Optimizing Multi-Agent Coordination via Hierarchical Graph Probabilistic Recursive Reasoning

Saar Cohen
Department of Computer Science
Bar Ilan University, Israel
saar30@gmail.com

Noa Agmon
Department of Computer Science
Bar Ilan University, Israel
agmon@cs.biu.ac.il

Abstract

Multi-agent reinforcement learning (MARL) requires coordination by some means of interaction between agents to efficiently solve tasks. Interaction graphs allow reasoning about joint actions based on the local structure of interactions, but they disregard the potential impact of an agent's action on its neighbors' behaviors, which could rapidly alter in dynamic settings. In this paper, we thus present a novel perspective on opponent modeling in domains with only local interactions using (level-1) *Graph Probabilistic Recursive Reasoning* (**GrPR2**). Unlike previous work on recursive reasoning, each agent iteratively best-responds to other agents' policies *over all possible local interactions*. Agents' policies are approximated via a variational Bayes scheme for capturing their uncertainties, and we prove that an induced variant of Q-learning converges under self-play when there exists only one Nash equilibrium. In *cooperative* settings, we further devise a variational lower bound on the likelihood of each agent's optimality. Opposed to other models, optimizing the resulting objective prevents each agent from attaining an unrealistic modelling of others, and yields an exact tabular Q-iteration method that holds convergence guarantees. Then, we deepen the recursion to level-k via *Cognitive Hierarchy GrPR2* (**GrPR2-CH**), which lets each level-k player best-respond to a *mixture* of strictly lower levels in the hierarchy. We prove that: (1) **level-3 reasoning is the optimal hierarchical level**, maximizing each agent's expected return; and (2) **the weak spot of the classical CH models is that 0-level is uniformly distributed**, as it *may* introduce policy bias. Finally, we propose a practical actor-critic scheme, and illustrate that GrPR2-CH outperforms strong MARL baselines in the particle environment.

Keywords

Multi-Agent Reinforcement Learning; Multi-Agent Coordination; Interaction Graphs; Cognitive Hierarchy; Variational Inference

ACM Reference Format:
Saar Cohen and Noa Agmon. 2022. Optimizing Multi-Agent Coordination via Hierarchical Graph Probabilistic Recursive Reasoning. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 10 pages.

1 Introduction

Humans are capable of wisely and *recursively* choosing when to convey latent mental contents (e.g., beliefs, desires, intentions [24, 54]) and whom to interact with. In the progression of multi-agent reinforcement learning (MARL) in mimicking such human intelligence, a milestone of an agent is its ability to understand and interact with other agents [40]. Naturally, MARL requires coordination by some means of interaction between agents to efficiently solve tasks. Interaction graphs allow reasoning about the joint action based on the local structure of interactions, yet prior studies disregard the potential impact of an agent's action on its neighbors' behaviors. In problems where the graph structure is dynamic due to changing coordination requirements as actions rapidly alter, such reasoning becomes even more crucial since each agent should dynamically select which agents to attend for querying relevant information.

In this paper, our main goal is modelling locality of interactions via a graph structure, while regarding the potential impact of an agent's action on its neighbors' behaviors. After augmenting decentralized partially observable MDPs (Dec-POMDPs) [52] to support interaction graphs, and inspired by *recursive reasoning* [11, 18, 19], we thus present a novel perspective on opponent modeling in domains with only local interactions using (level-1) *Graph Probabilistic Recursive Reasoning* (**GrPR2**), so as to incorporate both aspects into each agent's decision making. Namely, unlike prior work on adopting recursive reasoning into the MARL settings [68, 71, 72], according to our model each agent iteratively best-responds to other agents' policies *over all possible local interactions*. Agents' policies are approximated via a variational Bayes scheme for capturing their uncertainties, and we prove that an induced variant of Q-learning converges under self-play when there exists only one Nash equilibrium. In *cooperative* MARL, we further devise a variational lower bound on the likelihood of each agent's optimality. Opposed to other schemes [3, 28], optimizing the resulting objective prevents each agent from attaining an unrealistic modelling of others, and yields an exact tabular Q-iteration method with convergence guarantees.

In humans' decision making, instead of assuming people are perfectly rational, *bounded rationality* [61] enables to exhibit varying *hierarchical* levels of reasoning. Following the *cognitive hierarchy* (CH) model [5, 72], we thus deepen the recursion to level-k using CH by introducing **GrPR2-CH**, which lets each level-k player best-respond to a *mixture* of strictly lower levels in the hierarchy. We prove that **level-3 reasoning is the optimal hierarchical level, maximizing each agent's expected return**. Specifically, it yields the maximal monotonic improvement in the expected return between two subsequent levels of reasoning. We also demonstrate the ***weak spot*** of the classical Level-k [31] CH models: since an agent's 0-level policy is uniformly distributed, **it *may* be far from the optimal policy, thus biasing the final policy**. Finally, we propose a practical actor-critic scheme, and illustrate that GrPR2-CH outperforms strong MARL baselines in the particle environment.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

2 Related Work

Many problems in RL [66] (e.g., multi-robot control [48], analysis of social dilemmas [42]) operate as multi-agent systems. Notorious successes of RL are mostly restricted to *single* agent domains (e.g., robot manipulation [13, 53], Atari [50]), though transferring these methods into MARL is challenging. This stems from *independent learning*'s lack in information sharing, thus making it difficult to learn coordinated strategies that depend on interactions between multiple agents [15]. Such non-stationarity makes it incompatible with the experience replay memory on which deep RL relies [16].

