Constant-Memory Strategies in Stochastic Games: A Theoretical and Empirical Study

Fengming Zhu , **Fangzhen Lin**
Hong Kong University of Science and Technology
fzhuae@connect.ust.hk, flin@cs.ust.hk

Abstract

Stochastic games have become a prevalent framework for studying long-term multi-agent interactions, especially in the context of multi-agent reinforcement learning. In this work, we comprehensively investigate the concept of constant-memory strategies in stochastic games. We first establish some results on best responses and Nash equilibria for behavioral constant-memory strategies, followed by a discussion on the computational hardness of best responding to mixed constant-memory strategies. Those theoretic insights are later verified on several sequential decision-making testbeds, including the *Iterated Prisoner's Dilemma*, the *Iterated Traveler's Dilemma*, and the *Pursuit* domain. This work aims to enhance the understanding of theoretical issues in single-agent planning under multi-agent systems, and uncover the connection between decision models in single-agent and multi-agent contexts. The codebase and the full version of this paper is available at `github.com/Fernadoo/Const-Mem`.

1 Introduction

Various real-world situations that involve long-term interactions among a group of participants can be modeled as stochastic games, such as negotiation between multiple stakeholders [Baarslag *et al.*, 2016; Baarslag, 2024], bidding and mechanism design in repeated auctions [Iyer *et al.*, 2014; Balseiro *et al.*, 2015; Shen *et al.*, 2020; Su *et al.*, 2024], multi-agent teamwork [Stern, 2019; Mirsky *et al.*, 2022] and even human-robot collaboration [Puig *et al.*, 2023]. Stochastic games, also known as Markov games, model the interactions of these multi-agent systems as a Markov chain over a set of states, where the transitions are triggered by joint actions and are potentially stochastic.

The formalization of stochastic games was first proposed in Shapley's seminal work [Shapley, 1953]. A perfectly rational agent in a stochastic game is supposed to make use of all past histories to determine the next action, and therefore, the notion of strategies, defined as mappings from all possible histories to actions, is inherently complex. The fact that there are infinitely many strategies prohibits the direct application of Nash's theorem for establishing any existence result of equilibria. However, the stationary transitions of stochastic games inevitably draw attention to a highly specialized subclass of time-independent and memoryless strategies that only consider the current states while discarding all past histories, termed *stationary strategies*. Indeed, the existence of equilibria formed by stationary strategies in n-player general-sum stochastic games was later proven by Fink [1964] and Takahashi [1964], under mild assumptions. Despite its limited expressiveness, the notion of stationary strategies has enabled the community to practically investigate some complex real-world applications, particularly by resorting to multi-agent reinforcement learning (MARL) techniques, as advocated by Littman [1994] and implemented in a line of subsequent work [Lowe *et al.*, 2017; Rashid *et al.*, 2020; Yu *et al.*, 2022].

Notably, one would naturally expect strategies in other less restricted forms that can encode a broader class of behavioral patterns, hoping to achieve better payoff outcomes. For example, in the Iterated Prisoner's Dilemma (IPD), if only stationary strategies are considered, there is a unique Nash equilibrium where both players choose to *defect* all the time, resulting in the lowest overall payoff. However, even with the ability to remember only one past action played by the opponent, the well-known *Tit-For-Tat* (TFT) strategy (starting with *cooperation*) can be devised. One can easily see that if both players adopt the TFT strategy, they will follow a trajectory of both *cooperating* throughout the game, resulting in a Nash equilibrium with the highest possible payoff. Apart from other forms of representation, such as strategies represented as finite automata [Rubinstein, 1986; Ben-Porath, 1990; Zuo and Tang, 2015] and even Turing machines [Megiddo and Wigderson, 1986; Knoblauch, 1994; Nachbar and Zame, 1996; Chen *et al.*, 2017b], we focus our main effort on investigating the notion of *constant-memory strategies*, i.e., mappings from history segments of bounded lengths to actions, mainly because it directly relates to the concept of bounded rationality [Simon, 1990] in general, and highly implementable in practice. Note that this notion has been preliminarily investigated by [Chen *et al.*, 2017a] and [Wang and Lin, 2019]. However, they only focus on behavioral strategy best responses under repeated games, without further discussion on either Nash equilibria or mixed strategies under games with multiple states.

In this paper, we comprehensively study the theoretical properties associated with *constant-memory strategies* in *stochastic games*. We begin by presenting the following two results:

1. *A Characterization of Best Responses:* Given a constant-memory strategy profile adopted by the opponents, there always exists a best response in the form of a pure constant-memory strategy that makes use of the same length of memory.
2. *An Existence Result of Equilibria:* Given any finite length of memory, there always exists a Nash equilibrium where all agents adopt constant-memory (but not necessarily pure) strategies using that length of memory.

As a side benefit of using memories of constant lengths, any strategy that uses a shorter memory can always be implemented by one that uses a longer memory. Therefore, the above two results directly imply that any NE formed by shorter-length-memory strategies can be transformed into an NE formed by longer-length-memory strategies, suggesting that the longer the memory used by the strategies, the richer the equilibria one can potentially expect.

Additionally, we provide further results about best responses against mixed constant-memory strategies, mathematically defined as those sampled from a set of support strategies with certain probabilities. This is associated with broad applications in the domain of opponent modeling [Carmel and Markovitch, 1998; Albrecht and Ramamoorthy, 2015; Albrecht and Stone, 2018; Zhu and Lin, 2025], particularly for type-based methods [Albrecht and Ramamoorthy, 2015; Albrecht and Ramamoorthy, 2019; Albrecht *et al.*, 2016; Zhu and Lin, 2025]. However, we demonstrate that:

1. *A Negative Result on Strategy Equivalence:* An opponent with a mixed constant-memory strategy may not correspond to an equivalent opponent with a single (behavioral) constant-memory strategy in terms of resulting in the same payoff.
2. *A Negative Result on Best Responses:* The best response against a mixed constant-memory strategy is not necessarily constant-memory, and computing such best responses is computationally hard—possibly even not computable.

