

# Modelling Bounded Rationality in Multi-Agent Interactions by Generalized Recursive Reasoning

Ying Wen<sup>1,2\*</sup>, Yaodong Yang<sup>1,2\*</sup>, Jun Wang<sup>1,2</sup>

<sup>1</sup>University College London

<sup>2</sup>Huawei Research & Development U.K.

{ying.wen, yaodong.yang, jun.wang}@cs.ucl.ac.uk

## Abstract

Though limited in real-world decision making, most multi-agent reinforcement learning (MARL) models assume perfectly rational agents – a property hardly met due to individual’s cognitive limitation and/or the tractability of the decision problem. In this paper, we introduce generalized recursive reasoning (GR2) as a novel framework to model agents with different *hierarchical* levels of rationality; our framework enables agents to exhibit varying levels of “thinking” ability thereby allowing higher-level agents to best respond to various less sophisticated learners. We contribute both theoretically and empirically. On the theory side, we devise the hierarchical framework of GR2 through probabilistic graphical models and prove the existence of a perfect Bayesian equilibrium. Within the GR2, we propose a practical actor-critic solver, and demonstrate its convergent property to a stationary point in two-player games through Lyapunov analysis. On the empirical side, we validate our findings on a variety of MARL benchmarks. Precisely, we first illustrate the hierarchical thinking process on the Keynes Beauty Contest, and then demonstrate significant improvements compared to state-of-the-art opponent modeling baselines on the normal-form games and the cooperative navigation benchmark.

## 1 Introduction

In people’s decision making, rationality can often be compromised; it can be constrained by either the difficulty of the decision problem or the finite resources available to each individual’s mind. In behavioral game theory, instead of assuming people are perfectly rational, *bounded rationality* [Simon, 1972] serves as the alternative modeling basis by recognizing such cognitive limitations. One most-cited example that bounded rationality prescribes is Keynes Beauty Contest [Keynes, 1936]. In the contest, all players are asked to pick one number from 0 to 100, and the player whose guess is closest to  $1/2$  of the average number eventually becomes the winner. In this game, if all the players are perfectly rational, the only

choice is to guess 0 (the only Nash equilibrium) because each of them could reason as follows: “*if all players guess randomly, the average of those guesses would be 50 (level-0), I, therefore, should guess no more than  $1/2 \times 50 = 25$  (level-1), and then if the other players think similarly as me, I should guess no more than  $1/2 \times 25 = 13$  (level-2) ...*”. Such level of recursions can keep developing infinitely until all players guess the equilibrium 0. This theoretical result from the perfect rationality is however inconsistent with the experimental finding in psychology [Coricelli and Nagel, 2009] which suggests that most human players would choose between 13 and 25. In fact, it has been shown that human beings tend to reason only by 1-2 levels of recursions in strategic games [Camerer *et al.*, 2004]. In the Beauty Contest, players’ rationality is bounded and their behaviors are sub-optimal. As a result, it would be unwise to guess the Nash equilibrium 0 at all times.

In the multi-agent reinforcement learning (MARL), one common assumption is that all agents behave rationally [Albrecht and Stone, 2018] during their interactions. For example, we assume agents’ behaviors will converge to Nash equilibrium [Yang *et al.*, 2018]. However, in practice, it is hard to guarantee that all agents have the same level of sophistication in their abilities of understanding and learning from each other. With the development of MARL methods, agents could face various types of opponents ranging from naive independent learners [Bowling and Veloso, 2002], joint-action learners [Claus and Boutilier, 1998], to the complicated theory-of-mind learners [Rabinowitz *et al.*, 2018; Shum *et al.*, 2019]. It comes with no surprise that the effectiveness of MARL models decreases when the opponents act irrationally [Shoham *et al.*, 2003]. On the other hand, it is not desirable to design agents that can only tackle opponents that play optimal policies. Justifications can be easily found in modern AI applications including self-driving cars [Shalev-Shwartz *et al.*, 2016] or video game designs [Peng *et al.*, 2017; Hunicke, 2005]. Therefore, it becomes critical for MARL models to acknowledge different levels of bounded rationality.

In this work, we propose a novel framework – *Generalized Recursive Reasoning (GR2)* – that recognizes agents’ bounded rationality and thus can model their corresponding sub-optimal behaviors. GR2 is inspired by cognitive hierarchy theory [Camerer *et al.*, 2004], assuming that agents could possess different levels of reasoning rationality during the interactions. It begins with *level-0* (*L0* for short) non-strategic thinkers

---

\*First two authors contribute equally.

who do not model their opponents. L1 thinkers are more sophisticated than *level-0*; they believe the opponents are all at L0 and then act correspondingly. With the growth of  $k$ , L $k$  agents think in an increasing order of sophistication and then take the best response to all possible lower-level opponents. We immerse the GR2 framework into MARL through graphical models, and derive the practical GR2 soft actor-critic algorithm. Theoretically, we prove the existence of Perfect Bayesian Equilibrium in the GR2 framework, as well as the convergence of GR2 policy gradient methods on two-player normal-form games. Our proposed GR2 actor-critic methods are evaluated against multiple strong MARL baselines on Keynes Beauty Contest, normal-form games, and cooperative navigation. Results justify our theoretical findings and the effectiveness of bounded-rationality modeling.

## 2 Related Work

Modeling the opponents in a recursive manner can be regarded as a special type of opponent modeling [Albrecht and Stone, 2018]. Recently, studies on Theory of Mind (ToM) [Goldman and others, 2012; Rabinowitz *et al.*, 2018; Shum *et al.*, 2019] explicitly model the agent’s belief on opponents’ mental states in the reinforcement learning (RL) setting. The I-POMDP framework focuses on building the beliefs about opponents’ intentions into the planning and making agents acting optimally with respect to such predicted intentions [Gmytrasiewicz and Doshi, 2005]. GR2 is different in that it incorporates a hierarchical structure for opponent modeling; it can take into account opponents with different levels of rationality and therefore can conduct nested reasonings about the opponents (e.g. “I believe you believe that I believe ... ”). In fact, our method is most related to the probabilistic recursive reasoning (PR2) model [Wen *et al.*, 2019]. PR2 however only explores the *level-1* structure and it does not target at modeling the bounded rationality. Most importantly, PR2 does not consider whether an equilibrium exists in such sophisticated hierarchical framework at all. In this work, we extend the reasoning level to an arbitrary number, and theoretically prove the existence of equilibrium under the GR2 setting as well as the convergence of the subsequent learning algorithms.

Decision-making theorists have pointed out that the ability of thinking in a hierarchical manner is one direct consequence of the limitation in decision maker’s information-processing power; they demonstrate this result by matching real-world behavioral data with the model that trades off between utility maximization against information-processing costs (i.e. an entropy term applied on the policy) [Genewein *et al.*, 2015]. Interestingly, maximum-entropy framework has also been explored in the RL domain through inference on graphical models [Levine, 2018]; *soft* Q-learning [Haarnoja *et al.*, 2017] and *soft* actor-critic [Haarnoja *et al.*, 2018] methods were developed. Recently, *soft* learning has been further adapted into the context of MARL [Wei *et al.*, 2018; Tian *et al.*, 2019]. In this work, we bridge the gap by embedding the solution concept of GR2 into MARL, and derive the practical GR2 soft actor-critic algorithm. By recognizing bounded rationality, we expect the GR2 MARL methods to generalize across different types of opponents thereby showing robustness to their sub-optimal behaviors, which we believe is

a critical property for modern AI applications.

## 3 Preliminaries

*Stochastic Game* [Shapley, 1953] is a natural framework to describe the  $n$ -agent decision-making process; it is typically defined by the tuple  $\langle \mathcal{S}, \mathcal{A}^1, \dots, \mathcal{A}^n, r, \dots, r^n, \mathcal{P}, \gamma \rangle$ , where  $\mathcal{S}$  represents the state space,  $\mathcal{A}^i$  and  $r^i(s, a^i, a^{-i})$  denote the action space and reward function of agent  $i \in \{1, \dots, n\}$ ,  $\mathcal{P} : \mathcal{S} \times \mathcal{A}^1 \times \dots \times \mathcal{A}^n \rightarrow \mathcal{P}(\mathcal{S})$  is the transition probability of the environment, and  $\gamma \in (0, 1]$  a discount factor of the reward over time. We assume agent  $i$  chooses an action  $a^i \in \mathcal{A}^i$  by sampling its policy  $\pi_{\theta_i}^i(a^i|s)$  with  $\theta^i$  being a tuneable parameter, and use  $a^{-i} = (a^j)_{j \neq i}$  to represent actions executed by opponents. The trajectory  $\tau^i = [(s_1, a_1^i, a_1^{-i}), \dots, (s_T, a_T^i, a_T^{-i})]$  of agent  $i$  is defined as a collection of state-action triples over a horizon  $T$ .

### The Concept of Optimality in MARL

Analogous to standard reinforcement learning (RL), each agent in MARL attempts to determine an optimal policy maximizing its total expected reward. On top of RL, MARL introduces additional complexities to the learning objective because the reward now also depends on the actions executed by opponents. Correspondingly, the value function of the  $i_{th}$  agent in a state  $s$  is  $V^i(s; \boldsymbol{\pi}_{\theta}) = \mathbb{E}_{\boldsymbol{\pi}_{\theta}, \mathcal{P}} \left[ \sum_{t=1}^T \gamma^{t-1} r^i(s_t, a_t^i, a_t^{-i}) \right]$  where  $(a_t^i, a_t^{-i}) \sim \boldsymbol{\pi}_{\theta} = (\pi_{\theta_i}^i, \pi_{\theta^{-i}}^i)$  with  $\boldsymbol{\pi}_{\theta}$  denoting the joint policy of all learners. As such, *optimal* behavior in a multi-agent setting stands for acting in *best response* to the opponent’s policy  $\pi_{\theta^{-i}}$ , which can be formally defined as the policy  $\pi_*^i$  with  $V^i(s; \pi_*^i, \pi_{\theta^{-i}}) \geq V^i(s; \pi_{\theta_i}^i, \pi_{\theta^{-i}})$  for all valid  $\pi_{\theta_i}^i$ . If all agents act in best response to others, the game arrives at a Nash equilibrium [Nash and others, 1950]. Specifically, if agents execute the policy of the form  $\pi^i(a^i|s) = \frac{\exp(Q_{\boldsymbol{\pi}_{\theta}}^i(s, a^i, a^{-i}))}{\sum_{a'} \exp(Q_{\boldsymbol{\pi}_{\theta}}^i(s, a', a^{-i}))}$  – a standard type of policy adopted in RL literatures – with  $Q_{\boldsymbol{\pi}_{\theta}}^i(s, a^i, a^{-i}) = r^i(s, a^i, a^{-i}) + \gamma \mathbb{E}_{\mathcal{P}}[V^i(s'; \boldsymbol{\pi}_{\theta})]$  denoting agent  $i$ ’s Q-function and  $s'$  being a successor state, they reach a Nash-Quantal equilibrium [McKelvey and Palfrey, 1995].

