Generalized Recursive Reasoning for Bounded Rationality in Multi-Agent Interactions

Anonymous Author(s)
Affiliation
Address
email

Abstract

In real-world decision making, people's rationality can often be bounded by the tractability of the decision problem or by the cognitive limitations of individual's mind. In this paper, we consider computerized agents having bounded rationality toward their interactions with each other. We integrate generalized recursive reasoning (GR2) into multi-agent reinforcement learning (MARL), enabling agents to acknowledge their opponents' bounded rationality and to model the corresponding non-equilibrium behaviors. In particular, GR2 assumes a hierarchical level of sophistication on agents' reasoning capability that starts from level-0 non-strategic thinkers. Each agent in GR2 believes that his opponents are less sophisticated than him; agents with higher-level thinking take the best response to the opponents who are considered distributed over from level 0 to $k-1$. We prove in theory the existence of Perfect Bayesian Equilibrium on normal-form games as well as the convergence on two-player matrix games under the GR2 setting. Importantly, we formulate the multi-agent policy search problem into a hierarchy of nested single-agent policy search problems and derive the practical GR2 soft actor-critic algorithm. In three types of experiments, our methods exhibit excellent convergent behaviors against strong MARL baselines which justify our theoretical findings.

1 Introduction

In people's decision making, rationality can often be compromised; it can be constrained by either the tractability 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* [1] serves as the alternative modeling basis by recognizing the cognitive limitations of each individuals, and the fact that people seek the optimal decision based on the information and the resource for reasoning that are available. One most-cited example that bounded rationality prescribes is Keynes Beauty Contest [2]. In a simpler version, all players are asked to pick one number from 0 to 100; eventually, the player whose guess is closest to $1/2$ of the average number 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 $1/2 * 50 = 25$ (level-1), and then if the other players think similarly as me, I should guess no more than $1/2 * 25 = 13$ (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 findings in psychology [3] which suggests that most human players would choose between 13 and 25, and people with different education background could guess markedly different numbers. Evidently, not all players are able to behave in a perfectly rational manner in practice; different levels of reasoning capability lead to different guesses. Sticking to the equilibrium 0 is not guaranteed to win either.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

In multi-agent reinforcement learning (MARL), one common assumption is that all agents behave rationally [4] during their interactions. However, in practice it is hard to guarantee that all agents have the same level of sophistications in terms of their abilities in understanding and learning from each others. For instance, MARL agents could face independent learners [5], the joint-action learners [6], and the learners with a theory-of-mind module [7, 8]. The effectiveness of MARL models decreases especially when the opponents act irrationally [9]. Observations from the games on strategic thinking also suggests that assuming perfect rationality could harm behavior predictions [10–12]. Furthermore, it is not desirable to design agents that seek optimal behaviors only against rational opponents. Intelligent agents are supposed to be generalizable and show optimal behaviors to different types of adversarial opponents in the real-world applications. Such practical need can easily be found in self-driving cars [13] or video games design [14]. Therefore, it becomes critical for MARL models to incorporate the bounded rationality and enable agents to face the agents with different rationality.

In this work, we propose a novel learning protocol – *Generalized Recursive Reasoning* (GR2) – that recognizes agents' bounded rationality and their corresponding non-equilibrium behaviors. GR2 is inspired by cognitive hierarchy theory [15], assuming that agents could possess different levels of reasoning rationality during the interactions. It begins with level-0 (L0 for short) non-strategic thinkers who behave independently. L1 thinkers are more sophisticated; they model their opponents at L0 and then act correspondingly. With the growth of k, Lk agents think in an increasing order of sophistication, and take the best response to the lower-level opponents. We immerse the GR2 framework into MARL, and formulate the multi-agent policy search problem into a hierarchy of nested single-agent policy search problems. We prove the existence of Perfect Bayesian Equilibrium in the GR2 setting on normal-form games and the convergence on the two-player matrix game. Importantly, we develop the practical GR2 Soft Actor-Critic algorithm. Our methods are evaluated against multiple strong MARL baselines on Keynes Beauty Contest, matrix games, and cooperative navigation. Results justify the effective convergent behaviors of GR2 methods. In particular, GR2 methods are able to converge in the Beauty Contest game and matrix games where traditional level-0 methods fails. When facing two equilibriums, GR2 agents are able to pick the optimal one.

2 Related Work

Assigning agents with different levels of reasoning capability – *aka* the level-k model – was first introduced by Stahl in 1993. In the level-k model, agents anchor their beliefs through thought-experiments; they seek for the best response in an iterated chain style to the kth level, i.e., Lk agents believe its opponents are at the lower level $k-1$ and act with the best response. It is expected that agents would exhibit more complex behaviors as the recursion level increases [12]. Interestingly, human beings are shown to reason between 1 - 2 levels of recursions [15, 17] – "I consider how you would believe about how I believe, and then take the optimal responsive action" (L2 thinking) – which also explains the empirical findings in Keynes Beauty Contest (see the previous section).

Nonetheless, the level-k models are shown to exhibit stereotype bias due to the limited concentration on only one level below [18]. Camerer [15] extended the level-k model by making the Lk best respond to not only the level $k-1$, but a mixture of other lower levels as well, and named it cognitive hierarchy model (CHM). The mixture of the lower hierarchy is usually approximated by a discrete probability distribution. CHM presents distinct advantages over the level-k models on making robust behavior predictions [11]. Inspired by this, GR2 incorporates the idea of cognitive hierarchy to represent agent's different reasoning capabilities and embed it into the MARL framework.

Modeling the opponents in a recursive manner can be regarded as a special type of opponent modeling [4]. Recently, studies on Theory of Mind (ToM) [7, 8, 19] explicitly model the agent's belief on opponents' mental states in the reinforcement learning setting. The I-POMDP framework focuses on the planning tasks [20]. GR2 is different in that it incorporates a hierarchical structure of thinking in modeling the opponents rather than one single opponent model or state. Our method is most related to the probabilistic recursive reasoning (PR2) model [21] that also implements the idea of recursive reasoning for MARL tasks. However, PR2 only explores the level-1 reasoning; here we extend the recursion depth to an arbitrary number and design the learning solution so that agents can find the best response to a mixture of different agents. By incorporating the bounded rationality, we expect the GR2 methods to generalize well and show optimal behaviors to different type of agents, which we believe is a critical property for modern multi-agent AI applications.

We immerse GR2 into the RL framework through a probabilistic approach. It has been spelled out that there exists equivalence between solving the optimal policy in maximum entropy RL and the

inference problem on probabilistic graphical models [22, 23]. Based on the equivalence, the soft Q-learning [24] and soft actor-critic [25] methods were induced; very recently, they have been further generalized to the multi-agent scenarios [26–28]. In this work, we embed the recursive reasoning protocol into the multi-agent soft learning framework, and develop the GR2 soft actor-critic algorithm.

