

Article

Toward Convergence in Multi-Agent Reinforcement Learning: Best-Response Space Shrinking as a Sufficient Condition

Natchanon Manatphaiboon ^{1,*} , Akito Monden ^{1,*}  and Zeynep Yücel ² ¹ Graduate School of Environmental, Life, Natural Science and Technology, Okayama University, Okayama 700-8530, Japan² Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, 30123 Venice, Italy; zeynep.yucel@unive.it

* Correspondence: pb549wde@s.okayama-u.ac.jp (N.M.); monden@okayama-u.ac.jp (A.M.)

Abstract

We study convergence in multi-agent reinforcement learning (MARL) through the lens of sufficient conditions, using a single-point-of-failure analysis applied to multi-agent policy iteration integrated with linear programming (MAPI-LP), where results are proven for pure coordination games, and extension to broader settings is conjectured. We identify two sufficient conditions for convergence to Markov Perfect Equilibrium (MPE). The first is stability in best-response space that emerges from value monotonicity. The second, monotonic best-response space shrinking (MBRSS), is a novel condition requiring that each agent's best-response space contracts monotonically across iterations until it collapses to a stable space. Furthermore, we show that MBRSS does not necessarily imply monotonic improvement in this setting. However, value monotonicity and stability in best-response space imply each other when a complementary condition is applied. Building on this hierarchical relationship, we propose conjectures on sufficient condition relationships in both serial and parallel MARL. In addition, we propose conjectures on generalized MBRSS to arbitrary finite repeated games and validation of stability in best-response space. We further discuss connections between MBRSS and existing related frameworks, and outline directions toward a taxonomy of sufficient conditions for MARL convergence.

Keywords: sufficient condition; convergence analysis; Markov perfect equilibrium; multi-agent reinforcement learning; best-response dynamics

MSC: 68T05; 91A20; 91A26; 47H10; 90C05; 91A15



Academic Editor: Zhijia Zhao

Received: 17 April 2026

Revised: 3 June 2026

Accepted: 10 June 2026

Published: 15 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

MARL problems consist of three fundamental components: a game model defining the strategic environment, solution concepts specifying desired equilibria, and algorithms providing computational methods to reach these solutions [1]. From an algorithm development perspective, researchers typically pursue three primary directions: generalizing single-agent reinforcement learning algorithms to multi-agent settings [2–6], improving existing MARL algorithms for better reliability and stability [7–9], or adapting algorithms to new classes of game models or new solution concepts [10–12]. However, changing any component or a combination of them introduces non-stationarity or compatibility issues that prevent convergence. To address these problems, new sufficient conditions are needed to explain and predict when convergence will succeed or fail and to guide algorithm design and improvement.

To pursue this line of understanding, we begin from the generalization of policy iteration, the simplest extension of reinforcement learning to the multi-agent setting. We propose and analyze a multi-agent policy iteration with linear programming (MAPI-LP) in the context of repeated games, aimed at convergence to Markov Perfect Equilibria (MPE) under a sufficient condition and convergence analysis. The existence of MPE is guaranteed in any finite stochastic game, and thus also in repeated games [13]. However, MAPI-LP, a model of best-response dynamics that intuitively satisfies the definition of MPE, does not guarantee convergence (Section 3). Because of this, our approach adopts a single-point-of-failure perspective: by asking where and how MAPI-LP could fail, we identify sufficient conditions that ensure convergence.

In Section 4, first we explain why value monotonicity (Definition 12) is not generally sufficient for MPE convergence. Then we provide two main theoretical contributions. The first is stability in best-response space (Definition 13). The second is monotonic best-response space shrinking (MBRSS, Theorem 1), a condition requiring that each agent's best-response set decreases monotonically across iterations until it collapses to a stable space. We further show that MBRSS is not merely abstract, by showing that sequential MAPI-LP provably satisfies MBRSS in pure coordination games (Theorem 3), converging to a stable best-response space, where all of its members are MPE, in at most n iterations. Building on these conditions, we propose several conjectures related to MBRSS generalization and relationships between sufficient conditions, respectively (Section 5).

Finally, we discuss broader implications and future directions, including connections between MBRSS and existing frameworks, as well as directions for extending MBRSS to broader game classes and another application, and the development of a taxonomy of sufficient conditions for MARL convergence (Section 6). Together, these results provide both theoretical guarantees in simplified settings and a roadmap for investigating convergence in more complex multi-agent environments.

2. Preliminary

2.1. Sufficient Conditions and Where to Find Them

Sufficient conditions for convergence guarantee that a sequence, series, or iterative process will converge. In the context of MARL, such conditions typically involve assumptions, constraints, or specific properties of the components within the scope of learning towards a given solution concept. Examining these components through theoretical analysis or empirical observation can help identify sufficient conditions that may lead to significant improvements in algorithms.

For example, ref. [8] found that the full factorization of the action-value function used in VDN [14] is not necessary, and then proposed QMIX, which ensures a condition known as the Individual-Global-Max (IGM). Later, ref. [9] observed that both QMIX and VDN rely on two sufficient conditions for IGM: additivity and monotonicity. However, there exist tasks in which these conditions are not satisfied, motivating the development of QTRAN. Hence, depending on the interpretation of IGM, we can trace the evolution of algorithms within the value decomposition category. For example, ref. [15] stated that the sufficient condition for IGM sacrifices the completeness of the action-value function class, resulting in a parameter space structure that fails to guarantee optimality. Ref. [16] proposed a sufficient condition of identifiability on expected KL-divergence in an attempt to achieve equilibrium in a fully decentralized manner, and found that one of their algorithm's components shares similarities with IGM.

In contrast, sufficient conditions can also emerge from the properties of certain components. For example, ref. [17] developed an advantage estimation method based on a sufficient condition for policy invariance, a desirable property in which the optimal policy

remains unchanged under reward shaping. Ref. [18] investigated convergence probabilities with respect to the absorbing best-response graph, under the implication of a sufficient condition in weakly acyclic stochastic games (as studied by [19]). This graph corresponds to the best-response graph modeled as an absorbing Markov chain.

These works collectively highlight that sufficient conditions for convergence in MARL are often tied to specific structural assumptions, which may not hold across all tasks. When these assumptions fail, the algorithm in question can struggle to converge optimally, encouraging developments of alternative methods or improvements of existing methods. Examining single points of failure, where an algorithm is vulnerable to convergence failure, can offer valuable insight into identifying new sufficient conditions. In other words, by pinpointing the exact component or assumption that breaks down, we can better understand the necessities behind convergence and design more generalizable and robust algorithms. For example, along these lines, ref. [20] employed a strictly positive real property to establish a sufficient condition ensuring practical finite-time convergence for the stability of interconnected subsystems in a nonlinear control framework.

