u \geq \mathscr{Y}v$. Consequently, $\mathscr{W}^M u \geq \mathscr{Y}v$.

■

Lemma 5. *If $v \in \mathcal{V}_{\mathscr{B}} = \{\nu \in \mathcal{V} \mid \mathscr{B}\nu \geq 0\}$, then $\mathscr{W}^M v \in \mathcal{V}_{\mathscr{B}}$ for any $M \in \mathbb{N}$.*

Proof: Let $u = \mathscr{W}^M v$ and d_v be the v-improving decision rule. Then, from Proposition 1, there exists a $P'_{d_v} \in \mathcal{P}_{d_v}$ such that $\mathscr{B}u \geq \mathscr{B}v + ({Q'_{d_v}}^{-1} R'_{d_v} - I)(u - v)$, where (Q'_{d_v}, R'_{d_v}) is the GS regular splitting of $I - \lambda P'_{d_v}$. Moreover, from the definition of $\mathscr{W}^M$, we have $u - v = \mathscr{W}^M v - v = v + \sum_{\varsigma=0}^{M} \left({Q^*_{d_v}}^{-1} R^*_{d_v}\right)^{\varsigma} (\mathscr{B}v) - v = \sum_{\varsigma=0}^{M} \left({Q^*_{d_v}}^{-1} R^*_{d_v}\right)^{\varsigma} (\mathscr{B}v) \geq 0$, where $(Q^*_{d_v}, R^*_{d_v})$ is the GS regular splitting of $I - \lambda P^*_{d_v}$, and $P^*_{d_v} \in \mathcal{P}^*_{d_v} \subseteq \mathcal{P}_{d_v}$ is the d_v-decreasing transition probability matrix. Therefore, $\mathscr{B}u \geq \mathscr{B}v + ({Q'_{d_v}}^{-1} R'_{d_v} - I) \sum_{\varsigma=0}^{M} \left({Q^*_{d_v}}^{-1} R^*_{d_v}\right)^{\varsigma} (\mathscr{B}v) \geq \mathscr{B}v + \min_{P \in \mathcal{P}_{d_v}} \left(Q^{-1}R - I\right) \min_{P \in \mathcal{P}_{d_v}} \left\{\sum_{\varsigma=0}^{M} \left(Q^{-1}R\right)^{\varsigma} (\mathscr{B}v)\right\}$, where (Q, R) is the GS regular splitting of $I - \lambda P$ for $P \in \mathcal{P}_{d_v}$. Note that for any given $v' \in \mathcal{V}$, $\left(Q^{-1}R - I\right) v'$ and $\sum_{\varsigma=0}^{M} \left(Q^{-1}R\right)^{\varsigma} v'$ have the same minimum point subject to $P \in \mathcal{P}_{d_v}$. Thus, there exists a $\hat{P}_{d_v} \in \mathcal{P}_{d_v}$ such that $\min_{P \in \mathcal{P}_{d_v}} \left(Q^{-1}R - I\right) \min_{P \in \mathcal{P}_{d_v}} \left\{\sum_{\varsigma=0}^{M} \left(Q^{-1}R\right)^{\varsigma} (\mathscr{B}v)\right\} = \left(\hat{Q}_{d_v}^{-1} \hat{R}_{d_v} - I\right) \sum_{\varsigma=0}^{M} \left(\hat{Q}_{d_v}^{-1} \hat{R}_{d_v}\right)^{\varsigma} (\mathscr{B}v) = \left(\hat{Q}_{d_v}^{-1} \hat{R}_{d_v}\right)^{M+1} (\mathscr{B}v) - \mathscr{B}v$, where $(\hat{Q}_{d_v}, \hat{R}_{d_v})$ is the GS regular splitting of $I - \lambda \hat{P}_{d_v}$. As a result, $\mathscr{B}u \geq \mathscr{B}v + \left(\hat{Q}_{d_v}^{-1} \hat{R}_{d_v}\right)^{M+1} (\mathscr{B}v) - \mathscr{B}v \geq 0$.

■

With the help of these lemmas, we now show one of our main results that the GSraMPI algorithm converges to the robust team-optimal policy at a convergence rate of the integer power of λ in the exact case $\delta = 0$.

***Theorem* 3.** *For any nonnegative integer sequence $\{M_t\}_{t \in \mathbb{N}}$, if the approximation calculations are exact, i.e. $\delta = 0$, in Algorithm 1, then for any $v^0 \in \mathcal{V}_{\mathscr{B}}$, (a) the sequence $\{v^t\}$ generated by $v^{t+1} = \mathscr{W}^{M_t} v^t$ converges monotonically and in norm to the robust team-optimal value function v^*; (b) the GSraMPI algorithm terminates within a finite number of iterations with an ϵ-robust team-optimal policy $(d^\epsilon)^\infty$ and its corresponding value function v^ϵ satisfies $\|v^\epsilon - v^*\| < \epsilon$; and (c) let d_{v^t} and d_{v^*} be the v^t-improving and v^*-improving decision rules, respectively. Then, there exists a $P^*_{d_{v^t}} \in \mathcal{P}^*_{d_{v^t}}$ and a $P'_{d_{v^*}} \in \mathcal{P}_{d_{v^*}}$ such that $\|v^{t+1} - v^*\| \leq \left(\|{Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}} - {Q'_{d_{v^*}}}^{-1} R'_{d_{v^*}}\| \frac{1-\lambda^{M_t}}{1-\lambda} + \lambda^{M_t+1}\right) \|v^t - v^*\|$, where $(Q^*_{d_{v^t}}, R^*_{d_{v^t}})$ and $(Q'_{d_{v^*}}, R'_{d_{v^*}})$ are the GS regular splitting of $I - \lambda P^*_{d_{v^t}}$ and $I - \lambda P'_{d_{v^*}}$, respectively. In particular, if $\lim_{t\to\infty} \|P^*_{d_{v^t}} - P'_{d_{v^*}}\| = 0$, then there exists a $K \in \mathbb{N}$ for any $\epsilon > 0$ such that $\|v^{t+1} - v^*\| \leq \left(\epsilon + \lambda^{M_t+1}\right) \|v^t - v^*\|$ for any $t \geq K$.*