3 Preliminaries

A stochastic game [29] , denoted by $(\mathcal{S}, \mathcal{A}^1, \ldots, \mathcal{A}^n, r, \ldots, r^n, p, \gamma)$, is a natural framework to extend the Markov decision process to n agent, where $\mathcal{S}$ represents the state space, p is the transition probability of the environment, γ is the discount factor for future rewards, $\mathcal{A}^i$ and $r^i(s, a^i, a^{-i})$ are the action space and the reward function for agent $i \in \{1, \ldots, n\}$ respectively. Agent i chooses its action $a^i \in \mathcal{A}^i$ based on its policy function $\pi^i_{\theta^i}(a^i|s)$ parameterized by θ^i, and $a^{-i} = (a^j)_{j\neq i}$ represents its opponents' behaviors. Each agent tries to find the optimal policy that maximizes its own expected cumulative reward: $\max_{\pi^i_\theta} \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\pi_\theta, p}\left[r^i\left(s_t, a^i_t, a^{-i}_t\right)|s_0, \pi_\theta\right]$, with (a^i_t, a^{-i}_t) sampled from joint policy $\pi_\theta = (\pi^i_{\theta^i}, \pi^{-i}_{\theta^{-i}})$. We define the trajectory by $\tau = \left[(s_1, a^i_1, a^{-i}_1, r^i_1), \ldots, (s_T, a^i_T, a^{-i}_T, r^i_T)\right]$.

Searching the optimal policy for each agent can be treated as the inference problem on a probabilistic graphical model (Fig. 1) where a binary optimality variable $\mathcal{O}^i$ is introduced to represent how optimal agent i is performing at time t, i.e., $P(\mathcal{O}^i_t = 1|s_t, a^i_t, a^{-i}_t) \propto \exp(r^i(s_t, a^i_t, a^{-i}_t))$. With $\mathcal{O}^i_t$ defined, the optimal policy for agent i therefore becomes $\pi^i_* = p(a^i_t|s_t, a^{-i}_t, \mathcal{O}^{\{i,-i\}}_{t:T} = 1)$. However, unlike the single-agent case, now each agent has to *best respond* to the opponent's actions besides s_t.

Given the value function $V^i(s; \pi^i_{\theta^i}, \pi^{-i}_{\theta^{-i}}) = \sum_{t=1}^{\infty} \gamma^t \mathbb{E}_{\pi_\theta, p}\left[r^i\left(s_t, a^i_t, a^{-i}_t\right)|s_1, \pi_\theta\right]$ defined at state s, a **best response** to the opponent policy $\pi^{-i}_{\theta^{-i}}$ is the policy π^i_* s.t. $V^i(s; \pi^i_*, \pi^{-i}_{\theta^{-i}}) \geq V^i(s; \pi^i_{\theta^i}, \pi^{-i}_{\theta^{-i}})$ for all valid $\pi^i_{\theta^i}$. We mostly drop the value for $\mathcal{O}^{\{i,-i\}}_{t:T}$ since the optimum is our focus. When all the agents act in their best responses, $(\pi^1_*, \ldots, \pi^n_*)$ forms a Nash Equilibrium [30].

4 Generalized Recursive Reasoning

To embed the GR2 protocol into the MARL framework, we start from the fundamentals of multi-agent soft learning and level-1 probabilistic recursive reasoning (PR2). We then incorporate the solution of level-k reasoning solution (GR2-L) and finally propose the generalized GR2 model (GR2-M).

4.1 Multi-agent Soft Learning

In the single-agent case, optimal policy search in the case of stochastic dynamics turns out to be equivalent to solving trajectory optimization problem via variational inference [22]; this leads to the maximum-entropy variant of the policy iteration methods and induces the soft Q-learning/actor-critic algorithm [24, 25]. Very recently, it has been shown that such equivalence also holds in the multi-agent cooperative games [21, 26, 28], solving $p(a^i_t|s_t, a^{-i}_t, \mathcal{O}^{\{i,-i\}}_{t:T} = 1)$ is in fact equal to maximizing the $p(\tau|\mathcal{O}_{1:T})$ via approximate inference. Given a trajectory τ, its posterior distribution is $p(\tau|\mathcal{O}_{1:T}) \propto \left[p\left(s_1\right)\prod_{t=1}^{T} p\left(s_{t+1}|s_t, a^i_t, a^{-i}_t\right)\right]\exp\left(\sum_{t=1}^{T} r^i\left(s_t, a_t, a^{-i}_t\right)\right)$. With the proposal distribution $\hat{p}(\tau) = \left[q\left(s_1\right)\prod_{t=1}^{T} q\left(s_{t+1}|s_t, a^i_t, a^{-i}_t\right)\right]\pi_\theta\left(a^i_t, a^{-i}_t|s_t\right)$, we have the objective as

$$\mathcal{J}(\pi_\theta) = -\mathcal{D}_{\mathrm{KL}}\big(\hat{p}(\tau)||p(\tau|\mathcal{O}_{1:T})\big) = \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\hat{p}(\tau)}\left[r^i\left(s_t, a^i_t, a^{-i}_t\right) + \mathcal{H}\left(\pi_\theta\left(a^i_t, a^{-i}_t|s_t\right)\right)\right]. \quad (1)$$

On top of the reward maximization, Eq .1 has one additional entropy term $\mathcal{H}(\cdot)$ on the joint policy.

A variant of maximum-entropy policy iteration can still be applied to maximize $\mathcal{J}(\pi_\theta)$. For policy evaluation, Bellman expectation equation now holds on the *soft* value function $V^i(s) = \mathbb{E}_{a^i_t, a^{-i}_t}\left[Q^i(s_t, a^i_t, a^{-i}_t) - \log(\pi_\theta(a^i_t, a^{-i}_t|s_t))\right]$, with the new Bellman operator $\mathcal{T}^\pi$ defined by $\mathcal{T}^\pi Q^i(s_t, a^i_t, a^{-i}_t) \triangleq r^i(s_t, a^i_t, a^{-i}_t) + \gamma\mathbb{E}_{s_{t+1}}\left[\log\int\int \exp(Q^i(s_{t+1}, a^i_{t+1}, a^{-i}_{t+1}))\,\mathrm{d}a^i_{t+1}\,\mathrm{d}a^{-i}_{t+1}\right]$. Compared to the $\max$ operation in the normal Q-learning, $\mathcal{T}^\pi$ is *soft* because $\log\int\int \exp\left(Q(s_t, a_t, a^{-i}_t)\right)\mathrm{d}a_t\,\mathrm{d}a^{-i}_t \approx \max_{a_t, a^{-i}_t} Q\left(s_t, a_t, a^{-i}_t\right)$. 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; however, the exact opponent policy is usually unobservable. To solve this problem, different efforts have been put on factorizing the joint policy [21, 26, 28].