### The Graphical Model of MARL

Since GR2 is a probabilistic model, it is instructive to provide a brief review of graphical model for MARL. In single-agent RL, finding the optimal policy can be equivalently transferred into an inference problem on a graphical model [Levine, 2018]. Recently, it has been shown that such equivalence also holds in the multi-agent setting [Tian *et al.*, 2019; Wen *et al.*, 2019]. To illustrate, we first introduce a binary random variable  $\mathcal{O}_t^i \in \{0, 1\}$  (see Fig. 1) that stands for the optimality of agent  $i$ ’s policy at time  $t$ , i.e.,  $p(\mathcal{O}_t^i = 1 | \mathcal{O}_t^{-i} = 1, \tau_t^i) \propto \exp(r^i(s_t, a_t^i, a_t^{-i}))$ , which suggests that given a trajectory  $\tau_t^i$ , the probability of being optimal is proportional to the reward. In the fully-cooperative setting, if all agents play optimally, then agents receive the maximum reward that is also the Nash equilibrium; therefore, for agent  $i$ , it aims to maximize  $p(\mathcal{O}_{1:T}^i = 1 | \mathcal{O}_{1:T}^{-i} = 1)$  as this is the probability of obtaining the maximum cumulative reward/best response towards Nash equilibrium. For simplicity, we omit the value for  $\mathcal{O}_t$  hereafter.

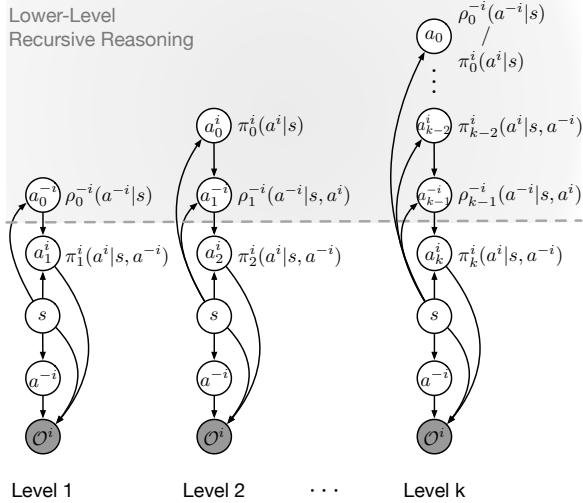


Figure 1: Graphical model of the *level-k* reasoning model. Subfix of  $a_*$  stands for the level of thinking not the timestep. The opponent policies are approximated by  $\rho^{-i}$ . The omitted *level-0* model considers opponents fully randomized. Agent  $i$  rolls out the recursive reasoning about opponents in its mind (grey area). In the recursion, agents with higher-level beliefs take the best response to the lower-level agents.

As we assume no knowledge of the optimal policies  $\pi_*$  and the model of the environment  $\mathcal{P}(\mathcal{S})$ , we treat them as latent variables and applied variational inference [2] to approximate such objective; using the variational form of  $\hat{p}(\tau^i|\mathcal{O}_{1:T}^i, \mathcal{O}_{1:T}^{-i}) = [\hat{p}(s_1) \prod_{t=1}^{T-1} \hat{p}(s_{t+1}|s_t, a_t^i, a_t^{-i})] \pi_\theta(a_t^i, a_t^{-i}|s_t)$  leads to

$$\begin{aligned} \max \mathcal{J}(\pi_\theta) &= \log p(\mathcal{O}_{1:T}^i = 1 | \mathcal{O}_{1:T}^{-i} = 1) \\ &\geq \sum_{\tau^i} \hat{p}(\tau^i | \mathcal{O}_{1:T}^i, \mathcal{O}_{1:T}^{-i}) \log \frac{p(\mathcal{O}_{1:T}^i, \tau^i | \mathcal{O}_{1:T}^{-i})}{\hat{p}(\tau^i | \mathcal{O}_{1:T}^i, \mathcal{O}_{1:T}^{-i})} \\ &= \sum_{t=1}^T \mathbb{E}_{\tau^i \sim \hat{p}(\tau^i)} \left[ r^i(s_t, a_t^i, a_t^{-i}) + \mathcal{H}(\pi_\theta(a_t^i, a_t^{-i}|s_t)) \right]. \end{aligned} \quad (1)$$

To maximize  $\mathcal{J}(\pi_\theta)$ , a variant of policy iteration called *soft learning* is applied. For policy evaluation, Bellman expectation equation now holds on the *soft* value function  $V^i(s) = \mathbb{E}_{\pi_\theta} [Q^i(s_t, a_t^i, a_t^{-i}) - \log(\pi_\theta(a_t^i, a_t^{-i}|s_t))]$ , with the updated Bellman operator  $\mathcal{T}^\pi Q^i(s_t, a_t^i, a_t^{-i}) \triangleq r^i(s_t, a_t^i, a_t^{-i}) + \gamma \mathbb{E}_\mathcal{P} [\text{soft } Q(s_t, a_t^i, a_t^{-i})]$ . Compared to the max operation in the normal Q-learning, soft operator is  $\text{soft } Q(s, a^i, a^{-i}) = \log \sum_a \sum_{a^{-i}} \exp(Q(s, a^i, a^{-i})) \approx \max_{a^i, a^{-i}} Q(s, a^i, a^{-i})$ . Policy improvement however becomes non-trivial because the Q-function now guides the improvement direction for the joint policy rather than for each single agent. Since the exact parameter of opponent policy is usually unobservable, agent  $i$  needs to approximate  $\pi_\theta^{-i}$ .

## 4 Generalized Recursive Reasoning

Recursive reasoning is essentially taking iterative best response to opponents' policies. *level-1* thinking is "I know you know how I know". We can represent such recursion by  $\pi(a^i, a^{-i}|s) = \pi^i(a^i|s)\pi^{-i}(a^{-i}|s, a^i)$  where  $\pi^{-i}(a^{-i}|s, a^i)$  stands for the opponent's consideration of agent  $i$ 's action  $a^i \sim \pi^i(a^i|s)$ . The unobserved opponent conditional policy  $\pi^{-i}$  can be approximated via a best-fit model

$\rho_{\phi^{-i}}^{-i}$  parameterized by  $\phi^{-i}$ . By adopting  $\pi_\theta(a^i, a^{-i}|s) = \pi_{\theta^i}(a^i|s)\rho_{\phi^{-i}}^{-i}(a^{-i}|s, a^i)$  in  $\hat{p}(\tau^i | \mathcal{O}_{1:T}^i, \mathcal{O}_{1:T}^{-i})$  in maximizing the Eq. 1, we can solve the best-fit opponent policy by

$$\rho_{\phi^{-i}}^{-i}(a^{-i}|s, a^i) \propto \exp(Q_{\pi_\theta}^i(s, a^i, a^{-i}) - Q_{\pi_\theta}^i(s, a^i)). \quad (2)$$

We provide the detailed derivation of Eq. 2 in *Appendix A*. Eq. 2 suggests that agent  $i$  believes his opponent will act in his interest in the cooperative games. Based on the opponent model in Eq. 2, agent  $i$  can learn the best response policy by considering all possible opponent agents' actions:  $Q^i(s, a^i) = \int_{a^{-i}} \rho_{\phi^{-i}}^{-i}(a^{-i}|s, a^i) Q^i(s, a^i, a^{-i}) da^{-i}$ , and then improve its own policy towards the direction of

$$\pi' = \arg \min_{\pi'} D_{\text{KL}} \left[ \pi'(\cdot|s_t) \middle\| \frac{\exp(Q_{\pi_\theta}^i(s_t, a^i, a^{-i}))}{\sum_{a'} \exp(Q^i(s_t, a', a^{-i}))} \right]. \quad (3)$$

### Level-k Recursive Reasoning – GR2-L

Our goal is to extend the recursion to the *level-k* ( $k \geq 2$ ) reasoning (see Fig. 1). In brief, each agent operating at level  $k$  assumes that other agents are using  $k-1$  level policies and then acts in best response. We name this approach **GR2-L**. In practice, the *level-k* policy can be constructed by integrating over all possible best responses from lower-level policies

$$\begin{aligned} \pi_k^i(a_k^i|s) &\propto \int_{a_{k-1}^{-i}} \left\{ \pi_k^i(a_k^i|s, a_{k-1}^{-i}) \right. \\ &\quad \left. \cdot \underbrace{\int_{a_{k-2}^i} \left[ \rho_{k-1}^{-i}(a_{k-1}^{-i}|s, a_{k-2}^i) \pi_{k-2}^i(a_{k-2}^i|s) \right] da_{k-2}^i \right\} da_{k-1}^{-i}. \end{aligned} \quad (4)$$

opponents of level k-1 best responds to agent  $i$  of level k-2

When the levels of reasoning develop, we could think of the marginal policies  $\pi_{k-2}^i(a^i|s)$  from lower levels as the *prior* and the conditional policies  $\pi_k^i(a^i|s, a^{-i})$  as the *posterior*. From agent  $i$ 's perspective, it believes that the opponents will take the best response to its own fictitious action  $a_{k-2}^i$  that are two levels below, i.e.,  $\rho_{k-1}^{-i}(a_{k-1}^{-i}|s) = \int \rho_{k-1}^{-i}(a_{k-1}^{-i}|s, a_{k-2}^i) \pi_{k-2}^i(a_{k-2}^i|s) da_{k-2}^i$ , where  $\pi_{k-2}^i$  can be further expanded by recursively using Eq. 4 until meeting  $\pi_0$  that is usually assumed uniformly distributed. Decisions are taken in a sequential manner. As such, *level-k* model transforms the multi-agent planning problem into a hierarchy of nested single-agent planning problems.

### Mixture of Hierarchy Recursive Reasoning – GR2-M

So far, *level-k* agent assumes all the opponents are at level  $k-1$  during the reasoning process. We can further generalize the model to let each agent believe that the opponents can be much less sophisticated and they are distributed over all lower hierarchies ranging from 0 to  $k-1$  rather than only the level  $k-1$ , and then find the corresponding best response to such mixed type of agents. We name this approach **GR2-M**.