Proof: We first prove part (a). Let $\{y^t\}$ and $\{\omega^t\}$ be the sequences generated by $y^{t+1} = \mathscr{Y} y^t$ and $\omega^{t+1} = \mathscr{U}^{M_t} \omega^t$, respectively, where $y^0 = \omega^0 = v^0$. In the following, we show by induction that $v^t \in \mathcal{V}_{\mathscr{B}}$, $v^{t+1} \geq v^t$, and $\omega^t \geq v^t \geq y^t$ for $\forall t \in \mathbb{N}$.

First, when $t = 0$, according to the assumption we have $v^0 \in \mathcal{V}_{\mathscr{B}}$ and $y^0 = \omega^0 = v^0$. Moreover, Lemma 4 implies that $v^1 = \mathscr{W}^{M_0} v^0 \geq \mathscr{Y} v^0 \geq v^0$. Consequently, the induction hypothesis holds for $t = 0$. Suppose now that it is satisfied when $t = \kappa$. Then, applying Lemma 5 leads to $v^{\kappa+1} = \mathscr{W}^{M_\kappa} v^\kappa \in \mathcal{V}_{\mathscr{B}}$. From the definition of $\mathscr{W}^{M_{\kappa+1}}$, furthermore, one can get $v^{\kappa+2} = v^{\kappa+1} + \sum_{\varsigma=0}^{M_{\kappa+1}} \left({Q^*_{d_{v^{\kappa+1}}}}^{-1} R^*_{d_{v^{\kappa+1}}}\right)^{\varsigma} (\mathscr{B} v^{\kappa+1}) \geq v^{\kappa+1}$. Since $\omega^\kappa \geq v^\kappa \geq y^\kappa$ according to the hypothesis, it follows from Lemma 4 that $\omega^{\kappa+1} = \mathscr{U}^{M_\kappa} \omega^\kappa \geq \mathscr{W}^{M_\kappa} v^\kappa = v^{\kappa+1} \geq \mathscr{Y} y^\kappa = y^{\kappa+1}$. Thus, the induction hypothesis is satisfied for $t = \kappa + 1$. On the other hand, since Lemma 1 and Lemma 3 show that both y^t and ω^t converge in norm to v^* when $t \to \infty$, it follows that v^t converges in norm to v^* when $t \to \infty$. This establishes part (a).

We proceed to prove part (b). Since the iterative scheme of GSraMPI is given by $v^{t+1} = \mathscr{W}^{M_t} v^t$ as shown in (24) and part (a) has shown that $\{v^t\}$ is convergent, the termination condition in step 3(a) will be satisfied after a finite number of iterations by noticing $v^{t+1} = (\mathscr{T}_{(d_{v^t}, P^*_{d_{v^t}})})^{M_t+1} v^t \geq u_0^t = \mathscr{T}_{(d_{v^t}, P^*_{d_{v^t}})} v^t = \mathscr{Y} v^t \geq v^t$. Suppose now that the GSraMPI algorithm terminates at $t = N$, and outputs a policy $(d^\epsilon)^\infty$. From the step 2(a) of Algorithm 1, we know that d^ϵ is the v^N-improving decision rule. Moreover, from (23) and (25), one can find that $u_0^N = \mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N$ and the value function v^ϵ corresponding to the policy $(d^\epsilon)^\infty$ is $(\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})})^\infty v^N$, where $P^*_{d^\epsilon} \in \mathcal{P}^*_{d^\epsilon}$ is the d^ϵ-decreasing transition probability matrix with respect to v^N. Since the algorithm terminates at $t = N$, i.e. $\|u_0^N - v^N\| = \|\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N - v^N\| < (1 - \lambda)\epsilon/2\lambda$, $\|\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N - v^*\| = \|\mathscr{Y} v^N - v^*\| \leq \|\mathscr{Y} v^N - (\mathscr{Y})^2 v^N\| + \|(\mathscr{Y})^2 v^N - v^*\| \leq \lambda \|v^N - \mathscr{Y} v^N\| + \lambda \|\mathscr{Y} v^N - v^*\| = \lambda \|v^N - \mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N\| + \lambda \|\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N - v^*\|$. After re-arrangement, it leads to $\|\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N - v^*\| \leq \lambda \|v^N - \mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N\| / (1 - \lambda) < \epsilon/2$. On the other hand, since $(\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})})^\infty v^N - v^N = \left(\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N - v^N\right) + \left((\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})})^2 v^N - \mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N\right) + \cdots$, applying the contraction property of $\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})}$ yields $\|(\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})})^\infty v^N - \mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N\| \leq \lambda \|(\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})})^\infty v^N - v^N\| \leq \lambda \sum_{\imath=0}^{\infty} \lambda^\imath \|\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N - v^N\| < \epsilon/2$. Therefore, $\|v^\epsilon - v^*\| \leq \|(\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})})^\infty v^N - \mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N\| + \|\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N - v^*\| < \epsilon$.

