Bayes-ToMoP: A Fast Detection and Best Response Algorithm Towards Sophisticated Opponents

Extended Abstract

Tianpei Yang, Jianye Hao*,
Zhaopeng Meng, Yan Zheng
College of Intelligence and
Computing, Tianjin University
Tianjin, China
{tpyang,jianye.hao,mengzp,
yanzheng}@tju.edu.cn

Chongjie Zhang
MMW, Tsinghua University
Beijing, China
chongjie@tsinghua.edu.cn

Ze Zheng
Beifang Investigation,Design &
Research CO.LTD
Tianjin, China
zezheng0618@gmail.com

ABSTRACT

Multiagent algorithms often aim to accurately predict the behaviors of other agents and find a best response accordingly. Previous works usually assume an opponent uses a stationary strategy or randomly switches among several stationary ones. However, an opponent may exhibit more sophisticated behaviors by adopting more advanced reasoning strategies, e.g., using a Bayesian reasoning strategy. This paper proposes a novel approach called Bayes-ToMoP which can efficiently detect the strategy of opponents using either stationary or higher-level reasoning strategies. Bayes-ToMoP also supports the detection of previously unseen policies and learning a best-response policy accordingly. We also propose a deep version of Bayes-ToMoP by extending Bayes-ToMoP with DRL techniques. Experimental results show both Bayes-ToMoP and deep Bayes-ToMoP outperform the state-of-the-art approaches when faced with different types of opponents in two-agent competitive games.

CCS CONCEPTS

• **Computing methodologies → Multi-agent systems**;

KEYWORDS

Multiagent learning; policy reuse; theory of mind

ACM Reference Format:
Tianpei Yang, Jianye Hao*, Zhaopeng Meng, Yan Zheng, Chongjie Zhang, and Ze Zheng. 2019. Bayes-ToMoP: A Fast Detection and Best Response Algorithm Towards Sophisticated Opponents. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

In multiagent systems, the ideal behavior of an agent is contingent on the behaviors of coexisting agents. However, agents may exhibit different behaviors adaptively depending on the contexts they encounter. Hence, it is critical for an agent to quickly predict or recognize the behaviors of other agents, and make a best response accordingly [1, 4, 9, 12].

Several works extended the idea of Bayesian Policy Reuse (BPR) [13], which was originally proposed to determine the best policy when faced with different tasks, to multiagent settings to predict the behaviors of the opponent [8, 10, 15]. However, all these approaches assume that an opponent randomly switches its policies among a class of stationary ones. In practice, an opponent can exhibit more sophisticated behaviors by adopting more advanced reasoning strategies, thus requiring more advanced techniques to beat such kinds of sophisticated opponents.

The above problems can be partially addressed by introducing the concept of Theory of Mind (ToM) [2, 6], which is a kind of recursive reasoning technique [1] describing a cognitive mechanism of explicitly attributing unobservable mental contents such as beliefs to other players. Previous methods often use nested beliefs and "simulate" the reasoning processes of other agents to predict their actions [3, 5, 14]. However, these approaches are designed for predicting the opponent's primitive actions instead of high-level strategies, resulting in slow adaptation to non-stationary opponents. Furthermore, they show poor performance against an opponent using previously unseen strategies.

To address the above challenges, we propose a novel algorithm, named Bayesian Theory of Mind on Policy (Bayes-ToMoP) to beat such sophisticated opponents. In contrast to BPR which is capable of detecting non-stationary opponents only, Bayes-ToMoP incorporates ToM into BPR to quickly and accurately detect not only non-stationary, and more sophisticated opponents and compute a best response accordingly. Besides, Bayes-ToMoP also supports detecting whether an opponent is using a previously unseen policy and learning an optimal response against it. We also propose a deep version of Bayes-ToMoP by extending Bayes-ToMoP with DRL techniques.

2 BAYES-TOMOP

We use the notation of Bayes-ToMoP$_k$ to denote an agent with the ability of using Bayes-ToMoP up to the k-th order. Intuitively, Bayes-ToMoP$_i$ with a higher-order theory of mind could take advantage of any Bayes-ToMoP$_j$ with a lower-order one ($i > j$). We focus on Bayes-Bayes-ToMoP$_1$ in the following section and Bayer-TomoP$_k$ ($k > 1$) can be naturally constructed by incorporating a higher-order ToM idea into our framework.