2.2. Multi-Agent Policy Iteration

The policy iteration is a classic reinforcement learning for Markov decision processes that alternates between two steps: policy evaluation ($\xrightarrow{E}$) and policy improvement ($\xrightarrow{I}$) [21]. Its greedy improvement step, which guarantees convergence to the optimal solution, naturally generalizes to best-response dynamics in multi-agent stochastic games, with Markov Perfect Equilibrium (MPE) [22] as the solution concept. For this generalization, we define $\xrightarrow{E}$, $\xrightarrow{I}$, and MPE as follows:

Definition 1. Given an input joint policy π , the process $\xrightarrow{E}$ computes the value function v^π using the Bellman operator. In other words, it finds v^π such that

$$\forall i, s, v_i^\pi(s) = \sum_a \pi(a | s) \sum_{s' \in S} \mathcal{P}(s' | s, a) [\mathcal{R}_i(s, a, s') + \gamma v_i^\pi(s')], \quad (1)$$

where $\mathcal{P}$ is the state transition probability and $\mathcal{R}_i$ is the reward function for agent i .

Definition 2. Given an input joint policy $\pi = (\pi_i)_i$, the process $\xrightarrow{I}$ outputs an updated joint policy π' , where each agent's policy π'_i is a state-wise best response to the others, i.e.,

$$\pi' \in \left\{ (\pi'_i)_i \mid \forall i, s, \forall \pi''_i, v_i^{(\pi'_i, \pi_{-i})}(s) \geq v_i^{(\pi''_i, \pi_{-i})}(s) \right\}. \quad (2)$$

Definition 3. A joint policy $\pi^* = (\pi_i^*, \pi_{-i}^*)$ is said to be (stationary) MPE if, for every agent i and state s ,

$$\forall \pi_i, v_i^{(\pi_i^*, \pi_{-i}^*)}(s) \geq v_i^{(\pi_i, \pi_{-i}^*)}(s), \quad (3)$$

where $v_i^\pi(s)$ denotes the expected return for agent i starting from state s under the joint policy π .

Following the proof structure of [13], the existence of an MPE in discounted finite stochastic games can be established using fixed-point theorems applied to the policy evaluation and improvement processes individually:

- Banach's fixed-point theorem ensures that $\xrightarrow{E}$ yields a unique value function v^π for each joint policy π .
- Kakutani's fixed-point theorem guarantees that $\xrightarrow{I}$ admits a fixed point, which corresponds to MPE under certain conditions on input joint policy π .

Therefore, starting from an arbitrary initial joint policy, alternating between $\xrightarrow{E}$ and $\xrightarrow{I}$ (as in [21]) can determine an MPE only in some cases, but does not guarantee convergence to an MPE.

2.3. Repeated Games

A repeated game (Definition 4) is a special class of stochastic games with n agents and a single state, where the state transitions back to itself at every time step, having only one value function to be calculated for each agent (Lemma 1).

Definition 4. A discounted finite repeated game with n agents and bounded rewards, where the state transitions back to itself every time step, is defined by $G = (I, \{A_i\}_i, \{\mathcal{R}_i\}_i, \gamma)$, which consists of:

- $I = \{1, \dots, n\}$, a finite set of agents,
- A_i , a finite set of actions available to agent i ; the set of joint actions is defined as $A = \prod_i A_i$.
- $\mathcal{R}_i : A \rightarrow [R_{\min}, R_{\max}]$, a deterministic reward function of agent i , where constants $R_{\min} \leq R_{\max}$.
- $\gamma \in [0, 1)$, a discount factor.

Since the only state always repeats to itself with a given joint policy, $\xrightarrow{E}$ (Definition 1) then evaluates the value function as follows.

Lemma 1. In a repeated game, from a given joint policy, if we follow process $\xrightarrow{E}$, the value function for agent i is

$$v_i^\pi = \frac{1}{1-\gamma} \sum_a \pi(a) \mathcal{R}_i(a). \quad (4)$$

Repeated games have been widely studied as a testbed for understanding the emergence of cooperative behavior in multi-agent systems. For example, ref. [23] investigates two-agent repeated games under Q-learning, demonstrating that the interplay between memory strength and future discounting determines whether agents converge to cooperative or defective policies. Their work highlights that even in the simplest repeated game setting, convergence behavior is highly sensitive to algorithm parameters and the structure of agent interactions. Hence, its simplicity makes it convenient and effective for analyzing the convergence of algorithms and provides a natural stepping stone toward generalizing to stochastic games with more steps. For these reasons, we choose the repeated game as our model of choice.

To support our analysis, Definitions 1–3 are generalized to the context of repeated games. In addition, we formally define the concepts of policy space and joint policy space as follows:

Definition 5. The policy space for agent i , denoted Δ_{A_i} , is the set of all possible probability distributions over actions in A_i that agent i can choose. Formally,

$$\Delta_i \doteq \left\{ (\pi_i(a_i))_{a_i \in A_i} \mid \sum_{a_i \in A_i} \pi_i(a_i) = 1 \text{ and } \forall a_i \in A_i, \pi_i(a_i) \geq 0 \right\}. \quad (5)$$

Thus, Δ_i is a standard simplex over the action set A_i .

Definition 6. The joint policy space, denoted Δ , is the set of all possible joint policies. Formally,

$$\Delta \doteq \prod_i \Delta_i. \quad (6)$$

3. Linear Programming Formulation

Solving $\xrightarrow{I}$ in Definition 2 is equivalent to solving a linear programming (LP) problem specified in Definition 7 for each agent i at iteration k . It seeks the best-response policy π_i^{k+1} to the opponents' policies π_{-i}^k , while ensuring a better response in expected value, and the output policy π_i^{k+1} lies within the policy space Δ_i .

Definition 7. For agent i , assuming that $v_i : \Delta \rightarrow \mathbb{R}$ is a continuous linear function, at iteration k , the LP problem, seeking best-response policy for agent i to π_{-i}^k , is defined as follows:

Maximize:

$$v_i^{(\pi_i^{k+1}, \pi_{-i}^k)} \quad (7)$$

Subject to:

$$v_i^{(\pi_i^{k+1}, \pi_{-i}^k)} \geq v_i^{\pi_i^k} \quad (8)$$

$$\sum_{a_i \in A_i} \pi_i^{k+1}(a_i) = 1 \quad (9)$$

$$\forall a_i \in A_i, \pi_i^{k+1}(a_i) \geq 0 \quad (10)$$

Remark 1. The LP defined in Definition 7 has the following properties:

- (a) The feasible region is a non-empty convex polytope defined by its linear constraints.
- (b) There exists an optimal solution at a vertex of the feasible region.
- (c) If multiple vertices yield the optimal solution, then all points within the convex hull of those vertices are also optimal.
- (d) The set of all optimal solutions is a non-empty compact convex set.

As a note for Remark 1, the feasible region is a closed subset of the probability simplex Δ_i , which is compact; hence the feasible region is compact. Since the objective function is continuous and linear, the set of all optimal policies is a closed convex subset of the feasible region; therefore it is compact and convex.