In spite of these negative results, we do provide a computational model for solving the best response against a mixed strategy, which (1) demonstrates that those computational models proposed in [Zhu and Lin, 2025] do not overcomplicate the problem; and (2) shows that their methods, which are originally designed for playing against opponents using stationary strategies, can be naturally extended to accommodate ones using constant-memory strategies.

2 Preliminaries

The whole system where the agents interact is modelled as a *stochastic game* (SG, also known as Markov games) [Shapley, 1953], which can be seen as an extension of both *normal-form games* (to dynamic situations with stochastic transitions) and *Markov decision processes* (to strategic situations with multiple agents). A stochastic game is a 5-tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, T, R \rangle$ given as follows,

1. $\mathcal{N}$ is a finite set of n agents.
2. $\mathcal{S}$ is a finite set of (environmental) states.
3. $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ is a set of joint actions, where $\mathcal{A}_i$ is the action set of agent i. In particular, we write a_i as the action of agent i and the one without any subscript $a = (a_i, a_{-i})$ as the joint action.
4. $T : \mathcal{S} \times \mathcal{A}_1 \times \cdots \mathcal{A}_n \mapsto \Delta(\mathcal{S})$ defines stochastic transitions among states.
5. $R_i : \mathcal{S} \times \mathcal{A}_1 \times \cdots \mathcal{A}_n \mapsto \mathbb{R}$ denotes the immediate rewards for agent i.

To define best responses and hence equilibria, we need to first define strategies and objectives.

Assuming complete observability and perfect recall, a perfectly rational agent should utilize the entire history, while in memory-restricted cases, an agent can only devise strategies based on past memories of finite lengths. We denote the space of all possible histories of length $K \in \mathbb{N}$ as $\mathcal{H}^K \triangleq (\mathcal{S} \times \mathcal{A})^K$. In particular, when $K = 0$, we have $\mathcal{H}^0 = \emptyset$ meaning that no history can be utilized. Then, given any non-negative integer K, a K-memory strategy for agent i is a mapping from all possible histories with lengths less then or equal to K and the current states to (possibly randomized) actions, mathematically denoted as $\pi_i : \mathcal{H}^{\leq K} \times \mathcal{S} \mapsto \Delta(\mathcal{A}_i)$ where $\mathcal{H}^{\leq K} \triangleq \cup_{k=0}^{K} \mathcal{H}^k$. Let Π_i^K denote the set of all such K-memory strategies for agent i. For convenience, we let $\mathcal{H}^\infty = (\mathcal{S} \times \mathcal{A})^*$ denote the set of complete histories that an agent with perfect recall can possibly memorize, and therefore, Π_i^∞ is the set of all possible infinite-memory strategies for agent i of the form $\pi_i : (\mathcal{S} \times \mathcal{A})^* \times \mathcal{S} \mapsto \Delta(\mathcal{A}_i)$. A direct consequence is that $\Pi_i^K \subseteq \Pi_i^{K'} \subseteq \Pi_i^\infty$ for any non-negative $K \leq K'$. Among them, one of the most popular class of strategies is Π_i^0, termed *stationary strategies*. To be clear, we use the term *constant-memory* to distinguish from those infinite-memory strategies, and use the term *K-memory* when this specific K needs to be emphasized.

The objective for each agent is to maximize its accumulated discounted rewards (a.k.a. the discounted-payoff scenario, as opposed to the average-payoff scenario). We let $R_{i,t}$ denote the reward signaled to agent i at step t, similarly for S_t and $a_{i,t}$, then the overall utility under a policy profile (π_i, π_{-i}) starting from any arbitrary state $S \in \mathcal{S}$ is

$$u_i(S; \pi_i, \pi_{-i}) = \mathbb{E}_{(\pi_i, \pi_{-i})}\Big[\sum_{t=0}^{\infty} \gamma^t R_{i,t} \Big| S_0 = S\Big] \quad (1)$$

π_i is said to be the best response of π_{-i}, denoted as $\pi_i \in BR(\pi_{-i})$, if

$$\forall S \in \mathcal{S}, \pi_i' \in \Pi_i^\infty, u_i(S; \pi_i, \pi_{-i}) \geq u_i(S; \pi_i', \pi_{-i}) \quad (2)$$

requiring that a π_i must outperform any other in Π_i^∞ to serve as the best response. Note that, to compare the values of two strategy profiles, one must ensure that the limit of the right-hand side (RHS) in Equation (1) exists in the first place. Also note that, some pairs of π_i and π_i' may not be comparable in the above sense, as this comparison requires value dominance across all possible states.

3 Best Responses and Nash Equilibria

We first characterize the best response of an agent when all the other opponents are equipped with constant-memory strategies with the same non-negative (and finite) memory length.

Theorem 1. *Given $\pi_j \in \Pi_j^K$ with $K \in \mathbb{Z}$ for all $j \neq i$, i.e., all the other agents are adopting constant-memory strategies with the same finite memory length K, there must exist a pure K-memory strategy for agent i as a best response.*

Proof sketch. As the full proof involves some fundamental (and probably tedious) derivations, we defer it to Appendix A.1. The main issue is that, in general, the decision process from the perspective of agent i is an infinite *Markov Decision Process* (MDP), where the state space encompasses infinitely many histories (of all possible lengths), and transitions/rewards are jointly controlled by the stochastic game dynamics as well as the other opponents. However, we can formally show that there exists a finite MDP that is strategically equivalent. Consequently, among all the optimal solutions of this constructed MDP, there must be a stationary and deterministic policy, which corresponds to a K-memory pure strategy of agent i. While it implies that using memory length up to K is sufficient, chances are that there may also exist best responses of a shorter memory $K' < K$, as later discussed in Section 5.1.

This proof is summarized in our code implementation.[1] □

One can immediately see the following corollary where the opponents use constant-memory strategies but with potentially different memory lengths. The justification is straightforward: all the opponents can be jointly viewed as a "super-agent", and consequently this "super-agent" is adopting a $(\max_{j\neq i}\{K_j\})$-memory strategy. Therefore, the best response for the pivotal agent shall also be of $(\max_{j\neq i}\{K_j\})$-memory. One may also note that it pertains to some recent work on asymmetric memory [Fujimoto *et al.*, 2024]; however, we leave it to future work.