Multi-agent coordination [4, 21, 22] is one of such challenging problems. Some works extend variants of actor-critic schemes to MARL settings and learn decentralized policies via centralized critics (e.g., DDPG [45], MADDPG [47], COMA [15]). Other studies focus on function decomposition schemes that gradually increase the representation ability of the global value function (e.g., QMIX [57], VDN [64]. Yet, they operate without explicit interactions. In fact, interactions can lead to increased exploration, higher rewards, and a higher diversity of solutions in both simulated high-dimensional optimization problems [1] and human experiments [41].

In contrast, [25] introduced *Coordination Graphs* (CG), which allow explicit modeling of the locality of interactions and formal reasoning about joint actions given the coordination graph structure. CGs were applied to tabular RL [38], and then extended to function approximation via neural networks (NNs) [2]. Though the CG terminology is focused on using a graph structure to decompose payoffs and utilities, we regard this structure as means for controlling when and whom to interact with for querying relevant information. Further, most of these approaches assume a *domain dependent static* interaction graph, where we assume it to be *dynamic* and *state dependent*. Another notable line of research is *MARL with communication*, which enables agents to communicate and exchange messages during execution (e.g., DIAL [17], IC3Net [63], FlowComm [12]). Although our work can be applied to such studies, we do not pose the restriction that agents have access to high-dimensional encodings of others' local information. Instead, agents learn a binary attention encapsulated by an adjacency matrix, which is used for sharing less-sophisticated information.

Despite recent attempts to learn such dynamic structure by domain heuristics [30, 37] and NNs [2, 44], all works mentioned so far disregard the potential impact of an agent's action on its neighbors' behaviors. Recently, following human intelligence [5, 6, 23, 55], *recursive reasoning* [11, 18, 19] about the potential impacts on *all* other agents' behaviors has become popular in opponent modelling. Studies on Theory of Mind (ToM) [20, 56, 60] are used for modelling an agent's belief on opponents' mental states in RL domains. Interactive POMDP (I-POMDP) [18] implements ToM by augmenting POMDPs with an extra space for building the beliefs about opponents' intentions during planning, and making agents acting optimally with respect to such predicted intentions. However, I-POMDP has limitations in its solvability due to its inherent complexity [59]. GrPR2 differs from I-POMDP as it employs a hierarchical structure for opponent modeling and does not adjust the MDP. Instead, the recursive reasoning is implemented via a probabilistic scheme and enables opponents with different hierarchical levels of thinking. In fact, our method is most related to Generalized Recursive Reasoning (GR2) [72]. Yet, GR2 does not target modelling locality of interactions, and does not explore the optimal hierarchical level of reasoning, as well as the drawbacks inherent in CH models [5, 72]. In this work, we extend GR2 to address locality of interactions and theoretically answer the later questions.

For modelling optimality, a common approach is casting RL into an inference problem by introducing a binary random variable representing "optimality" [43, 58, 70]. In single-agent and Bayesian RL, maximizing entropy improves the diversity [14], and presents in the evidence lower bound for the log-likelihood of optimality [7, 26]. In MARL, the existence of other agents increases uncertainties in the environment, and thus [68] attempt to address this issue by reformulating the MARL problem into Bayesian inference. However, local interactions pose an additional challenge. We thus bridge this gap by embedding the graph structure into the inference problem.

