EasyChair Preprint

№ 6939

Estimating the Level of Inference Using an Order-Mimic Agent

Haram Joo, Inhyeok Jeong and Sang Wan Lee

October 26, 2021

Estimating the level of inference using an order-mimic agent

Haram Joo[1][0000−0002−0164−6941], Inhyeok Jeong[2][0000−0002−8710−4683], and Sang Wan Lee[1][0000−0001−6266−9613]

[1] Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea
{haramjoo,sangwan}@kaist.ac.kr
http://aibrain.kaist.ac.kr
[2] School of Freshman, Korea Advanced Institute of Science and Technology, Daejeon, South Korea
jih7368@kaist.ac.kr

Abstract. Multi-agent reinforcement learning (RL) considers problems of learning policies and predicting values through interactions with multiple opponents. To make the solutions feasible, one assumes single-type opponents. However, this may not hold in most real-world situations. Interactions with a mixture of different types of agents make it extremely hard to learn. This study examines the hypothesis that when the potential types of agents are unknown, the level of agent inference can act as a proxy for characterizing the opponents. We present a computational framework to estimate the level of agent's inference using a deterministic and stochastic order-mimic agent. We then propose a calibration method for unbiased estimation, which offsets the adverse effect of order-mimic agents on the environment's order estimation. Finally, to generalize the method to a wide range of contexts, we proposed iterative inference level estimation. We demonstrate the feasibility of the proposed method in computer simulations with agents mimicking agents' behavior with various inference levels. Our framework can estimate the learning capacity of various algorithms and humans; therefore it can be used to design high-level inference models that can effectively handle the complexity of multi-agent learning problems.

Keywords: Multi-agent reinforcement learning · Keynesian beauty contest · Level of inference.

1 Introduction

The problem of reinforcement learning (RL) is defined based on the Markov decision process (MDP). The basic idea is to capture the most important features that predict future rewards [1]. In traditional RL, the agent interacts with the environment, or only one opponent exists and is thought of as part of the environment [2–4]. Recently, multi-agent reinforcement learning (MARL) concerns the learning problem in which one agent interacts with other opponents [5–9]. It

is difficult to deal with a mixture of different types of agents, so it is generally assumed that the opponents are single-type. However, this may not generalize to real-world situations. It will be effective if we can estimate the level of inference and use it to understand the behavior of others.

In this study, we propose a method to estimate an agent's level of inference. In doing so, we defined an order-mimic agent, and combined it with the Keynesian beauty contest environment, a generic task for evaluating the ability to infer public perception [10–14], as a multi-agent simulation scenario. Simulations demonstrated the validity of the proposed method. To the best of our knowledge, this is the first study to open up the possibility of estimating the level of inference of humans and algorithms.

2 Keynesian Beauty Contest

Keynesian beauty contest was designed by Keynes [15] to explain stock market price fluctuations. To be precise, we quote the description of the Keynesian beauty contest from his book:

> It is not a case of choosing those [faces] that, to the best of one's judgment, are really the prettiest, nor even those that average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practice the fourth, fifth and higher degrees [15].

This implies that participants want to choose a face they think other people will choose a lot, and this is affected by their level of inference. The Keynesian contest can be also used to simulate a stock market choice between what is considered an asset's fundamental value and the value it expects to appreciate by other investors.

Nagel [16] has formulated the Keynesian beauty contest mathematically to use it in a multi-agent simulation scenario. Each participant chooses a number from 0 to 100. The winner of p-beauty contest is the one who picks a number close to p times the average number chosen by all participants. In this study, the same reward was given to all winners, regardless of the number of winners. To implement a task with a high degree of freedom, the choice was defined in *(python)float64* range.