Since more computational resources are required with increasing  $k$ , e.g., human beings show limited amount of working memory (1–2 levels on average) in strategic thinkings [Devetag and Warglien, 2003], it is reasonable to restrict the reasoning so that fewer agents are willing to conduct the reasoning beyond  $k$  when  $k$  grows large. We thus assume that

**Assumption 1.** With increasing  $k$ , level- $k$  agents have an accurate guess about the relative proportion of agents who are doing lower-level thinking than them.

The motivation of such assumption is to ensure that when  $k$  is large, there is no benefit for level- $k$  thinkers to reason even harder to higher levels (e.g. level  $k+1$ ), as they will almost have the same belief about the proportion of lower level thinkers, and subsequently make similar decisions. In order to meet Assumption 1, we choose to model the distribution of reasoning levels by the Poisson distribution  $f(k) = \frac{e^{-\lambda} \lambda^k}{k!}$  where  $\lambda$  is the mean. A nice property of Poisson is that  $f(k)/f(k-n)$  is inversely proportional to  $k^n$  for  $1 \leq n < k$ , which satisfies our need that high-level thinkers should have no incentives to think even harder. We can now mix all  $k$  levels' thinkings  $\{\hat{\pi}_k^i\}$  into agent's belief about its opponents at lower levels by

$$\pi_k^{i,\lambda}(a_k^i | s, a_{0:k-1}^{-i}) := \frac{e^{-\lambda}}{Z} \left( \frac{\lambda^0}{0!} \hat{\pi}_0^i(a_0^i | s) + \dots + \frac{\lambda^k}{k!} \hat{\pi}_k^i(a_k^i | s, a_{0:k-1}^{-i}) \right), \quad (5)$$

where the term  $Z = \sum_{n=1}^k e^{-\lambda} \lambda^n / n!$ . In practice,  $\lambda$  can be set as a hyper-parameter, similar to TD- $\lambda$  [Tesauro, 1995].

Note that GR2-L is a special case of GR2-M. As the mixture in GR2-M is Poisson distributed, we have  $\frac{f(k-1)}{f(k-2)} = \frac{\lambda}{k-1}$ ; the model will bias towards the  $k-1$  level when  $\lambda \gg k$ .

### Theoretical Guarantee of GR2 Methods

Recursive reasoning is essentially to let each agent take the best response to its opponents at different hierarchical levels. A natural question to ask is does the equilibrium ever exist in GR2 settings? If so, will the learning methods ever converge?

Here we demonstrate our **theoretical contributions** that 1) the dynamic game induced by GR2 has Perfect Bayesian Equilibrium; 2) the learning dynamics of policy gradient in GR2 is asymptotically stable in the sense of Lyapunov.

**Theorem 1.** GR2 strategies extend a norm-form game into extensive-form game, and there exists a Perfect Bayesian Equilibrium (PBE) in that game.

*Proof (of sketch).* See Appendix C for the full proof. We can extend the level- $k$  reasoning procedures at one state to an extensive-form game with perfect recall. We prove the existence of PBE by showing both the requirements of *sequentially rational* and *consistency* are met. ■

**Theorem 2.** In two-player normal-form games, if these exist a mixed strategy equilibrium, under mild conditions, the convergence of GR2 policy gradient to the equilibrium is asymptotic stable in the sense of Lyapunov.

*Proof (of sketch).* See Appendix D for the full proof. In the two-player normal-form game, we can treat the policy gradient update as a dynamical system. Through Lyapunov analysis, we first show why the convergence of level-0 method, i.e. **independent learning**, is not stable. Then we show that the level- $k$  method's convergence is asymptotically stable as it accounts for opponents' steps before updating the policy. ■

**Proposition 1.** In both the GR2-L & GR2-M model, if the agents play pure strategies, once level- $k$  agent reaches a Nash Equilibrium, all higher-level agents will follow it too. ■

*Proof.* See Appendix E for the full proof. ■

**Corollary 1.** In the GR2 setting, higher-level strategies weakly dominate lower-level strategies. ■

## 5 Practical Implementations

Computing the recursive reasoning is computational expensive. Here we first present the GR2 soft actor-critic algorithm with the pseudo-code in Algo. 1, and then introduce the compromises we make to afford the implementation.

---

### Algorithm 1 GR2 Soft Actor-Critic Algorithm

---

```

1: Init:  $\lambda, k$  and  $\psi$  (learning rates).
2: Init:  $\theta^i, \phi^{-i}, \omega^i$  for each agent  $i$ .  $\bar{\omega}^i \leftarrow \omega^i, \mathcal{D}^i \leftarrow \emptyset$ .
3: for each episode do
4:   for each step  $t$  do
5:     Agents take a step according to  $\pi_{\theta^i, k}^i(s)$  or  $\pi_{\theta^i, k}^{i, \lambda}(s)$ .
6:     Add experience  $(s, a^i, a^{-i}, r^i, s')$  to  $\mathcal{D}^i$ .
7:   for each agent  $i$  do
8:     Sample a batch  $\{(s'_j, a_j^i, a_j^{-i}, r_j^i, s'_j)\}_{j=0}^M \sim \mathcal{D}^i$ .
9:     Roll out policy to level  $k$  via GR2-L/M to get  $a_j^{i'}$  and record inter-level results  $(a_{j,k}^{i'}, a_{j,k-1}^{-i'}, \dots)$ .
10:    Sample  $a_j^{-i'} \sim \rho_{\phi^{-i}}^{-i}(\cdot | s'_j, a_j^{i'})$ .
11:     $\omega^i \leftarrow \omega^i - \psi_{Q^i} \hat{\nabla}_{\omega^i} J_{Q^i}(\omega^i)$ .
12:     $\theta^i \leftarrow \theta^i - \psi_{\pi^i} \hat{\nabla}_{\theta^i} \left( J_{\pi_k^i}(\theta^i) + J_{\pi_k^i}(\theta^i) \right)$ .
13:     $\phi^{-i} \leftarrow \phi^{-i} - \psi_{\rho^{-i}} \hat{\nabla}_{\phi^{-i}} J_{\rho^{-i}}(\phi^{-i})$ .
14:  end for
15:   $\bar{\omega}^i \leftarrow \psi_{\bar{\omega}} \omega^i + (1 - \psi_{\bar{\omega}}) \bar{\omega}^i$ .
16: end for
17: end for

```

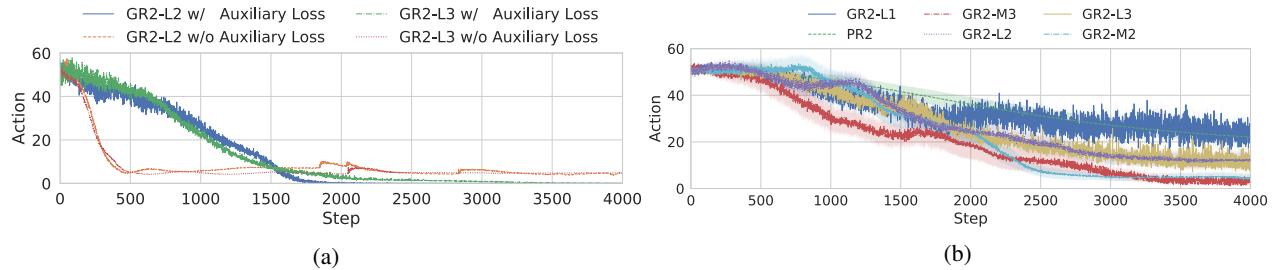
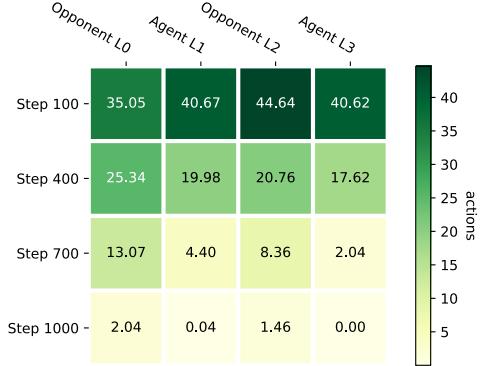
---

**GR2 Soft Actor-Critic.** For policy evaluation, each agent rolls out the reasoning policies recursively to level  $k$  by either Eq. 4 or Eq. 5, the parameter  $\omega^i$  of the joint soft  $Q$ -function is then updated via minimizing the soft Bellman residual  $J_{Q^i}(\omega^i) = \mathbb{E}_{\mathcal{D}^i} [\frac{1}{2} (Q_{\omega^i}^i(s, a^i, a^{-i}) - \hat{Q}^i(s, a^i, a^{-i}))^2]$  where  $\mathcal{D}^i$  is the replay buffer storing trajectories, and the target  $\hat{Q}^i$  goes by  $\hat{Q}^i(s, a^i, a^{-i}) = r^i(s, a^i, a^{-i}) + \gamma \mathbb{E}_{s' \sim \mathcal{P}} [V^i(s')]$ . In computing  $V^i(s')$ , since agent  $i$  has no access to the exact opponent policy  $\pi_{\theta^{-i}}$ , we instead compute the soft  $Q^i(s, a^i)$  by marginalizing the joint  $Q$ -function via the estimated opponent model  $\rho_{\phi^{-i}}^{-i}$  by  $Q^i(s, a^i) = \log \int \rho_{\phi^{-i}}^{-i}(a^{-i} | s, a^i) \exp(Q^i(s, a^i, a^{-i})) da^{-i}$ ; the value function of the level- $k$  policy  $\pi_k^i(a^i | s)$  then comes as  $V^i(s) = \mathbb{E}_{a^i \sim \pi_k^i} [Q^i(s, a^i) - \log \pi_k^i(a^i | s)]$ . Note that  $\rho_{\phi^{-i}}^{-i}$  at the same time is also conducting recursive reasoning against agent  $i$  in the format of Eq. 4 or Eq. 5. From agent  $i$ 's perspective however, the optimal opponent model  $\rho^{-i}$  still follows Eq. 2 in the multi-agent soft learning setting. We can therefore update  $\phi^{-i}$  by minimizing the KL,  $J_{\rho^{-i}}(\phi^{-i}) = \mathcal{D}_{KL}[\rho_{\phi^{-i}}^{-i}(a^{-i} | s, a^i) \| \exp(Q_{\omega^i}^i(s, a^i, a^{-i}) - Q_{\omega^i}^i(s, a^i))]$ . We maintain two approximated  $Q$ -functions of  $Q_{\omega^i}^i(s, a^i, a^{-i})$  and  $Q_{\omega^i}^i(s, a^i)$  separately for robust training, and the gradient of  $\phi^{-i}$  is computed via SVGD [Liu and Wang, 2016].