Finally, part (c) can be obtained by the following derivation. First, from Proposition 1, we know that there exists a $P'_{d_{v^*}} \in \mathcal{P}_{d_{v^*}}$ such that $\mathscr{B} v^t \geq \mathscr{B} v^* + ({Q'_{d_{v^*}}}^{-1} R'_{d_{v^*}} - I)(v^t - v^*)$, where $d_{v^*} \in \mathcal{D}_{v^*}$ is the v^*-improving decision rule and $(Q'_{d_{v^*}}, R'_{d_{v^*}})$ is the GS regular splitting of $I - \lambda P'_{d_{v^*}}$. Note that $\mathscr{B}v^* = 0$. Then, from the definition of $\mathscr{W}^{M_t}$, we have $0 \le v^* - v^{t+1} = v^* - \mathscr{W}^{M_t} v^t = v^* - v^t - \sum_{\varsigma=0}^{M_t} \left({Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}} \right)^{\varsigma} (\mathscr{B}v^t) \le v^* - v^t + \sum_{\varsigma=0}^{M_t} \left({Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}} \right)^{\varsigma} \left({Q'_{d_{v^*}}}^{-1} R'_{d_{v^*}} - I \right) (v^* - v^t) = \left({Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}} - {Q'_{d_{v^*}}}^{-1} R'_{d_{v^*}} \right) \sum_{\varsigma=0}^{M_t - 1} \left({Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}} \right)^{\varsigma} (v^t - v^*) - \left({Q'_{d_{v^*}}}^{-1} R'_{d_{v^*}} \right) \left({Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}} \right)^{M_t} (v^t - v^*)$. Taking norms on both sides leads to $\|v^* - v^{t+1}\| \le \|{Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}} - {Q'_{d_{v^*}}}^{-1} R'_{d_{v^*}}\| \frac{1-\lambda^{M_t}}{1-\lambda} \|v^t - v^*\| + \lambda^{M_t+1} \|v^t - v^*\| = \left(\|{Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}} - {Q'_{d_{v^*}}}^{-1} R'_{d_{v^*}}\| \frac{1-\lambda^{M_t}}{1-\lambda} + \lambda^{M_t+1} \right) \|v^t - v^*\|$. In particular, if $\lim_{t\to\infty} \|P^*_{d_{v^t}} - P'_{d_{v^*}}\| = 0$, then $\lim_{t\to\infty} \|{Q^*_{d_{v^t}}}^{-1} R^*_{d_{v^t}} - {Q'_{d_{v^*}}}^{-1} R'_{d_{v^*}}\| = 0$. This implies that there exists a $K \in \mathbb{N}$ for any $\epsilon > 0$ such that $\|v^{t+1} - v^*\| \le \left(\epsilon + \lambda^{M_t+1}\right) \|v^t - v^*\|$ for any $t \ge K$. ∎

Next, we extend these convergence results to the inexact case where the tolerance of approximation calculations in Algorithm 1 is within $\lambda\delta$ for $0 \le \delta < (1-\lambda)^2\epsilon/2\lambda(1+\lambda)$. Similarly, to distinguish from the exact case $\delta = 0$, we denote the sequence generated by Algorithm 1 in the inexact case by $\{\tilde{v}^t\}$. Moreover, for those intermediate variables given in (6) and (8), we use $\tilde{u}_0^t(s)$ and $\tilde{u}_{\varsigma+1}^t(s)$ to represent $\tilde{u}_0^t(s) = \max_{a\in\mathcal{A}} \tilde{\rho}_{(s,a)}(\tilde{v}^t)$ and $\tilde{u}_{\varsigma+1}^t(s) = \tilde{\rho}_{(s,\tilde{d}_{t+1}(s))}(\tilde{u}_\varsigma^t)$ in the inexact case, respectively, where $\tilde{d}_{t+1}(s) \in \arg\max_{a\in\mathcal{A}} \tilde{\rho}_{(s,a)}(\tilde{v}^t)$ for $\forall s \in \mathcal{S}$; while in the exact case we still use the notations $\{v^t\}$, $u_0^t(s) = \max_{a\in\mathcal{A}} \tilde{\rho}_{(s,a)}(v^t)$, and $u_{\varsigma+1}^t(s) = \tilde{\rho}_{(s,d_{t+1}(s))}(u_\varsigma^t)$ for $\forall s \in \mathcal{S}$, where $d_{t+1}(s) \in \arg\max_{a\in\mathcal{A}} \tilde{\rho}_{(s,a)}(v^t)$ and $\rho_{(s,a)}(\cdot) = \tilde{\rho}_{(s,a)}(\cdot)$ for $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$ in view of $\delta = 0$. Then, the convergence of the GSraMPI algorithm in the inexact case can be given by the following theorem.