3 Method

3.1 Order-mimic agent

If all agents in Keynesian beauty contest randomly pick a number, the average of all submitted numbers is 50. If the level-1 participant know p=2/3, it will choose 50*2/3=33.33. Similarly, the level-2 participant will choose 50*4/9=22.22

by assuming opponents are level-1. The level-k assume opponents are level-(k-1), and the Keynesian beauty contest used in our study repeats multiple rounds, so it can be expanded as follows:

$$\begin{aligned} a_t^{(1)} &= \overline{x_{t-1}} \times p \\ a_t^{(2)} &= \overline{x_{t-1}} \times p^2 \\ &\cdots \\ a_t^{(k)} &= \overline{x_{t-1}} \times p^k \end{aligned} \tag{1}$$

where $a_t^{(i)}$ refers to the level-i participant's action in the current round, and $\overline{x_{t-1}}$ represents the average of all submitted numbers in the previous round.

Equation (1) assumes that the participant knows p. Therefore, we generalize the formula with the information available in the actual contest only:

$$a_t^{(k)} = \overline{x_{t-1}} \times (\frac{a_{t-1}^*}{\overline{x_{t-1}}})^k, \tag{2}$$

where a_{t-1}^* refers to the number that was associated with the reward in the previous round. In this study, the agent following Eq. (2) is named as order-k mimic agent, M_k. For example, the order-3 mimic agent, M_3, chooses $\overline{x_{t-1}} \times (\frac{a_{t-1}^*}{\overline{x_{t-1}}})^3$ as the action every round. All order-mimic agents' first round actions are chosen randomly.

3.2 Inference level estimation method

The order-k agent assumes that opponents are order-(k-1) agents and chooses the best action. Therefore, if the order-k agent and the order-(k-1) agent confront in a non-probabilistic environment, theoretically the order-k agent will always win. This means that if the order-k agent performs the task where order-(k-1) agents are dominant, the order-k agent can obtain a higher reward.

The proposed algorithm makes the best use of this characteristic. If a population of target agents of which we want to estimate the order is dominant, we can compute the order by using the order-mimic agent. The order of the target agent can be estimated by averaging the number of rewards that order-mimic agents earned:

$$ORD(T) = \frac{\sum_{A_i \neq T} ORD(A_i) R_i}{\sum_{A_i \neq T} R_i} - 1 \tag{3}$$

where T refers to the target agent, and A_i, R_i represents the i-th agent and its cumulative reward. Rewards are accumulated each round until convergence. $ORD(x)$ stands for the order of agent x. For example, the order of the order-mimic agent is $ORD(M_k) = k$.

All agents except target agents, $\sum_{A_i \neq T}$, are order-mimic agents. By calculating the average based on the reward ratio of the order-mimic agent, we can

approximate the order that performs best in a given environment. Furthermore, subtracting one from this value is the order of target agents.

While Eq. (3) is potentially a good order estimator, it does not take into account the effect of order-mimic agents on the environment's order, which is used to estimate the order of target agent. To tackle this issue, we propose a method for order calibration:

$$y = \frac{n_t y^{'} + \sum_{A_i \neq T} ORD(A_i)}{n_t + n_m}$$

$$y' = \frac{(n_t + n_m)y - \sum_{A_i \neq T} ORD(A_i)}{n_t} \tag{4}$$

where y refers to the previously estimated order by Eq. (3), $ORD(T)$, and y' represents the calibrated order.

In the default setting, order-mimic agents are used one by one from order-1 to order-n_m, so Eq. (4) can be organized as follows:

$$y' = \frac{(n_t + n_m)y - \frac{n_m(n_m+1)}{2}}{n_t}$$

$$y' = \frac{2(n_t + n_m)y - n_m(n_m+1)}{2n_t} \tag{5}$$

Additionally, in $n_t \gg n_m$ environment where the target agent is highly dominant, we can confirm that $y' \approx y$.

$$y' = (1 + \frac{n_m}{n_t})y - \frac{{n_m}^2}{2n_t} - \frac{n_m}{2n_t} \approx y$$

Using Eq. (3) and Eq. (5), we propose the inference level estimation algorithm:

Algorithm 3.1: InferenceLevelEstimationAlgorithm(T, n_t, n_m, p)