Finally, the policy parameter  $\theta^i$  for agent  $i$  can be learned by improving towards what the current  $Q$ -function  $Q_{\omega^i}^i(s, a^i)$  suggests, as shown in Eq. 3. By applying the reparameterization trick  $a^i = f_{\theta^i}(\epsilon; s)$  with  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , we have

Table 1: The Converging Equilibrium on Keynes Beauty Contest.

RECURSIVE DEPTH		LEVEL 3		LEVEL 2		LEVEL 1		LEVEL 0			
EXP. SETTING	NASH	GR2-L3	GR2-L2	GR2-L1	PR2	DDPG-ToM	MADDPG	DDPG-OM	MASQL	DDPG	
$p = 0.7, n = 2$	0.0	<b>0.0</b>	<b>0.0</b>	<b>0.0</b>	4.4	7.1	10.6	8.7	8.3	18.6	
$p = 0.7, n = 10$	0.0	<b>0.0</b>	0.1	0.3	9.8	13.2	18.1	12.0	8.7	30.2	
$p = 1.1, n = 10$	100.0	99.0	94.2	92.2	64.0	63.1	68.2	61.7	87.5	52.2	


 Figure 2: Beauty Contest of  $p = 0.7, n = 2$ . (a) Learning curves w/ or w/o the auxiliary loss of Eq. 6. (b) Average learning curves of each GR2 method against the other six baselines (round-robin style).

 Figure 3: The guessing number of both agents during the training of the GR2-L3 model in the Beauty Contest setting ( $n = 2, p = 0.7$ ).

$J_{\pi_k^i}(\theta^i) = \mathbb{E}_{s, a_k^i, \epsilon} [\log \pi_{\theta^i, k}^i(f_{\theta^i}(\epsilon; s)|s) - Q_{\omega^i}^i(s, f_{\theta^i}(\epsilon; s))]$ . Note that as the agent' final decision comes from the best response to all lower levels, we would expect the gradient of  $\partial J_{\pi_k^i} / \partial \theta^i$  propagated from all higher levels during training.

**Approximated Best Response via Deterministic Policy.** As the reasoning process of GR2 methods involves iterated usages of  $\pi_k^i(a^i|s, a^{-i})$  and  $\rho_k^{-i}(a^{-i}|s, a^i)$ , should they be stochastic, the cost of integrating possible actions from lower-level agents would be unsustainable for large  $k$ . Besides, the reasoning process is also affected by the environment where stochastic policies could further amplify the variance. Considering such computational challenges, we approximate by deterministic policies throughout the recursive rollouts, e.g., the mean of Gaussian policy. However, note that the highest-level agent policy  $\pi_k^i$  that interacts with the environment is still stochastic. To mitigate the potential weakness of deterministic policies, we enforce the inter-level policy improvement. The intuition comes from the Corollary. 1 that higher-level policy should perform better than lower-level policies against the opponents. To maintain this property, we introduce an auxiliary loss  $J_{\pi_k^i}(\theta^i)$  in training  $\pi_{\theta^i}^i$  (see Fig. 5 in Appendix B), with  $s \sim \mathcal{D}^i, a_{\tilde{k}}^i \sim \pi_{\theta^i}^i, a_{\tilde{k}}^{-i} \sim \rho_{\phi^{-i}}^i$  and  $\tilde{k}, \geq 2$ , we have

$$J_{\pi_k^i}(\theta^i) = \mathbb{E}_{s, a_{\tilde{k}}^i, a_{\tilde{k}}^{-i}} [Q^i(s, a_{\tilde{k}}^i, a_{\tilde{k}-1}^{-i}) - Q^i(s, a_{\tilde{k}-2}^i, a_{\tilde{k}-1}^{-i})]. \quad (6)$$

As we later show in Fig. 2a, such auxiliary loss plays a critical role in improving the performance.

**Parameter Sharing across Levels.** We further assume parameter sharing for each agent during the recursive rollouts, i.e.,  $\theta^k = \theta^{k+2}$  for all  $\pi_{\theta^k}^i$  and  $\rho_{\theta^k}^{-i}$ . However, note that the policies that agents take at different levels are still **different** because the inputs in computing high-level policies depend on integrating different outputs from low-level policies as shown in Eq. 4. In addition, we have the constraint in Eq. 6 that enforces the inter-policy improvement. Finally, in the GR2-M setting, we also introduce different mixing weights for each lower-level policy in the hierarchy (see Eq. 5).

## 6 Experiments

We start the experiments by elaborating how the GR2 model works on Keynes Beauty Contest, and then move onto the normal-form games that have non-trivial equilibria where common MARL methods fail to converge. Finally, we test on the navigation task that requires effective opponent modeling.

We compare the GR2 methods with six types of baselines including Independent Learner via DDPG [Lillicrap et al., 2015], PR2 [Wen et al., 2019], multi-agent soft-Q (MASQL) [Wei et al., 2018], and MADDPG [Lowe et al., 2017]. We also include the opponent modeling [He et al., 2016] by augmenting DDPG with an opponent module (DDPG-OM) that predicts the opponent behaviors in future states, and a theory-of-mind model [Rabinowitz et al., 2018] that captures the dependency of agent's policy on opponents' mental states (DDPG-ToM). We denote  $k$  as the *highest* level of reasoning in GR2-L/M, and adopt  $k = \{1, 2, 3\}, \lambda = 1.5$ . All results are reported with 6 random seeds. We leave the detailed hyper-parameter settings and ablation studies in *Appendix F* due to space limit.

**Keynes Beauty Contest.** In Keynes Beauty Contest ( $n, p$ ), all  $n$  agents pick a number between 0 and 100, the winner is the agent whose guess is closest to  $p$  times of the average number. The reward is set as the absolute difference.

In reality, higher-level thinking helps humans to get close to the Nash equilibrium of Keynes Beauty Contest (see *Introduction*). To validate if higher *level-k* model would make multi-agent learning more effective, we vary different  $p$  and  $n$  values and present the self-play results in Table. 1. We can tell that the GR2-L algorithms can effectively approach the equilibrium while the other baselines struggle to reach. The only exception is 99.0 in the case of ( $p = 1.1, n = 10$ ), which we believe is because of the saturated gradient from the reward.

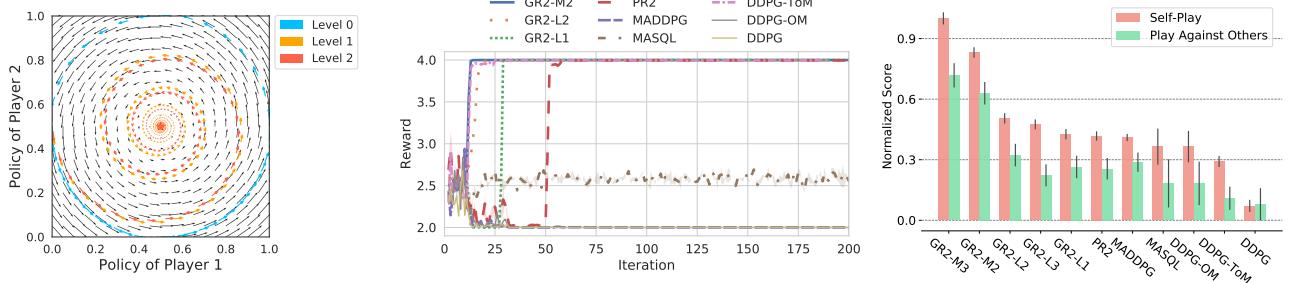


Figure 4: (a) Learning dynamics of GR2-L on Rotational Game. (b) Average reward on Stag Hunt. (c) Performance on Coop. Navigation.

We argue that the synergy of agents’ reaching the equilibria in this game only happens when the learning algorithm is able to make agents acknowledge different levels of rationality. For example, we visualize the step-wise reasoning outcomes of GR2-L3 in Fig. 3. During training, the agent shows ability to respond to his estimation of the opponent’s action by guessing a smaller number, e.g., in step 400,  $19.98 < 25.34$  and  $17.62 < 20.76$ . Even though the opponent estimation is not be accurate yet ( $20.76 \neq 19.98 \times 1.1$ ), the agent still realizes that, with the recursive level increases, the opponent’s guessing number will become smaller, in this case,  $20.76 < 25.34$ . Following this logic, both agents finally reach to 0. In addition, we find that in  $(p = 0.7, n = 2)$ , GR2-L1 is soon followed by the other higher-level GR2 models once it reaches the equilibria; this is in line with the Proposition 1.

To evaluate the robustness of GR2 methods outside the self-play context, we make each GR2 agent play against all the other six baselines by a round-robin style and present the averaged performance in Fig. 2b. GR2-M models outperform all the rest models by successfully guessing the right equilibrium, which is expected since the GR2-M is by design capable of considering different types of opponents.

Finally, we justify the necessity of adopting the auxiliary loss of Eq. 6 by Fig. 2a. As we simplify the reasoning rollouts by using deterministic policies, we believe adding the auxiliary loss in the objective can effectively mitigate the potential weakness of policy expressiveness and guide the joint Q-function to a better direction to improve the policy  $\pi_k^i$ .

**Normal-form Games.** We further evaluate the GR2 methods on two normal-form games: Rotational Game (RG) and Stag Hunt (SH). The reward matrix of RG is  $R_{RG} = \begin{bmatrix} 0, 3 & 3, 2 \\ 1, 0 & 2, 1 \end{bmatrix}$ , with the only equilibria at  $(0.5, 0.5)$ . In SH, the reward matrix is  $R_{SH} = \begin{bmatrix} 4, 4 & 1, 3 \\ 3, 1 & 2, 2 \end{bmatrix}$ . SH has two equilibria  $(S, S)$  that is Pareto optimal and  $(P, P)$  that is deficient.

In RG, we examine the effectiveness that *level-k* policies can converge to the equilibrium but *level-0* methods cannot. We plot the gradient dynamics of RG in Fig. 4a. *level-0* policy, represented by independent learners, gets trapped into the looping dynamics that never converges, while the GR2-L policies can converge to the center equilibrium, with higher-level policy allowing faster speed. These empirical findings in fact match the theoretical results on different learning dynamics demonstrated in the **proof of Theorem 2**.