***Theorem* 4.** *For any nonnegative integer sequence $\{M_t\}_{t\in\mathbb{N}}$, if the tolerance of approximation calculations is within $\lambda\delta$ for $\delta \in [0, (1-\lambda)^2\epsilon/2\lambda(1+\lambda))$ in Algorithm 1, then, for any $v^0 \in \mathcal{V}_{\mathscr{B}}$, (a) if $\tilde{v}^0 = v^0$, $v^t - \theta\mathbf{1} \le \tilde{v}^t \le v^t + \theta\mathbf{1}$ and $u_\varsigma^t - \theta\mathbf{1} \le \tilde{u}_\varsigma^t \le u_\varsigma^t + \theta\mathbf{1}$, $\varsigma = 0, 1, \ldots, M_t$, hold for $\forall t \in \mathbb{N}$, where $\theta := \lambda\delta/(1-\lambda) \in [0, \epsilon/2)$; (b) the GSraMPI algorithm terminates within a finite number of iterations with an ϵ-robust team-optimal policy $(d^\epsilon)^\infty$ and its corresponding value function v^ϵ satisfies $\|v^\epsilon - v^*\| < \epsilon$.*

Proof: We prove part (a) by using a similar argument to Lemma 2. First, note that $v^0 = \tilde{v}^0$ according to the assumption, and $u_0^0(s)$ and $\tilde{u}_0^0(s)$ for $\forall s \in \mathcal{S}$ are calculated by the same formulas (5), (6), and (9) as in the degenerated case GSraVI. Then, applying Lemma 2 leads to $u_0^0 - \theta\mathbf{1} \le \tilde{u}_0^0 \le u_0^0 + \theta\mathbf{1}$. Moreover, since from (9), $\rho_{(s,a)}(v^0) = \rho_{(s,a)}(\tilde{v}^0)$ for $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$ in view of $v^0 = \tilde{v}^0$, and from (5), for any given $s \in \mathcal{S}$, the approximation error between $\tilde{\rho}_{(s,a)}(\tilde{v}^0)$ and $\rho_{(s,a)}(\tilde{v}^0)$ for $\forall a \in \mathcal{A}$ does not affect the selection of the maximizing actions, $\arg\max_{a\in\mathcal{A}} \tilde{\rho}_{(s,a)}(\tilde{v}^0) = \arg\max_{a\in\mathcal{A}} \rho_{(s,a)}(v^0)$ for $\forall s \in \mathcal{S}$. Without loss of generality, we select $\tilde{d}_1(s) = d_1(s)$ for $\forall s \in \mathcal{S}$. Then, from (9), one can get that the set of minimizing transition probability distributions are the same for $\rho_{(s,a)}(v^0)$ and $\rho_{(s,a)}(\tilde{v}^0)$, $\forall s \in \mathcal{S}$ and $a = \tilde{d}_1(s) = d_1(s)$. As such, it follows from (10) that $\rho_{(s^1,a)}(u_0^0 - \theta\mathbf{1}) = \rho_{(s^1,a)}(u_0^0) - \lambda\theta \le \rho_{(s^1,a)}(\tilde{u}_0^0) \le \rho_{(s^1,a)}(u_0^0) + \lambda\theta = \rho_{(s^1,a)}(u_0^0 + \theta\mathbf{1})$ for $a = \tilde{d}_1(s^1) = d_1(s^1)$. In view of $u_1^0(s^1) = \rho_{(s^1,a)}(u_0^0)$ for $a = d_1(s)$ from (8) in the exact case, it yields $u_1^0(s^1) - \lambda\theta \le \rho_{(s^1,a)}(\tilde{u}_0^0) \le u_1^0(s^1) + \lambda\theta$. On the other hand, from (7), we have $\tilde{\rho}_{(s^1,a)}(\tilde{u}_0^0) - \lambda\delta \le \rho_{(s^1,a)}(\tilde{u}_0^0) \le \tilde{\rho}_{(s^1,a)}(\tilde{u}_0^0) + \lambda\delta$ for $a = \tilde{d}_1(s)$, which implies that $\tilde{u}_1^0(s^1) - \lambda\delta \le \rho_{(s^1,a)}(\tilde{u}_0^0) \le \tilde{u}_1^0(s^1) + \lambda\delta$ in view of $\tilde{u}_1^0(s^1) = \tilde{\rho}_{(s^1,a)}(\tilde{u}_0^0)$ from (8) in the inexact case. Consequently, we have $u_1^0(s^1) - \lambda(\theta+\delta) = u_1^0(s^1) - \theta \le \tilde{u}_1^0(s^1) \le u_1^0(s^1) + \theta = u_1^0(s^1) + \lambda(\theta+\delta)$. Using this inequality and $u_0^0(s^j) - \theta \le \tilde{u}_0^0(s^j) \le u_0^0(s^j) + \theta$, $j = 2, 3, \ldots, m$, one can similarly get $u_1^0(s^2) - \theta \le \tilde{u}_1^0(s^2) \le u_1^0(s^2) + \theta$ from (7), (8), and (10). By recursion, applying the same argument for s^j, $j = 3, 4, \ldots, m$, will lead to $u_1^0(s^j) - \theta \le \tilde{u}_1^0(s^j) \le u_1^0(s^j) + \theta$. Then, $u_1^0 - \theta\mathbf{1} \le \tilde{u}_1^0 \le u_1^0 + \theta\mathbf{1}$. Leveraging this inequality and applying the same argument for $\varsigma = 2, 3, \ldots, M_0$ by recursion, one can further get $u_\varsigma^0 - \theta\mathbf{1} \le \tilde{u}_\varsigma^0 \le u_\varsigma^0 + \theta\mathbf{1}$. Therefore, the induction hypothesis is satisfied for $t = 0$.