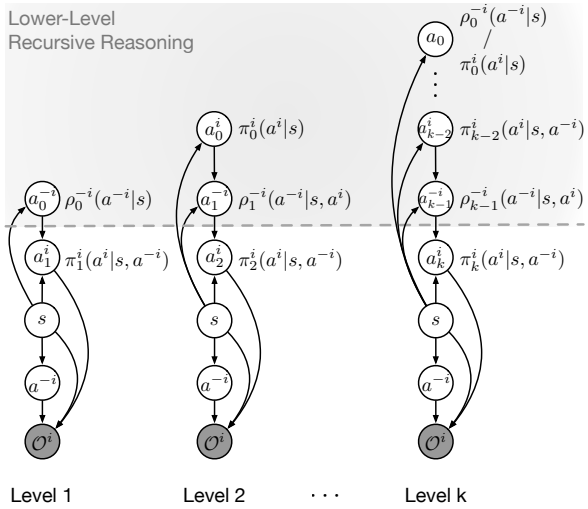


Figure 1: Graphical model of the level-k reasoning model. Subfix of a_* stands for the level of thinking not the timestep. The opponent policies are approximated by ρ^{-i}. The omitted level-0 model considers opponents fully randomized. Agent i rolls out the recursive reasoning about opponents in its mind (grey area). In the recursion, agents with higher-level beliefs take the best response to the lower-level thinkers' actions. Higher-level models would conduct all the computations that the lower-level models have done, e.g. level-2 model contains level-1 model by integrating out $\pi_0^i(a^i|s)$.

4.2 Probabilistic Recursive Reasoning

PR2 [21] adopted the so-called *correlated factorization* on the joint policy, i.e., $\pi(a^i, a^{-i}|s) = \pi^i(a^i|s)\pi^{-i}(a^{-i}|s, a^i)$. PR2 is regarded as a level-1 model because $\pi^{-i}(a^{-i}|s, a^i)$ stands for the opponent's consideration of $a^i \sim \pi^i(a^i|s)$. As π^{-i} may not be directly observable, PR2 approximates the actual opponent conditional policy π^{-i} via a best-fit model $\rho^{-i}_{\phi^{-i}}$ parameterized by ϕ^{-i}. By adopting $\pi_\theta(a^i, a^{-i}|s) = \pi^i_{\theta^i}(a^i|s)\rho^{-i}_{\phi^{-i}}(a^{-i}|s, a^i)$ in $\hat{p}(\tau)$ in the approximate inference on Eq. 1, the optimal opponent policy can be solved as (Theorem 1 in [21])

$$\rho^{-i}_{\phi^{-i}}(a^{-i}|s, a^i) \propto \exp\left(Q^i_{\pi_\theta}(s, a^i, a^{-i}) - Q^i_{\pi_\theta}(s, a^i)\right). \tag{2}$$

Based on the optimal opponent model in Eq. 2, agent i can learn the best response policy by considering all possible opponent agents' actions: $Q^i(s, a^i) = \int_{a^{-i}} \rho^{-i}_{\phi^{-i}}(a^{-i}|s, a^i)Q^i(s, a^i, a^{-i})\,\mathrm{d}a^{-i}$, and then improve its own policy towards this direction,

$$\pi_{\text{new}} = \arg\min_{\pi'} \mathcal{D}_{\text{KL}}\left(\pi'(\cdot|s_t)\middle\|\frac{\exp\left(Q^{i,\pi_{\text{old}}}(s_t, \cdot)\right)}{Z^{\pi_{\text{old}}}(s_t)}\right). \tag{3}$$

Interestingly, Tian et al. adopted the way of factorization by $\pi(a^i, a^{-i}|s) = \pi^{-i}(a^{-i}|s)\pi^i(a^i|s, a^{-i})$, and Grau-Moya et al. adopted $\pi(a^i, a^{-i}|s) = \pi^{-i}(a^{-i}|s)\pi^i(a^i|s)$; they both derived similar results.

4.3 Level-k Recursive Reasoning (GR2-L)

Previous work has limited the exploration to level-1 model for describing agents's reasoning ability. Now we extend 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 reacts with the best response. The level-k policy can be constructed by integrating over all possible best responses from lower-level policies

$$\pi^i_k(a^i_k|s) \propto \int_{a^{-i}_{k-1}} \left\{\pi^i_k(a^i_k|s, a^{-i}_{k-1}) \cdot \underbrace{\int_{a^i_{k-2}} \left[\rho^{-i}_{k-1}(a^{-i}_{k-1}|s, a^i_{k-2})\pi^i_{k-2}(a^i_{k-2}|s)\right]\mathrm{d}a^i_{k-2}}_{\text{opponents of level k-1 best responds to agent } i \text{ of level k-2}}\right\}\mathrm{d}a^{-i}_{k-1}. \tag{4}$$

When the levels of reasoning develop, we could think of the marginal policies $\pi^i_{k-2}(a^i|s)$ from lower levels as the *prior* and the conditional policies $\rho^{-i}_{k-1}(a^{-i}|s, a^i)$, $\pi^i_k(a^i|s, a^{-i})$ as the *posterior* so that decisions are taken in a sequential manner. From agent i's perspective, it believes that the opponents will take the best response to its own fictitious action a^i_{k-2} that are two levels below: $\rho^{-i}_{k-1}(a^{-i}_{k-1}|s) = \int \rho^{-i}_{k-1}(a^{-i}_{k-1}|s, a^i_{k-2})\pi^i_{k-2}(a^i_{k-2}|s)\,\mathrm{d}a^i_{k-2}$, and π^i_{k-2} can be further expanded by recursively using Eq. 4 until meeting π_0 that is usually assumed uniformly distributed. As such, it maps multi-agent planning problem into a hierarchy of nested single-agent planning problems. Considering the computational feasibility, we assume that each of the agents adopts the same functional form for its policy throughout different recursive levels: $\pi^i_{\theta^k} = \pi^i_{\theta^{k+2}}, \rho^{-i}_{\theta^k} = \rho^{-i}_{\theta^{k+2}}$.

4.4 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.

Since more computational resources are required with increasing k, e.g., human beings show limited amount of working memory in strategic thinkings [31], it is reasonable to restrict the reasoning to the cases where fewer agents are willing to conduct the reasoning beyond k when k grows large.