Remark 1(c) suggests that the convergence behavior depends not only on the trajectory of the update chain of $\xrightarrow{E}$ and $\xrightarrow{I}$, but also on the choice of best responses. We then separate policy improvement into two parts for each agent: identifying the set of best responses ($\xrightarrow{I_1}$) and selecting a specific best response ($\xrightarrow{I_2}$).

Definition 8. The best-response function solving the LP problem (Definition 7) for agent i is denoted by a set-valued function

$$BR_i : \Delta_{-i} \rightrightarrows \Delta_i, \quad BR_i(\pi_{-i}) = \operatorname{argmax}_{\pi_i} v_i^{(\pi_i, \pi_{-i})}, \quad (11)$$

Definition 9. At iteration k , consider agent i and a given joint policy π^k . The process $\xrightarrow{I_1}$ finds all vertices of optimal solution for LP defined in Definition 7 by vertex testing to obtain the set of best responses $BR_i(\pi_{-i}^k)$, which is a convex hull of these optimal vertices. In other words, we can call $BR_i(\pi_{-i}^k)$ the best-response space of agent i , at iteration k . By this process, it is guaranteed that $BR_i(\pi_{-i}^k)$ is convex and compact (Remark 1(d))

Definition 10. At iteration k , considering agent i , the process $\xrightarrow{I_2}$ arbitrarily selects a policy

$$\pi_i^{k+1} \in BR_i(\pi_{-i}^k). \quad (12)$$

By combining $\xrightarrow{E}$, $\xrightarrow{I_1}$, and $\xrightarrow{I_2}$, we obtain a version of multi-agent policy iteration integrated with LP, called MAPI-LP, that simultaneously solves LP for each agent.

Definition 11. MAPI-LP process at iteration k proceeds as follows:

$$\pi^k \xrightarrow{E} v^{\pi^k} \rightarrow \left\{ \begin{array}{l} (\pi_{-1}^k, v_1^{\pi^k}) \xrightarrow{I_1} BR_1(\pi_{-1}^k) \xrightarrow{I_2} \pi_1^{k+1} \\ \vdots \\ (\pi_{-i}^k, v_i^{\pi^k}) \xrightarrow{I_1} BR_i(\pi_{-i}^k) \xrightarrow{I_2} \pi_i^{k+1} \\ \vdots \\ (\pi_{-n}^k, v_n^{\pi^k}) \xrightarrow{I_1} BR_n(\pi_{-n}^k) \xrightarrow{I_2} \pi_n^{k+1} \end{array} \right\} \rightarrow \pi^{k+1}. \quad (13)$$

MAPI-LP generates a sequence of joint policies $\{\pi^k\}_{k \geq 0}$.

Since $\xrightarrow{I}$ is equivalent to the combination of $\xrightarrow{I_1}$ and $\xrightarrow{I_2}$, we know that it is possible to find an MPE with MAPI-LP, although convergence is not guaranteed, as stated in Section 2.2.

4. Sufficient Conditions for MPE Convergence

To address sufficient conditions, one can examine the outcome of each step in the process. If any step tends to violate the solution concept, in this case the MPE, then some form of sufficient condition must exist to prevent such violations and ensure convergence. In other words, it must be possible to find sufficient conditions on single points of convergence failure. For completeness, we assume the following:

Assumption 1. The joint policy sequence $\{\pi^k\}_{k \geq 0}$ is an infinite sequence and is generated by MAPI-LP, where the notion of $\{\pi^k\}$ refers to this.

Assumption 2. The function $v_i : \Delta \rightarrow \mathbb{R}$ is linear continuous for all i .

4.1. Monotonic Improvement

In our representation of best-response dynamics, each agent proposes a new policy that is a best response to the others, in terms of value function. Hence, one of the desired properties is value-monotonic (Definition 12).

Definition 12 (Value Monotonicity). The sequence $\{\pi^k\}_k$ is improving its value monotonically, if

$$\forall i, k \geq 0, v_i^{\pi^k} \leq v_i^{(\pi_i^{k+1}, \pi_{-i}^k)} \leq v_i^{\pi^{k+1}}. \quad (14)$$

Corollary 1. If $\{\pi^k\}$ is value-monotonic, $\{v_i^{\pi^k}\}_{k \geq 0}$ is a non-decreasing monotone sequence. By monotone convergence theorem, for all i , the sequence of value functions converges to a constant c_i , that is

$$\lim_{k \rightarrow \infty} v_i^{\pi^k} = \sup_k v_i^{\pi^k} = c_i. \quad (15)$$

Corollary 2. If $\{\pi^k\}$ is value-monotonic, it may hold that two successive joint policies π^k and π^{k+1} are distinct, i.e., $\pi^k \neq \pi^{k+1}$, while their value functions coincide, $v^{\pi^k} = v^{\pi^{k+1}}$.

By Corollary 1, value monotonicity guarantees that each agent value sequence $\{v_i^{\pi^k}\}_{k \geq 0}$ is monotone non-decreasing and converges to a constant limit c_i . In particular, Corollary 2 suggests that the sequence of joint policies may continue to move within a

set of value-equivalent policies even after the values have stabilized, and potentially cause a cycle. Therefore, value monotonicity establishes convergence in value space, but not sufficient convergence of the policies as an MPE.

4.2. Stability in Best-Response Space

Considering $\xrightarrow{I_1}$, we can collectively assemble its results as a sequence of best-response spaces. The idea of MPE can derive stability in best-response space (Definition 13) as another sufficient condition for MAPI-LP.

Definition 13. The sequence $\{\pi^k\}$ is said to be stable in best-response space Δ^* at iteration K , if there exists $\Delta^* = \prod_i \Delta_i^* \subseteq \Delta$, such that

$$\forall i, \Delta_i^* = BR_i(\pi_{-i}^K) = BR_i(\pi_{-i}^{K+1}) = \dots, \quad (16)$$

Since Δ_i^* does not change from $k \geq K$, then we can alternatively define it without iteration index as

$$\Delta_i^* = \{\pi_i \mid \pi_i \in BR_i(\pi_{-i}), \pi_{-i} \in \Delta_{-i}^*\}, \quad (17)$$

where $\Delta_{-i}^* = \prod_{j \in I \setminus \{i\}} \Delta_j^*$.

To see why Equation (17) is well-defined, note that once stability is reached at iteration K , the best-response space of each agent no longer changes. Therefore, any arbitrary selection of π_i^k where $k \geq K$ produces the same stable space Δ_i^* , and so $\Delta^* = \prod_i \Delta_i^*$. Since $\{\pi^k\}$ is infinite by Assumption 1, there exists an infinite subsequence of best responses beyond iteration K for each i , all mapping within Δ_i^* , which exhaustively generates Δ_i^* itself. This gives Equation (17) its self-referential character: the stable space is exactly the set of all best responses that agents can play against one another within Δ^* , without reference to any specific iteration index.