To further evaluate the superiority of *level-k* models, we present Fig. 4b that compares the average reward on the

SH game where two equilibria exist. GR2 models, together with PR2 and DDPG-ToM, can reach the Pareto optima with the maximum reward 4, whereas other models are either fully trapped in the deficient equilibrium or mix in the middle. SH is a coordination game with no dominant strategy; agents choose between self-interest  $(P, P)$  and social welfare  $(S, S)$ . Without knowing the opponent’s choice, GR2 has to first anchor the belief that the opponent may choose the social welfare to maximize its reward, and then reinforce this belief by passing it to the higher-level reasonings so that finally the trust between agents can be built. The *level-0* methods cannot develop such synergy because they cannot discriminate the self-interest from the social welfare as both equilibria can saturate the value function. On the convergence speed in Fig. 4b, as expected, higher-level models are faster than lower-level methods, and GR2-M models are faster than GR2-L models.

**Cooperative Navigation.** We test the GR2 methods in more complexed Particle World environments [Lowe *et al.*, 2017] for the high-dimensional control task of *Cooperative Navigation* with 2 agents and 2 landmarks. Agents are collectively rewarded based on the proximity of any one of the agent to the closest landmark while penalized for collisions. The comparisons are shown in Fig. 4c where we report the averaged minimax-normalized score. We compare both the self-play performance and the averaged performance of playing with the rest 10 baselines one on one. We notice that the GR2 methods achieve critical advantages over traditional baselines in both the scenarios of self-play and playing against others; this is inline with the previous findings that GR2 agents are good at managing different levels of opponent rationality (in this case, each opponent may want to go to a different landmark) so that collisions are avoided at maximum. In addition, we can find that all the listed models show better self-play performance than that of playing with the others; intuitively, this is because the opponent modeling is more accurate during self-plays.

## 7 Conclusion

We have proposed a new solution concept to MARL – generalized recursive reasoning (GR2) – that enables agents to recognize opponents’ bounded rationality and their corresponding sub-optimal behaviors. GR2 establishes a reasoning hierarchy among agents, based on which we derive the practical GR2 soft actor-critic algorithm. Importantly, we prove in theory the existence of Perfect Bayesian Equilibrium under the GR2 setting as well as the convergence of the policy gradient methods on the two-player normal-form games. Series of experimental results have justified the advantages of GR2 methods over strong MARL baselines on modeling different opponents.

## References

- [Abdallah and Lesser, 2008] Sherief Abdallah and Victor Lesser. A multiagent reinforcement learning algorithm with non-linear dynamics. *JAIR*, 33:521–549, 2008.
- [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.
- [Bowling and Veloso, 2001] Michael Bowling and Manuela Veloso. Convergence of gradient dynamics with a variable learning rate. In *ICML*, pages 27–34, 2001.
- [Bowling and Veloso, 2002] Michael Bowling and Manuela Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136(2):215–250, 2002.
- [Camerer *et al.*, 2004] Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 2004.
- [Claus and Boutilier, 1998] Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI*, 1998:746–752, 1998.
- [Coricelli and Nagel, 2009] Giorgio Coricelli and Rosemarie Nagel. Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, 106(23):9163–9168, 2009.
- [Devetag and Warglien, 2003] Giovanna Devetag and Massimo Warglien. Games and phone numbers: Do short-term memory bounds affect strategic behavior? *Journal of Economic Psychology*, 24(2):189–202, 2003.
- [Genewein *et al.*, 2015] Tim Genewein, Felix Leibfried, Jordi Grau-Moya, and Daniel Alexander Braun. Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, 2:27, 2015.
- [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.
- [Goldman and others, 2012] Alvin I Goldman et al. Theory of mind. *The Oxford handbook of philosophy of cognitive science*, pages 402–424, 2012.
- [Haarnoja *et al.*, 2017] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. *NIPS*, 2017.
- [Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.
- [He *et al.*, 2016] He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. Opponent modeling in deep reinforcement learning. In *ICML*, 2016.
- [Hunicke, 2005] Robin Hunicke. The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI ACE*. ACM, 2005.
- [Keynes, 1936] J. M. Keynes. *The General Theory of Employment, Interest and Money*. Macmillan, 1936. 14th edition, 1973.
- [Kreps and Wilson, 1982] David M Kreps and Robert Wilson. Sequential equilibria. *Econometrica: Journal of the Econometric Society*, pages 863–894, 1982.
- [Levin and Zhang, 2019] Dan Levin and Luyao Zhang. Bridging level-k to nash equilibrium. *Available at SSRN 2934696*, 2019.
- [Levine, 2018] Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.
- [Lillicrap *et al.*, 2015] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.
- [Liu and Wang, 2016] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *NIPS*, pages 2378–2386, 2016.
- [Lowe *et al.*, 2017] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multiagent actor-critic for mixed cooperative-competitive environments. In *NIPS*, pages 6379–6390, 2017.
- [Marquez, 2003] Horacio J Marquez. *Nonlinear control systems: analysis and design*, volume 1. Wiley-Interscience Hoboken, 2003.
- [McKelvey and Palfrey, 1995] Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 1995.
- [Nash and others, 1950] John F Nash et al. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.
- [Peng *et al.*, 2017] Peng Peng, Quan Yuan, Ying Wen, Yaodong Yang, Zhenkun Tang, Haitao Long, and Jun Wang. Multiagent bidirectionally-coordinated nets for learning to play starcraft combat games. *arXiv preprint arXiv:1703.10069*, 2, 2017.
- [Rabinowitz *et al.*, 2018] Neil C Rabinowitz, Frank Perbet, H Francis Song, Chiyuan Zhang, SM Eslami, and Matthew Botvinick. Machine theory of mind. *ICML*, 2018.
- [Shalev-Shwartz *et al.*, 2016] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.
- [Shapley, 1953] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.
- [Shoham *et al.*, 2003] Yoav Shoham, Rob Powers, and Trond Grenager. Multi-agent reinforcement learning: a critical survey. In *Technical report*, 2003.
- [Shum *et al.*, 2019] Michael Shum, Max Kleiman-Weiner, Michael L Littman, and Joshua B Tenenbaum. Theory of

- minds: Understanding behavior in groups through inverse planning. *AAAI*, 2019.
- [Simon, 1972] Herbert A Simon. Theories of bounded rationality. *Decision and organization*, 1(1):161–176, 1972.
- [Taylor *et al.*, 2018] Adrien Taylor, Bryan Van Scoy, and Laurent Lessard. Lyapunov functions for first-order methods: Tight automated convergence guarantees. *arXiv preprint arXiv:1803.06073*, 2018.
- [Tesauro, 1995] Gerald Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.
- [Tian *et al.*, 2019] Zheng Tian, Ying Wen, Zhicheng Gong, Faiz Punakkath, Shihao Zou, and Jun Wang. A regularized opponent model with maximum entropy objective. *IJCAI*, 2019.
- [Wei *et al.*, 2018] Ermo Wei, Drew Wicke, David Freelan, and Sean Luke. Multiagent soft q-learning. *AAAI*, 2018.
- [Wen *et al.*, 2019] Ying Wen, Yaodong Yang, Rui Luo, Jun Wang, and Wei Pan. Probabilistic recursive reasoning for multi-agent reinforcement learning. In *ICLR*, 2019.
- [Yang *et al.*, 2018] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *ICML*, 2018.
- [Zhang and Lesser, 2010] Chongjie Zhang and Victor Lesser. Multi-agent learning with policy prediction. In *AAAI*, 2010.

## Appendix

### A Maximum Entropy Multi-Agent Reinforcement Learning

We give the overall optimal distribution  $p(\tau^i) = p(a_{1:T}^i, a_{1:T}^j, s_{1:T})$  of agent  $i$  at first:

$$p(a_{1:T}^i, a_{1:T}^j, s_{1:T}) = [p(s_1) \prod_{t=1}^T p(s_{t+1}|s_t, a_t^i, a_t^{-i})] \exp\left(\sum_{t=1}^T r^i(s_t, a_t, a_t^{-i})\right). \quad (7)$$

Analogously, we factorize empirical trajectory distribution  $\hat{p}(\tau^i)$  as:

$$\hat{p}(\tau^i) = p(s_1) \prod_t p(s_{t'}|s_t, a_t) \pi^i(a_t^i|s_t) \rho^{-i}(a_t^{-i}|s_t, a_t^i), \quad (8)$$

where  $\rho^{-i}(a_t^{-i}|s_t, a_t^i)$  is agent  $i$ 's model about the opponent's conditional policy, and  $\pi^i(a_t^i|s_t)$  marginal policy of agent  $i$ . With fixed dynamics assumption, we can minimize the KL-divergence as follow:

$$\begin{aligned} -D_{\text{KL}}(\hat{p}(\tau^i)\|p(\tau^i)) &= \mathbb{E}_{\tau^i \sim \hat{p}(\tau^i)} \left[ \log p(s_1) + \sum_{t=1}^T \left( \log p(s_{t+1}|s_t, a_t, a_t^{-i}) + r^i(s_t, a_t^i, a_t^{-i}) \right) \right. \\ &\quad \left. - \log p(s_1) - \sum_{t=1}^T \left( \log p(s_{t+1}|s_t, a_t^i, a_t^{-i}) + \log (\pi^i(a_t^i|s_t) \rho^{-i}(a_t^{-i}|s_t, a_t^i)) \right) \right] \\ &= \mathbb{E}_{\tau^i \sim \hat{p}(\tau^i)} \left[ \sum_{t=1}^T r^i(s_t, a_t^i, a_t^{-i}) - \log (\pi^i(a_t^i|s_t) \rho^{-i}(a_t^{-1}|s_t, a_t^i)) \right] \\ &= \sum_{t=1}^T \mathbb{E}_{(s_t, a_t^i, a_t^{-i}) \sim \hat{p}(s_t, a_t^i, a_t^{-i})} \left[ r^i(s_t, a_t^i, a_t^{-i}) - \log (\pi^i(a_t^i|s_t) \rho^{-i}(a_t^{-1}|s_t, a_t^i)) \right] \\ &= \sum_{t=1}^T \mathbb{E}_{(s_t, a_t^i, a_t^{-i}) \sim \hat{p}(s_t, a_t^i, a_t^{-i})} \left[ r^i(s_t, a_t^i, a_t^{-i}) + \mathcal{H}(\rho^{-i}(a_t^{-i}|s_t, a_t^i)) + \mathcal{H}(\pi^i(a_t^i|s_t)) \right], \end{aligned} \quad (9)$$

where  $\mathcal{H}$  is entropy term, and the objective is to maximize reward and polices' entropy.