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.*

This assumption suggests that even though the lower-k agents could maintain an inaccurate belief on the probability distribution of the other agents' recursive depths; however, as k increases, the deviation between level-k agent's belief and the actual distribution profile of lower k will shrink. The motivation of this 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 model the proportion of recursive depth by the Poisson distribution $f(k) = \frac{e^{-\lambda}\lambda^k}{k!}$, where the mean, also the variance, of the distribution is λ. A nice property of Poisson is that $f(k)/f(k-n)$ is inversely proportional to k^n for $1 \leq n < k$, which also satisfies our previous proposal that high-level thinkers should have no incentives to think even harder. Interestingly, a study on humans suggests that on average people tend to act with $\lambda \approx 1.5$ [15]. With the mixture of hierarchy modeled by Poisson, we can now mix all k levels' thinkings $\{\hat{\pi}_k^i\}$ into one perception, and build each agent's belief on its opponents who could stay at any possible lower recursive levels by

$$\pi_k^{i,\lambda}(a_k^i|s, a_{0:k-1}^{-i}) := \frac{e^{-\lambda}}{Z}\left(\frac{\lambda^0}{0!}\hat{\pi}_0^i(a_0^i|s) + \frac{\lambda^1}{1!}\hat{\pi}_1^i(a_1^i|s, a_0^{-i}) + \cdots + \frac{\lambda^k}{k!}\hat{\pi}_k^i(a_k^i|s, a_{0:k-1}^{-i})\right), \tag{5}$$

where the normalization term $Z = \sum_{n=1}^{k} e^{-\lambda}\lambda^n/n!$. In the reinforcement learning setting, λ can be set as a hyper-parameter, similar to the idea of TD-λ [32] (note that the λ have different meanings).

GR2-L is in fact a special case of GR2-M. When the mixture of depth is Poisson distributed, we have $f(k-1)/f(k-2) = \lambda/(k-1)$; the model will put heavy weights on the $k-1$ level if $\lambda \gg k$; that is to say, a level-k agent will act as if all the opponents are reasoning at level $k-1$.

4.5 Theoretical Convergence

Recursive reasoning is essentially to let each agent take the best response to its opponents with the opponents' actions being the best response to the agent's best-responded action. A natural question to ask is then under the GR2 setting, does the equilibrium ever exist? If so, will GR2 ever converge?

We show that on two-player games, the learning dynamics of GR2 is asymptotically stable in the sense of Lyapunov. In addition, the dynamic game induced by GR2 has Perfect Bayesian Equilibrium.

Theorem 1. *In two-player two-action games, if these 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. See Appendix B. ■

Theorem 2. *GR2 strategies extend a norm-form game into extensive-form game, and there exists a Perfect Bayesian Equilibrium in that game.*

Proof. See Appendix C. ■

Apart from the main theorems, we could also have two interesting propositions.

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 D. ■

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

Proof. By considering all possible actions from lower-level agents, higher-level thinkers will always conduct at least the same amount of computations as the lower-level thinkers as shown in Fig. 1. ■

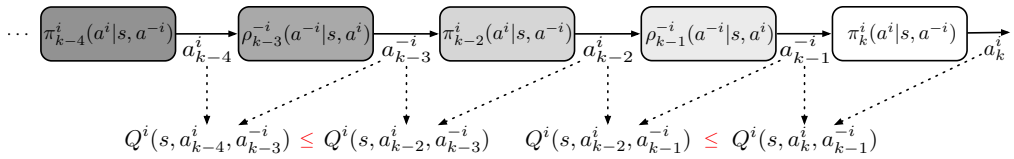


Figure 2: Inter-level policy improvement. Higher-level policy weakly dominate lower-level policies.

5 GR2 Reinforcement Learning in Practice

We now propose the practical GR2 soft actor-critic algorithm. To make the computation feasible for high-level reasoning, we adopt the deterministic policy during the recursive rollouts; to mitigate the loss of generality, we maintain the inter-level policy improvement through an auxiliary objective.

5.1 GR2 Soft Actor-Critic Algorithm

For policy evaluation, each agent rolls out the recursive reasoning policies to level k by either Eq. 4 or Eq. 5. We can train the parameter ω^i of the joint soft Q-function via minimizing the soft Bellman residual

$$J_{Q^i}(\omega^i) = \mathbb{E}_{(s,a^i,a^{-i})\sim\mathcal{D}^i}\left[\frac{1}{2}\left(Q^i_{\omega^i}(s,a^i,a^{-i}) - \hat{Q}^i(s,a^i,a^{-i})\right)^2\right], \tag{6}$$

where s' is next state, $\mathcal{D}^i$ is the replay buffer storing trajectories, and the target is $\hat{Q}^i(s,a^i,a^{-i}) = r^i(s,a^i,a^{-i}) + \gamma\mathbb{E}_{s'\sim p}\left[V^i(s')\right]$. For $V^i(s')$, since agent i has no access to the opponent policy $\pi^{-i}_{\theta^{-i}}$, we instead compute the soft $Q^i(s,a^i)$ by marginalizing the joint Q-function via the estimated opponent model $\rho^{-i}_{\phi^{-i}}$, that is, $Q^i(s,a^i) = \log\int \rho^{-i}_{\phi^{-i}}(a^{-i}|s,a^i)\exp\left(Q^i(s,a^i,a^{-i})\right)\mathrm{d}a^{-i}$, then the value function of the level-k policy $\pi^i_k(a^i|s)$ comes as $V^i(s) = \mathbb{E}_{a^i\sim\pi^i_k}\left[Q^i(s,a^i) - \log\pi^i_k(a^i|s)\right]$.

Note that $\rho^{-i}_{\phi^{-i}}$ 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 ρ^{-i} still follows Eq. 2 in the multi-agent soft learning setting. We can therefore update ϕ^{-i} by minimizing the KL of

$$J_{\rho^{-i}}(\phi^i) = \mathcal{D}_{\mathrm{KL}}\left[\rho^{-i}_{\phi^{-i}}(a^{-i}|s,a^i)\middle\|\exp\left(Q^i_{\omega^i}(s,a^i,a^{-i}) - Q^i_{\omega^i}(s,a^i)\right)\right]. \tag{7}$$

In practice, we maintain two approximated Q-functions of $Q^i_{\omega^i}(s,a^i,a^{-i})$ and $Q^i_{\omega^i}(s,a^i)$ separately for robust training, and the gradient of ϕ^{-i} is computed via SVGD [33] for continuous control tasks.