Corollary 1. *Given $\pi_j \in \Pi_j^{K_j}$ with each $K_j \in \mathbb{Z}$ for all $j \neq i$, i.e., all the other agents are adopting constant-memory strategies but with varying memory lengths, there must exist a pure $(\max_{j\neq i}\{K_j\})$-memory strategy for agent i as a best response.*

As best responses are well established, we now examine whether an equilibrium exists when everyone best responds to each other.

Definition 1 (Nash Equilibrium)**.** *A strategy profile $\{\pi_i^*\}_{i\in\mathcal{N}}$ is a Nash equilibrium (NE) if*

$$\forall i \in \mathcal{N}, \pi_i^* \in BR(\pi_{-i}^*)$$

Theorem 2. *There exists an NE when the agents are all adopting K-memory strategies, for any arbitrary non-negative finite K.*

Proof sketch. As previously, we also summarize this proof into a ready-to-run code implementation.[2]

We postpone the full proof to Appendix A.2. Similarly to Nash's proof of his well-known existence theorem, here we construct a mapping to iteratively refine those agents' strategies, which operates by adjusting the probability of each action at each state based on its non-negative advantage, followed by a normalization step. It can be formally shown that there is a bijection between the fixed points of this refinement mapping and the solutions to the system of equations that characterizes NEs with K-memory strategy profiles under a given stochastic game. □

The above existence result implies the following. Consider two agents playing a stochastic game, where agent 1 employs a two-memory strategy and agent 2 uses a three-memory strategy. If agent 1 asserts that it will adhere to its two-memory strategy, agent 2 may identify another two-memory strategy as a best response, potentially yielding the same payoff but allowing for memory saving. Conversely, if agent 2 can convince agent 1 that it will maintain its three-memory strategy, agent 1 may find it advantageous to switch to a three-memory strategy as a better response.

Note that the above theorem is not a direct consequence of Nash's existence theorem, as we only discuss randomizing actions within a single strategy, rather than randomizing across multiple strategies, which we will refer to as *mixed strategies* in the next section.

Another benefit of constant-memory strategies is that any K-memory strategy can be represented as a K'-memory strategy, provided that $K' \geq K$, by simply utilizing the most recent K records. Thus, we arrive at the following corollary.

Corollary 2. *Any payoff profile that is reached by an NE under a K-memory strategy profile can also be realized by another NE under a K'-memory strategy profile, if $K' \geq K$.*

4 Best Responses to Mixed Strategies: A Tournament Perspective

We have established that the best response to a single (possibly randomized) constant-memory strategy results in another constant-memory strategy. The next natural question is: what is the best response to a set of constant-memory strategies played according to a specified distribution? A further related question is: can a mixed strategy be converted into a singleton constant-memory strategy? If this is feasible, then the best response must also be a constant-memory strategy.

In the following two subsections, we will show:

1. In repeated games, if an agent encounters an opponent using a mixed zero-memory strategy, it will yield the same expected utility for this agent to play against an opponent with a single induced zero-memory strategy.
2. In general, when the game involves multiple states or the opponent employs a nonzero-memory strategy, then the best response will be hard to compute (possibly not computable) and time-dependent, which may not be encoded as a finite-memory strategy. Therefore, it implies that the

[1] Please refer to `kMemBR.ipynb` in our code.

[2] Please refer to `kMemNE_full.ipynb` in our code.

opponent's mixed strategy cannot be equivalently transformed into a singleton constant-memory strategy.

4.1 Mixed Strategies vs Behavioral Strategies

We first emphasize the notion of *match*. When we say an agent i adopts a K-memory strategy, it means that agent i will select one strategy $\pi_i \in \Pi_i^K$ just before a match begins. Once the agent has "confirmed" its strategy, it will **not** deviate to any other strategies during the match until the termination. Note that some strategies, especially those in Π_i^∞, may be semantically interpreted as "learning" or "evolving" strategies, as they gradually modify the decisions based on accumulated observations; however, each of them remains a singleton strategy within the strategy space Π_i^∞. From the perspective of a single agent, we may also use the term *episode* interchangeably with *match*, as commonly done in the context of MDPs. The *overall utility* will be calculated as the expected payoff over all possible matches.

Now we are ready to explain the difference between a *behavioral* strategy and a *mixed* strategy. Recall that a K-memory strategy of agent i is defined as $\pi_i : \mathcal{H}^{\leq K} \times \mathcal{S} \mapsto \Delta(\mathcal{A}_i)$; it is also referred to as a *behavioral* strategy as it can randomize over actions. By definition, a pure strategy that performs deterministic actions is also considered a behavioral strategy. A *mixed* strategy (for agent i and of K-memory) first specifies its support set $\Pi_i^{K+} \subseteq \Pi_i^K$, where each behavioral strategy $\pi_i^\iota \in \Pi_i^{K+}$ will be selected with a positive probability p_ι, before each match begins. Thus, we use a tuple $(\Pi_i^{K+}, \vec{p})$ to denote a mixed strategy for agent i. Intuitively, when an agent is playing against a mixed strategy $(\Pi_i^{K+}, \vec{p})$, it simply means this particular agent will encounter an opponent using the behavioral strategy $\pi_i^\iota \in \Pi_i^{K+}$ for a fraction p_ι of the whole time.

One may be particularly interested in a specific type of strategies, namely the behavioral strategy obtained by state-wise randomization over the actions according to the probability distribution provided by the mixed strategy.

