Appendix

A Maximum Entropy Multi-Agent Reinforcement Learning

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

$$p(a^i_{1:T}, a^j_{1:T}, s_{1:T}) = [p(s_1)\prod_{t=1}^{T} p(s_{t+1}|s_t, a^i_t, a^{-i}_t)]\exp(\sum_{t=1}^{T} r^i(s_t, a_t, a^{-i}_t)). \tag{7}$$

Analogously, we factorize empirical trajectory distribution $q(\tau^i)$ as:

$$\hat{p}(\tau^i) = p(s_1)\prod_{t} p(s_{t'}|s_t, a_t)\pi^i(a^i_t|s_t)\rho^{-i}(a^{-i}_t|s_t, a^i_t), \tag{8}$$

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

$$\begin{aligned}
-D_{\mathrm{KL}}(\hat{p}(\tau^i)\|p(\tau^i)) &= \mathbb{E}_{\tau^i\sim\hat{p}(\tau^i)}\Big[\log p(s_1) + \sum_{t=1}^{T}\Big(\log p(s_{t+1}|s_t, a_t, a^{-i}_t) + r^i(s_t, a^i_t, a^{-i}_t)\Big) \\
&\qquad - \log p(s_1) - \sum_{t=1}^{T}\Big(\log p(s_{t+1}|s_t, a^i_t, a^{-i}_t) + \log\big(\pi^i(a^i_t|s_t)\rho^{-i}(a^{-i}_t|s_t, a^i_t)\big)\Big)\Big] \\
&= \mathbb{E}_{\tau^i\sim\hat{p}(\tau^i)}\Big[\sum_{t=1}^{T} r^i(s_t, a^i_t, a^{-i}_t) - \log\big(\pi^i(a^i_t|s_t)\rho^{-i}(a^{-1}_t|s_t, a^i_t)\big)\Big] \\
&= \sum_{t=1}^{T}\mathbb{E}_{(s_t, a^i_t, a^{-i}_t)\sim\hat{p}(s_t, a^i_t, a^{-i}_t)}\Big[r^i(s_t, a^i_t, a^{-i}_t) - \log\big(\pi^i(a^i_t|s_t)\rho^{-i}(a^{-1}_t|s_t, a^i_t)\big)\Big] \\
&= \sum_{t=1}^{T}\mathbb{E}_{(s_t, a^i_t, a^{-i}_t)\sim\hat{p}(s_t, a^i_t, a^{-i}_t)}\Big[r^i(s_t, a^i_t, a^{-i}_t) + \mathcal{H}\big(\rho^{-i}(a^{-i}_t|s_t, a^i_t)\big) + \mathcal{H}\big(\pi^i(a^i_t|s_t)\big)\Big],
\end{aligned} \tag{9}$$

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

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

$$\mathbb{E}_{(s_t, a^i_t, a^{-i}_t)\sim\hat{p}(s_t, a^i_t, a^{-i}_t)}\Big[-D_{\mathrm{KL}}\Big(\rho^{-i}_t(a^{-i}_t|s_t, a^i_t)\Big\|\frac{\exp\big(Q^i(s_t, a^i_t, a^{-i}_t)\big)}{\exp\big(Q^i(s_t, a^i_t)\big)}\Big) + Q^i(s_t, a^i_t)\Big], \tag{10}$$

where we define:

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

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

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

B Algorithm Implementations

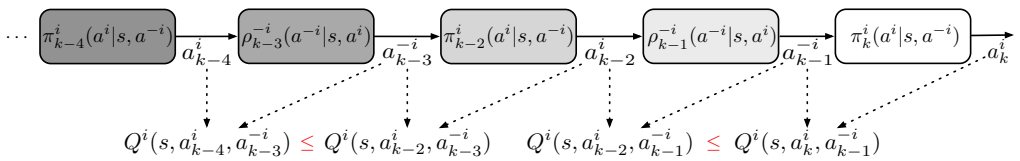


Figure 5: Inter-level policy improvement maintained by Eq. 6 so that higher-level policy weakly dominates lower-level policies.

C Proof of Theorem 1

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

Proof. Consider an extensive game, which is extended from a normal form game by $level\text{-}k$ 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 the lower-level reasoning trajectory/path so far. An intermediate reasoning action/node in the $level\text{-}k$ 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 $level\text{-}\tilde{k}$ player and $\tilde{k} \in \{0, 1, 2 \cdots k\}$. A $level\text{-}k$ player holds a prior belief that the opponent is a $level\text{-}\tilde{k}$ 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 $level\text{-}\tilde{k}$ player as $p_{\tilde{k}}^i(h^t)$. In equilibrium, a $level\text{-}k$ player chooses an optimal strategy according to her belief at every decision node, which implies choice is sequentially rational as following defined:

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

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

Based on Definition 1, 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 $level\text{-}k$ player are consistent. Let $p^i(h_t | \pi^i, \pi^{-i})$ denote the probability that reasoning action h_t is reached according to the strategy pair, $\{\pi^i, \pi^{-i}\}$. Then we have the consistency definition:

Definition 2. *(Consistency). The belief pair $\{\rho_*^{-i}, \rho_*^i\}$ 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 $h_t^i \in H^i$:*

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

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 [Levin and Zhang, 2019]. Consistency ensures correctness of the beliefs.