Suppose now that the induction hypothesis is satisfied for $t = \kappa$, and we show that it holds for $t = \kappa+1$. First, according to the hypothesis when $t = \kappa$, we have $u_\varsigma^\kappa - \theta\mathbf{1} \le \tilde{u}_\varsigma^\kappa \le u_\varsigma^\kappa + \theta\mathbf{1}$, $\varsigma = 0, 1, \ldots, M_\kappa$. Since the setup in step 3(f) implies that $v^{\kappa+1} = u_{M_\kappa}^\kappa$ in the exact case and $\tilde{v}^{\kappa+1} = \tilde{u}_{M_\kappa}^\kappa$ in the inexact case, $v^{\kappa+1} - \theta\mathbf{1} \le \tilde{v}^{\kappa+1} \le v^{\kappa+1} + \theta\mathbf{1}$. Adopting a similar argument process to the above case for $t = 0$, one can first obtain $u_0^{\kappa+1} - \theta\mathbf{1} \le \tilde{u}_0^{\kappa+1} \le u_0^{\kappa+1} + \theta\mathbf{1}$ by applying Lemma 2, and then by recursion, one can further show that $u_\varsigma^{\kappa+1} - \theta\mathbf{1} \le \tilde{u}_\varsigma^{\kappa+1} \le u_\varsigma^{\kappa+1} + \theta\mathbf{1}$ holds for $\varsigma = 1, 2, \ldots, M_{\kappa+1}$. As a result, part (a) is established.

We next prove part (b). First, when $\delta = 0$, from the proof of Theorem 3(b), one can obtain that there exists a positive integer N for any $\epsilon_1 > 0$ such that $\|u_0^N - v^N\| < (1-\lambda)\epsilon_1/2\lambda$ and $\|v^{N+1} - v^*\| = \|(\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})})^{M_N+1} v^N - v^*\| \le \|\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N - v^*\| < \epsilon_1/2$. Therein, $v^N \le \mathscr{Y} v^N = \mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})} v^N \le (\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})})^2 v^N \le \cdots \le (\mathscr{T}_{(d^\epsilon, P^*_{d^\epsilon})})^{M_N+1} v^N = v^{N+1} \le v^*$ is applied in view of $v^N \in \mathcal{V}_{\mathscr{B}}$. Moreover, from part (a), we have $\|\tilde{u}_0^N - \tilde{v}^N\| \le \|\tilde{u}_0^N - u_0^N\| + \|u_0^N - v^N\| + \|v^N - \tilde{v}^N\| \le \|u_0^N - v^N\| + 2\theta < (1-\lambda)\epsilon_1/2\lambda + 2\lambda\delta/(1-\lambda)$. Given that ϵ_1 is arbitrary, we therefore select a specific ϵ_1 such that $\epsilon_1 \le \epsilon - 2\lambda(1+\lambda)\delta/(1-\lambda)^2$ for any $\epsilon > 0$. As a result, $\|\tilde{u}_0^N - \tilde{v}^N\| < (1-\lambda)\epsilon/2\lambda - \delta$. It implies that the GSraMPI algorithm will terminate at $t = N$ in the inexact case. Suppose that the algorithm terminates with outputting a policy $(\tilde{d}^\epsilon)^\infty$, and denote its corresponding value function under the worst-case transition probability matrix by $\tilde{v}^\epsilon$. From (25), one can find that $\tilde{v}^\epsilon$ can be regard as the value $\tilde{v}^{N+1}$ generated by Algorithm 1 at $t = N+1$ by setting $M_N \to \infty$ and $\delta = 0$ when $t = N$. Therefore, from the result in part (a), one can get $\|v^\epsilon - v^*\| = \|\tilde{v}^{N+1} - v^*\| \le \|\tilde{v}^{N+1} - v^{N+1}\| + \|v^{N+1} - v^*\| < \theta + \epsilon_1/2 < \epsilon$. ∎

In the following section, we illustrate our theoretical results by simulations.

