Constant-Memory Strategies in Stochastic Games: Best Responses and Equilibria

Extended Abstract

Fengming Zhu
Hong Kong University of Science and Technology
Hong Kong SAR, China
fzhuae@connect.ust.hk

Fangzhen Lin
Hong Kong University of Science and Technology
Hong Kong SAR, China
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 some discussion on the computational hardness of best responding to *mixed* constant-memory strategies.[1]

KEYWORDS

Stochastic games; Bounded rationality; Best response; Agent memory; Reinforcement learning

ACM Reference Format:
Fengming Zhu and Fangzhen Lin. 2026. Constant-Memory Strategies in Stochastic Games: Best Responses and Equilibria: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/SQOI3711

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 [1, 2, 8], bidding and mechanism design in repeated auctions [3, 11, 12, 14, 27, 30], multi-agent teamwork [20, 26, 29], and even human-robot collaboration [22, 34]. Stochastic games, also known as Markov games, model the interactions under 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 [25]. The existence of equilibria formed by stationary strategies in n-player general-sum stochastic games was later proven by Fink [9] and Takahashi [31], under mild assumptions. Despite being highly restricted in terms of 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 [16] and implemented in a line of subsequent work [10, 17, 23, 33].

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), with the ability to remember only one past action played by the opponent, the well-known *Tit-For-Tat* (TFT) strategy (start 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 [4, 24, 36] and even Turing machines [7, 15, 19, 21], 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 [28] in general, and is highly implementable in practice. Note that this notion has been preliminarily investigated by Chen et al. [6] and Wang and Lin [32]. 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 start by presenting a characterization of best responses to *behavioral* constant-memory strategies, as well as the existence result of Nash equilibria. Then, we provide some negative results for computing best responses to *mixed* constant-memory strategies.

While not elaborated in this short paper, we report some experimental results in the full paper on several sequential decision-making testbeds, including the *Iterated Prisoner's Dilemma*, the *Iterated Traveler's Dilemma*, and the *Pursuit* domain, aiming 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 code is available at `https://github.com/Fernadoo/Const-Mem`.

2 PRELIMINARIES

The long-horizon interaction of multiple agents is modelled as a *stochastic game* (SG). An SG is a 5-tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, T, R \rangle$,

(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.
(4) $T : \mathcal{S} \times \mathcal{A}_1 \times \cdots \mathcal{A}_n \mapsto \Delta(\mathcal{S})$ defines stochastic transitions.
(5) $R_i : \mathcal{S} \times \mathcal{A}_1 \times \cdots \mathcal{A}_n \mapsto \mathbb{R}$ denotes the rewards for agent i.

[1] The full version can be found at `https://arxiv.org/abs/2505.07008`.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/SQOI3711

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. Given any non-negative integer K, a *behavioral* 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 defined 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 *behavioral* 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 as the *stationary strategies*.

The objective for each agent is to maximize its accumulated discounted rewards. 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 strategy profile (π_i, π_{-i}) starting from any state $S \in \mathcal{S}$ is

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

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

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

requiring that a π_i must outperform any other in Π_i^∞ to serve as the best response.

3 MAIN RESULTS

3.1 Best Responses to Behavioral Strategies

We first characterize the best responses of an agent when all the other opponents are using *behavioral* constant-memory strategies, followed by the existence result of equilibria.

Theorem 3.1. *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.*

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.*

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

Theorem 3.2. *There exists an NE when the agents are all adopting K-memory strategies, for any arbitrary non-negative finite K.*

3.2 Best Responses to Mixed Strategies

When an agent i is said to adopt a behavioral 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. Slightly different from a behavioral strategy, a *mixed* strategy (for agent i and of K-memory) 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. The *overall utility* will be calculated as the expected payoff over all possible matches.

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. Unfortunately, this induced strategy may not lead to the same outcome as the original mixed strategy in general, although the equivalence happens to hold for repeated games with stationary strategies.

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 $(H, S) \in \mathcal{H}^{\leq K} \times \mathcal{S}$,*

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

Theorem 3.3 (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 $\omega_{(\Pi_{-i}^{0+},\vec{p})}$.*

We also show that computing the best response against a mixed K-memory strategy is as hard as optimally solving infinite-horizon *Contextual MDPs* (CMDPs) [5, 13]. 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 [18]. We prove the above result by presenting a two-way reduction.

Theorem 3.4. *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 a special case of infinite-horizon POMDPs.*

Theorem 3.5. *Optimally solving a CMDP can be reduced to computing the best response for an agent i against a profile of mixed zero-memory strategies $(\Pi_{-i}^{0+}, \vec{p})$ adopted by its opponents.*

4 CONCLUSION

In this work, we develop a theoretic framework to study constant-memory strategies. The notion of best responses and equilibria are well-established. In particular, we highlight that best responding to mixed constant-memory strategies may be computationally hard, possibly even not computable. These results can be seen as an extension of both [6, 32] (from repeated games to stochastic games) and [35] (from stationary strategies to K-memory ones). In the full paper, we also report some experimental results conducted on two well-known social dilemmas as well as a multi-robot domain to verify those theoretic insights.

REFERENCES

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

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

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

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

[5] Carolin Benjamins, Theresa Eimer, Frederik Schubert, Aditya Mohan, Sebastian Döhler, André Biedenkapp, Bodo Rosenhahn, Frank Hutter, and Marius Lindauer. 2023. Contextualize Me – The Case for Context in Reinforcement Learning. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=Y42xVBQusn

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

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

[8] Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. 2017. Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems* 31 (2017), 250–287.

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

[10] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[11] Jiayan Guo, Yusen Huo, Zhilin Zhang, Tianyu Wang, Chuan Yu, Jian Xu, Bo Zheng, and Yan Zhang. 2024. Generative Auto-bidding via Conditional Diffusion Modeling. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5038–5049.

[12] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. 2019. Learning mean-field games. *Advances in neural information processing systems* 32 (2019).

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

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

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

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

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

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

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

[20] Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V. Albrecht. 2022. A Survey of Ad Hoc Teamwork Research. arXiv:2202.10450 [cs.MA]

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

[22] 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. 2023. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724* (2023).

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

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

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

[26] Guni Sharon, Roni Stern, Ariel Felner, and Nathan R Sturtevant. 2015. Conflict-based search for optimal multi-agent pathfinding. *Artificial intelligence* 219 (2015), 40–66.

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

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

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

[30] Kefan Su, Yusen Huo, Zhilin Zhang, Shuai Dou, Chuan Yu, Jian Xu, Zongqing Lu, and Bo Zheng. 2024. AuctionNet: A Novel Benchmark for Decision-Making in Large-Scale Games. *Advances in Neural Information Processing Systems* 37 (2024), 94428–94452.

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

[32] Shiheng Wang and Fangzhen Lin. 2019. Pure Strategy Best Responses to Mixed Strategies in Repeated Games. *arXiv preprint arXiv:1902.09066* (2019).

[33] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=YVXaxB6L2Pl

[34] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. 2023. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485* (2023).

[35] Fengming Zhu and Fangzhen Lin. 2025. 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*. 2382–2391.

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