These lemmas establish the connection between stability in the best-response space and MPE. Lemma 2 shows that once the learning dynamics enter a stable best-response space Δ^* , all subsequent iterations remain within that space. Building on this property, Lemma 3 proves that every joint policy in Δ^* is MPE.

Lemma 2. If $\{\pi^k\}$ is stable in Δ^* at iteration K , $\forall i, k \geq K$, $\pi_i^k \in \Delta_i^*$.

Proof. Immediate from the definition of stability in best-response space, assuming $\{\pi^k\}$ is generated by MAPI-LP. $\square$

Lemma 3. All π in a stable best-response space Δ^* are MPEs.

Proof. By Definition 8, we know that

$$\forall i, \forall \pi_{-i} \in \Delta_{-i}, \forall \pi', \pi'' \in BR_i(\pi_{-i}), v_i^{(\pi'_i, \pi_{-i})} = v_i^{(\pi''_i, \pi_{-i})}. \quad (18)$$

If $\pi_{-i} \in \Delta_{-i}^*$, then $BR_i(\pi_{-i}) \in \Delta_i^*$ by Definition 13. Hence, in this case,

$$\forall i, \forall \pi_{-i} \in \Delta_{-i}^*, \forall \pi', \pi'' \in \Delta_i^*, v_i^{(\pi'_i, \pi_{-i})} = v_i^{(\pi''_i, \pi_{-i})}. \quad (19)$$

Since quantifiers of (19) can exhaustively compare any combination of joint policies in Δ^* , then it is equivalent to

$$\forall i, \forall \pi, \pi' \in \Delta^*, v_i^\pi = v_i^{\pi'}. \quad (20)$$

Since $\pi \in \Delta^*$ is no better than other outside policies, then all $\pi \in \Delta^*$ are MPEs. $\square$

The key observation is that stability implies mutual best-response consistency among all policies in Δ^* , while the best-response function guarantees value indifference between all admissible best responses. Consequently, no agent can improve its value by deviating from any policy in Δ^* , establishing the entire stable space as an equilibrium set (Theorem 1).

Theorem 1. *If $\{\pi^k\}$ is stable in best-response space Δ^* at K , then, for all $k \geq K$, π^k is an MPE.*

4.3. Monotonic Best-Response Space Shrinking

Another issue is that the choice of the best response is significant in $\xrightarrow{I_2}$, according to the possibility of equally effective policies (Remark 1(c)). Since the process in an ideal trajectory ultimately converges to a single joint policy corresponding to MPE, iteratively reducing the size of the set of possible best responses can theoretically lead to the desired outcome, possibly as a contraction mapping on policy space itself that converges to a stable best-response space. Therefore, with this idea, we define the following sufficient condition:

Theorem 2 (Monotonic Best-Response Space Shrinking, MBRSS). *If $\{\pi^k\}_{k \geq 0}$ satisfies*

$$\forall i, \forall k \geq 0, BR_i(\pi_{-i}^{k+1}) \subseteq BR_i(\pi_{-i}^k), \quad (21)$$

then

$$\lim_{k \rightarrow \infty} BR_i(\pi_{-i}^k) = \Delta_i^* \quad (22)$$

and all $\pi \in \prod_i \Delta_i^ = \Delta^*$ are MPEs.*

Proof. Assuming that (21) holds, BR is a sequence of compact best-response spaces, and each of them is a subset of its prior best-response spaces. Therefore,

$$\lim_{k \rightarrow \infty} BR_i(\pi_{-i}^k) = \bigcap_{k \geq 1} BR_i(\pi_{-i}^k) = \Delta_i^*. \quad (23)$$

Since BR_i will always contain at least one member, $BR_i(\pi_{-i}^0)$ will never be empty, and neither will Δ_i^* .

When the best-response space of each agent converges to Δ_i^* , $\{\pi^k\}$ becomes stable in $\Delta^* = \prod_i \Delta_i^*$. Thus, by Lemma 3, for all $\pi \in \Delta^*$. $\square$

In addition, this implies that the joint policy space, formed as the Cartesian product of the best-response sets of each agent i , also shrinks monotonically over iterations.

4.4. Satisfiability of MBRSS in Pure Coordination Games

MBRSS is defined over the policy space and becomes particularly useful in games where multiple joint actions yield the same reward, such as in pure coordination games [24]. This condition can help anchor the update direction, reduce variability, and narrow the trajectory toward a best-response path. In fact, we can view this sufficient condition as a set-valued contraction mapping. Accordingly, we demonstrate that sequential MAPI-LP satisfies MBRSS in the context of pure coordination games (Definition 14), as shown in Theorem 3.

Definition 14. *A pure coordination game is a repeated game in which all agents share the following reward structure:*

$$\mathcal{R}_i(a) = \begin{cases} r, & a \in A^\ell, \\ 0, & a \notin A^\ell, \end{cases} \quad (24)$$

where the alignment set $A^\ell \subseteq A$ is not empty, and r is a positive constant. For each agent i , we define the individual alignment set as $A_i^\ell = \{a_i \mid \exists a_{-i}, (a_i, a_{-i}) \in A^\ell\} = \{a_i^1, \dots, a_i^{|A_i^\ell|}\}$, representing the set of alignment actions available to agent i .

Definition 15. We define a count function C to return the total number of unique alignments in the alignment set A^ℓ that contain its input. For example, if $A^\ell = \{(a_1^1, a_2^1), (a_1^1, a_2^2)\}$, then $C(a_1^1) = 2, C(a_2^1) = 1$ and $C(a_1^1, a_2^1) = 1$.

Lemma 4. If $C((a_i)_{i \in I'}) \geq 1$ where $I' \subset I$, then there exists $a_j \in A_j^\ell$ where $j \notin I'$ that satisfies $C((a_i)_{i \in I'}, a_j) \geq 1$.

Proof. This is a direct implication when $|A^\ell| \geq 1$. $\square$

Definition 16. Sequential MAPI-LP is a modified MAPI-LP, where at iteration k , only agent $i = (k \bmod n) + 1$ selects a policy which is a vertex of its best-response space $BR_i(\pi_{-i}^k)$, while for other agents the policies remain unchanged.

Theorem 3. Assuming that the initial policies π_i^0 are uniformly distributed for all agents i and all agents follow sequential MAPI-LP (Definition 16), in any pure coordination game, MBRSS is satisfied and is stable in some best-response space at iteration $k = n$. Therefore, for all $k \geq n$, π^k is an MPE.