IV. Numerical studies: sequential robust social dilemmas

To demonstrate the effectiveness of GSraMPI numerically, we here generalize the stochastic game model of social dilemmas in [45] to the scenario of incomplete information, and we refer to it as sequential robust social dilemmas. Formally, we consider an n-player stochastic game where every player can only select one of the two actions, cooperation (C) and defection (D), from the action set $\mathcal{A} = \{\mathrm{C}, \mathrm{D}\}$. As the consequence of the action choices of players, the game environment will transit from the current state to a new one at the next time. More specifically, when there are $\hbar \in \{0, 1, 2, \ldots, n\}$ players choosing action C in the game-environmental state $s^k \in \mathcal{S}$, at the next time, the game-environmental state will change to $s^l \in \mathcal{S}$ with probability $p(s^l \mid s^k, \hbar)$, where $p(s^l \mid s^k, \hbar)$ lies in a discrete finite set $\mathcal{P}_\hbar(l \mid k)$. Subsequently, those players who choose action C (resp. D) will get a payoff $\mathfrak{a}_\hbar(s^l \mid s^k) \in \mathbb{R}$ (resp. $\mathfrak{b}_\hbar(s^l \mid s^k) \in \mathbb{R}$) as a function of $\hbar$, s^k, and s^l. To adhere to the existence of dilemmas, we assume as in the canonical multi-player social dilemmas [46], [47] that (i) $\mathfrak{a}_{\hbar+1}(s^l \mid s^k) \geq \mathfrak{a}_\hbar(s^l \mid s^k)$ and $\mathfrak{b}_{\hbar+1}(s^l \mid s^k) \geq \mathfrak{b}_\hbar(s^l \mid s^k)$; (ii) $\mathfrak{b}_\hbar(s^l \mid s^k) > \mathfrak{a}_\hbar(s^l \mid s^k)$; (iii) $\mathfrak{a}_n(s^l \mid s^k) > \mathfrak{b}_0(s^l \mid s^k)$ for all $\hbar$ and $\forall s^k, s^l \in \mathcal{S}$. Condition (i) states that players' payoffs increase with the number of C players in the group, whereas condition (ii) implies that within any mixed group, those C players always have a strictly lower payoff than that of those D players. These two conditions indicate that taking action C is an altruistic behavior, based on the 'individual-centered' interpretation of altruism [48]. By contrast, condition (iii) shows that mutual cooperation is more beneficial than mutual defection. Hence, for maximizing the gains of the whole group, all players should uniformly choose C. However, under the hypothesis of Homo economicus, each rational player will be tempted by myopic interests to take action D, thereby leading to the existence of social dilemmas. By extending the game model of social dilemmas to incomplete information, the current model incorporates the prototypical multi-player social dilemmas [46], [47] as an extreme case in which the set of the game-environmental states is a singleton. Also, since the uncertainty of transition probabilities will result in the ambiguity of players' payoffs, it generalizes the social dilemmas to the scenario of robust games [35], [36]

Using this extended model, we apply the GSraMPI algorithm to seek the robust team-optimal policy for a sequential robust social dilemma with three game-environmental states, $\mathcal{S} = \{s^1, s^2, s^3\}$. Specifically, in the state s^1, players play a public goods game (PGG) and the payoffs of players are calculated by $\mathfrak{a}_\hbar(s^l \mid s^1) = \hbar r_{s^l} c/n - c$ and $\mathfrak{b}_\hbar(s^l \mid s^1) = \hbar r_{s^l} c/n$ for $\forall s^l \in \mathcal{S}$, where c is the cost of cooperation and $r_{s^l} \in (c, n)$ is the synergy factor dependent on the state s^l at the next time. In the state s^2, an n-player stag hunter game (nSH) is played, where the payoffs of players are calculated by $\mathfrak{a}_\hbar(s^l \mid s^2) = \hbar r_{s^l} c/n - c$ and $\mathfrak{b}_\hbar(s^l \mid s^2) = \hbar r_{s^l} c/n$ if $\hbar$ is no less than the threshold Z; and otherwise $\mathfrak{a}_\hbar(s^l \mid s^2) = -c$ and $\mathfrak{b}_\hbar(s^l \mid s^2) = 0$. In the state s^3, an n-player snowdrift game (nSD) is played, and players' payoffs are calculated by $\mathfrak{a}_\hbar(s^l \mid s^3) = \vartheta_{s^l} - c/\hbar$ and $\mathfrak{b}_\hbar(s^l \mid s^3) = \vartheta_{s^l}$ if $\hbar > 0$, and otherwise $\mathfrak{b}_\hbar(s^l \mid s^3) = 0$, where ϑ_{s^l}, as a function of the state s^l at the next time, is the benefit of players when there exists at least one player choosing action C in the group. (see [45] and references therein for more details of these three games.) The simulation results are shown in Fig. 1 and Table I, where the model parameters of games are given by $r_{s^1} = 1.5$, $r_{s^2} = 1.8$, $r_{s^3} = 2.2$, $\vartheta_{s^l} = r_{s^l}$ for $\forall s^l \in \mathcal{S}$, $n = 3$, $c = 1.0$, and $\mathcal{P}_\hbar(l \mid k) = \{1 - \mu\hbar\}_{\mu \in \{0.1, 0.2, 0.3\}}$ if $k = l$ and otherwise $\mathcal{P}_\hbar(l \mid k) = \{\mu\hbar/2\}_{\mu \in \{0.1, 0.2, 0.3\}}$. Moreover, the parameters in Algorithm 1 are $\epsilon = 10^{-5}$ and $\delta \approx (1-\lambda)^2\epsilon/2\lambda(1+\lambda)$, where a fixed $\lambda = 0.97$ is adopted in Fig. 1 while different λ values are used in Table I. By analyzing the results in Fig. 1 and Table I, one can find that both GSraMPI and GSraVI are able to find the robust team-optimal policy effectively, and their convergence rates are faster than rVI and rMPI, respectively.