Definition 2 (Mixed-Strategy-Induced Behavioral Strategy)**.** *Given a mixed strategy $(\Pi_i^{K+}, \vec{p})$, we define $\omega_{(\Pi_i^{K+}, \vec{p})}$ as the behavior strategy induced by this mixed strategy. Formally, for each $(H, S) \in \mathcal{H}^{\leq K} \times \mathcal{S}$,*

$$\omega_{(\Pi_i^{K+}, \vec{p})}(a_i|H, S) \triangleq \sum_\iota p_\iota \cdot \pi_i^\iota(a_i|H, S)$$

The underlying intuition is that, instead of randomly selecting one of the support strategies at the beginning and sticking to it, we also allow an agent to switch to another strategy within the same probability distribution at each timestep during play, resulting in a single behavioral strategy that randomizes over each support strategy at every state. One can see that if the original strategy is a mixed one over a set of K-memory support strategies, then its induced behavioral strategy, according to Definition 2, will still be a K-memory strategy, and its best response will also be a K-memory strategy, as stated in Theorem 1.

We will first demonstrate that in a special case where a stochastic game is reduced to a repeated game and the agents use stationary strategies, a mixed strategy has the same effect as its induced behavioral strategy. However, in general, if a game involves transitions across multiple states or the opponents adopt nonzero-memory strategies, such equivalence does not necessarily hold.

Theorem 3 (Utility equivalence for repeated games)**.** *If the stochastic game is merely a repeated game, i.e. $\mathcal{S}$ is a singleton, then an agent i's overall utility when playing against a mixed strategy $(\Pi_{-i}^{0+}, \vec{p})$ will be the same as the utility against the induced behavioral strategy $\omega_{(\Pi_{-i}^{0+}, \vec{p})}$.*

We omit the elaboration here; one may refer to the detailed proof in Appendix A.3. The intuition behind is quite clear: if it is stateless and memoryless, then mixing over strategies is the same as mixing over action.

Theorem 4 (Utility Equivalence Does Not Hold for General Stochastic Games)**.** *In general, when a stochastic game involves multiple states, an agent i's overall utility when playing against a mixed stationary strategy $(\Pi_{-i}^{0+}, \vec{p})$ is not necessarily the same as the utility against the induced behavioral strategy $\omega_{(\Pi_{-i}^{0+}, \vec{p})}$.*

Proof sketch. We here provide some intuitions, while the full proof is deferred to Appendix A.4. The key issue is as follows. Even when an agent plays against a mixed stationary strategy, its overall utility is the expectation of the returns of playing against each of the support strategies (each corresponding to a multi-step MDP), involving $|\Pi_{-i}^{0+}|$ contraction mappings. However, when it plays against the induced behavioral strategy, its utility is computed by evaluating only one MDP induced by $\omega_{(\Pi_{-i}^{0+}, \vec{p})}$, involving a contraction mapping that differs from any of the aforementioned $|\Pi_{-i}^{0+}|$. A counter-example is also provided in Appendix A.5. □

As increasing either the length of memory or the number of environmental states results in a multi-state MDP from an individual agent's perspective, a natural implication is that utility equivalence between a mixed strategy and its induced behavioral strategy does not necessarily hold for K-memory strategies once K is positive, even in repeated games.

4.2 Computing BR to Mixed Strategies is Hard

We will first show that computing the best response against a mixed K-memory strategy can be reduced to optimally solving an infinite-horizon *partially observable MDPs* (POMDPs) [Sondik, 1978; Kaelbling *et al.*, 1998]. It turns out the reduced ones belong to a subclass of generic POMDPs, namely *Contextual MDP* (CMDPs), although it may not necessarily imply less challenging computation. To show that this reduction does not complicate the original problem, we also construct a reduction from the problem of optimally solving CMDPs back to that of computing best responses against mixed strategies in stochastic games.

Theorem 5. *Given a mixed strategy profile $(\Pi_{-i}^{K+}, \vec{p})$ of the opponents, computing the best response for agent i can be reduced to optimally solving an infinite-horizon POMDP.*

Proof sketch. Here we only provide some intuition, while the full reduction to POMDPs and the proof of correctness are postponed to Appendix A.6. Basically, a POMDP *state* consists of the history segment, the environmental state of the

stochastic game, and the unobservable opponent strategies. *Observations* are made by simply discarding the third component of the POMDP state. *Transitions* across those states that are associated with different opponent strategies are assigned zero probabilities, similarly for rewards. An *optimal solution* for such a POMDP, in terms of maximizing infinite-horizon discounted rewards, is typically obtained as a mapping from all possible histories (or equivalently, from beliefs over states) to potentially randomized actions [Sondik, 1978; Kaelbling *et al.*, 1998], thus, may not correspond to any finite-memory strategy in general. □

One can see that the constructed POMDPs in the above theorem belong to a subclass of generic POMDPs, where a state is composed of directly observable variables and other hidden ones, and there are no transitions among the hidden state variables. This subclass is specially termed as *Contextual MDPs* (CMDPs) [Hallak *et al.*, 2015; Benjamins *et al.*, 2023]. While CMDPs are special cases of POMDPs, the complexity/computability results for CMDPs remain unresolved. So far, the common conjecture is that CMDPs are **not** significantly easier to solve than POMDPs in general, and it is proven that optimally solving infinite-horizon POMDPs is undecidable [Madani *et al.*, 2003].

This result highly pertains to discussions on type-based methods for single-agent planning in the presence of multiple other agents [Albrecht and Ramamoorthy, 2015; Albrecht and Ramamoorthy, 2019; Albrecht *et al.*, 2016; Zhu and Lin, 2025]. Albrecht and Ramamoorthy [2015] characterized the general problem from a conceptual standpoint, where each opponent's strategy serves as an oracle that can be queried via Monte Carlo simulations; however, certain intermediate cases were not well formalized. As a supplement, Zhu and Lin [2025] offered a fine-grained spectrum of implementable planners for the stationary base case, where each strategy within the support of the opponent's mixed strategy is stationary. Here, Theorem 5 further generalizes this to constant-memory strategies, enabling all the formulations in [Zhu and Lin, 2025] to be extended to the entire family of constant-memory strategies.

We will also present a reduction in the reversed direction.