Proof. At $k = 0$, for each vertex of policy space $\pi_1^1(a_1) = 1$ where $a_1 \in A_1$,

$$v_i^{\pi_1^1, \pi_{-1}^0} = \begin{cases} \frac{r}{1-\gamma} \cdot \frac{C(a_1)}{\prod_{j \in I \setminus \{1\}} |A_j^\ell|}, & a_1 \in A_1^\ell, \\ 0, & a_1 \notin A_1^\ell. \end{cases} \quad (25)$$

By process $\xrightarrow{I_1}$ the best-response space is

$$BR_1(\pi_{-1}^0) = \text{Hull} \left\{ \pi_1^1(a_1) = 1 \mid a_1 \in \underset{a_1 \in A_1^\ell}{\operatorname{argmax}} \frac{C(a_1)}{\prod_{j \in I \setminus \{1\}} |A_j^\ell|} \right\}, \quad (26)$$

where $\text{Hull}\{\cdot\}$ is the convex hull of the given set. Since $|A^\ell|$ is not empty, $BR_i(\pi_i^0)$ is not empty, so we can arbitrarily select a vertex of it as π_1^1 where $\pi_1^1(a_1) = 1$, for the next iteration.

At the iteration $k < n$, agents $i \in 1, \dots, k-1$ already select $\pi_i^i = \pi_i^{i+1} = \dots = \pi_i^k$ where some $\pi_i^k(a_i) = 1$ by sequential constraints on $\xrightarrow{I_2}$. This implies that $C(a_1, a_2, \dots, a_{k-1}) \geq 1$. By Lemma 4, there exists $a_k \in A_k^\ell$ such that $C(a_1, a_2, \dots, a_{k-1}, a_k) \geq 1$, which means

$$BR_k(\pi_{-k}^k) = \text{Hull} \left\{ \pi_k^k(a_k) = 1 \mid a_k \in \underset{a_k \in A_k^\ell}{\operatorname{argmax}} \frac{C(a_1, a_2, \dots, a_k)}{\prod_{j \in I \setminus \{1, \dots, k\}} |A_j^\ell|} \right\} \quad (27)$$

is subset or equal to $BR_k(\pi_{-k}^{k-1})$ due to the restrictions of positive $C(\cdot)$ imposed on each agent's alignment set A_i^ℓ .

Hence, at the smallest iteration $k \geq n$, all agents already choose some alignment $a^\ell = (a_i)_i \in A^\ell$ where $C(a^\ell) \geq 1$, which is an optimal choice. So agent $i = (k \bmod n) + 1$ can only arbitrarily choose

$$\pi_i^{k+1} \in BR_i(\pi_{-i}^k) = \text{Hull} \left\{ \pi_i^{k+1}(a_i) = 1 \mid \exists a_i \in A_i^\ell, C((a_j)_{j \in I \setminus \{i\}}, a_i) = 1 \right\}, \quad (28)$$

which does not improve value function and does not change the best-response space. So, the sequence of best-response spaces satisfies MBRSS (Theorem 2) and is stable in

$$\Delta^* = \prod_i \text{Hull} \left\{ \pi_i^{k+1}(a_i) = 1 \mid \exists a_i \in A_i^\ell, \mathcal{C}((a_j^\ell)_{j \in I \setminus \{i\}}, a_i) = 1 \right\}, \quad (29)$$

for some alignment $a^\ell = (a_i^\ell)_i$ that all agents previously resemble at iteration k . Since we can identify Δ^* at iteration $k = n$, by Lemma 2, for all $k \geq n$, π^k is an MPE. $\square$

However, in this setting, MBRSS does not imply value monotonicity.

Corollary 3. *Under the same assumption of Theorem 3, the generated sequence of joint policies does not necessarily satisfy value monotonicity (Definition 12).*

Proof. Let $\{\pi^k\}$ be the generated sequence, then for some agent i at iteration $k \leq n$, $v_i^{(\pi_i^k, \pi_{-i}^{k-1})}$ is not necessarily less than or equal to $v_i^{(\pi_i^{k+1}, \pi_{-i}^k)}$, because the ratio of count and fixed chosen policy is the same as in (27). So, $\{\pi^k\}$ also does not satisfy value monotonicity. $\square$

Please note that Theorem 3 has very narrow settings, but it is a sufficient illustrative case that there exists a class of repeated games that naturally satisfies MBRSS.

5. Conjectures on Sufficient Conditions

5.1. Existence of Best-Response Dynamics for Generalized MBRSS

Sequential MAPI-LP (Theorem 3) induces a dimensional contraction mapping on the best-response space of the joint policy, where, for all i , the nested sequence $\text{BR}_i^1(\pi_{-i}^0) \supseteq \text{BR}_i^1(\pi_{-i}^1) \supseteq \dots \supseteq \Delta_i^*$ contracts monotonically in dimension, with the specific trajectory determined by agents' best-response selection.

For example, consider a pure coordination game with two agents with three available actions each that consist of: $I = \{1, 2\}$, $\forall i, A_i = \{a_i^1, \dots, a_i^3\}$ and alignment set $A^\ell = \{(a_1^1, a_2^1), (a_1^1, a_2^2), (a_2^2, a_2^3)\}$. By using sequential MAPI-LP (Theorem 3), the sequence of generated best-response sets is illustrated in Figure 1, showing a sufficient case that all possible trajectories satisfy MBRSS under sequential policy selection.

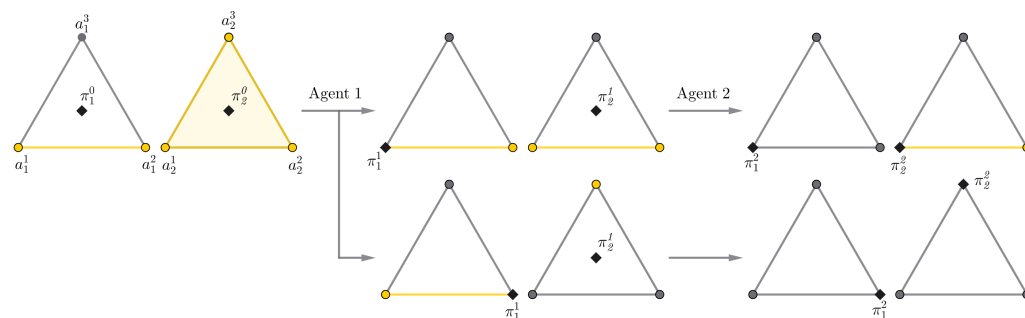


Figure 1. Example of trajectories that satisfy MBRSS according to Theorem 3. The yellow area shows the best-response space corresponding to each iteration, and the black diamond shows the current policy for each agent. In this scenario, after agent 2 selects a policy, agent 1's best response remains constant, and agent 1's value function also remains unchanged (Lemma 2). Similarly, agent 2's best response and value function do not change afterward.