Bayes-ToMoP$_1$ Algorithm A Bayes-ToMoP$_1$ agent maintains its zero-order belief which is a probability that its opponent may

[*] Corresponding author.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

adopt each strategy, and first-order belief, which is a probability distribution that describes the probability that an agent believes his opponent believes it will choose a policy. Bayes-ToMoP$_1$ agent first predicts the policy $\hat{j}$ of its opponent assuming the opponent maximizes its own utility under its first-order belief. Then, an integration function I is introduced to compute the final prediction results following Equation 1 [3].

$$I(\beta^{(0)}, \hat{j}, c_1)(j) = \begin{cases} (1 - c_1)\beta^{(0)}(j) + c_1 & \text{if } j = \hat{j} \\ (1 - c_1)\beta^{(0)}(j) & \text{otherwise} \end{cases} \tag{1}$$

where, $0 \geq c_1 \leq 1$ is the first-order confidence for the first-order prediction. Next, Bayes-ToMoP$_1$ agent computes the optimal policy based on the integrated belief. The next issue is how to update the value of c_1 which can be seen as the exploration rate of using first-order belief to predict the opponent's strategies. We propose an adaptive and generalized mechanism to adjust the value of c_1, formally we have:

$$c_1 = \begin{cases} ((1 - \lambda)c_1 + \lambda)\mathbf{F}(v_i) & \text{if } v_i \geq v_{i-1} \\ (\frac{lgv_i}{lg(v_i - \delta)}c_1)\mathbf{F}(v_i) & \text{if } \delta < v_i < v_{i-1} \\ \lambda \mathbf{F}(v_i) & \text{if } v_i \leq \delta \end{cases} \tag{2}$$

where v_i is the winning rate during most recent l episodes, δ is the threshold of the winning rate v_i, and $\mathbf{F}(v_i)$ is an indicator function to control the direction of adjusting the value of c_1:

$$\mathbf{F}(v_i) := \begin{cases} 1 & \text{if } (v_i \leq \delta \ \& \ \mathbf{F}(v_i) = 0) \\ 0 & \text{if } (v_i \leq \delta \ \& \ \mathbf{F}(v_i) = 1) \end{cases} \tag{3}$$

Bayes-ToMoP$_1$ agent also detects whether the opponent is using an unknown strategy, which is achieved by recording a fixed length of game outcomes and using the winning rate θ_i over the most recent h episodes as the signal indicating the average performance over all policies till the current episode i. If the winning rate θ_i is lower than a given threshold δ ($\theta_i < \delta$), it indicates that all existing policies show poor performance against the current opponent strategy, in this way Bayes-ToMoP$_k$ agent infers that the opponent is using a previously unseen policy outside the current policy library.

After detecting the opponent is using a new strategy, the agent begins to learn the best-response policy against it. Since our learning framework is general, various RL approaches can be applied to learn the optimal policy. Finally, the new policy and the estimated opponent policy are added to the policy library and its opponent's policy library respectively.

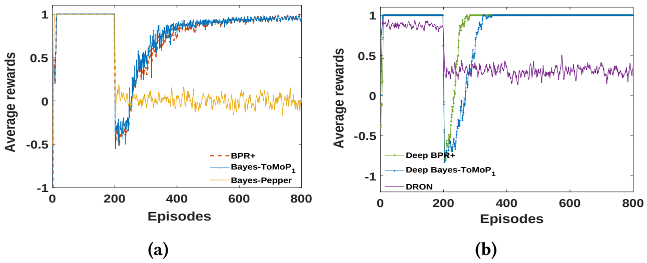
3 SIMULATIONS

Bayes-ToMoP is evaluated on soccer [7, 11] compared with state-of-the-art tabular approaches (BPR+ [10] and Bayes-Pepper [8]). For deep Bayes-ToMoP, we compare with DRON [7] and deep BPR+ [15]. Three kinds of opponents are considered: (1) a Bayes-ToMoP$_0$ opponent (O$_{ToMoP_0}$); (2) an opponent that randomly switches its policy among stationary strategies and lasts for an unknown number of episodes (O$_{ns}$) and (3) an opponent switching its strategy between stationary strategies and Bayes-ToMoP$_0$ (O$_{ToMoP_0}$-s).

Table 1 shows the average performance of various approaches against different opponents. We can see that only Bayes-ToMoP$_1$ can beat all kinds of opponents. However, BPR+, Bayes-Pepper, and deep BPR+ can only detect the non-stationary opponents but fail when the opponent switches its strategy to Bayes-ToMoP$_0$. DRON performs similarly against three kinds of opponents. However, it fails to achieve the highest average winning rate against each particular type of opponents. Figure 1 (a-b) shows the dynamics of average rewards of different approaches with the opponent starting with one of the known strategies and switching to the new one at the 200th episode. We can observe that Bayes-ToMoP$_1$, BPR+ and Deep BPR+ can quickly detect the new opponent strategy, and finally learn an optimal policy. However, Bayes-Pepper and DRON fail.