In multi-agent cooperation case, the agents work on a shared reward, which implies  $\rho^{-i}(a_t^{-i}|s_t, a_t^i)$  would help to maximize the shared reward. It does not mean that the agent can control the others, just a reasonable assumption that the others would coordinate on the same objective. As before, we can find the optimal  $\rho^i(a_t^j|s_t, a_t^i)$  by recursively maximizing:

$$\mathbb{E}_{(s_t, a_t^i, a_t^{-i}) \sim \hat{p}(s_t, a_t^i, a_t^{-i})} \left[ -D_{\text{KL}} \left( \rho_t^{-i}(a_t^{-i}|s_t, a_t^i) \middle\| \frac{\exp(Q^i(s_t, a_t^i, a_t^{-i}))}{\exp(Q^i(s_t, a_t^i))} \right) + Q^i(s_t, a_t^i) \right], \quad (10)$$

where we define:

$$Q^i(s, a^i) = \log \sum_{a^{-i}} \exp(Q^i(s, a^i, a^{-i})), \quad (11)$$

which corresponds to a bellman backup with a soft maximization. And optimal opponent conditional policy is given as:

$$\rho^{-i}(a^{-i}|s, a^i) \propto \exp(Q^i(s, a^i, a^{-i}) - Q^i(s, a^i)). \quad (12)$$

### B Algorithm Implementations

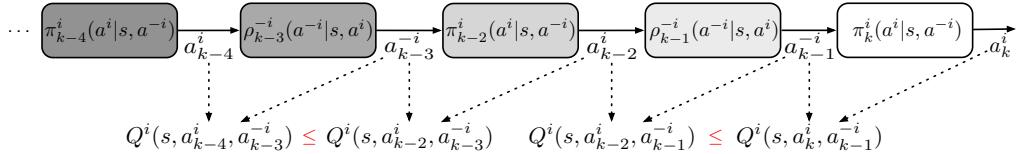


Figure 5: Inter-level policy improvement maintained by Eq. 6 so that higher-level policy weakly dominates lower-level policies.

## C Proof of Theorem 1

**Theorem 1.** GR2 strategies extend a norm-form game into extensive-form game, and there exists a Perfect Bayesian Equilibrium in that game.

*Proof.* Consider an extensive game, which is extended from a normal form game by *level-k* strategies, with perfect information and recall played by two players  $(i, -i)$ :  $(\pi^i, \pi^{-i}, u^i, u^{-i}, \Lambda)$ , where  $\pi^{i/-i}$  and  $u^{i/-i}$  are strategy pairs and payoff functions for player  $i, -i$  correspondingly.  $\Lambda$  denotes the lower-level reasoning trajectory/path so far. An intermediate reasoning action/node in the *level-k* reasoning procedure is denoted by  $h_t$ . The set of the intermediate reasoning actions at which player  $i$  chooses to move is denoted  $H^i$  (a.k.a information set). Let  $\pi_{\tilde{k}}^{i/-i}$  denote the strategies of a *level- $\tilde{k}$*  player and  $\tilde{k} \in \{0, 1, 2 \dots k\}$ . A *level-k* player holds a prior belief that the opponent is a *level- $\tilde{k}$*  player with probability  $\lambda_{\tilde{k}}$ , where  $\lambda_{\tilde{k}} \in [0, 1]$  and  $\sum_{\tilde{k}=0}^k \lambda_{\tilde{k}} = 1$ . We denote the belief that the opponent is a *level- $\tilde{k}$*  player as  $p_{\tilde{k}}^i(h^t)$ . In equilibrium, a *level-k* player chooses an optimal strategy according to her belief at every decision node, which implies choice is sequentially rational as following defined:

**Definition 1.** (*Sequential Rationality*). A strategy pair  $\{\pi_*^i, \pi_*^{-i}\}$  is sequentially rational with respect to the belief pair  $\{p^i, p^{-i}\}$  if for both  $i, -i$ , all strategy pairs  $\{\pi^i, \pi^{-i}\}$  and all nodes  $h_t^i \in H^i$ :

$$\sum_{\tilde{k}=0}^k \lambda_{\tilde{k}} p_{\tilde{k}}^i(h_t^i) u^i(\pi_*, \pi_*^{-i} | h_t^i) \geq \sum_{\tilde{k}=0}^k \lambda_{\tilde{k}} p_{\tilde{k}}^i(h_t^i) u^i(\pi^i, \pi_*^{-i} | h_t^i),$$

Based on Definition 1, we have the strategy  $\pi^i$  is **sequentially rational** given  $p^i$ . It means strategy of player  $i$  is optimal in the part of the game that follows given the strategy profile and her belief about the history in the information set that has occurred.

In addition, we also require the beliefs of an *level-k* player are consistent. Let  $p^i(h_t | \pi^i, \pi^{-i})$  denote the probability that reasoning action  $h_t$  is reached according to the strategy pair,  $\{\pi^i, \pi^{-i}\}$ . Then we have the consistency definition:

**Definition 2.** (*Consistency*). The belief pair  $\{\rho_*^{-i}, p_*^i\}$  is consistent with the subjective prior  $\lambda_{\tilde{k}}$ , and the strategy pair  $\{\pi^i, \pi^{-i}\}$  if and only if for  $i, -i$  and all nodes  $h_t^i \in H^i$ :

$$p_{k,*}^i(h_t^i) = \frac{\lambda_{\tilde{k}} p_{\tilde{k}}^i(h_t^i | \pi^i, \pi^{-i})}{\sum_{\tilde{k}=0}^k \lambda_{\tilde{k}} p_{\tilde{k}}^i(h_t^i | \pi^i, \pi^{-i})},$$

where there is at least one  $\tilde{k} \in \{0, 1, 2 \dots k\}$  and  $p_{\tilde{k}}^i(h_t^i | \pi^i, \pi^{-i}) > 0$ .

The belief  $p^i$  is **consistent** given  $\pi^i, \pi^{-i}$  indicates that for every intermediate reasoning actions reached with positive probability given the strategy profile  $\pi^i, \pi^{-i}$ , the probability assigned to each history in the reasoning path by the belief system  $p^i$  is given by Bayes' rule. In summary, sequential rationality implies each player's strategy optimal at the beginning of the game given others' strategies and beliefs [Levin and Zhang, 2019]. Consistency ensures correctness of the beliefs.

Although the game itself has perfect information, the belief structure in our *level-k* thinking makes our solution concept an analogy of a Perfect Bayesian Equilibrium. Based on above two definitions, we have the existence of Perfect Bayesian Equilibrium in *level-k* thinking game.

**Proposition.** For any  $\lambda_{\tilde{k}}$ , where  $\lambda_{\tilde{k}} \in [0, 1]$  and  $\sum_{\tilde{k}=0}^k \lambda_{\tilde{k}} = 1$ , there is a Perfect Bayesian Equilibrium exists.

Now, consider an extensive game of incomplete information,  $(\pi^i, \pi^{-i}, u^i, u^{-i}, p^i, p^{-i}, \lambda_k, \Lambda)$ , where  $\lambda_k$  denotes the possible levels/types for agents, which can be arbitrary *level-k* player. Then, according to Kreps and Wilson [1982], for every finite extensive form game, there exists at least one sequential equilibrium should satisfy Definition. 1 and 2 for sequential rationality and consistency, and the details proof as following:

We use  $E^i(\pi, p, \lambda_k, h^i) = \sum_{\tilde{k}=0}^k \lambda_{\tilde{k}} p_{\tilde{k}}^i(h_t^i) u^i(\pi^i, \pi^{-i} | h_t^i)$  as expected payoff for player  $i$ , for every player  $i$  and each reasoning path  $h_t^i$ . Choose a large integer  $m (m > 0)$  and consider the sequence of strategy pairs and consistent belief pairs  $\{\pi_m, p_m\}_m$ , there exists a  $(\pi_m, p_m)$ :

$$E^i(\pi_m, p_m, \lambda_k, h_{t^i}^i) \geq E^i((\pi_m^{-i}, \pi^i), p_n(\pi_m^{-i}, \pi^i), \lambda_k, h_{t^i}^i),$$

for any strategy  $\pi^i$  with induced probability distributions in  $\Pi_{t^i=1}^T = \Delta^{\frac{1}{m}}(p(h_{t^i}^i))$ .

Then, consider the strategy and belief pair  $(\hat{\pi}, \hat{p})$  given by:

$$(\hat{\pi}, \hat{p}) = \lim_{m \rightarrow \infty} (\pi_m, p_m).$$

Such a limit exists because  $\Pi_{j=1}^m \Pi_{t_j=1}^{T_j} \Delta^{\frac{1}{m}}(p(h_{t_j}^j))$  forms a compact subset of a Euclidean space, and every sequence  $\{\pi_m, p_m\}_m$  has a limit point. We claim that for each player  $i$  and each reasoning path  $h_{t^i}^i$ :

$$E^i(\hat{\pi}_m, \hat{p}_m, \lambda_k, h_{t^i}^i) \geq E^i((\hat{\pi}_m^{-i}, \pi^i), p(\hat{\pi}_m^{-i}, \pi^i), \lambda_k, h_{t^i}^i), \quad (13)$$

for any strategy  $\pi^i$  of player  $i$ .

If not, then for some player  $i$  and some strategy  $\pi^i$  of player  $i$ , we have:

$$E^i(\hat{\pi}_m, \hat{p}_m, \lambda_k, h_{t^i}^i) < E^i((\hat{\pi}_m^{-i}, \lambda_k, \pi^i), p(\hat{\pi}_m^{-i}, \pi^i), \lambda_k, h_{t^i}^i).$$

Then, we let

$$E^i((\hat{\pi}_m^{-i}, \pi^i), p(\hat{\pi}_m^{-i}, \pi^i), \lambda_k, h_{t^i}^i) - E^i(\hat{\pi}_m, \hat{p}_m, \lambda_k, h_{t^i}^i) = b > 0.$$

Now as the expected payoffs are continuous in the probability distributions at the reasoning paths and the beliefs, it follows that there is an  $m_0$  sufficiently large such that for all  $m \geq m_0$ ,

$$|E^i(\pi_m, p_m, \lambda_k, h_{t^i}^i) - E^i(\hat{\pi}_m, \hat{p}_m, \lambda_k, h_{t^i}^i)| \leq \frac{b}{4},$$

and

$$E^i((\hat{\pi}_m^{-i}, \pi^i), p_n(\hat{\pi}_m^{-i}, \pi^i), \lambda_k, h_{t^i}^i) - E^i(\hat{\pi}_m, \hat{p}_m, \lambda_k, h_{t^i}^i) \leq \frac{b}{4}.$$