Theorem 6. *Optimally solving a CMDP can be reduced to computing the best response for an agent i against a profile of mixed zero-memory (i.e., stationary) strategies $(\Pi^{0+}_{-i}, \vec{p})$ adopted by its opponents.*

Proof sketch. A CMDP is formally defined as a tuple $\langle \mathcal{C}, \mathcal{S}, \mathcal{A}, f_T, f_R, \gamma \rangle$, where

1. $\mathcal{C}$ is a finite set of unobservable contexts, one of which will be selected at the beginning of each episode.
2. f_T and f_R take a context $c \in \mathcal{C}$ and output a transition function $T^c : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ as well as a reward function $R^c : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, respectively. The tuple $\langle \mathcal{S}, \mathcal{A}, T^c, R^c, \gamma \rangle$ then constitutes an MDP.

One can then construct an SG with two players

$$\langle \{1, 2\}, \mathcal{S}, \{\mathcal{A}, \mathcal{A}'\}, T_0 : \mathcal{S} \times \mathcal{A} \times \mathcal{A}' \mapsto \Delta(\mathcal{S}), \{R_1, R_2\} \rangle,$$

where agent 1 with action set $\mathcal{A}$ is playing against agent 2 (as a "context controller"), as if the latter holds a set of stationary strategies $\{\pi_c : \mathcal{S} \mapsto \Delta(\mathcal{A}')\}_{c \in \mathcal{C}}$. It can be shown that such a T_0 and $\{\pi_c\}_{c \in \mathcal{C}}$ can be factorized out of f_T, i.e., the following system of equations holds simultaneously

$$\forall c \in \mathcal{C},\ T^c(S'|S, a) = \sum_{a' \in \mathcal{A}'} T_0(S'|S, a, a') \cdot \pi_c(a'|S) \quad (3)$$

We omit the analogous factorization of R^c. It turns out that by an algorithm that resembles the *Gram-Schmidt Orthogonalization* [Cheney and Kincaid, 2009], one can obtain a feasible factorization of minimum size within polynomial time. The full proof is deferred to Appendix A.7. □

To clearly position our results, we highlight several key connections between our findings and the existing literature.

1. If the opponent is allowed to switch between support strategies independently at each state (i.e., a scenario of using a mixed-strategy-induced behavioral strategy), then our construction in Theorem 1 for computing the best response corresponds to solving the *belief-induced MDP* described in [Zhu and Lin, 2025].
2. Wang and Lin [2019] observed that there may not exist a pure one-memory strategy as a best response against a population of one-memory opponents, each potentially adopting a different strategy (e.g., a tournament). Here, Theorem 5 and 6 provide some formal evidence: in general, the best response is not even within constant-memory; instead, it may incorporate infinite memory.
3. Computing best responses to mixed strategies is a building block of the recursive formulation in the I-POMDP framework [Gmytrasiewicz and Doshi, 2005]. Thus, Theorem 5 and 6 can serve as the missing justification for why recursively incorporating POMDPs or CMPDs, rather than vanilla MDPs, is essential for I-POMDPs.

One may also notice a possible correspondence to *Epistemic Game Theory* [Dekel and Siniscalchi, 2015], as the latter also investigates type and belief structures, albeit through a different formalism; we intend to establish more rigorous connections in future work.

One may further wonder whether a group of agents can form an NE if all of them play mixed strategies, i.e.,

$$\forall i \in \mathcal{N},\ (\Pi^{K+}_i, \vec{p}_i) \in BR((\Pi^{K+}_{-i}, \vec{p}_{-i})).$$

If the support sets $\{\Pi^{K+}_i\}_{i \in \mathcal{N}}$ are exogenously specified, one can invoke Nash's existence theorem, as the game becomes finite. Due to the page limit, we defer the elaboration to Appendix B for interested readers; however, the application of such theoretic results remains unclear. Of greater significance is the following **open problem**: *In general, given a mixed constant-memory strategy, does there exist a best response that is also a mixed constant-memory strategy, rather than an infinite-memory behavioral one?*

5 Empirical Study

The purpose of these empirical studies is not to benchmark the RL algorithms mentioned in this section; rather, it is to present an intuitive illustration of the effects of memory. The evaluated domains include two matrix games played sequentially, and one domain borrowed from robotics with raw image inputs.

5.1 Sequential Matrix Games

We consider two matrix games that are played in a repeated manner, namely the *Iterated Prisoner's Dilemma* (IPD), and the *Iterated Traveler's Dilemma* (ITD).

The Iterated Prisoner's Dilemma

The payoff matrix is shown in the table below. We also remind the readers of the `Axelrod` library[3] that implements most of the strategies from the well-known Axelrod's IPD tournament. *Our approach can compute the best response of any constant-memory strategy in this library, whether deterministic or randomized.* In particular, we would like to highlight a family of strategies, called N*-Tit(s)-for-*M*-Tat(s)* (originally named by Harper *et al.* [2017]), which is a parameterized version of the classic *Tit-for-Tat*. An agent adopting N*-Tit(s)-for-*M*-Tat(s)* will retaliate immediately after it has been *defected* M times, by responding with *defection* in the next N rounds. Thus, it is a $\max(N, M)$-memory strategy.

	C	D
C	(1, 1)	(-1, 2)
D	(2, -1)	(0, 0)

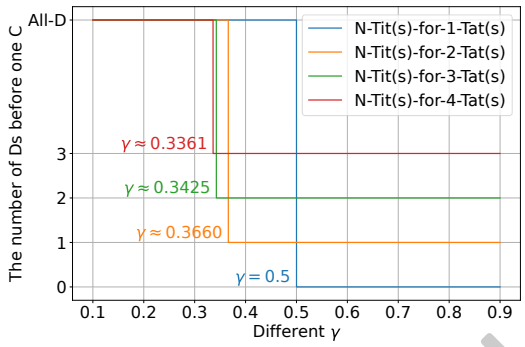

Figure 1: Experimental results of the IPD.