Finally, the level-k policy parameter θ^i can be learned by improving towards the current Q-function $Q^i_{\omega^i}(s,a^i)$ similar to Eq. 3. By applying the reparameterization trick $a^i = f_{\theta^i}(\epsilon;s)$, we have

$$J_{\pi^i_k}(\theta^i) = \mathbb{E}_{s\sim\mathcal{D}^i, a^i_k\sim\pi^i_{\theta^i,k},\epsilon\sim\mathcal{N}}\left[\log\pi^i_{\theta^i,k}(f_{\theta^i}(\epsilon;s)|s) - Q^i_{\omega^i}(s,f_{\theta^i}(\epsilon;s))\right]. \tag{8}$$

Note that as we turn the multi-agent problem into nested single-agent problems, and reuse the policy networks across different levels of the recursive rollouts, we would expect the gradient of $\partial\pi^i_{\theta^i,k}/\partial\theta^i$ propagated back from all different higher levels, similar to back-propagation through time in RNN.

5.2 Deterministic Strategy as Approximated Best Response

Considering the computational feasibility, we approximate the best response in the form of deterministic strategy throughout the recursive rollouts, for example, using the mean in the case of Gaussian policy. As the reasonings process involves iterated usages of $\pi^i(a^i|s,a^{-i})$ and $\rho^{-i}(a^{-i}|s,a^i)$, should they be stochastic policies, the cost of integrating over actions from all possible lower-level agents would be unsustainable when k increases. Besides, the reasoning process is affected by the stochasticity of the environment, the variance will be further amplified by the stochastic policies.

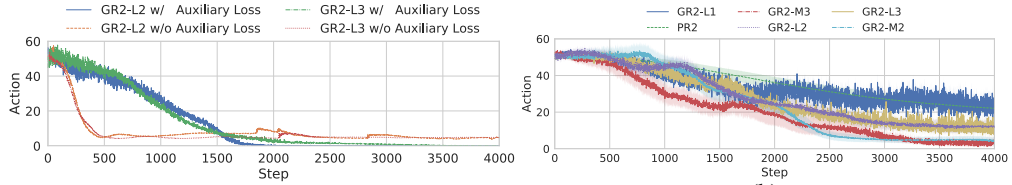
To mitigate the weakening effect of applying deterministic strategy during the recursive rollouts, we enforce the policy improvement between different levels. It is also pointed out by Corollary. 1 that higher-level policy should perform at least as good as lower-level policy against the opponents. To maintain this property, we introduce an auxiliary loss on top of Eq. 8 (see Fig. 2),

$$J_{\pi^i_{\tilde{k}}}(\theta^i) = \mathbb{E}_{s\sim\mathcal{D}^i,(a^i_{\tilde{k}},a^{-i}_{\tilde{k}})\sim\pi^i_{\theta^i},\rho^{-i}_{\phi^{-i}}}\left[Q^i(s,a^i_{\tilde{k}},a^{-i}_{\tilde{k}-1}) - Q^i(s,a^i_{\tilde{k}-2},a^{-i}_{\tilde{k}-1})\right], \tilde{k}\ge 2. \tag{9}$$

Note that the highest-level agent policy π^i_k is still stochastic by design; it receives the reward directly from the environment then propagates the feedback to itself at lower levels as we reuse the policy throughout different recursive levels for π^i_k and ρ^i_k respectively. We summarize the pseudo-code of the whole algorithm in Appendix A.

Table 1: The converged 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	**0.0**	**0.0**	**0.0**	4.4	7.1	10.6	8.7	8.3	18.6
$p = 0.7, n = 10$	0.0	**0.0**	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



(a) (b)

Figure 3: Beauty Contest of $p = 0.7$, $n = 2$. (a) Learning curves w/ or w/o the auxiliary loss of Eq. 9. (b) Average learning curves of each GR2 method against the other six baselines (round-robin style).

6 Experiments

We start the experiments on Keynes Beauty Contest to investigate the convergence property in self-plays. We then move on to the matrix games with non-trivial equilibria where common level-0 gradient-based methods fail to find the optimal solution. Finally, we test on the high-dimensional cooperative navigation task that requires effective coordinations among agents to avoid collisions.

We compare the GR2 methods with six MARL baselines including PR2 [21], MASQL [27], MADDPG [34] and independent learner via DDPG [35]. We also include the opponent modeling [36] by augmenting DDPG with an opponent module (DDPG-OM) that predicts the opponent behaviors in the future states, and a Theory of Mind model [7] 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 $k = \{1, 2, 3\}, \lambda = 1.5$. Due to the space limit, we leave the detailed hyper-parameter settings and the ablation study in *Appendix E*. All results are reported with 6 random seeds.

6.1 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 to the winner guess.

Higher level thinking helps humans to get close to the equilibrium in Keynes Beauty Contest. 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. At first glance, we can tell that the GR2-L algorithms can effectively approach the equilibrium while the other baselines struggle to reach. The only exception is the case of $(p = 1.1, n = 10)$ which we think is because of the saturated gradient from the reward.

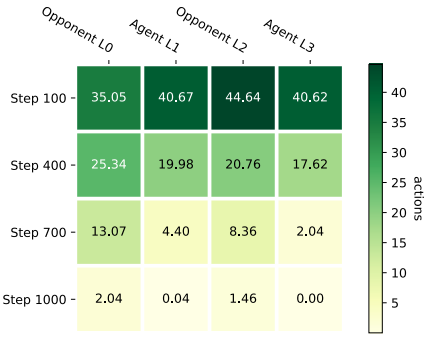
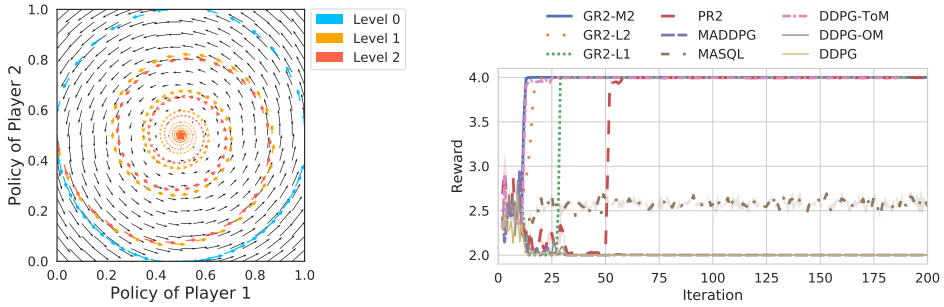


Figure 4: The reasoning path during the training of Level-3 policy in the Beauty Contest $(n = 2, p = 0.7)$.

