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

Ying Wen[1*], **Yaodong Yang**[1,2*], **Jun Wang**[1]
[1]University College London
[2]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 individuals' 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. Keynes Beauty Contest [Keynes, 1936] is one of the most-cited examples prescribe bounded rationality. 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 levels 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 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.*, 2018a]. 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 joint-action learners [Claus and Boutilier, 1998], factorized Q-learners [Zhou *et al.*, 2019], to the complicated theory-of-mind learners [Rabinowitz *et al.*, 2018]. It comes as 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 studying population dynamics [Yang *et al.*, 2018b] 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 interactions. It begins with $level$-0 (L0 for short) non-strategic thinkers who do not model their opponents. L1 thinkers are

*First two authors contribute equally.

more sophisticated than $level$-0; they believe the opponents are all at L0 and then act correspondingly. With the growth of k, Lk 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 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 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] 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 act 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 bounded rationality. Importantly, PR2 does not consider whether an equilibrium exists in such a sophisticated hierarchical framework at all. In this work, we extend the reasoning level to an arbitrary number, and theoretically prove the existence of equilibrium 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 makers' 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, the 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 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

The *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 $i \in \{1, \dots, n\}$, $\mathcal{P} : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^n \to \mathcal{P}(\mathcal{S})$ 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 $\pi^i_{\theta^i}(a^i|s)$ with θ^i being a tuneable parameter, and use $a^{-i} = (a^j)_{j \neq i}$ to represent actions executed by opponents. The trajectory $\tau^i = \left[(s_1, a^i_1, a^{-i}_1), \dots, (s_T, a^i_T, a^{-i}_T)\right]$ of agent i is defined as a collection of state-action triples over a horizon T.

3.1 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 $V^i(s; \boldsymbol{\pi_\theta}) = \mathbb{E}_{\boldsymbol{\pi_\theta}, \mathcal{P}} \left[\sum_{t=1}^{T} \gamma^{t-1} r^i \left(s_t, a^i_t, a^{-i}_t\right)\right]$ where $(a^i_t, a^{-i}_t) \sim \boldsymbol{\pi_\theta} = (\pi^i_{\theta^i}, \pi^{-i}_{\theta^{-i}})$ 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 $\pi^{-i}_{\theta^{-i}}$, which can be formally defined as the policy π^i_* with $V^i(s; \pi^i_*, \pi^{-i}_{\theta^{-i}}) \geq V^i(s; \pi^i_{\theta^i}, \pi^{-i}_{\theta^{-i}})$ for all valid $\pi^i_{\theta^i}$. 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 $\pi^i(a^i|s) = \frac{\exp\left(Q^i_{\boldsymbol{\pi_\theta}}(s, a^i, a^{-i})\right)}{\sum_{a'} \exp\left(Q^i_{\boldsymbol{\pi_\theta}}(s, a', a^{-i})\right)}$ – a standard type of policy adopted in RL literatures – with $Q^i_{\boldsymbol{\pi_\theta}}(s, a^i, a^{-i}) = r^i(s, a^i, a^{-i}) + \gamma \mathbb{E}_{\mathcal{P}}[V^i(s'; \boldsymbol{\pi_\theta})]$ denoting agent i's Q-function and s' being a successor state, they reach a Nash-Quantal equilibrium [McKelvey and Palfrey, 1995].

3.2 The Graphical Model of MARL

Since GR2 is a probabilistic model, it is instructive to provide a brief review of graphical models 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 [Wen *et al.*, 2019; Grau-Moya *et al.*, 2018]. To illustrate, we first introduce a binary random variable $\mathcal{O}^i_t \in \{0, 1\}$ (see Fig. 1) that stands for the optimality of agent i's policy at time t, i.e., $p\left(\mathcal{O}^i_t = 1 | \mathcal{O}^{-i}_t = 1, \tau^i_t\right) \propto \exp\left(r^i\left(s_t, a^i_t, a^{-i}_t\right)\right)$, which suggests that given a trajectory τ^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 $p(\mathcal{O}^i_{1:T} = 1 | \mathcal{O}^{-i}_{1:T} = 1)$ 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. As we assume no