However, MAPI-LP is heavily restricted by both the policy selection mechanism and the class of games in which MBRSS is proven to hold, which is the class of pure coordination games. Hence, a natural question is to ask whether there exist best-response dynamics beyond MAPI-LP that satisfy a generalized notion of MBRSS. This motivates the following conjecture, which relaxes MBRSS while keeping the property of stable best-response space:

Conjecture 1. For any finite repeated game $G = (I, \{A_i\}_i, \{R_i\}_i, \gamma)$, if stable best-response space Δ^* exists for this game, there exists a set-valued best-response map $\mathcal{M} : \Delta_A \rightrightarrows \Delta_A$ that generates $\{\pi^k\}_{k \geq 0}$ that satisfies the following:

- (a) $\forall k \geq 0, \exists k' > k, \mathcal{M}(\pi^{k'}) \subseteq \mathcal{M}(\pi^k)$
- (b) $\lim_{k \rightarrow \infty} \mathcal{M}(\pi^k) = \Delta^*$

In addition, from the characteristic of stability in best-response space, by infinite iteration and arbitrary best-response selection, Δ_i^* can be expressed by (17), which is equivalent to checking if all joint policies in $\Delta^* = \prod_i \Delta_i^*$ can map back to themselves, that is

$$\forall \pi \in \Delta^*, \Delta^* = \prod_i BR_i(\pi_{-i}) = \prod_i \Delta_i^*. \quad (30)$$

In case that $\{\pi^k\}_{k \geq 0}$ is stable in Δ^* at iteration k , we can test only all combinations of vertices in Δ^* to confirm this property, so Conjecture 2 is true. However, it will fail if at least one joint policy is within the convex hull and does not map back to itself in terms of best-response space.

Conjecture 2. Consider the sequence $\{\pi^k\}$. Let $BR_i^V(\pi_{-i}^k)$ be a set of vertices of $BR_i(\pi_{-i}^k)$ and $BR^V(\pi^k) = \prod_i BR_i^V(\pi_{-i}^k)$. At some iteration k , if

$$\forall \pi \in BR^V(\pi^k), BR^V(\pi) = BR^V(\pi^k), \quad (31)$$

then

$$\Delta^* = \text{Hull}(BR^V(\pi)) = \text{Hull}(BR^V(\pi^k)) \quad (32)$$

and $\{\pi^k\}$ is stable in Δ^* at iteration $k = n$, which indicates convergence to an MPE-consistent space.

5.2. Relationship Between Sufficient Conditions Within the Same Algorithm

5.2.1. In Serial MARL

Both value monotonicity (Definition 12) and stability in best-response space (Definition 13) are sufficient conditions for MARL. However, the value monotonicity is defined in terms of value function improvement and does not involve details about how individual policies are chosen. In contrast, stability in best-response space is defined based on the selection of best responses, which depends on only the value function determined by $\xrightarrow{I_1}$, that is the best better-responses. This implies that a sufficient condition defined on the sequence $\{v_i^{\pi^k}\}_k$ has broader scope than one defined on the sequence $\{BR_i(\pi_{-i}^k)\}_k$.

Corollary 4. Suppose that $\{\pi^k\}$ is value-monotonic and $\lim_{k \rightarrow \infty} v_i^{\pi^k} = c_i$ for all i . If there exists K such that, for all $i, k \geq K, v_i^{\pi^k} = c_i$ and $BR_i(\pi_{-i}^k) = BR_i(\pi_{-i}^{k+1})$, then $\{\pi^k\}$ is stable in best-response space at iteration K .

Corollary 5. Suppose that $\{\pi^k\}$ is stable in best-response space Δ^* at iteration K . If, for all $i, 0 \leq k \leq K, v_i^{\pi^k} \leq v_i^{(\pi_i^{k+1}, \pi_{-i}^k)} \leq v_i^{\pi^{k+1}}$, then $\{\pi^k\}$ is also value-monotonic.

As in Corollaries 4 and 5, we see the potential relationship between sufficient condition, which leads to the following conjecture:

Conjecture 3. Suppose that C_1 and C_2 are sufficient conditions for the convergence of a given MARL algorithm, that serially depend on each other, sharing some scopes of computation. Then there exists a complementary condition A such that

$$C_1 \wedge A \implies C_2. \quad (33)$$

5.2.2. In Parallel MARL

Conjecture 3 is formulated from the perspective of serial processes, where algorithm components operate sequentially. This raises a question: does this hierarchical relationship between sufficient conditions also apply to parallel processes commonly found in modern MARL architectures?

Consider the actor-critic model [7,25], where the actor network focuses on policy updates, while the critic network learns to estimate value functions or similar functions. Both components take the state as input, but the actor requires additional feedback from the critic's value estimation. This creates a parallel yet interdependent computational structure that differs fundamentally from our serial policy iteration framework.

In such parallel architectures, we hypothesize that sufficient conditions might exhibit more complex relationships than simple hierarchical implications. The complementary condition A in Conjecture 3 may need to account for synchronization between parallel processes. The relationship between C_1 and C_2 may involve dependencies rather than a nested scope. Hence, we conjecture that the hierarchical relationship can be extended to parallel processes as follows:

Conjecture 4. Let C_1 and C_2 be sufficient conditions for convergence in parallel components of a MARL algorithm, where C_2 is parallelly dependent on C_1 . Then, there exists a synchronization condition S , ensuring proper coordination between the parallel components, and a complementary condition A , such that

$$C_1 \wedge A \wedge S \implies C_2. \quad (34)$$

Recent advances in decentralized multi-agent actor-critic models demonstrate the importance of agent synchronization within the communication network as a fundamental assumption for convergence guarantees [26], highlighting the critical role of coordination mechanisms in parallel multi-agent systems. Hence, this extended conjecture opens several promising directions for formal verification of sufficient synchronization conditions. Examples include identifying the minimal amount of information required for communication, understanding the structure of information exchange, and characterizing process dependencies that affect convergence.

6. Discussion

6.1. Connections of MBRSS to Existing Frameworks

We clarify that MBRSS is not a reformulation of existing set-based convergence notions, but operates on a different axis. The following two comparisons make this precise.

6.1.1. Iterated Elimination of Dominated Strategies

Iterated elimination of dominated strategies (IEDS) is a classical approach to finding equilibria by iteratively removing dominated strategies, which are strategies that always result in a worse outcome for an agent compared to another available strategy, regardless of what the other agents do. This is based on the idea that rational agents are expected to avoid dominated strategies, as there is always a better alternative, making such strategies unlikely to be played. In this sense, each iteration of elimination shrinks the feasible joint policy space.

More precisely, IEDS removes strategies that are dominated with respect to all possible opponent strategies across the entire joint action space, independent of the learning trajectory. MBRSS, by contrast, tracks the shrinkage of each agent's best-response set relative to the current opponent policy profile at each iteration. A strategy that survives all rounds of IEDS may still be pruned from the best-response space under MBRSS at a later iteration, and conversely, a strategy eliminated by IEDS could remain in an early MBRSS best-response set if it happens to be a best response to the current profile. The two processes therefore operate on orthogonal axes: IEDS is a static equilibrium-refinement procedure, while MBRSS is a dynamic convergence condition on the algorithm's trajectory.

6.1.2. Absorbing Best-Response Graphs

A related treatment of convergence within the best-response space appears in [18] (called the best-reply set in [18]). Their work proposes a decentralized MARL algorithm for stochastic games with general state spaces, where each agent learns an estimated best-response space by running quantized Q-learning during exploration phases. The global policy-updating dynamics is then modeled as a Markov chain over the joint policy space, and convergence to equilibrium is characterized in closed form via the theory of absorbing Markov chains. Their framework shares a common object with MBRSS: both operate on a best-response space that collects all policies a rational agent might adopt given the current behavior of others. However, the two frameworks use this object in structurally different ways.