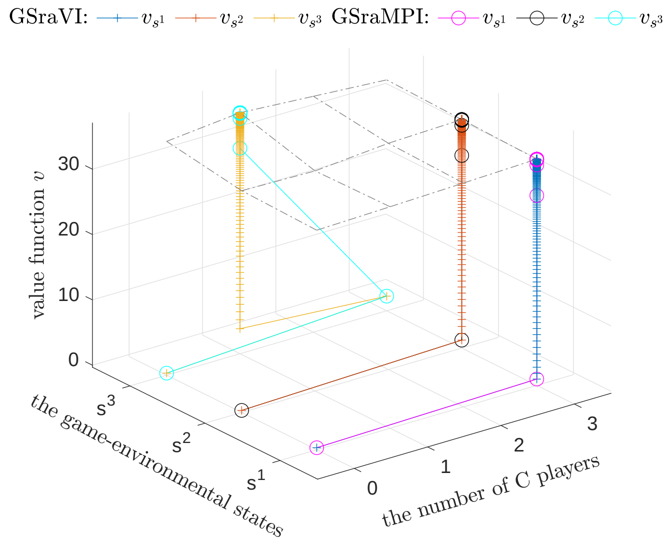

Fig. 1: The iterative process of the value functions generated by GSraVI and GSraMPI. The dashed lattice illustrates the values of $\rho_{(s,a)}(v^t)$ for $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$ at the termination time.

TABLE I: The number of iterations for different algorithms to find the ϵ-robust team-optimal policy.

Algorithms	the magnitude of λ				
	0.95	0.96	0.97	0.98	0.99
rVI	298	380	519	802	1679
GSraVI	**258**	**328**	**446**	**690**	**1442**
rMPI	7	9	12	17	34
GSraMPI	**7**	**8**	**10**	**15**	**30**

V. Concluding Remarks

In this paper, by relaxing the complete information assumption of team stochastic games, we have proposed a distribution-free model of robust team stochastic games, in which players do not know the true transition probability distributions and/or players' payoffs, but rather are commonly aware of an uncertainty set. Without assuming a prior probability distribution over the uncertainty set, we have considered that players adopt a robust optimization method to make decisions. That is, players resort to seek a joint policy such that the expected total discounted return of the team average payoffs can be maximized with respect to the worst case of the uncertain parameters. To characterize such a policy, we have offered an alternative solution concept for the robust team stochastic games, and meanwhile to find the solution, we have developed a learning algorithm named Gauss-Seidel robust approximate modified policy iteration. Under mild conditions, the convergence of the algorithm has been proven as well. By calculating the convergence rates, moreover, we have demonstrated that the algorithm could be faster than those robust dynamic programming algorithms. Finally, by generalizing the game model of social dilemmas to the incomplete-information scenario, the effectiveness of the algorithm has been verified via numerical simulations. A potential and significant research direction in the future is to develop decentralized learning algorithms to solve the optimization problems associated with robust stochastic games in the context of networked multi-agent systems [49], [50].

REFERENCES

[1] L. S. Shapley, "Stochastic games," *Proc. Natl Acad. Sci. USA*, vol. 39, no. 10, pp. 1095–1100, 1953.

[2] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proc. 11th Int. Conf. Mach. Learn.*, pp. 157–163, 1994.

[3] D. P. Bertsekas, *Dynamic programming and optimal control*, vol. I. Belmont, USA: Athena Scientific, 4th ed., 2017.

[4] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst. Man Cybernet. C*, vol. 38, no. 2, pp. 156–172, 2008.

[5] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *arXiv preprint arXiv:1911.10635*, 2019.

[6] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, "A survey and critique of multiagent deep reinforcement learning," *Auton. Agents Multi-Agent Syst.*, vol. 33, no. 6, pp. 750–797, 2019.

[7] T. Başar and G. J. Olsder, *Dynamic noncooperative game theory*. SIAM, 1998.

[8] M. J. Osborne and A. Rubinstein, *A course in game theory*. Cambridge, MA: MIT Press, 1994.

[9] J. M. Smith, *Evolution and the Theory of Games*. Cambridge, UK: Cambridge University Press, 1982.

[10] J. R. Riehl, P. Ramazi, and M. Cao, "A survey on the analysis and control of evolutionary matrix games," *Annu. Rev. Control*, vol. 45, pp. 87–106, 2018.

[11] E. Solan and N. Vieille, "Stochastic games," *Proc. Natl Acad. Sci. USA*, vol. 112, no. 45, pp. 13743–13746, 2015.

[12] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[13] X. Wang and T. Sandholm, "Reinforcement learning to play an optimal Nash equilibrium in team Markov games," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 15, pp. 1603–1610, 2002.

[14] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proc. 15th Nat. Conf. Artif. Intell. 10th Conf. Innov. Appl. Artif. Intell.*, pp. 746–752, 1998.

[15] L. Panait and S. Luke, "Cooperative multi-agent learning: The state of the art," *Auton. Agents Multi-Agent Syst.*, vol. 11, no. 3, pp. 387–434, 2005.

[16] P. Stone and M. Veloso, "Multiagent systems: A survey from a machine learning perspective," *Auton. Robots*, vol. 8, no. 3, pp. 345–383, 2000.

[17] H. S. Witsenhausen, "On information structures, feedback and causality," *SIAM J. Control*, vol. 9, no. 2, pp. 149–160, 1971.

[18] Y.-C. Ho, "Team decision theory and information structures in optimal control problems–part I," *IEEE Trans. Autom. Control*, vol. 17, no. 1, pp. 15–22, 1972.

[19] J. R. Marden and J. S. Shamma, "Game theory and control," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 1, pp. 105–134, 2018.

[20] Y. Shoham, R. Powers, and T. Grenager, "If multi-agent learning is the answer, what is the question?," *Artif. Intell.*, vol. 171, no. 7, pp. 365–377, 2007.