INPUT T : Target agent
n_t : Number of target agents
n_m : Number of order-mimic agents
p : Keynesian beauty contest hyperparameter
OUTPUT y' : Esimated (calibrated) order of target agent
$A, a \leftarrow InitializeAgent(T, n_t, n_m)$
$R, r \leftarrow InitializeReward(n_t, n_m)$
$C \leftarrow CreateContest(A, p)$
while *TRUE*
do
$a \leftarrow A(a, r)$
$r \leftarrow C(a)$
if $\sum r = n_t + n_m$
then *break*
$R \leftarrow R + r$
$y \leftarrow \frac{\sum_{A_i \neq T} ORD(A_i) R_i}{\sum_{A_i \neq T} R_i} - 1$
$y' \leftarrow \frac{2(n_t+n_m)y - n_m(n_m+1)}{2n_t}$
return (y')

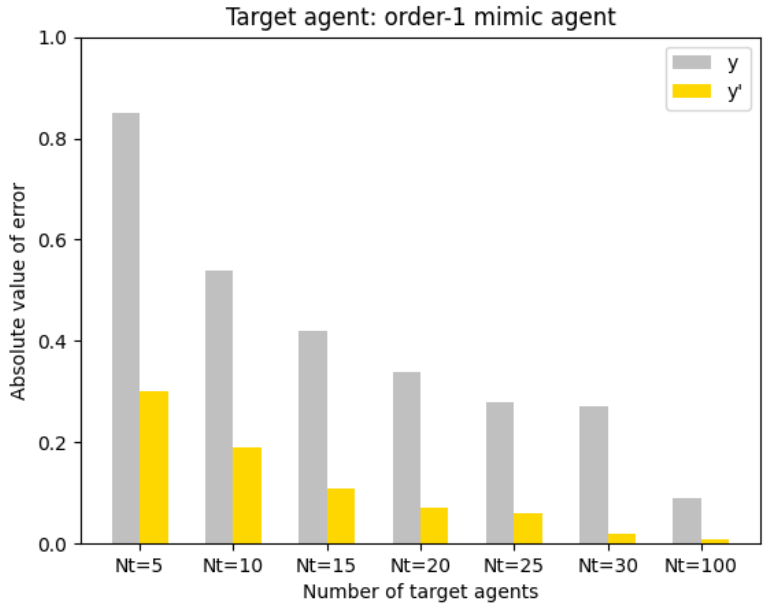
4 Result and Discussion

4.1 Performance and efficiency of the proposed method

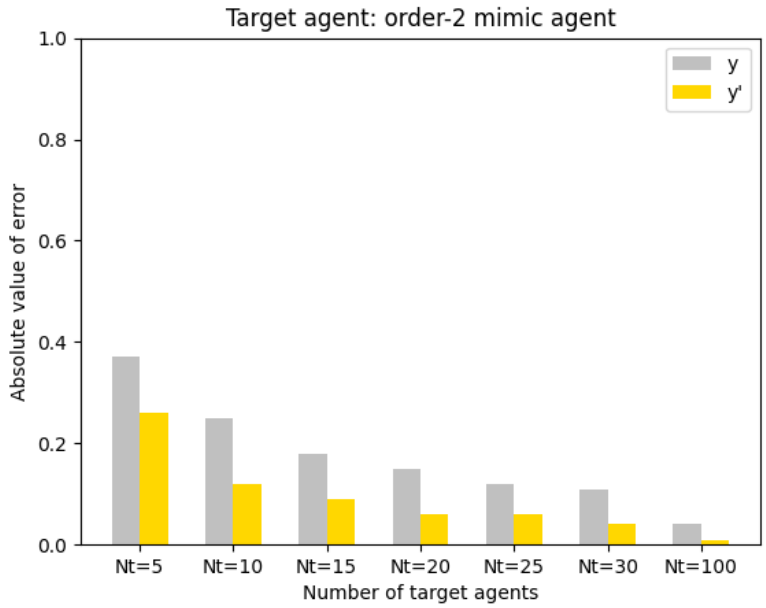
First, we ran in the situation with n_m=5 and p=2/3 while varying the n_t from 5 to 100. Order-mimic agents were used one each from order-1 mimic to order-5 mimic. Fig. 1 shows the results when using an order-1 mimic agent and order-2 mimic agent as target agents. We applied noise to target agents (range of $\times 0.95$ to $\times 1.05$). Note that y is the order without calibration, and y' is the calibrated order. As shown in Fig. 1, the error of y decreases as the number of target agents increases. Because, y does not take into account the changing of the environment's order due to order-mimic agents, as n_t increases, the order of the target agent and the order of the environment become similar.