In [18], the best-response space enters as a static graph structure that encodes which joint policies are reachable from which other joint policies under one round of best-reply updates. Convergence is therefore a probabilistic property of how agents sample from their best-response spaces. In particular, when multiple best responses exist, an agent may sample among them at random, and the probability of reaching any given equilibrium is derived from the canonical form of the absorbing Markov chain.

Instead, MBRSS requires that the best-response space itself contracts monotonically across iterations. Convergence under MBRSS is therefore deterministic. There is no role for a sampling distribution, and no need to characterize absorption probabilities, because shrinkage of the space itself guarantees the outcome. Consequently, the two frameworks are complementary: ref. [18] asks with what probability does a best-response-based dynamics reach equilibrium within a induced best-response graph, while MBRSS asks under what set-contraction condition does the algorithm necessarily converge.

6.2. Future Directions

Our analysis of MBRSS in repeated games opens several promising avenues for further research. We outline three directions that extend both the theoretical scope and the practical implications of this sufficient condition.

6.2.1. Finding Game Classes and Algorithms That Satisfy MBRSS

Introduced by [27], the ratio-bounded game is a class of two-player non-cooperative games in which the strategy sets are Hilbert spaces and the defining parameters are derived from the game itself. Within this class, three of the four subclasses (distinguished by parameter configurations) admit a unique Nash equilibrium and can be solved via an affine-relaxed best-response algorithm. A key property enabling this result is that the best-response correspondence is single-valued.

In contrast, MBRSS relaxes this requirement and assumes only non-empty correspondence, due to $\xrightarrow{I_1}$. As a consequence, the class of games and algorithms that satisfy MBRSS is fundamentally different, raising natural questions about their characterization and prop-

erties. Identifying such games therefore becomes an important future direction of study. At minimum, there exists at least one class of games that satisfies MBRSS, providing a starting point for investigation.

A natural first step is to examine the implications of this class by considering simple canonical examples, beginning with pure coordination games (as discussed in Section 4.3). From there, one can seek to identify broader classes of games that meet the MBRSS criteria. By doing so, we can deepen our understanding of the dynamics at play and potentially uncover new strategies that could be applied in various game-theoretic contexts.

6.2.2. A Possible Application in Search Problem

The structural properties of MBRSS suggest a natural connection to search problems in which the feasible solution space can be reduced iteratively. Specifically, the monotone contraction of each agent's best-response set in Theorem 2 resembles a set-valued pruning process, that stabilizes at Δ^* (Definition 13). This dynamic is analogous to the progressive narrowing of a search space in constraint-satisfaction, branch-and-bound, and version-space learning frameworks, where feasibility conditions are applied iteratively to eliminate candidate solutions.

More precisely, the stability condition in Definition 13 ensures that the search has effectively terminated when no further pruning is possible or necessary. This mirrors the role of a feasibility check in combinatorial search, where a node is pruned precisely when no extension of the current partial solution can yield an improvement.

This analogy motivates a conjecture that MBRSS-type conditions could serve as a principled basis for search space reduction in multi-agent settings. Consider, for example, a cooperative task-assignment problem in which agents must jointly select actions from a large combinatorial space. If the agents' best-response spaces satisfy MBRSS, then the effective search space contracts monotonically across iterations, and the algorithm is guaranteed to concentrate its computation within a progressively smaller region of policy space. This property would be practically valuable in settings where exhaustive search is intractable and heuristic pruning lacks formal guarantees.

A formal development of this connection would require translating MBRSS from the continuous policy simplex into discrete or mixed search domains, characterizing the rate of contraction as a function of game structure, and identifying the classes of search problems for which the stability condition in stability in best-response space (Definition 13) can be verified efficiently (when Conjecture 2 is true, or more effective approximation approach). We leave these questions as directions for future work, noting that the single-point-of-failure methodology introduced here may offer a useful lens for analyzing when and why search procedures succeed or fail to concentrate on high-quality solutions.

6.2.3. Sufficient Condition Taxonomy

Conjectures 3 and 4 point toward a broader investigation into the landscape of sufficient conditions for MARL convergence, suggesting future work on classifying and unifying such conditions. As noted in Section 2.1, existing sufficient conditions suggest several different directions for improvement. A systematic taxonomy of these conditions organized by properties, such as static versus dynamic, serial versus parallel, or local versus global, could clarify their interrelationships and relative strengths. Such a taxonomy would help us understand the connections between various MARL categories, identify missing links, and inspire new conditions that may lead to novel MARL models. Since sufficient conditions indirectly indicate how corresponding MARL algorithms work, it would be desirable to identify appropriate algorithms based on the characteristics of the game and the desired solution concept.

7. Conclusions

We have studied convergence in multi-agent reinforcement learning through the lens of sufficient conditions, using MAPI-LP in repeated games as a concrete framework for analysis. By adopting a single-point-of-failure perspective, we identified two sufficient conditions for monotonic convergence to MPE. The first is stability in best-response space (Definition 13). The second, monotonic best-response space shrinking (Theorem 1), is a novel set-valued condition requiring that each agent's best-response set contracts monotonically across iterations until collapsing to a stable best-response space. In contrast, value monotonicity establishes convergence in value space but does not guarantee convergence to an MPE. And, we further proved that sequential MAPI-LP satisfies MBRSS in pure coordination games, converging to a stable best-response space, where all of its members are MPE, in at most n iterations (Theorem 3).

Building on these results, we proposed four conjectures: the first two are on MBRSS and stability in best-response space, while the others point to relationships between sufficient conditions within the same system (Section 5). Conjecture 1 asks whether generalized best-response dynamics satisfying MBRSS exist for arbitrary finite repeated games, motivated by the restriction of Theorem 3 to pure coordination games. Conjecture 2 proposes a more effective way to check stability in best-response space, instead of exhaustively checking every iteration infinitely. Conjecture 3 extends the sequential relationship, inspired by Corollaries 4 and 5. Conjecture 4 provisions the parallel relationship between components.

We also discussed connections between MBRSS and classical frameworks, including iterated elimination of dominated strategies and absorbing best-response graphs, highlighting both the similarities and the structural distinctions that make MBRSS a complementary perspective on convergence.

Together, these results offer a theoretical foundation and a structured roadmap for understanding convergence in multi-agent settings. The most immediate open problems are the proof or disproof of Conjectures 1 and 2 for the extension and the application of MBRSS in other game classes, from other repeated games to stochastic games where the policy space becomes high-dimensional and coupled through transition dynamics. We would like to develop a taxonomy of sufficient conditions that systematically organizes existing and future conditions by their scope, structural assumptions, and relationships to one another. We hope that the single-point-of-failure methodology introduced here, and the sufficient conditions and conjectures yielded in this analysis, provide a useful starting point for theoretical investigations in the future.