From above equations and for all  $m \geq m_0$ , we have

$$\begin{aligned} E^i((\pi_m^{-i}, \pi^i), p(\pi_m^{-i}, \pi^i), \lambda_k, h_{t^i}^i) &\geq E^i((\hat{\pi}_m^{-i}, \pi^i), p(\hat{\pi}_m^{-i}, \pi^i), \lambda_k, h_{t^i}^i) - \frac{b}{4} \\ &= E^i(\hat{\pi}, \hat{p}, \lambda_k, h_{t^i}^i) + \frac{3b}{4} \\ &\geq E^i((\pi_m^{-i}, \pi^i), p(\pi_m^{-i}, \pi^i), \lambda_k, h_{t^i}^i) + \frac{b}{2}. \end{aligned}$$

for a given sequential game, there is a  $T > 0$  such that

$$\left| E^i\left[(\pi_\xi^{-i}, \pi^i), p_n(\pi_\xi^{-i}, \pi^i), \lambda_k, h_{t^i}^i\right] - E^i(\hat{\pi}_\xi, \hat{p}_\xi, \lambda_k, h_{t^i}^i) \right| < \frac{T}{\xi},$$

where  $\pi^i = \lim_{\xi \rightarrow \infty} \pi_\xi^i$  of a sequence  $\{\pi_\xi^i\}_\xi$  of  $\frac{1}{\xi}$  bounded strategies of player  $i$ . For the sequence  $\{\pi_m, p_m\}$  we now choose an  $m_1$  sufficiently large such that  $\frac{T}{m} < \frac{b}{4}$ . Therefore, for any strategy  $\pi^i$  of player  $i$ , we have

$$\begin{aligned} E^i((\pi_m^{-i}, \pi^i), p_n(\pi_m^{-i}, \pi^i), \lambda_k, h_{t^i}^i) &\geq E^i((\pi_m^{-i}, \pi^i), p(\pi_m^{-i}, \pi^i), \lambda_k, h_{t^i}^i) - \frac{T}{m} \\ &= E^i(\pi_m, p_m, \lambda_k, h_{t^i}^i) + \frac{b}{4}. \end{aligned}$$

But this result contradicts the previous claim in Eq. 13, which indicates the claim must hold. In other words, Perfect Bayesian Equilibrium must exist. ■

**Remark.** When  $\lambda_k = 1$ , it is the special case where the policy is level- $k$  strategy, and it coincides with Perfect Bayesian Equilibrium.

## D Proof of Theorem 2

**Theorem 2.** In two-player two-action games, if there exist a mixed strategy equilibrium, under mild conditions, the learning dynamics of GR2 methods to the equilibrium is asymptotic stable in the sense of Lyapunov.

*Proof.* We start by defining the matrix game that a mixed-strategy equilibrium exists, and they we show that on such game level-0 independent learner through iterated gradient ascent will not converge, and finally derive why the level- $k$  methods would converge in this case. Our tool is Lyapunov function and its stability analysis.

Lyapunov function is used to verify the stability of a dynamical system in control theory, here we apply it in convergence proof for level- $k$  methods. It is defined as following:

**Definition 3. (Lyapunov Function.)** Give a function  $F(x, y)$  be continuously differentiable in a neighborhood  $\sigma$  of the origin. The function  $F(x, y)$  is called the Lyapunov function for an autonomous system if that satisfies the following properties:

1. (nonnegative)  $F(x, y) > 0$  for all  $(x, y) \in \sigma \setminus (0, 0)$ ;

2. (zero at fixed-point)  $F(0, 0) = 0$ ;
3. (decreasing)  $\frac{dF}{dt} \leq 0$  for all  $(x, y) \in \sigma$ .

**Definition 4.** (*Lyapunov Asymptotic Stability*.) For an autonomous system, if there is a Lyapunov function  $F(x, y)$  with a negative definite derivative  $\frac{dF}{dt} < 0$  (strictly negative, negative definite LaSalle's invariance principle) for all  $(x, y) \in \sigma \setminus (0, 0)$ , then the equilibrium point  $(x, y) = (0, 0)$  of the system is asymptotically stable [Marquez, 2003].

### Single State Game

Given a two-player, two-action matrix game, which is a single-state stage game, we have the payoff matrices for row player and column player as follows:

$$\mathbf{R}_r = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{R}_c = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}.$$

Each player selects an action from the action space  $\{1, 2\}$  which determines the payoffs to the players. If the row player chooses action  $i$  and the player 2 chooses action  $j$ , then the row player and column player receive the rewards  $r_{ij}$  and  $c_{ij}$  respectively. We use  $\alpha \in [0, 1]$  to represent the strategy for row player, where  $\alpha$  corresponds to the probability of player 1 selecting the first action (action 1), and  $1 - \alpha$  is the probability of choosing the second action (action 2). Similarly, we use  $\beta$  to be the strategy for column player. With a joint strategy  $(\alpha, \beta)$ , the expected payoffs of players are:

$$\begin{aligned} V_r(\alpha, \beta) &= \alpha\beta r_{11} + \alpha(1 - \beta)r_{12} + (1 - \alpha)\beta r_{21} + (1 - \alpha)(1 - \beta)r_{22}, \\ V_c(\alpha, \beta) &= \alpha\beta c_{11} + \alpha(1 - \beta)c_{12} + (1 - \alpha)\beta c_{21} + (1 - \alpha)(1 - \beta)c_{22}. \end{aligned}$$

One crucial aspect to the learning dynamics analysis are the points of zero-gradient in the constrained dynamics, which they show to correspond to the equilibria which is called the center and denoted  $(\alpha^*, \beta^*)$ . This point can be found mathematically  $(\alpha^*, \beta^*) = \left( \frac{-b_c}{u_c}, \frac{-b_r}{u_r} \right)$ , where  $u_r = r_{11} - r_{12} - r_{21} + r_{22}$ ,  $b_r = r_{12} - r_{22}$ ,  $u_c = c_{11} - c_{12} - c_{21} + c_{22}$ , and  $b_c = c_{21} - c_{22}$ .

Here we are more interested in the case that there exists a mixed strategy equilibrium, i.e., the location of the equilibrium point  $(\alpha^*, \beta^*)$  is in the interior of the unit square, equivalently, it means  $u_r u_c < 0$ . In other cases where the Nash strategy on the boundary of the unit square [Marquez, 2003; Bowling and Veloso, 2001], we are not going to discuss in this proof.

### Learning in level-0 Gradient Ascent

One common *level-0* policy is Infinitesimal Gradient Ascent (IGA), which assumes independent learners and is a *level-0* method, a player increases its expected payoff by moving its strategy in the direction of the current gradient with fixed step size. The gradient is then computed as the partial derivative of the agent's expected payoff with respect to its strategy, we then have the policies dynamic partial differential equations:

$$\frac{\partial V_r(\alpha, \beta)}{\partial \alpha} = u_r \beta + b_r, \quad \frac{\partial V_c(\alpha, \beta)}{\partial \beta} = u_c \alpha + b_c.$$

In the gradient ascent algorithm, a player will adjust its strategy after each iteration so as to increase its expected payoffs. This means the player will move their strategy in the direction of the current gradient with some step size. Then we can have dynamics are defined by the differential equation at time  $t$ :

$$\begin{bmatrix} \frac{\partial \alpha}{\partial t} \\ \frac{\partial \beta}{\partial t} \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & u_r \\ u_c & 0 \end{bmatrix}}_U \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} b_r \\ b_c \end{bmatrix}.$$

By defining multiplicative matrix term  $U$  above with off-diagonal values  $u_r$  and  $u_c$ , we can classify the dynamics of the system based on properties of  $U$ . As we mentioned, we are interested in the case that the game has just one mixed center strategy equilibrium point (not saddle point) that in the interior of the unit square, which means  $U$  has purely imaginary eigenvalues and  $u_r u_c < 0$  [Zhang and Lesser, 2010].

Consider the quadratic Lyapunov function which is continuously differentiable and  $F(0, 0) = 0$ :

$$F(x, y) = \frac{1}{2}(u_c x^2 - u_r y^2),$$

where we suppose  $u_c > 0, u_r < 0$  (we can have identity case when  $u_c < 0, u_r > 0$  by switching the sign of the function). Its derivatives along the trajectories by setting  $x = \alpha - \alpha^*$  and  $y = \beta - \beta^*$  to move the the equilibrium point to origin can be calculated as:

$$\frac{dF}{dt} = \frac{\partial F}{\partial x} \frac{dx}{dt} + \frac{\partial F}{\partial y} \frac{dy}{dt} = xy(u_r u_c - u_r u_c) = 0,$$

where the derivative of the Lyapunov function is identically zero. Hence, the condition of asymptotic stability is not satisfied [Marquez, 2003; Taylor *et al.*, 2018] and the IGA *level-0* dynamics is unstable. There are some IGA based methods (WoLF-IGA, WPL etc. [Bowling and Veloso, 2002; Abdallah and Lesser, 2008]) with varying learning step, which change the  $U$  to  $\begin{bmatrix} 0 & l_r(t)u_r \\ l_c(t)u_c & 0 \end{bmatrix}$ . The time dependent learning steps  $l_r(t)$  and  $l_c(t)$  are chose to force the dynamics would converge. Note that diagonal elements in  $U$  are still zero, which means a player's personal influences to the system dynamics are not reflected on its policy adjustment.

### Learning in *level-k* Gradient Ascent

Consider a *level-1* gradient ascent, where agent learns in term of  $\pi_r(\alpha)\pi_c^1(\beta|\alpha)$ , the gradient is computed as the partial derivative of the agent's expected payoff after considering the opponent will have *level-1* prediction to its current strategy. We then have the *level-1* policies dynamic partial differential equations:

$$\frac{\partial V_r(\alpha, \beta_1)}{\partial \alpha} = u_r(\beta + \zeta \partial_\beta V_c(\alpha, \beta)) + b_r, \quad \frac{\partial V_c(\alpha_1, \beta)}{\partial \beta} = u_c(\alpha + \zeta \partial_\alpha V_r(\alpha, \beta)) + b_c,$$

where  $\zeta$  is short-term prediction of the opponent's strategy. Its corresponding *level-1* dynamic partial differential equations:

$$\begin{bmatrix} \frac{\partial \alpha}{\partial t} \\ \frac{\partial \beta}{\partial t} \end{bmatrix} = \underbrace{\begin{bmatrix} \zeta u_r u_c & u_r \\ u_c & \zeta u_r u_c \end{bmatrix}}_U \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \zeta u_r b_c + b_r \\ \zeta u_c b_r + b_c \end{bmatrix}.$$

Apply the same quadratic Lyapunov function:  $F(x, y) = 1/2(u_c x^2 - u_r y^2)$ , where  $u_c > 0, u_r < 0$ , and its derivatives along the trajectories by setting  $x = \alpha - \alpha^*$  and  $y = \beta - \beta^*$  to move the coordinates of equilibrium point to origin:

$$\frac{dF}{dt} = \zeta u_r u_c (u_c x^2 - u_r y^2) + xy(u_r u_c - u_r u_c) = \zeta u_r u_c (u_c x^2 - u_r y^2),$$

where the conditions of asymptotic stability is satisfied due to  $u_r u_c < 0, u_c > 0$  and  $u_r < 0$ , and it indicates the derivative  $\frac{dF}{dt} < 0$ . In addition, unlike the *level-0*'s case, we can find that the diagonal of  $U$  in this case is non-zero, it measures the mutual influences between players after *level-1* looks ahead and helps the player to update it's policy to a better position.