Next, to simulate a stochastic environment, we implemented the order-1.5 mimic agent. The order-1.5 mimic agent has a 50% chance to act as an order-1 mimic agent and a 50% chance to act as an order-2 mimic agent. The order-2.5 mimic agent can be defined in the same way.

Fig. 2 shows the results when using an order-1.5 mimic agent and order-2.5 mimic agent as target agents in n_m=5 and p=2/3. When the target agent is stochastic, the error is larger than the case with the deterministic target agent. Despite stochastic settings, the calibrated method outperforms the original version.

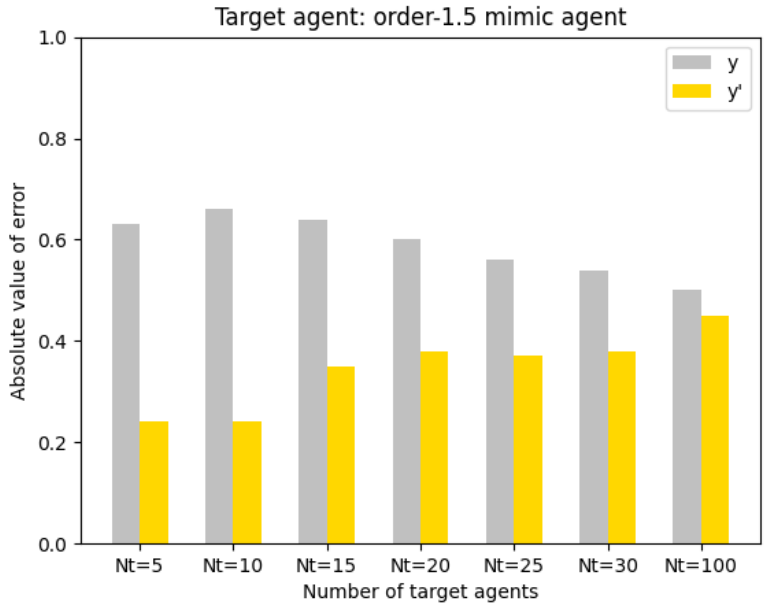


(a) T=order-1 mimic agent, n_m=5, p=2/3

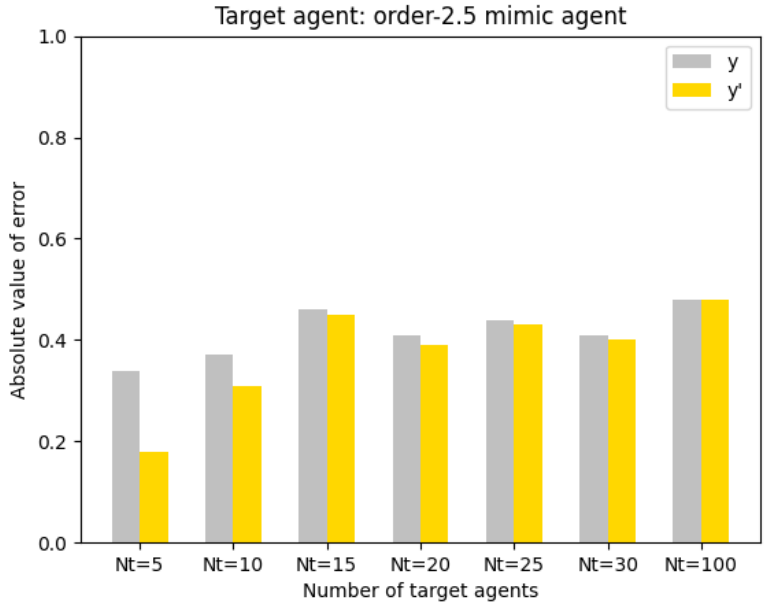


(b) T=order-2 mimic agent, n_m=5, p=2/3

Fig. 1. Absolute value of error of two methods in the p-beauty contest. Silver bar (y) represents the uncalibrated method, and gold bar (y') represents the calibrated method. Order-mimic agents are used one each from order-1 mimic to order-n_m mimic. The noise was applied in the range of $\times 0.95$ to $\times 1.05$.