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 reasoning path of GR2-L3 in Fig. 4. 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 $(20.76 \neq 19.98 * 1.1)$, 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. Interestingly, we also 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 one on one (round-robin style) and present the averaged performance in Fig. 3b. GR2-M models in general out-perform the rest, which is expected since a mixture model is supposed to be more generalizable to model a wide range of different agents.

Finally, we justify the necessity of adopting the auxiliary loss of Eq. 9 by Fig. 3a. As we simplify the reasoning rollouts by using deterministic policies, we believe adding the auxiliary loss in the objective can effectively mitigate the negative issue of policy expressiveness and guide the joint Q-function to a better direction to finally improve the policy π_k^i, which leads to a better decision.



(a) (b)

Figure 5: (a) Learning dynamics on Rotational Game. (b) Average reward on the Stag Hunt.

6.2 The Matrix Games

We further evaluate the GR2 methods on two matrix games: Rotational Game (RG) [5] and Stag Hunt (SH). The reward matrix of RG is $R_{\text{RG}} = \begin{bmatrix} 0,3 & 3,2 \\ 1,0 & 2,1 \end{bmatrix}$, with the only equilibria at $(0.5, 0.5)$. In SH, the reward matrix is $R_{\text{SH}} = \begin{bmatrix} 4,4 & 1,3 \\ 3,1 & 2,2 \end{bmatrix}$. SH comes with two equilibria (S, S) that is Pareto optimal and the (P, P) that is deficient. We define the state at time t to be $s_t = (a^1_{t-1}, a^2_{t-1})$.

In RG, we examine the effectiveness that level-k policies can converge to the equilibrium but level-0 cannot. We plot the gradient dynamics of RG in Fig. 5a. 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 in the proof of Theorem 1.

In order to further evaluate the superiority of level-k policies, which could achieve a better equilibrium when two equilibria exist, we present Fig. 5b that compares the average reward on SH. 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 equilibriums can saturate the value function. On the convergence speed, as expected, higher-level models are faster, and GR2-L mixture models are faster than GR2-L models.

6.3 Cooperative Navigation

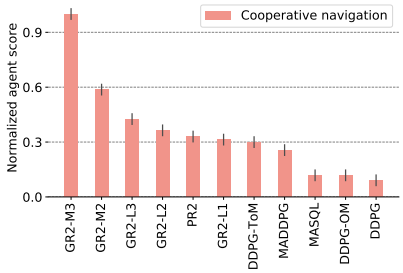


Figure 6: Performance comparison.

We test the GR2 methods in more complexed Particle World Environments [34] for the high-dimensional control task of *Cooperative Navigation* with 3 agents and 3 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. 6 where we report the minimax-normalized score. We notice that the GR2 methods achieve critical advantages over traditional baselines, 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 to the fullest extent.

7 Conclusion

Acknowledging the bounded rationality is critical for many real-world AI applications. We propose a new solution to multi-agent RL – generalized recursive reasoning (GR2) – that enables agents to handle the opponents with different levels of rationality. In GR2, each agent believes that the opponents are less sophisticated than him and then takes the corresponding best response. We prove in theory the existence of Perfect Bayesian Equilibrium in normal-form games as well as the convergence on matrix games under the GR2 setting. Importantly, we formulate the multi-agent problem into a hierarchy of nested single-agent policy search problems and derive the practical GR2 soft actor-critic algorithm. Experimental results justify the effective convergence of GR2 methods.

Acknowledgements

We thank Dr. Haitham Bou Ammar and Dr. Aivar Sootla for comments that greatly improved the manuscript. Ying Wen was partially funded by MediaGamma Ltd.

References

[1] Herbert A Simon. Theories of bounded rationality. *Decision and organization*, 1(1):161–176, 1972.

[2] J. M. Keynes. *The General Theory of Employment, Interest and Money*. Macmillan, 1936. 14th edition, 1973.

[3] 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.

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

[5] Michael Bowling and Manuela Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136(2):215–250, 2002.

[6] Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998:746–752, 1998.

[7] Neil C Rabinowitz, Frank Perbet, H Francis Song, Chiyuan Zhang, SM Eslami, and Matthew Botvinick. Machine theory of mind. *arXiv preprint arXiv:1802.07740*, 2018.

[8] Michael Shum, Max Kleiman-Weiner, Michael L Littman, and Joshua B Tenenbaum. Theory of minds: Understanding behavior in groups through inverse planning. *arXiv preprint arXiv:1901.06085*, 2019.

[9] Yoav Shoham, Rob Powers, and Trond Grenager. Multi-agent reinforcement learning: a critical survey. Technical report, Technical report, Stanford University, 2003.

[10] Herve Moulin. *Game theory for the social sciences*. NYU press, 1986.

[11] Vincent P Crawford, Miguel A Costa-Gomes, and Nagore Iriberri. Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1):5–62, 2013.

[12] Carlos Gracia-Lázaro, Luis Mario Floría, and Yamir Moreno. Cognitive hierarchy theory and two-person games. *Games*, 8(1):1, 2017.

[13] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

[14] Robin Hunicke. The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pages 429–433. ACM, 2005.

[15] Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

[16] Dale O Stahl. Evolution of smartn players. *Games and Economic Behavior*, 5(4):604–617, 1993.

[17] Volker Benndorf, Dorothea Kübler, and Hans-Theo Normann. Depth of reasoning and information revelation: An experiment on the distribution of k-levels. *International Game Theory Review*, 19(04):1750021, 2017.

[18] Juin-Kuan Chong, Teck-Hua Ho, and Colin Camerer. A generalized cognitive hierarchy model of games. *Games and Economic Behavior*, 99:257–274, 2016.

[19] Alvin I Goldman et al. Theory of mind. *The Oxford handbook of philosophy of cognitive science*, pages 402–424, 2012.

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

[21] Ying Wen, Yaodong Yang, Rui Luo, Jun Wang, and Wei Pan. Probabilistic recursive reasoning for multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rkl6As0cF7.

[22] Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

[23] Marc Peter Deisenroth, Gerhard Neumann, Jan Peters, et al. A survey on policy search for robotics. *Foundations and Trends® in Robotics*, 2(1–2):1–142, 2013.

[24] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. *arXiv preprint arXiv:1702.08165*, 2017.

[25] 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.

[26] Jordi Grau-Moya, Felix Leibfried, and Haitham Bou-Ammar. Balancing two-player stochastic games with soft q-learning. *IJCAI*, 2018.

[27] Ermo Wei, Drew Wicke, David Freelan, and Sean Luke. Multiagent soft q-learning. *AAAI*, 2018.

[28] Zheng Tian, Ying Wen, Zhicheng Gong, Faiz Punakkath, Shihao Zou, and Jun Wang. A regularized opponent model with maximum entropy objective. *International Joint Conferences on Artificial Intelligence*, 2019.