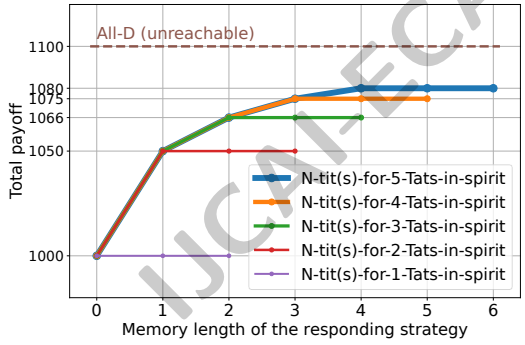

Figure 2: Experimental results of the ITD.

We compute the best responses for various settings of (N, M) and the discount factor γ, and illustrate our findings in Figure 1. We utilize `MDPtoolbox` [Chadès *et al.*, 2014] to compute the exact solutions of the formulated MDPs, where the resulted policies are all deterministic ones. One can observe clear phase transitions in the best responses. Reading the figure from right to left, it indicates that when the discount factor is sufficiently large, the best response to N*-Tit(s)-for-*M*-Tat(s)* is to *defect* (M-1) times followed by one *cooperation*, and to repeat this pattern periodically, regardless of N. However, when the agent is less patient by placing less value on future reciprocity, it will consider taking one additional round of *defection*, leading to permanent mutual defection from the (M+1)-th round onwards.

This finding is summarized in a formal theorem that can be mathematically justified. We also compute the closed-form solutions for those values of γ that trigger the phase transition. *Please refer to Appendix C for details.* Please note that while this paper focuses on the discounted-payoff setting, our code also includes the computation of best responses under the average-payoff setting, although a detailed discussion is omitted here.

We note that readers in the IPD community may be particularly interested in *Zero Determinant* (ZD) strategies [Press and Dyson, 2012]. Originally derived under the scenario of average payoffs, ZD strategies were later generalized to settings with discounted payoffs. These strategies enable a player to unilaterally enforce a linear relationship between its own payoff and that of its opponent, which does **not** generally imply anything strong about best responses. That said, certain ZD strategies can constitute NEs: for instance, *Tit-for-Tat* is both a ZD strategy and a best response to itself.

The Iterated Traveler's Dilemma

One may notice that the aforementioned *(M-1)-D-before-One-C* strategy against N*-Tit(s)-for-*M*-Tat(s)* can actually be implemented using only $(M-1)$ memory instead of $\max(N, M)$ memory. Specifically, if there has been no *cooperation* played by itself in the previous $(M-1)$ rounds, it should *cooperate* in the current round; otherwise, it should continue *defecting*. However, we have not addressed the theoretic case of computing the best response against a K-memory strategy using only K'-memory with $K' < K$. Note that one cannot formulate a K-memory MDP but compute its optimal policy in the form of K'-memory using dynamic programming, as this would result in inconsistent policy updates. To circumvent this issue, we can run model-free algorithms, e.g., Q-learning which only requires a form of the policy in advance, to see the outcome K'-memory strategy.

Therefore, we generalize our previous findings to a broader class of games, called the *Iterated Traveler's Dilemma* (ITD) [Dasler and Tosic, 2010; Tošic, 2016], which repeats over the matrix game of traveler's dilemma [Basu, 1994] with payoffs given as,

$$u_i(a_i, a_{-i}) \triangleq \min(a_i, a_{-i}) + 2 \cdot sign(a_{-i} - a_i) \qquad (4)$$

where each action a_i is an integer variable, also known as a *bid*. Please note that when $i \in \{1, 2\}$ and $a_i \in \{0, 1\}$ for each i, it reduces to the aforementioned prisoner's dilemma. In our study, we consider $i \in \{1, 2\}$ with a more fine-grained action space $a_i \in \{0, 1, \cdots 10\}$ to render a harder computational problem. As the investigation of ITD is still relatively nascent, we also provide additional background as well as some justification for its significance in Appendix D.

We implement *N-Tit(s)-for-M-Tat(s)* in spirit, as there are no longer well-defined notions of *defections* or *cooperations*.

[3] https://github.com/Axelrod-Python/Axelrod

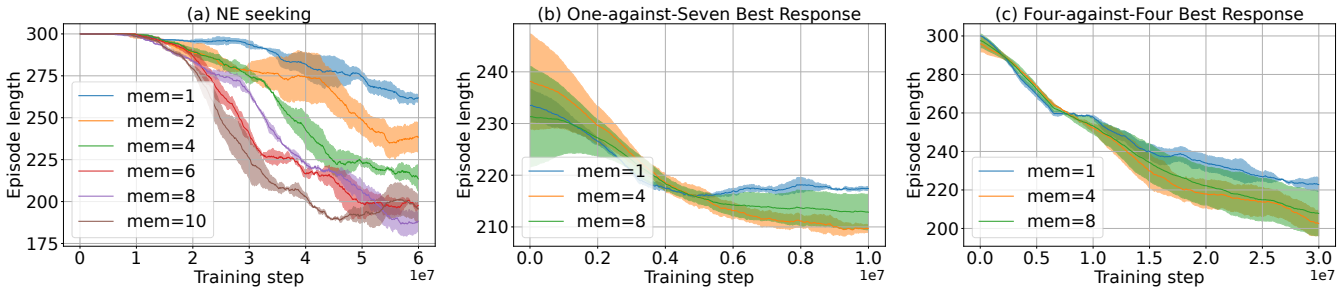

Figure 3: Experimental results of the Pursuit domain. (a) NE seeking; (b) Single agent best responding to the rest 7 agents; (c) One team of four agents best responding to the other team of four agents.

In this context, if an agent finds that its opponent's last bid is smaller than its own last bid, this will be interpreted as a *defection*; otherwise, it is treated as a *cooperation*. Tabular Q-learning is leveraged to compute the best responses with various memory length against *N-Tit(s)-for-M-Tat(s)* for different values of M. For a specific value of M, we compute the best response using memory lengths ranging from 0 to $(M+1)$. We illustrate the total (undiscounted) payoff for 100 rounds in Figure 2. It turns out that, when restricted to M'-memory, the best response against *N-Tit(s)-for-M-Tat(s)* is exactly M'*-D-before-One-C*, given any $M' < M$. For example, playing against *N-Tit(s)-for-5-Tat(s)*, the best response restricted to only 3-memory will be *Three-D-before-One-C*, resulting in the agent exploiting its opponent for 3/4 of the time, with a total payoff $= 10 \times 25 + 11 \times 75 = 1075$.