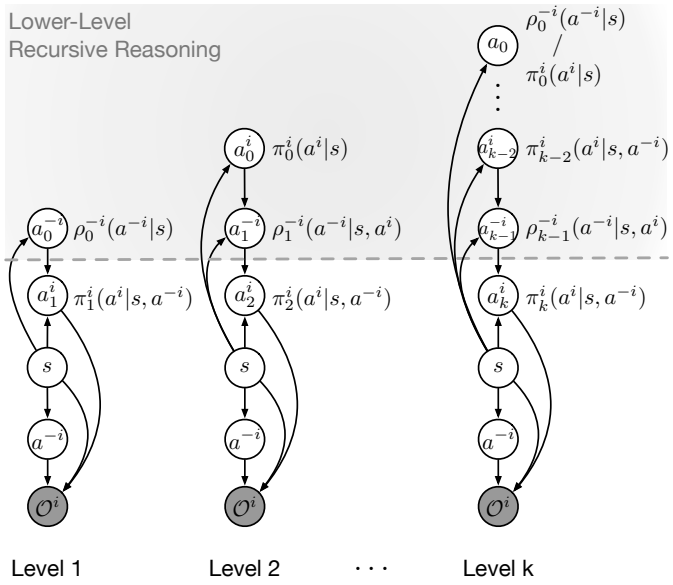


Figure 1: Graphical model of the $level$-k reasoning model. Subfix of a_* stands for the level of thinking not the timestep. The opponent policies are approximated by ρ^{-i}. The omitted $level$-0 model considers opponents fully randomized. Agent i rolls out the recursive reasoning about opponents in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

knowledge of the optimal policies $\boldsymbol{\pi}_*$ and the model of the environment $\mathcal{P}(\mathcal{S})$, we treat them as latent variables and apply variational inference [Blei *et al.*, 2017] to approximate such objective; using the variational form of $\hat{p}(\tau^i|\mathcal{O}^i_{1:T}, \mathcal{O}^{-i}_{1:T}) = [\hat{p}(s_1)\prod_{t=1}^{T-1}\hat{p}(s_{t+1}|s_t, a^i_t, a^{-i}_t)]\boldsymbol{\pi}_{\boldsymbol{\theta}}(a^i_t, a^{-i}_t|s_t)$ leads to

$$\begin{aligned}\max \mathcal{J}(\boldsymbol{\pi}_{\boldsymbol{\theta}}) &= \log p(\mathcal{O}^i_{1:T} = 1|\mathcal{O}^{-i}_{1:T} = 1) \\ &\geq \sum_{\tau^i} \hat{p}(\tau^i|\mathcal{O}^i_{1:T}, \mathcal{O}^{-i}_{1:T}) \log \frac{p(\mathcal{O}^i_{1:T}, \tau^i|\mathcal{O}^{-i}_{1:T})}{\hat{p}(\tau^i|\mathcal{O}^i_{1:T}, \mathcal{O}^{-i}_{1:T})} \\ &= \sum_{t=1}^{T} \mathbb{E}_{\tau^i\sim\hat{p}(\tau^i)}\Big[r^i\big(s_t, a^i_t, a^{-i}_t\big) + \mathcal{H}\big(\boldsymbol{\pi}_{\boldsymbol{\theta}}(a^i_t, a^{-i}_t|s_t)\big)\Big].\end{aligned} \tag{1}$$

To maximize $\mathcal{J}(\boldsymbol{\pi}_{\boldsymbol{\theta}})$, a variant of policy iteration called *soft learning* is applied. For policy evaluation, the Bellman expectation equation now holds on the *soft* value function $V^i(s) = \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\big[Q^i(s_t, a^i_t, a^{-i}_t) - \log(\boldsymbol{\pi}_{\boldsymbol{\theta}}(a^i_t, a^{-i}_t|s_t))\big]$, with the updated Bellman operator $\mathcal{T}^\pi Q^i(s_t, a^i_t, a^{-i}_t) \triangleq r^i(s_t, a^i_t, a^{-i}_t) + \gamma\mathbb{E}_{\mathcal{P}}\big[\text{soft}\, Q(s_t, a^i_t, a^{-i}_t)\big]$. Compared to the $\max$ operation in the normal Q-learning, soft operator is $\text{soft}\, Q(s, a^i, a^{-i}) = \log\sum_a\sum_{a^{-i}}\exp\big(Q(s, a^i, a^{-i})\big) \approx \max_{a^i, a^{-i}} Q\left(s, a^i, a^{-i}\right)$. 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. Since the exact parameter of the opponent policy is usually unobservable, agent i needs to approximate $\pi^{-i}_{\theta^{-i}}$.