(a) T=order-1.5 mimic agent, n_m=5, p=2/3



(b) T=order-2.5 mimic agent, n_m=5, p=2/3

Fig. 2. Absolute value of error when target agent is stochastic. Silver bar (y) represents the uncalibrated method, and gold bar (y') represents the calibrated method. Order-mimic agents are used one each from order-1 mimic to order-n_m mimic. The noise was applied in the range of $\times 0.95$ to $\times 1.05$.

Table 1. Sum of error absolute values.

Nt	5	10	15	20	25	30	100
y	2.19	1.82	1.70	1.50	1.40	1.33	1.11
y'	0.98	0.86	1.00	0.90	0.92	0.84	0.95

Table 1 represents the sum of error absolute values for the four cases executed above, from order-1 mimic agent to order-2.5 mimic agent. The error of y decreases as the number of target agents increases, but the error of y' is less affected by n_t. This is because y' cancels out the adverse effect of order-mimic agents on the environment's order. Another important implication is that we do not necessarily have to simulate in huge n_t to get smaller errors. Note that the calibrated method always performs better than the original method.

4.2 Iterative inference level estimation

Theoretically, the proposed method requires that the order of target agent should be less than the order of the highest order-mimic agent minus one. This is because the proposed method is based on the idea that an order-mimic agent, close to the target agent's order plus one, obtains a larger amount of reward where the target agent is dominant.

To investigate the effect of the measurable range on estimation performance, we ran the simulation where n_t=25, p=2/3, and order-5 mimic agent as target agent. As shown in Fig. 3, the estimation error increases when the target agent's order is out of the measurable range (n_m=4 and n_m=5). As long as it remains within the measurable range, the error appears to be small regardless of n_m.

The fact that the true order of an arbitrary agent is not known may impede the ability of estimation. However, this issue can be solved by the following iterative inference level estimation, which uses the results in Fig. 3:

(1) Start with small n_m and gradually increase to large n_m (not necessarily in increments of one), if values are similar in several consecutive intervals, it can be considered as the order of the target agent.
or
(2) The order of the target agent can be obtained by simulating only once with a sufficiently large n_m.

5 Conclusion

In this study, we proposed a method to estimate the level of arbitrary agent's inference. For this, we defined a deterministic and stochastic order-mimic agent. We also propose a calibration method for unbiased estimation, which offsets the adverse effect of order-mimic agents on the environment's order estimation. Simulation results show that the unbiased estimation outperforms the basic method

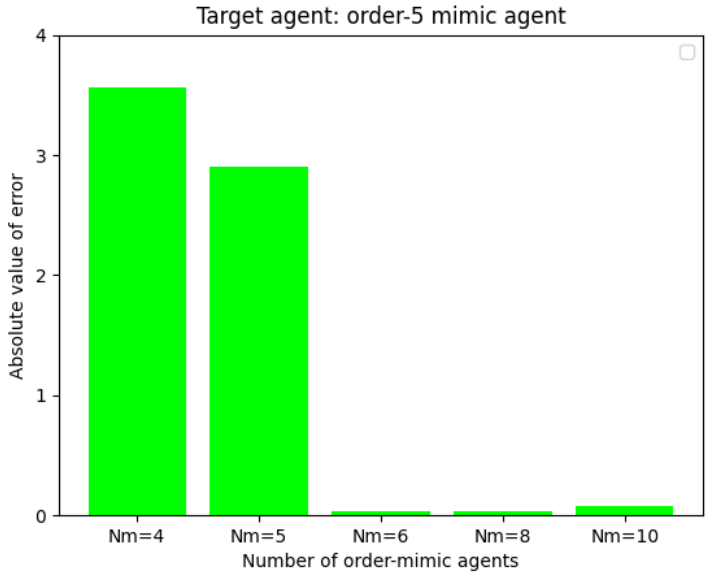


Fig. 3. The range of inference level of the proposed method. The y-axis represents the absolute value of error of calibrated method. The measurable range of the proposed method is 0 to highest order-mimic agent's order minus one. Order-mimic agents are used one each from order-1 mimic to order-n_m mimic. T=order-5 mimic agent, n_t=25, p=2/3, and the noise was applied in the range of $\times 0.95$ to $\times 1.05$.