5.2 The Pursuit Domain

The aforementioned two games present clear social dilemmas, such that the technique used in the proof of Theorem 2 may only find NEs where both players *defect* from the very beginning, regardless of the memory length utilized (cf. the aforementioned notebook[4]). Therefore, we also conduct some experiments on a more intricate testbed borrowed from the robotics community, namely the *Pursuit* domain [Gupta *et al.*, 2017]. In this task, 8 pursuer agents attempt to catch 20 random walkers (also called evaders). Each pursuer agent can only observe a limited local range, and once 4 pursuers simultaneously overlap with the same evader, this evader will be removed from the game. An episode terminates immediately after all the evaders are removed. Ideally, these 8 agents will devide into two teams for evader hunting. It is actually a Partially Observable SG (POSG) rather than strictly an SG. We aim to investigate: 1) whether longer memories will lead to improved NEs; 2) how well one agent can respond to the 7 others; 3) how well one team (four agents) can respond to the other team (the rest four).

The main results are presented in Figure 3 focusing solely on the results obtained with DQN [Mnih *et al.*, 2013], as it significantly outperforms other algorithms in this task. *Additional benchmarking results using other algorithms, e.g., A2C [Mnih* et al.*, 2016] and PPO [Schulman* et al.*, 2017], along with the detailed hyper-parameters, are provided in Appendix E for the reader's reference.* To increase the memory length, we simply stack the historical observations and actions. For multi-agent learning in search for NEs, we equip each agent with an identical network and train them to learn independently. As shown in Figure 3(a), utilizing longer memory indeed helps the pursuers catch the evaders faster, indicating a better NE. We extract the eventual strategy trained with 8-memory as it appears to be the best. In the experiments depicted in Figure 3(b), 7 agents are equipped with this pretrained 8-memory strategy, while the remaining agent learns from scratch to find the best response. In the experiments shown in Figure 3(c), one team of 4 agents are equipped with this pretrained 8-memory strategy, leaving the remaining team of the rest 4 agents learning from scratch to find a best "team" response. As a result, using memories of length 4 and 8 is clearly better than using memories of length one. However, 8-memory responses are not significantly distinguishable from 4-memory responses, which may be attributed to the fact that 4-memory strategies are already sufficient to serve as the best response, or possibly due to some representation error introduced by the deep neural networks.

It is also interesting to note that, the improvement, which is reflected by the episode length, made by one agent with the other 7 agents fixed (as shown in Figure 3(b)) is clearly less then that made by a team of agents with the other team fixed (as shown in Figure 3(c)). As we examined, in the former case, there is typically one of the 7 fixed agents who occasionally collaborates with three of them and at other times with the remaining three, creating the pseudo-effect of two four-agent teams. This observation may potentially explain why adding one more learning agent only leads to incremental improvement.

6 Conclusion

In this work, we formally study constant-memory strategies in stochastic games. Some theoretic results for best responses and equilibria in the form of *behavioral* constant-memory strategies are developed. Moreover, we highlight that best responding to *mixed* constant-memory strategies is a significantly harder computational problem—possibly even not computable. These results extend prior work in two main aspects: they generalize 1) the findings of [Chen *et al.*, 2017a; Wang and Lin, 2019] from repeated games to stochastic games; and 2) the planning framework in [Zhu and Lin, 2025] from stationary opponents to K-memory ones. We also conduct experiments on well-known social dilemmas as well as a multi-robot domain to verify those theoretic insights.

[4]Please refer to `kMemNE_full.ipynb` in our code.

References

[Albrecht and Ramamoorthy, 2015] Stefano V Albrecht and Subramanian Ramamoorthy. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems. *arXiv:1506.01170*, 2015.

[Albrecht and Ramamoorthy, 2019] Stefano V Albrecht and Subramanian Ramamoorthy. On convergence and optimality of best-response learning with policy types in multiagent systems. *arXiv:1907.06995*, 2019.

[Albrecht and Stone, 2018] Stefano V Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

[Albrecht *et al.*, 2016] Stefano V Albrecht, Jacob W Crandall, and Subramanian Ramamoorthy. Belief and truth in hypothesised behaviours. *Artificial Intelligence*, 235:63–94, 2016.

[Baarslag *et al.*, 2016] Tim Baarslag, Mark JC Hendrikx, Koen V Hindriks, and Catholijn M Jonker. Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques. *Autonomous Agents and Multi-Agent Systems*, 30:849–898, 2016.

[Baarslag, 2024] Tim Baarslag. Multi-deal negotiation. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 2668–2673, 2024.

[Balseiro *et al.*, 2015] Santiago R Balseiro, Omar Besbes, and Gabriel Y Weintraub. Repeated auctions with budgets in ad exchanges: Approximations and design. *Management Science*, 61(4):864–884, 2015.

[Basu, 1994] Kaushik Basu. The traveler's dilemma: Paradoxes of rationality in game theory. *The American Economic Review*, 84(2):391–395, 1994.

[Ben-Porath, 1990] Elchanan Ben-Porath. The complexity of computing a best response automaton in repeated games with mixed strategies. *Games and Economic Behavior*, 2(1):1–12, 1990.

[Benjamins *et al.*, 2023] Carolin Benjamins, Theresa Eimer, Frederik Schubert, Aditya Mohan, Sebastian Döhler, André Biedenkapp, Bodo Rosenhahn, Frank Hutter, and Marius Lindauer. Contextualize me – the case for context in reinforcement learning. *Transactions on Machine Learning Research*, 2023.

[Carmel and Markovitch, 1998] David Carmel and Shaul Markovitch. How to explore your opponent's strategy (almost) optimally. In *Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160)*, pages 64–71. IEEE, 1998.