4 Generalized Recursive Reasoning

Recursive reasoning is essentially taking an iterative best response to opponents' policies. $level$-1 thinking is "I know you know how I know". We can represent such recursion by $\boldsymbol{\pi}(a^i, a^{-i}|s) = \pi^i(a^i|s)\pi^{-i}(a^{-i}|s, a^i)$ 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 π^{-i} can be approximated via a best-fit model $\rho^{-i}_{\phi^{-i}}$ parameterized by ϕ^{-i}. By adopting $\boldsymbol{\pi}_{\boldsymbol{\theta}}(a^i, a^{-i}|s) = \pi^i_{\theta^i}(a^i|s)\rho^{-i}_{\phi^{-i}}(a^{-i}|s, a^i)$ in $\hat{p}(\tau^i|\mathcal{O}^i_{1:T}, \mathcal{O}^{-i}_{1:T})$ in maximizing the Eq. 1, we can solve the best-fit opponent policy by

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

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: $Q^i(s, a^i) = \int_{a^{-i}} \rho^{-i}_{\phi^{-i}}(a^{-i}|s, a^i)Q^i(s, a^i, a^{-i})\,\mathrm{d}a^{-i}$, and then improve its own policy towards the direction of

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

4.1 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

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

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 $\pi^i_k(a^i|s, a^{-i})$ as the *posterior*. From agent i's perspective, it believes that the opponents will take the best response to its own fictitious action a^i_{k-2} that are two levels below, i.e., $\rho^{-i}_{k-1}(a^{-i}_{k-1}|s) = \int\rho^{-i}_{k-1}(a^{-i}_{k-1}|s, a^i_{k-2})\pi^i_{k-2}(a^i_{k-2}|s)\,\mathrm{d}a^i_{k-2}$, where π^i_{k-2} can be further expanded by recursively using Eq. 4 until meeting π_0 that is usually assumed uniformly distributed. Decisions are taken in a sequential manner. As such, a $level$-k model transforms the multi-agent planning problem into a hierarchy of nested single-agent planning problems.

4.2 Mixture of Hierarchy Recursive Reasoning – GR2-M

So far, a $level$-k agent assumes all 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 types 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 λ 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

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

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

4.3 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 [Levin and Zhang, 2019]; 2) the learning dynamics of policy gradient in GR2 is asymptotically stable in the sense of Lyapunov [Marquez, 2003].

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.*

Algorithm 1 GR2 Soft Actor-Critic Algorithm

Init: λ, k and ψ (learning rates).
Init: $\theta^i, \phi^{-i}, \omega^i$ for each agent i. $\bar{\omega}^i \leftarrow \omega^i$, $\mathcal{D}^i \leftarrow \emptyset$.
for each episode **do**
  for each step t **do**
    Agents take a step according to $\pi^i_{\theta^i,k}(s)$ or $\pi^{i,\lambda}_{\theta^i,k}(s)$.
    Add experience $(s, a^i, a^{-i}, r^i, s')$ to $\mathcal{D}^i$.
    for each agent i **do**
      Sample a batch $\{(s_j, a_j^i, a_j^{-i}, r_j^i, s_j')\}_{j=0}^M \sim \mathcal{D}^i$.
      Roll out policy to level k via GR2-L/M to get $a_j^{i\prime}$ and record inter-level results $(a_{j,k}^{i\prime}, a_{j,k-1}^{-i\prime}, \cdots)$.
      Sample $a_j^{-i\prime} \sim \rho^{-i}_{\phi^{-i}}(\cdot|s_j', a_j^{i\prime})$.
      $\omega^i \leftarrow \omega^i - \psi_{Q^i}\hat{\nabla}_{\omega^i} J_{Q^i}(\omega^i)$.
      $\theta^i \leftarrow \theta^i - \psi_{\pi^i}\hat{\nabla}_{\theta^i}\left(J_{\pi_k^i}(\theta^i) + J_{\pi_{\bar{k}}^i}(\theta^i)\right)$.
      $\phi^{-i} \leftarrow \phi^{-i} - \psi_{\rho^{-i}}\hat{\nabla}_{\phi^{-i}} J_{\rho^{-i}}(\phi^{-i})$.
    end for
    $\bar{\omega}^i \leftarrow \psi_{\bar{\omega}}\omega^i + (1-\psi_{\bar{\omega}})\bar{\omega}^i$.
  end for
end for

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 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 GR2 setting, higher-level strategies weakly dominate lower-level strategies.*