[21] F. Salehisadaghiani and L. Pavel, "Distributed Nash equilibrium seeking: A gossip-based algorithm," *Automatica*, vol. 72, pp. 209–216, 2016.

[22] K. Lu, G. Li, and L. Wang, "Online distributed algorithms for seeking generalized Nash equilibria in dynamic environments," *IEEE Trans. Autom. Control*, vol. 66, no. 5, pp. 2289–2296, 2021.

[23] Y. Tian, P. Jia, A. Mirtabatabaei, L. Wang, N. E. Friedkin, and F. Bullo, "Social power evolution in influence networks with stubborn individuals," *IEEE Trans. Autom. Control*, 2021.

[24] A. M. Fink, "Equilibrium in a stochastic n-person game," *J. Sci. Hiroshima Univ. Ser. A-I Math.*, vol. 28, no. 1, pp. 89–93, 1964.

[25] A. G. Barto, "Reinforcement learning: Connections, surprises, and challenge," *AI Mag.*, vol. 40, no. 1, pp. 3–15, 2019.

[26] A. Nilim and L. El Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Oper. Res.*, vol. 53, no. 5, pp. 780–798, 2005.

[27] G. N. Iyengar, "Robust dynamic programming," *Math. Oper. Res.*, vol. 30, no. 2, pp. 257–280, 2005.

[28] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.

[29] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability," *Comput. Sci. Rev.*, vol. 37, p. 100270, 2020.

[30] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin, "A general safety framework for learning-based control in uncertain robotic systems," *IEEE Trans. Autom. Control*, vol. 64, no. 7, pp. 2737–2752, 2019.

[31] S. Russell, D. Dewey, and M. Tegmark, "Research priorities for robust and beneficial artificial intelligence," *AI Mag.*, vol. 36, no. 4, pp. 105–114, 2015.

[32] J. C. Harsanyi, "Games with incomplete information played by "Bayesian" players, I–III," *Manage. Sci.*, vol. 14, no. 3,5,7, pp. 159–182,320–334,486–502, 1967,1968.

[33] B. Holmström and R. B. Myerson, "Efficient and durable decision rules with incomplete information," *Econometrica*, vol. 51, no. 6, pp. 1799–1819, 1983.

[34] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust optimization*. Princeton, NJ: Princeton University Press, 2009.

[35] M. Aghassi and D. Bertsimas, "Robust game theory," *Math. Program., Ser. B*, vol. 107, no. 1, pp. 231–273, 2006.

[36] E. Kardeş, F. Ordóñez, and R. W. Hall, "Discounted robust stochastic games and an application to queueing control," *Oper. Res.*, vol. 59, no. 2, pp. 365–382, 2011.

[37] K. Zhang, T. Sun, Y. Tao, S. Genc, S. Mallya, and T. Basar, "Robust multi-agent reinforcement learning with model uncertainty," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.

[38] D. L. Kaufman and A. J. Schaefer, "Robust modified policy iteration," *INFORMS J. Comput.*, vol. 25, no. 3, pp. 396–410, 2013.

[39] A. Tamar, S. Mannor, and H. Xu, "Scaling up robust MDPs using function approximation," in *Proc. 31st Int. Conf. Mach. Learn.*, pp. 181–189, 2014.

[40] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. MA, Belmont: Athena Scientific, 1996.

[41] H. S. Witsenhausen, "A counterexample in stochastic optimum control," *SIAM J. Control*, vol. 6, no. 1, pp. 131–147, 1968.

[42] D. Bertsekas, "Multiagent reinforcement learning: Rollout and policy iteration," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 2, pp. 249–272, 2021.

[43] R. Munos and C. Szepesvári, "Finite-time bounds for fitted value iteration," *J. Mach. Learn. Res.*, vol. 9, no. 5, pp. 815–857, 2008.

[44] B. Scherrer, M. Ghavamzadeh, V. Gabillon, B. Lesner, and M. Geist, "Approximate modified policy iteration and its application to the game of Tetris," *J. Mach. Learn. Res.*, vol. 16, no. 49, pp. 1629–1676, 2015.

[45] F. Huang, M. Cao, and L. Wang, "Learning enables adaptation in cooperation for multi-player stochastic games," *J. R. Soc. Interface*, vol. 17, no. 172, p. 20200639, 2020.

[46] C. Hilbe, B. Wu, A. Traulsen, and M. A. Nowak, "Cooperation and control in multiplayer social dilemmas," *Proc. Natl Acad. Sci. USA*, vol. 111, no. 46, pp. 16425–16430, 2014.

[47] A. Govaert and M. Cao, "Zero-determinant strategies in repeated multiplayer social dilemmas with discounted payoffs," *IEEE Trans. Autom. Control*, 2020.

[48] B. Kerr, P. Godfrey-Smith, and M. W. Feldman, "What is altruism?," *Trends Ecol. Evol.*, vol. 19, no. 3, pp. 135–140, 2004.

[49] J. R. Riehl and M. Cao, "Towards optimal control of evolutionary games on networks," *IEEE Trans. Autom. Control*, vol. 62, no. 1, pp. 458–462, 2016.

[50] F. Huang, X. Chen, and L. Wang, "Evolutionary dynamics of networked multi-person games: mixing opponent-aware and opponent-independent strategy decisions," *New J. Phys.*, vol. 21, no. 6, p. 063013, 2019.