[Chadès *et al.*, 2014] Iadine Chadès, Guillaume Chapron, Marie-Josée Cros, Frédérick Garcia, and Régis Sabbadin. Mdptoolbox: a multi-platform toolbox to solve stochastic dynamic programming problems. *Ecography*, 37(9):916–920, 2014.

[Chen *et al.*, 2017a] Lijie Chen, Fangzhen Lin, Pingzhong Tang, Kangning Wang, Ruosong Wang, and Shiheng Wang. K-memory strategies in repeated games. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1493–1498, 2017.

[Chen *et al.*, 2017b] Lijie Chen, Pingzhong Tang, and Ruosong Wang. Bounded rationality of restricted turing machines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[Cheney and Kincaid, 2009] Elliott Ward Cheney and David Ronald Kincaid. *Linear Algebra: Theory and Applications*. Jones & Bartlett Learning, 2009.

[Dasler and Tosic, 2010] P Dasler and P Tosic. The iterated traveler's dilemma: Finding good strategies in games with "bad" structure: Preliminary results and analysis. In *Proc of the 8th Euro. Workshop on Multi-Agent Systems, EUMAS*, volume 10, 2010.

[Dekel and Siniscalchi, 2015] Eddie Dekel and Marciano Siniscalchi. Epistemic game theory. In *Handbook of game theory with economic applications*, volume 4, pages 619–702. Elsevier, 2015.

[Fink, 1964] Arlington M Fink. Equilibrium in a stochastic n-person game. *Journal of science of the hiroshima university, series ai (mathematics)*, 28(1):89–93, 1964.

[Fujimoto *et al.*, 2024] Yuma Fujimoto, Kaito Ariu, and Kenshi Abe. Memory asymmetry creates heteroclinic orbits to nash equilibrium in learning in zero-sum games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17398–17406, 2024.

[Gmytrasiewicz and Doshi, 2005] Piotr J Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005.

[Gupta *et al.*, 2017] Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 66–83. Springer, 2017.

[Hallak *et al.*, 2015] Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

[Harper *et al.*, 2017] Marc Harper, Vincent Knight, Martin Jones, Georgios Koutsovoulos, Nikoleta E Glynatsi, and Owen Campbell. Reinforcement learning produces dominant strategies for the iterated prisoner's dilemma. *PloS one*, 12(12):e0188046, 2017.

[Iyer *et al.*, 2014] Krishnamurthy Iyer, Ramesh Johari, and Mukund Sundararajan. Mean field equilibria of dynamic auctions with learning. *Management Science*, 60(12):2949–2970, 2014.

[Kaelbling *et al.*, 1998] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[Knoblauch, 1994] Vicki Knoblauch. Computable strategies for repeated prisoner' s dilemma. *Games and Economic Behavior*, 7(3):381–389, 1994.

[Littman, 1994] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[Lowe *et al.*, 2017] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[Madani *et al.*, 2003] Omid Madani, Steve Hanks, and Anne Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, 147(1-2):5–34, 2003.

[Megiddo and Wigderson, 1986] Nimrod Megiddo and Avi Wigderson. On play by means of computing machines: preliminary version. In *Theoretical aspects of reasoning about knowledge*, pages 259–274. Elsevier, 1986.

[Mirsky *et al.*, 2022] Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V. Albrecht. A survey of ad hoc teamwork research, 2022.

[Mnih *et al.*, 2013] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[Mnih *et al.*, 2016] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PmLR, 2016.

[Nachbar and Zame, 1996] John H Nachbar and William R Zame. Non-computable strategies and discounted repeated games. *Economic theory*, 8:103–122, 1996.

[Press and Dyson, 2012] William H Press and Freeman J Dyson. Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences*, 109(26):10409–10413, 2012.

[Puig *et al.*, 2023] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023.

[Rashid *et al.*, 2020] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

[Rubinstein, 1986] Ariel Rubinstein. Finite automata play the repeated prisoner's dilemma. *Journal of economic theory*, 39(1):83–96, 1986.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

[Shapley, 1953] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[Shen *et al.*, 2020] Weiran Shen, Binghui Peng, Hanpeng Liu, Michael Zhang, Ruohan Qian, Yan Hong, Zhi Guo, Zongyao Ding, Pengjun Lu, and Pingzhong Tang. Reinforcement mechanism design: With applications to dynamic pricing in sponsored search auctions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2236–2243, 2020.

[Simon, 1990] Herbert A Simon. Bounded rationality. *Utility and probability*, pages 15–18, 1990.

[Sondik, 1978] Edward J Sondik. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations research*, 26(2):282–304, 1978.

[Stern, 2019] Roni Stern. Multi-agent path finding–an overview. *Artificial Intelligence*, pages 96–115, 2019.

[Su *et al.*, 2024] Kefan Su, Yusen Huo, Zhilin Zhang, Shuai Dou, Chuan Yu, Jian Xu, Zongqing Lu, and Bo Zheng. Auctionnet: A novel benchmark for decision-making in large-scale games. *Advances in Neural Information Processing Systems*, 37:94428–94452, 2024.

[Takahashi, 1964] Masayuki Takahashi. Equilibrium points of stochastic non-cooperative n-person games. *Journal of Science of the Hiroshima University, Series AI (Mathematics)*, 28(1):95–99, 1964.

[Tošic, 2016] Predrag T Tošic. On learning and co-learning effective strategies in iterated travelers' dilemma. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 674–677. IEEE, 2016.

[Wang and Lin, 2019] Shiheng Wang and Fangzhen Lin. Pure strategy best responses to mixed strategies in repeated games. *arXiv preprint arXiv:1902.09066*, 2019.

[Yu *et al.*, 2022] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[Zhu and Lin, 2025] Fengming Zhu and Fangzhen Lin. Single-agent planning in a multi-agent system: A unified framework for type-based planners. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 2382–2391, 2025.

[Zuo and Tang, 2015] Song Zuo and Pingzhong Tang. Optimal machine strategies to commit to in two-person repeated games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.