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

[30] John F Nash et al. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

[31] 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.

[32] Gerald Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

[33] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2378–2386, 2016.

[34] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pages 6379–6390, 2017.

[35] 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.

[36] He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. Opponent modeling in deep reinforcement learning. In *International Conference on Machine Learning*, pages 1804–1813, 2016.

[37] Wassim M Haddad and VijaySekhar Chellaboina. *Nonlinear dynamical systems and control: a Lyapunov-based approach*. Princeton University Press, 2011.

[38] Horacio J Marquez. *Nonlinear control systems: analysis and design*, volume 1. Wiley-Interscience Hoboken, 2003.

[39] Michael Bowling and Manuela Veloso. Convergence of gradient dynamics with a variable learning rate. In *ICML*, pages 27–34, 2001.

[40] Chongjie Zhang and Victor Lesser. Multi-agent learning with policy prediction. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[41] Adrien Taylor, Bryan Van Scoy, and Laurent Lessard. Lyapunov functions for first-order methods: Tight automated convergence guarantees. *arXiv preprint arXiv:1803.06073*, 2018.

[42] Sherief Abdallah and Victor Lesser. A multiagent reinforcement learning algorithm with non-linear dynamics. *Journal of Artificial Intelligence Research*, 33:521–549, 2008.

[43] Dan Levin and Luyao Zhang. Bridging level-k to nash equilibrium. *Available at SSRN 2934696*, 2019.

[44] David M Kreps and Robert Wilson. Sequential equilibria. *Econometrica: Journal of the Econometric Society*, pages 863–894, 1982.

[45] SK CHAKRABARTI and I TOPOLYAN. A direct proof of the existence of sequential equilibrium and a backward induction characterization. 2011.

Appendix

A Generalized Recursive Reasoning Algorithm

Algorithm 1 GR2 Soft Actor-Critic Algorithm

Set hyper-parameter λ, k and ψ (learning rates). Initialize $\theta^i, \phi^{-i}, \omega^i$ for each agent i.
Assign target parameters: $\bar{\omega}^i \leftarrow \omega^i$. $\mathcal{D}^i \leftarrow$ empty replay buffer for each agent i.
for each episode **do**
  for each step t **do**
    For each agent i, sample action a^i according to $\pi^i_{\theta^i,k}(s)$ in Eq. 4 or $\pi^{i,\lambda}_{\theta^i,k}(s)$ in Eq. 5.
    Sample next state: $s' \sim p\left(s'|s, a^i, a^{-i}\right)$.
    Add the tuple $(s, a^i, a^{-i}, r^i, s')$ to $\mathcal{D}^i$.
    for each agent i **do**
      Sample $\{(s'_j, a^i_j, a^{-i}_j, r^i_j, s'_j)\}^M_{j=0} \sim \mathcal{D}^i$.
      Roll out policy from level $0 \rightarrow k$ to get $a^{i\prime}_j$, with each level take the greedy best response.
      Record inter-level results $(a^{i\prime}_{j,k}, a^{-i\prime}_{j,k-1}, \cdots)$ for auxiliary objective in Eq. 9.
      Sample $a^{-i\prime}_j \sim \rho^{-i}_{\phi^{-i}}(\cdot|s'_j, a^{i\prime}_j)$ for each $a^{i\prime}_j$, s'_j.
      $\omega^i \leftarrow \omega^i - \psi_{Q^i} \hat{\nabla}_{\omega^i} J_{Q^i}(\omega^i)$.
      $\theta^i \leftarrow \theta^i - \psi_{\pi^i} \hat{\nabla}_{\theta^i} \left(J_{\pi^i_k}(\theta^i) + J_{\pi^i_{\tilde{k}}}(\theta^i)\right)$.
      $\phi^{-i} \leftarrow \phi^{-i} - \psi_{\rho^{-i}} \hat{\nabla}_{\phi^{-i}} J_{\rho^{-i}}(\phi^{-i})$.
    end for
    Update target network for each agent i:

$$\bar{\omega}^i \leftarrow \psi_{\text{target}} \omega^i + (1 - \psi_{\text{target}}) \bar{\omega}^i.$$

  end for
end for

B Proof of Theorem 1

Theorem 1. *In two-player two-action games, if these 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 [37] 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 1. *(Lyapunov Function.) Give a function $F(x, y)$ be continuously differentiable in a neighborhood σ 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 $(x, y) \in \sigma \backslash (0, 0)$;*
2. *(zero at fixed-point) $F(0, 0) = 0$;*
3. *(decreasing) $\frac{\mathrm{d}F}{\mathrm{d}t} \leq 0$ for all $(x, y) \in \sigma$.*

Definition 2. *(Lyapunov Asymptotic Stability.) For an autonomous system, if there is a Lyapunov function $F(x, y)$ with a negative definite derivative $\frac{\mathrm{d}F}{\mathrm{d}t} < 0$ (strictly negative, negative definite LaSalle's invariance principle [37]) for all $(x, y) \in \sigma \backslash (0, 0)$, then the equilibrium point $(x, y) = (0, 0)$ of the system is asymptotically stable [37, 38].*

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 α 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 β to be the strategy for column player. With a joint strategy (α, β), the expected payoffs of players are:

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

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 (α^*, β^*). This point can be found mathematically $(\alpha^*, \beta^*) = \left(\frac{-b_c}{u_c}, \frac{-b_r}{u_r}\right)$.

Here we are more interested in the case that there exists a mixed strategy equilibrium, i.e., the location of the equilibrium point (α^*, β^*) 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 [38, 39], 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:

$$\partial V_r(\alpha, \beta)/\partial\alpha = u_r\beta + b_r, \quad \partial V_c(\alpha, \beta)/\partial\beta = u_c\alpha + b_c,$$

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 $b_c = c_{21} - c_{22}$. 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:

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

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 $u_r u_c < 0$ [40]. Consider the quadratic Lyapunov function which is continuously differentiable and $F(0, 0) = 0$:

$$F(x, y) = 1/2(u_c x^2 - u_r y^2),$$

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:

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \frac{\partial F}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial F}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t} = xy(u_r u_c - u_r u_c) = 0,$$

where the derivative of the Lyapunov function is identically zero. Hence, the condition of asymptotic stability is not satisfied [38, 41] and the IGA level-0 dynamics is unstable. There are some IGA based methods (WoLF-IGA, WPL etc. [5, 42]) 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:

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

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

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

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:

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \zeta u_r u_c(u_c x^2 - u_r y^2) + xy(u_r u_c - u_r u_c) = \zeta u_r u_c(u_c x^2 - u_r y^2),$$