5 Practical Implementations

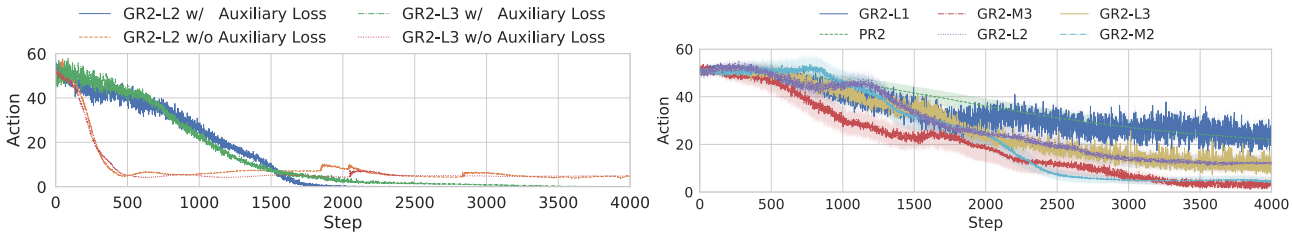
Computing the recursive reasoning is computationally expensive. Here we first present the GR2 soft actor-critic algorithm with the pseudo-code in Algo. 1, and then introduce the compromises we make to afford the implementation.

5.1 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 ω^i of the joint soft Q-function is then updated via minimizing the soft Bellman residual $J_{Q^i}(\omega^i) = \mathbb{E}_{\mathcal{D}^i}[\frac{1}{2}(Q^i_{\omega^i}(s, a^i, a^{-i}) - \hat{Q}^i(s, a^i, a^{-i}))^2]$ 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 $\pi^{-i}_{\theta^{-i}}$, we instead compute the soft $Q^i(s, a^i)$ by marginalizing the joint Q-function via the estimated opponent model $\rho^{-i}_{\phi^{-i}}$ by

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

Table 1: The Converging Equilibrium on Keynes Beauty Contest.



(a) (b)

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).

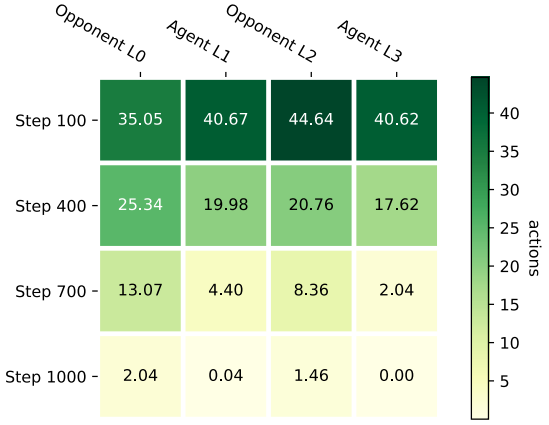


Figure 3: The guessing number of both agents during the training of the GR2-L3 model in the Beauty Contest setting ($n = 2, p = 0.7$).

$Q^i(s, a^i) = \log \int \rho^{-i}_{\phi^{-i}}(a^{-i}|s, a^i) \exp\left(Q^i(s, a^i, a^{-i})\right) \mathrm{d}a^{-i}$; the value function of the *level-k* policy $\pi^i_k(a^i|s)$ then 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, $J_{\rho^{-i}}(\phi^i) = \mathcal{D}_{\mathrm{KL}}\left[\rho^{-i}_{\phi^{-i}}(a^{-i}|s, a^i) \| \exp\left(Q^i_{\omega^i}(s, a^i, a^{-i}) - Q^i_{\omega^i}(s, a^i)\right)\right]$. 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 [Liu and Wang, 2016].

Finally, the policy parameter θ^i for agent i can be learned by improving towards what the current Q-function $Q^i_{\omega^i}(s, a^i)$ suggests, as shown in Eq. 3. By applying the reparameterization trick $a^i = f_{\theta^i}(\epsilon; s)$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, we have $J_{\pi^i_k}(\theta^i) = \mathbb{E}_{s, a^i_k, \epsilon}[\log \pi^i_{\theta^i, k}(f_{\theta^i}(\epsilon; s)|s) - Q^i_{\omega^i}(s, f_{\theta^i}(\epsilon; s))]$.

Note that as the agents' final decision comes from the best response to all lower levels, we would expect the gradient of $\partial J_{\pi^i_k} / \partial \theta^i$ to be propagated from all higher levels during training.