regardless of the number of target agents. We also analyzed the measurable range of the proposed method, and confirmed that the estimation error is small as long as the target agent's order is within the measurable range. Finally, to generalize the method to a wide range of contexts, we proposed iterative inference level estimation.

In future research, it is possible to design an agent that better deals with opponents by using their level of inference. It would also be interesting to examine the relationship between the level of inference and task performance of various algorithms and humans. These will contribute to the development of high-level inference models.

Acknowledgements This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2019-0-01371, Development of brain-inspired AI with human-like intelligence)

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (NRF-2019M3E5D2A01066267)

This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-TC1603-52

References

1. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, MA: (2018)
2. Tsitsiklis, J.N., van Roy, B.: An analysis of temporal-difference learning with function approximation. IEEE Transactions on Automatic Control. 42, 674–690 (1997). https://doi.org/10.1109/9.580874
3. Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y., others: Policy gradient methods for reinforcement learning with function approximation. In: NIPs. pp. 1057–1063 (1999)
4. Kakade, S.: A natural policy gradient. Advances in Neural Information Processing Systems. (2002)
5. Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., Aru, J., Vicente, R.: Multiagent cooperation and competition with deep reinforcement learning. PLoS ONE. 12, 1–12 (2017). https://doi.org/10.1371/journal.pone.0172395
6. Garant, D., da Silva, B.C., Lesser, V., Zhang, C.: Accelerating multi-agent reinforcement learning with dynamic co-learning. Technical report, Tech. Rep. (2015)
7. Leibo, J.Z., Zambaldi, V., Lanctot, M., Marecki, J., Graepel, T.: Multi-agent reinforcement learning in sequential social dilemmas. Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS. 1, 464–473 (2017)
8. Harper, M., Knight, V., Jones, M., Koutsovoulos, G., Glynatsi, N.E., Campbell, O.: Reinforcement learning produces dominant strategies for the Iterated Prisoner's Dilemma. (2017). https://doi.org/10.1371/JOURNAL.PONE.0188046
9. Cao, K., Lazaridou, A., Lanctot, M., Leibo, J.Z., Tuyls, K., Clark, S.: Emergent communication through negotiation. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings. 1–15 (2018)
10. Crawford, V.P.: Boundedly rational versus optimization-based models of strategic thinking and learning in games. Voprosy Ekonomiki. 2014, 27–44 (2014). https://doi.org/10.32609/0042-8736-2014-5-27-44
11. García-Schmidt, M., Woodford, M.: Are low interest rates deflationary? A paradox of perfect-foresight analysis†. American Economic Review. 109, 86–120 (2019). https://doi.org/10.1257/aer.20170110
12. Cornand, C., dos Santos Ferreira, R.: Cooperation in a differentiated duopoly when information is dispersed: A beauty contest game with endogenous concern for coordination. Mathematical Social Sciences. 106, 101–111 (2020). https://doi.org/10.1016/J.MATHSOCSCI.2020.02.003
13. Coricelli, G., Nagel, R.: Neural correlates of depth of strategic reasoning in medial prefrontal cortex. Proceedings of the National Academy of Sciences of the United States of America. 106, 9163–9168 (2009). https://doi.org/10.1073/pnas.0807721106
14. Pantelis, P.C., Kennedy, D.P.: Autism does not limit strategic thinking in the "beauty contest" game. Cognition. 160, 91–97 (2017). https://doi.org/10.1016/j.cognition.2016.12.015
15. Keynes, J.M.: The General Theory of Employment, Interest, and Money. Palgrave Macmillan, London (1936)
16. Nagel, R.: Unraveling in Guessing Games: An Experimental Study. The American Economic Review. 85, 1313–1326 (1995). https://doi.org/10.2307/2950991