Although the game itself has perfect information, the belief structure in our $level\text{-}k$ 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 $level\text{-}k$ 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 $level\text{-}k$ player. Then, according to Kreps and Wilson [1982], for every finite extensive form game, there exists at least one sequential equilibrium should satisfy Definition. 1 and 2 for sequential rationality and consistency, and the details proof as following:

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

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

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

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

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

Such a limit exists because $\Pi_{j=1}^{m} \Pi_{t_j=1}^{T_j} \Delta^{\frac{1}{m}}\left(p\left(h_{t_j}^j\right)\right)$ 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 $h_{t^i}^i$:

$$E^i\left(\hat{\pi}_m, \hat{p}_m, \lambda_k, h^i_{t^i}\right) \geq E^i\left((\hat{\pi}^{-i}_m, \pi^i), p(\hat{\pi}^{-i}_m, \pi^i), \lambda_k, h^i_{t^i}\right), \tag{13}$$

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 $m \geq m_0$,

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

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

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

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

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

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

■

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

D Proof of Theorem 2

Theorem 2. *In two-player two-action games, if 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 is used to verify the stability of a dynamical system in control theory, here we apply it in convergence proof for $level$-k methods. It is defined as following:

Definition 3. *(Lyapunov Function.) Give a function $F(x, y)$ be continuously differentiable in a neighborhood 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 4. *(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) for all* $(x,y) \in \sigma \backslash (0,0)$*, then the equilibrium point* $(x,y) = (0,0)$ *of the system is asymptotically stable [Marquez, 2003].*

Single State Game

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

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

Each player selects an action from the action space $\{1,2\}$ which determines the payoffs to the players. If the row player chooses action i and the player 2 chooses action j, then the row player and column player receive the rewards r_{ij} and c_{ij} respectively. We use $\alpha \in [0,1]$ to represent the strategy for row player, where 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 $(\alpha^*, \beta^*) = \left(\frac{-b_c}{u_c}, \frac{-b_r}{u_r}\right)$, where $u_r = r_{11} - r_{12} - r_{21} + r_{22}$, $b_r = r_{12} - r_{22}$, $u_c = c_{11} - c_{12} - c_{21} + c_{22}$, and $b_c = c_{21} - c_{22}$.

Here we are more interested in the case that there exists a mixed strategy equilibrium, i.e., the location of the equilibrium point (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 [Marquez, 2003; Bowling and Veloso, 2001], we are not going to discuss in this proof.

Learning in $level$**-0 Gradient Ascent**

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

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

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

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

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 $u_r u_c < 0$ [Zhang and Lesser, 2010].

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

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

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

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

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

Learning in $level$-k Gradient Ascent

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

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

where 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. ■

E Proof of Proposition 1

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. 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 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 GR2-M model, if the $level$-k step agent play the equilibrium strategy $\pi^i_{k,*}$, 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 $\pi^i_{k,*}$ 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 $\pi^i_{k,*}$ is already the best response by definition. For $level$-k, $\pi^{i,*}_k$ is still the best response due to the conclusion in the above GR2-L. Considering the linearity of the expected reward for GR2-M:

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

where λ_k is $level$-k policy's proportion. Therefore, $\pi^{i,*}_k$ 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 $\pi^{i,*}_k$ is the best response to both $level$-k and all lower levels from 0 to $(k-1)$, it is therefore the best response of the $level$-$(k+1)$ thinker.

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

■

F Detailed Settings for Experiments

The Recursive Level

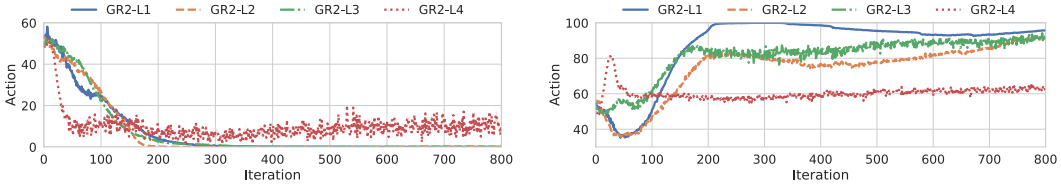
We regard DDPG, DDPG-OM, MASQL, MADDPG as $level$-0 reasoning models because from the policy level, they do not explicitly model the impact of one agent's action on the other agents or consider the reactions from the other agents. Even though the value function of the joint policy is learned in MASQL and MADDPG, but they conduct a *non-correlated factorization* [Wen *et al.*, 2019] when it comes to each individual agent's policy. PR2 and DDPG-ToM are in fact the $level$-1 reasoning model, but note that the $level$-1 model in GR2 stands for $\pi^i_1(a^i|s)=\int_{a^{-i}}\pi^i_1(a^i|s,a^{-i})\rho^{-i}_0(a^{-i}|s)\,\mathrm{d}a^{-i}$, while the $level$-1 model in PR2 starts from the opponent's angel, that is $\rho^{-i}_1(a^{-i}|s)=\int_{a^i}\rho^i_1(a^{-i}|s,a^i)\pi^i_0(a^i|s)\,\mathrm{d}a^i$.

Hyperparameter Settings

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

Ablation Study

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

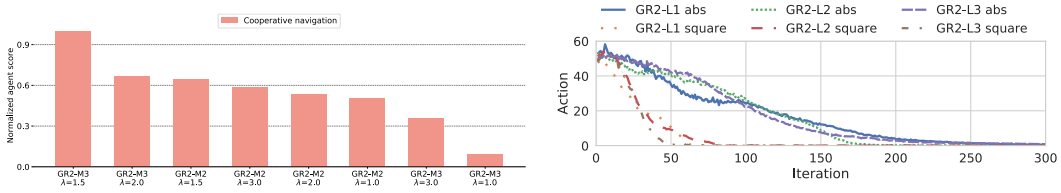


(a) $p=0.7, n=10$ (b) $p=1.1, n=10$

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

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

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



(a) (b)

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

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

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