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 $\frac{\mathrm{d}F}{\mathrm{d}t} < 0$. 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:

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \zeta u_r u_c(u_c x^2 - u_r y^2) + xy\left(1 + \zeta^2 u_r u_c\right)(u_r u_c - u_r u_c) = \zeta u_r u_c(u_c x^2 - u_r y^2),$$

and level-3 dynamics:

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \zeta u_r u_c(2 + \zeta^2 u_r u_c)(u_c x^2 - u_r y^2).$$

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

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \zeta u_r u_c(k - 1 + \cdots + \zeta^{k-1}(u_r u_c)^{k-2})(u_c x^2 - u_r y^2),$$

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

C Proof of the Existence of Perfect Bayesian Equilibrium

Theorem 2. *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. Λ denotes a level-k reasoning path/tree. 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 $\tilde{k} \in \{0, 1, 2 \cdots k\}$. 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 3. *(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 $h_t^i \in H^i$:*

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

Based on Definition 3, we have the strategy π^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 4. *(Consistency). The belief pair $\{\rho_*^{-i}, \rho_*^i\}$ 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 $h_t^i \in H^i$:*

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

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

The belief p^i is **consistent** given π^i, π^{-i} indicates that for every intermediate reasoning actions reached with positive probability given the strategy profile π^i, π^{-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 [43]. 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 λ_k denotes the possible levels/types for agents, which can be arbitrary level-k player. Then, according to Kreps and Wilson [44], for every finite extensive form game, there exists at least one sequential equilibrium should satisfy Definition. 3 and 4 for sequential rationality and consistency, and the details proof as following:

We use $E^i(\pi, p, \lambda_k, h^i) = \sum_{\tilde{k}=0}^{k} \lambda_{\tilde{k}} p^i_{\tilde{k}}(h^i_t) u^i(\pi^i, \pi^{-i}|h^i_t)$ as expected payoff for player i, for every player i and each reasoning path h^i_t. 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 (π_m, p_m):

$$E^i\left(\pi_m, p_m, \lambda_k, h^i_{t^i}\right) \geq E^i\left((\pi^{-i}_m, \pi^i), p_n(\pi^{-i}_m, \pi^i), \lambda_k, h^i_{t^i}\right),$$

for any strategy π^i with induced probability distributions in $\Pi^T_{t^i=1} = \Delta^{\frac{1}{m}}(p(h^i_{t^i}))$ [45]. Then, consider the strategy and belief pair $\hat{\pi}, \hat{p}$ given by:

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

Such a limit exists because $\Pi^m_{j=1}\Pi^{T_j}_{t_j=1}\Delta^{\frac{1}{m}}\left(p\left(h^j_{t_j}\right)\right)$ 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 $h^i_{t^i}$:

$$E^i\left(\hat{\pi}_m, \hat{p}_m, \lambda_k, h^i_{t^i}\right) \geq E^i\left((\hat{\pi}^{-i}_m, \pi^i), p(\hat{\pi}^{-i}_m, \pi^i), \lambda_k, h^i_{t^i}\right), \tag{10}$$

for any strategy π^i of player i.

If not, then for some player i and some strategy π^i of player i, we have:

$$E^i\left(\hat{\pi}_m, \hat{p}_m, \lambda_k, h^i_{t^i}\right) < E^i\left((\hat{\pi}^{-i}_m, \lambda_k, \pi^i), p(\hat{\pi}^{-i}_m, \pi^i), \lambda_k, \lambda_k, h^i_{t^i}\right).$$

Then, we let

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

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 $m \geq m_0$ [45],

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

and

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

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

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

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

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

where $\pi^i = \lim_{\xi\to\infty} \pi^i_\xi$ of a sequence $\{\pi^i_\xi\}_\xi$ 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 π^i of player i, we have

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

But this result contradicts the previous claim in Eq. 10, 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 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:

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

where λ_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 level 0- $(k-1)$, 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.

■

E More Details for Experiments

E.1 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* [21] 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 $\pi_1^i(a^i|s)=\int_{a^{-i}}\pi_1^i(a^i|s,a^{-i})\rho_0^{-i}(a^{-i}|s)\,\mathrm{d}a^{-i}$, while the level-1 model in PR2 starts from the opponent's angel, that is $\rho_1^{-i}(a^{-i}|s)=\int_{a^i}\rho_1^i(a^{-i}|s,a^i)\pi_0^i(a^i|s)\,\mathrm{d}a^i$.

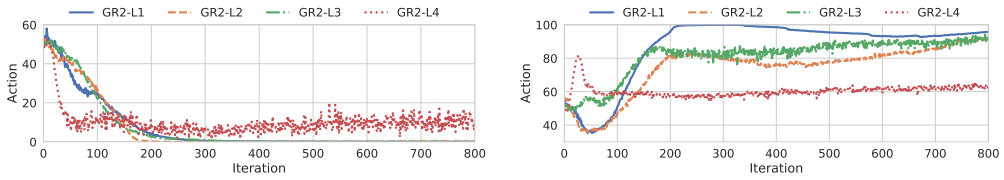
E.2 Hyperparameter Setting.

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.

E.3 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 λ in GR2-M methods, as well as the influences of incentive intensity in the games.

E.3.1 Choice of Level-k



(a) $p = 0.7, n = 10$ (b) $p = 1.1, n = 10$

Figure 7: Learning curves on Keynes Beauty Contest game with GR2-L policies from level-1 to level-4.

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. 7, 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.

E.3.2 Choice of λ of Poisson distribution in GR2-M

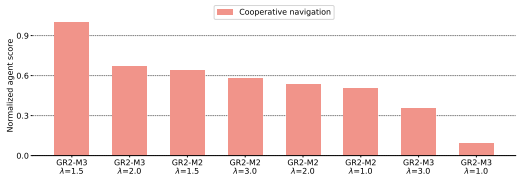


Figure 8: Effect of varying λ in GR2-M methods, the score is normalized to $0 - 1$.

We investigate the effect of hyper-parameter λ 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.

E.3.3 Choice of Reward Function in Keynes Beauty Contest

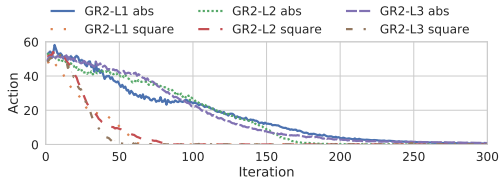


Figure 9: Learning curves with two reward schemes: absolute difference (default) and squared absolute difference.

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. 9 that the amplified reward can significantly speed up the convergence for GR2-L methods.