This conclusion can be easily extended and proved in *level-k* gradient ascent policy ( $k > 1$ ). In *level-k* gradient ascent policy, we can have the derivatives of same quadratic Lyapunov function in *level-2* dynamics:

$$\frac{dF}{dt} = \zeta u_r u_c (u_c x^2 - u_r y^2) + xy(1 + \zeta^2 u_r u_c)(u_r u_c - u_r u_c) = \zeta u_r u_c (u_c x^2 - u_r y^2),$$

and *level-3* dynamics:

$$\frac{dF}{dt} = \zeta u_r u_c (2 + \zeta^2 u_r u_c)(u_c x^2 - u_r y^2).$$

Repeat the above procedures, we can easily write the general derivatives of quadratic Lyapunov function in *level-k* dynamics:

$$\frac{dF}{dt} = \zeta u_r u_c (k - 1 + \dots + \zeta^{k-1} (u_r u_c)^{k-2})(u_c x^2 - u_r y^2),$$

where  $k \geq 3$ . These *level-k* policies still owns the asymptotic stability property when  $\zeta^2$  is sufficiently small (which is trivial to meet in practice) to satisfy  $k - 1 + \dots + \zeta^{k-1} (u_r u_c)^{k-2} > 0$ , which meets the asymptotic stability conditions, therefore coverages. ■

## E Proof of Proposition 1

**Proposition 1.** *In both the GR2-L & GR2-M model, if the agents play pure strategies, once level-k agent reaches a Nash Equilibrium, all higher-level agents will follow it too.*

*Proof.* Consider the following two cases GR2-L and GR2-M.

**GR2-L.** Since agents are assumed to play pure strategies, if a *level-k* agent reaches the equilibrium,  $\pi_{k,*}^i$ , in the GR2-L model, then all the higher-level agents will play that equilibrium strategy too, i.e.  $\pi_{k+1,*}^{-i} = \pi_{k,*}^i$ . The reason is because high-order thinkers will conduct at least the same amount of computations as the lower-order thinkers, and *level-k* model only needs to best respond to *level-(k-1)*. On the other hand, as it is showed by the Eq. 3 in the main paper, higher-level recursive model contains the lower-level models by incorporating it into the inner loop of the integration.

**GR2-M.** In the GR2-M model, if the *level- $k$*  step agent play the equilibrium strategy  $\pi_{k,*}^i$ , it means the agent finds the best response to a mixture type of agents that are among *level-0* to *level- $(k - 1)$* . Such strategy  $\pi_{k,*}^i$  is at least weakly dominant over other pure strategies. For *level- $(k + 1)$*  agent, it will face a mixture type of *level-0* to *level- $(k - 1)$* , plus *level- $k$* .

For mixture of *level-0* to *level- $(k - 1)$* , the strategy  $\pi_{k,*}^i$  is already the best response by definition. For *level- $k$* ,  $\pi_{k,*}^{i,*}$  is still the best response due to the conclusion in the above GR2-L. Considering the linearity of the expected reward for GR2-M:

$$\mathbb{E}[\lambda_0 V^i(s; \pi_{0,*}^i, \pi^{-i}) + \dots + \lambda_k V^i(s; \pi_{k,*}^i, \pi^{-i})] = \lambda_0 \mathbb{E}[V^i(s; \pi_{0,*}^i, \pi^{-i})] + \dots + \lambda_k \mathbb{E}[V^i(s; \pi_{k,*}^i, \pi^{-i})],$$

where  $\lambda_k$  is *level- $k$*  policy's proportion. Therefore,  $\pi_{k,*}^{i,*}$  is the best response to the mixture of *level-0* to *level- $k$*  agent, i.e. the best response for *level- $(k + 1)$*  agent. Given that  $\pi_{k,*}^{i,*}$  is the best response to both *level- $k$*  and all lower levels from 0 to  $(k - 1)$ , it is therefore the best response of the *level- $(k + 1)$*  thinker.

Combining the above two results, therefore, in GR2, once a *level- $k$*  agent reaches a pure Nash strategy, all higher-level agents will play it too. ■

## F Detailed Settings for Experiments

### The Recursive Level

We regard DDPG, DDPG-OM, MASQL, MADDPG as *level-0* reasoning models because from the policy level, they do not explicitly model the impact of one agent's action on the other agents or consider the reactions from the other agents. Even though the value function of the joint policy is learned in MASQL and MADDPG, but they conduct a *non-correlated factorization* [Wen et al., 2019] when it comes to each individual agent's policy. PR2 and DDPG-ToM are in fact the *level-1* reasoning model, but note that the *level-1* model in GR2 stands for  $\pi_1^i(a^i|s) = \int_{a^{-i}} \pi_1^i(a^i|s, a^{-i}) \rho_0^{-i}(a^{-i}|s) da^{-i}$ , while the *level-1* model in PR2 starts from the opponent's angel, that is  $\rho_1^{-i}(a^{-i}|s) = \int_{a^i} \rho_1^i(a^{-i}|s, a^i) \pi_0^i(a^i|s) da^i$ .

### Hyperparameter Settings

In all the experiments, we have the following parameters. The Q-values are updated using Adam with learning rate  $10^{-4}$ . The DDPG policy and soft Q-learning sampling network use Adam with a learning rate of  $10^{-4}$ . The methods use a replay pool of size  $100k$ . Training does not start until the replay pool has at least  $1k$  samples. The batch size 64 is used. All the policies and  $Q$ -functions are modeled by the MLP with 2 hidden layers followed by ReLU activation. In matrix games and Keynes Beauty Contest, each layer has 10 units and 100 units are set in cooperative navigation's layers. In the actor-critic setting, we set the exploration noise to 0.1 in the first  $1k$  steps. The annealing parameter in soft algorithms is decayed in linear scheme with training step grows to balance the exploration. Deterministic policies additional OU Noise to improve exploration with parameters  $\theta = 0.15$  and  $\sigma = 0.3$ . We update the target parameters softly by setting target smoothing coefficient to 0.001. We train with 6 random seeds for all environments. In Keynes Beauty Contest, we train all the methods for 400 iterations with 10 steps per iteration. In the matrix games, we train the agents for 200 iterations with 25 steps per iteration. For the cooperative navigation, all the models are trained up to  $300k$  steps with maximum 25 episode length.

### Ablation Study

The results in the experiment section of the main paper suggest that GR2 algorithms can outperform other multi-agent RL methods various tasks. In this section, we examine how sensitive GR2 methods is to some of the most important hyper-parameters, including the *level- $k$*  and the choice of the poisson mean  $\lambda$  in GR2-M methods, as well as the influences of incentive intensity in the games.

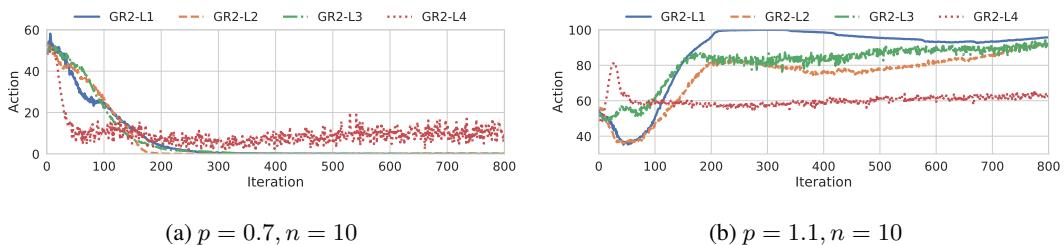


Figure 6: Learning curves on Keynes Beauty Contest game with GR2-L policies from *level-1* to *level-4*.

**Choice of  $k$  in *level- $k$*  Models.** First, we investigate the choice of *level- $k$*  by testing the GR2-L models with various  $k$  on Keynes Beauty Contest. According to the Fig. 6, in both setting, the GR3-L with level form 1 – 3 can converge to the

equilibrium, but the GR3-L4 cannot. The learning processes show that the GR3-L4 models have high variance during the learning. This phenomenon has two reasons: with  $k$  increases, the reasoning path would have higher variance; and in GR2-L4 policy, it uses the approximated opponent conditional policy  $\rho^{-i}(a^{-i}|s, a^i)$  twice (only once in GR2-L2/3), which would further amplify the variance.

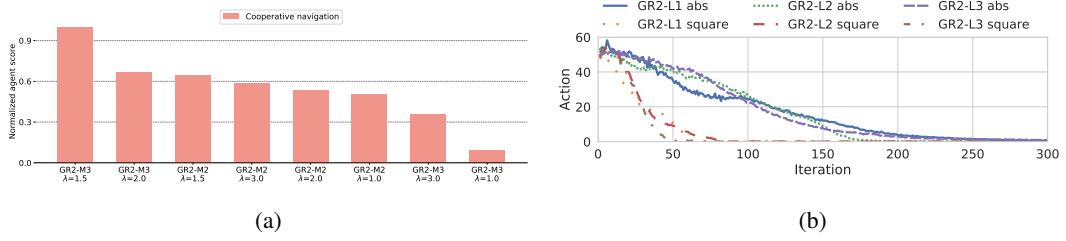


Figure 7: (a) Effect of varying  $\lambda$  in GR2-M methods, the score is normalized to 0 – 1. (b) Learning curves with two reward schemes: absolute difference (default) and squared absolute difference.

**Choice of  $\lambda$  of Poisson Distribution in GR2-M.** We investigate the effect of hyper-parameter  $\lambda$  in the GR2-M models. We test the GR2-M model on the cooperative navigation game; empirically, the test selection of  $\lambda = 1.5$  on both GR2-M3 and GR2-M2 would lead to best performance. We therefore use  $\lambda = 1.5$  in the experiment section in the main paper.

**Choice of Reward Function in Keynes Beauty Contest.** One sensible finding from human players suggests that when prize of winning gets higher, people tend to use more steps of reasoning and they may think others will think harder too. We simulate a similar scenario by reward shaping. We consider two reward schemes of absolute difference and squared absolute difference. Interestingly, we find similar phenomenon in Fig. 7b that the amplified reward can significantly speed up the convergence for GR2-L methods.