Table 1: Average winning rates with std.dev.(±) in soccer.

Opponents / Approaches	O$_{ToMoP_0}$	O$_{ns}$	O$_{ToMoP_0}$-s
BPR+	49.78%±1.71%	99.37%±0.72%	66.8%±0.57%
Bayes-Pepper	49.45%±2.32%	99.47%±0.35%	66.2%±0.61%
DRON	74.75%±0.19%	76.54%±0.16%	75.22%±0.18%
Deep BPR+	71.57%±1.26%	**99.49%±0.51%**	78.6%±0.76%
Bayes-ToMoP$_1$	**99.82%±0.18%**	98.21%±0.37%	**98.48%±0.54%**



(a) (b)

Figure 1: Different approaches against an opponent using an unknown policy in soccer game.

4 CONCLUSION

This paper presents a novel algorithm called Bayes-ToMoP to handle not only switching, non-stationary opponents and also more sophisticated ones (e.g., BPR-based). Bayes-ToMoP also enables an agent to learn a new optimal policy when encountering a previously unseen strategy. Extensive simulations show Bayes-ToMoP outperforms the state-of-the-art approaches both in tabular and deep learning environments. As future work, it is worth investigating how to accelerate the online new policy learning phase and how to extend Bayes-ToMoP to multi-opponent scenarios. Furthermore, higher order of Bayes-ToMoP is worth investigating to handle more sophisticated opponents and apply to large scale, real scenarios.

ACKNOWLEDGMENTS

The work is supported by the National Natural Science Foundation of China (Grant Nos.: 61702362, U1836214), Special Program of Artificial Intelligence, Tianjin Research Program of Application Foundation and Advanced Technology (No.: 16JCQNJC00100), and Special Program of Artificial Intelligence of Tianjin Municipal Science and Technology Commission (No.: 569 17ZXRGGX00150).

REFERENCES

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

[2] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the Annual Meeting of the Cognitive Science Society* 33 (2011), 2469–2474.

[3] Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. 2013. How much does it help to know what she knows you know? An agent-based simulation study. *Artificial Intelligence* 199 (2013), 67–92.

[4] Fernando Fernández and Manuela Veloso. 2006. Probabilistic policy reuse in a reinforcement learning agent. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. ACM, 720–727.

[5] Piotr J Gmytrasiewicz and Prashant Doshi. 2005. A framework for sequential planning in multi-agent settings. *J. Artif. Intell. Res.(JAIR)* 24 (2005), 49–79.

[6] Alvin I. Goldman. 2012. Theory of mind. *The Oxford handbook of philosophy of cognitive science* (2012), 402–424.

[7] He He and Jordan L. Boyd-Graber. 2016. Opponent Modeling in Deep Reinforcement Learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*. 1804–1813.

[8] Pablo Hernandez-Leal and Michael Kaisers. 2017. Towards a fast detection of opponents in repeated stochastic games. In *Proceedings of the 2017 International Conference on Autonomous Agents and Multiagent Systems*. Springer, 239–257.

[9] Pablo Hernandez-Leal, Michael Kaisers, Tim Baarslag, and Enrique Munoz de Cote. 2017. A Survey of Learning in Multiagent Environments: Dealing with Non-Stationarity. *CoRR* abs/1707.09183 (2017).

[10] Pablo Hernandez-Leal, Matthew E. Taylor, Benjamin Rosman, Luis Enrique Sucar, and Enrique Munoz de Cote. 2016. Identifying and Tracking Switching, Non-Stationary Opponents: A Bayesian Approach. In *Multiagent Interaction without Prior Coordination, Papers from the 2016 AAAI Workshop*.

[11] Michael L. Littman. 1994. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Proceedings of the Eleventh International Conference on Machine Learning*. 157–163.

[12] Rob Powers and Yoav Shoham. 2005. Learning against opponents with bounded memory. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*. 817–822.

[13] Benjamin Rosman, Majd Hawasly, and Subramanian Ramamoorthy. 2016. Bayesian policy reuse. *Machine Learning* 104, 1 (2016), 99–127.

[14] Michael Wunder, John Robert Yaros, Michael Kaisers, and Michael Littman. 2012. A framework for modeling population strategies by depth of reasoning. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. International Foundation for Autonomous Agents and Multiagent Systems, 947–954.

[15] YAN ZHENG, Zhaopeng Meng, Jianye Hao, Zongzhang Zhang, Tianpei Yang, and Changjie Fan. 2018. A Deep Bayesian Policy Reuse Approach Against Non-Stationary Agents. In *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 954–964.