5.2 Approximated Best Response via Deterministic Policy

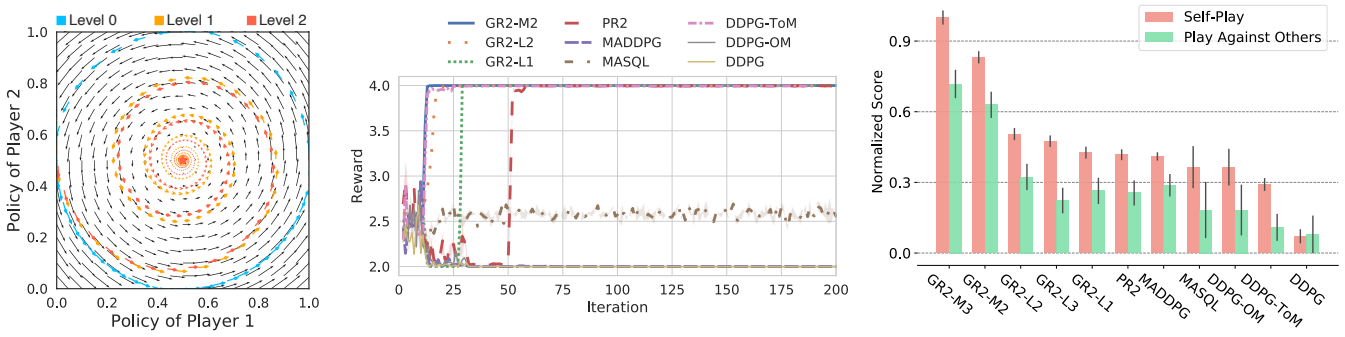
As the reasoning process of GR2 methods involve iterated usages of $\pi^i_k(a^i|s, a^{-i})$ and $\rho^{-i}_k(a^{-i}|s, a^i)$, should they be stochastic, the cost of integrating possible actions from lower-level agents would be unsustainable for large k. Besides, the reasoning process is also affected 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 a Gaussian policy. However, note that the highest-level agent policy π^i_k 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 policies should perform better than lower-level policies against the opponents. To maintain this property, we introduce an auxiliary loss $J_{\pi^i_{\tilde{k}}}(\theta^i)$ in training $\pi^i_{\theta^i}$ (see Fig. 5 in *Appendix B*), with $s \sim \mathcal{D}^i$, $a^i_{\tilde{k}} \sim \pi^i_{\theta^i}$, $a^{-i}_{\tilde{k}} \sim \rho^{-i}_{\phi^{-i}}$ and $\tilde{k}, \geq 2.$, we have

$$J_{\pi^i_{\tilde{k}}}(\theta^i) = \mathbb{E}_{s, a^i_{\tilde{k}}, a^{-i}_{\tilde{k}}}\left[Q^i(s, a^i_{\tilde{k}}, a^{-i}_{\tilde{k}-1}) - Q^i(s, a^i_{\tilde{k}-2}, a^{-i}_{\tilde{k}-1})\right]. \quad (6)$$

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

5.3 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 $\pi^i_{\theta^k}$ and $\rho^{-i}_{\theta^k}$. 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



(a) (b) (c)

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

constraint in Eq. 6 that enforces the inter-policy improvement. Finally, in 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[1] 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.

6.1 Keynes Beauty Contest

In a 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 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 it. 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.

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 accurate yet $(20.76 \neq 19.98 \times 1.1)$, the agent performance can still be improved as 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 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 equilibrium; 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 other models by successfully guessing the right equilibrium, which is expected since 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 π_k^i.

6.2 Normal-form Games

We further evaluate GR2 methods on two normal-form games: Rotational Game (RG) 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 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 GR2-L policies can converge to the center equilibrium, with higher-level pol-

[1]The experiment code and appendix are available at https://github.com/ying-wen/gr2

icy 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.

6.3 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 agents 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 other 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 line 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 GR2 setting as well as the convergence of the policy gradient methods on the two-player normal-form games. A series of experimental results justified the advantages of GR2 methods over strong MARL baselines on modeling different opponents.

References

[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.

[Blei *et al.*, 2017] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, pages 859–877, 2017.

[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. *JAIR*, 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.

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

[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.

[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. Multi-agent 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, 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.

[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.

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

[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.*, 2018a] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *ICML*, 2018.

[Yang *et al.*, 2018b] Yaodong Yang, Lantao Yu, Yiwei Bai, Ying Wen, Weinan Zhang, and Jun Wang. A study of ai population dynamics with million-agent reinforcement learning. In *17th AAMAS*, pages 2133–2135, 2018.

[Zhou *et al.*, 2019] Ming Zhou, Yong Chen, Ying Wen, Yaodong Yang, Yufeng Su, Weinan Zhang, Dell Zhang, and Jun Wang. Factorized q-learning for large-scale multi-agent systems. In *Proceedings of the First International Conference on Distributed Artificial Intelligence*, pages 1–7, 2019.