Author Contributions: Conceptualization, N.M.; methodology, N.M.; validation, N.M.; formal analysis, N.M.; investigation, N.M.; resources, A.M.; data curation, N.M.; writing—original draft preparation, N.M.; writing—review and editing, N.M. and Z.Y.; visualization, N.M.; supervision, A.M. and Z.Y.; project administration, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors would like to thank the reviewers for their time and valuable comments, which helped us to improve the quality of this manuscript. We have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MAPI-LP	Multi-Agent Policy Iteration Integrated with Linear Programming
MARL	Multi-Agent Reinforcement Learning
MBRSS	Monotonic Best-Response Space Shrinking
MPE	Markov Perfect Equilibrium

References

1. Albrecht, S.V.; Christianos, F.; Schäfer, L. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*; MIT Press: Cambridge, MA, USA, 2024.
2. Littman, M.L. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Machine Learning Proceedings 1994*; Elsevier: Amsterdam, The Netherlands, 1994; pp. 157–163. [\[CrossRef\]](#)
3. Foerster, J.; Nardelli, N.; Farquhar, G.; Afouras, T.; Torr, P.H.; Kohli, P.; Whiteson, S. Stabilising Experience Replay for Deep Multi-Agent Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017), Sydney, Australia, 6–11 August 2017*; PMLR: New York, NY, USA, 2017; pp. 1146–1155.
4. Bertsekas, D. Multiagent Reinforcement Learning: Rollout and Policy Iteration. *IEEE/CAA J. Autom. Sin.* **2021**, *8*, 249–272. [\[CrossRef\]](#)
5. Yu, C.; Velu, A.; Vinitzky, E.; Gao, J.; Wang, Y.; Bayen, A.; Wu, Y. The Surprising Effectiveness of Ppo in Cooperative Multi-Agent Games. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2022**, *35*, 24611–24624.
6. Kuba, J.; Chen, R.; Wen, M.; Wen, Y.; Sun, F.; Wang, J.; Yang, Y. Trust Region Policy Optimisation in Multi-Agent Reinforcement Learning. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Virtually, 25–29 April 2022*; The International Conference on Learning Representations (ICLR): Amherst, MA, USA, 2022; p. 1046.
7. Lowe, R.; Wu, Y.L.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; Mordatch, I. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2017**, *30*, 6379–6390.
8. Rashid, T.; Samvelyan, M.; De Witt, C.S.; Farquhar, G.; Foerster, J.; Whiteson, S. Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *J. Mach. Learn. Res.* **2020**, *21*, 1–51.
9. Son, K.; Kim, D.; Kang, W.J.; Hostallero, D.E.; Yi, Y. QTRAN: Learning to Factorize With Transformation for Cooperative Multi-Agent Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019), Long Beach, CA, USA, 9–15 June 2019*; PMLR: New York, NY, USA, 2019; pp. 5887–5896.
10. Yang, Y.; Ma, C.; Ding, Z.; McAleer, S.; Jin, C.; Wang, J.; Sandholm, T. Game-Theoretic Multiagent Reinforcement Learning. *arXiv* **2025**, arXiv:2011.00583.
11. He, S.; Han, S.; Su, S.; Han, S.; Zou, S.; Miao, F. Robust Multi-Agent Reinforcement Learning with State Uncertainty. *Trans. Mach. Learn. Res.* **2023**.
12. Christoffersen, P.J.; Haupt, A.A.; Hadfield-Menell, D. Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, UK, 29 May–2 June 2023*; pp. 448–456.
13. Fink, A.M. Equilibrium in a Stochastic n -Person Game. *J. Sci. Hiroshima Univ. Ser. AI (Math.)* **1964**, *28*, 89–93. [\[CrossRef\]](#)
14. Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W.M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J.Z.; Tuyls, K.; et al. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2018), Stockholm, Sweden, 10–15 July 2018*; pp. 2085–2087.
15. Ye, J.; Li, C.; Wang, J.; Zhang, C. Towards Global Optimality in Cooperative MARL With the Transformation and Distillation Framework. *arXiv* **2022**, arXiv:2207.11143.
16. Li, Z.; Zhao, W.; Wu, L.; Pajarinen, J. AgentMixer: Multi-Agent Correlated Policy Factorization. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI 2025); AAAI: Washington, DC, USA, 2025; Volume 39*, pp. 18611–18619. [\[CrossRef\]](#)
17. Kim, S.; Kim, W.; Jeon, J.; Sung, Y.; Han, S. Off-Policy Multi-Agent Policy Optimization with Multi-Step Counterfactual Advantage Estimation. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, UK, 29 May–2 June 2023*.
18. Altabaa, A.; Yongacoglu, B.; Yüksel, S. Decentralized Multi-Agent Reinforcement Learning for Continuous-Space Stochastic Games. In *Proceedings of the 2023 American Control Conference (ACC), San Diego, CA, USA, 31 May–2 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 72–77. [\[CrossRef\]](#)
19. Arslan, G.; Yüksel, S. Decentralized Q-Learning for Stochastic Teams and Games. *IEEE Trans. Autom. Control* **2016**, *62*, 1545–1558. [\[CrossRef\]](#)

20. Bey, O.; Chemachema, M. Finite-time event-triggered output-feedback adaptive decentralized echo-state network fault-tolerant control for interconnected pure-feedback nonlinear systems with input saturation and external disturbances: A fuzzy control-error approach. *Inf. Sci.* **2024**, *669*, 120557. [[CrossRef](#)]
21. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; A Bradford Book: Cambridge, MA, USA, 2018.
22. Maskin, E.; Tirole, J. Markov Perfect Equilibrium: I. Observable Actions. *J. Econ. Theory* **2001**, *100*, 191–219. [[CrossRef](#)]
23. Ding, Z.W.; Zheng, G.Z.; Cai, C.R.; Cai, W.R.; Chen, L.; Zhang, J.Q.; Wang, X.M. Emergence of cooperation in two-agent repeated games with reinforcement learning. *Chaos Solitons Fractals* **2023**, *175*, 114032. [[CrossRef](#)]
24. Raducha, T.; San Miguel, M. Coordination and Equilibrium Selection in Games: The Role of Local Effects. *Sci. Rep.* **2022**, *12*, 3373. [[CrossRef](#)] [[PubMed](#)]
25. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347. [[CrossRef](#)]
26. Trivedi, P.; Hemachandra, N. Multi-Agent Natural Actor-Critic Reinforcement Learning Algorithms. *Dyn. Games Appl.* **2023**, *13*, 25–55. [[CrossRef](#)]
27. Caruso, F.; Ceparano, M.C.; Morgan, J. Affine Relaxations of the Best Response Algorithm: Global Convergence in Ratio-Bounded Games. *SIAM J. Optim.* **2023**, *33*, 1914–1942. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.