3 Problem Setup

3.1 Graphical Decentralized POMDPs

We consider the MARL problem as a decentralized partially observable Markov game (Dec-POMDPs) [52], where agents perform selective interactions based an interaction graph. Formally, a *Graphical Dec-POMDP* (GDec-POMDP) can be described by a tuple $M = \langle \mathcal{S}, \mathcal{N}, \{\mathcal{U}^i\}_{i\in\mathcal{N}}, \mathcal{P}, \{r^i\}_{i\in\mathcal{N}}, \{\Omega^i\}_{i\in\mathcal{N}}, p_0, \gamma\rangle$, in which n agents $\mathcal{N} := \{1, \dots, n\}$ perform sequential actions with a state space $\mathcal{S}$. At time t, interactions are restricted to a *interaction graph* depicted by an adjacency matrix $\mathcal{A}_t \in \mathbb{B} \subseteq \{0,1\}^{n\times n}$, with $\mathcal{A}_t^{ij} = 1$ if and only if agent j interacts directly with agent i, and $\mathcal{A}_t^{ij} = 0$ otherwise. Being in state $s_t \in \mathcal{S}$, each agent $i \in \mathcal{N}$ executes an action $u_t^i \in \mathcal{U}^i$, forming a joint action $u_t \in \prod_{i\in\mathcal{N}} \mathcal{U}^i =: \mathcal{U}$ which induces a transition in the environment via the transition function $\mathcal{P} : \mathcal{S} \times \mathcal{U} \times \mathcal{S} \to [0,1]$, where p_0 is the distribution of the initial state. Correspondingly, agent i determines its individual reward r_t^i via $r^i : \mathcal{S} \times \mathcal{U} \to \mathbb{R}$, and receives a private observation $\omega_t^i := o^i(s_t, \mathcal{A}_t) \in \Omega^i$ correlated with the state and the interaction graph by $o^i : \mathcal{S} \times \mathbb{B} \to \Omega^i$. Each agent has a local stochastic policy $\pi_{\theta^i}^i(u_t^i|\omega_t^i)$ with parameters θ^i, specifying the probability of taking an action. Denoting $\theta^{-i} = (\theta^i)_{j\neq i}$, $u_t^{-i} = (u_t^j)_{j\neq i}$, $\omega_t^{-i} = (\omega_t^j)_{j\neq i}$, we let $\pi_{\theta^{-i}}^{-i}(u_t^{-i}|\omega_t^{-i})$ be a latent representation of the joint policy of all complementary agents of i. Letting $R_t^i(s_t, u_t^i, u_t^{-i}) = \sum_{\ell=0}^{\infty} \gamma^\ell r_{t+\ell}^i(s_t, u_t^i, u_t^{-i})$, each agent i is presumed to pursue the maximal cumulative reward [65]:

$$\eta^i(\pi_{\theta^i}^i, \pi_{\theta^{-i}}^{-i}) = \mathbb{E}_{(s_t, u_t^i, u_t^{-i})\sim\mathcal{P}, \pi_{\theta^i}^i, \pi_{\theta^{-i}}^{-i}} [R_t^i(s_t, u_t^i, u_t^{-i})] \qquad (1)$$

Similarly, let $V_{\pi_\theta}^i(s) = \mathbb{E}[R_t^i|s_t = s]$ and $Q_{\pi_\theta}^i(s,u) = \mathbb{E}[R_t^i|s_t = s, u_t = u]$ be the *local* state- and action-value (resp.) functions of agent i, where $\pi_\theta = (\pi_{\theta^i}^i, \pi_{\theta^{-i}}^{-i})$ denotes the joint policy.

3.2 Correlated Interaction Topology

Let $\rho : \mathbb{B} \times \mathcal{S} \to [0,1]$ be the *true* posterior distribution over adjacency matrices conditioned on the agents' *global* state. Du et al. [12] approximate ρ via learning a *centralized* graph reasoning policy ρ_φ with parameters φ given by a normalizing flow (See Appendix B [10]), and thus follow the *centralized training with decentralized execution* (CTDE) paradigm [39, 52]. Further, instead

of assuming the graph is *symmetric* and *undirected* [36, 62], we build a *directed* graph to allow each agent to *dynamically* select the agents for coordination. As such, [12] factorize the joint policy while assuming conditional independence of actions from different agents, i.e., $\pi_\theta(u^i, u^{-i}, \mathcal{A}|\omega) = \pi^i_{\theta^i}(u^i|\omega^i)\pi^{-i}_{\theta^{-i}}(u^{-i}|\omega^{-i})\rho_\varphi(\mathcal{A}|s)$. In Appendix A [10] we provide the following policy gradient:

$$\nabla_{\theta^i}\eta^i = \mathbb{E}[\log \rho_\varphi(\mathcal{A}|s)\nabla_{\theta^i}\log \pi^i_{\theta^i}(u^i|o^i(s,\mathcal{A}))\cdot \cdot \int_{u^{-i}} \pi^{-i}_{\theta^{-i}}(u^{-i}|o^{-i}(s,\mathcal{A}))Q^i(s,u^i,u^{-i})\mathrm{d}u^{-i}] \quad (2)$$

where $s \sim p, u^i \sim \pi^i_{\theta^i}, \mathcal{A} \sim \rho_\varphi$. Opposed to [12], we make the novel insight that (2) states that each agent should improve its policy toward the direction of its best response to other agents' strategies over *all* possible interactions. This indicates the vulnerability of π_θ's non-correlated factorization: *it ignores impacts of one agent's action on others, and their subsequent reactions*. For instance, consider a two-player zero-sum differential game, where two agents act in x and y with the reward functions defined by $(xy, -xy)$ and $\rho_\varphi(\mathcal{J}_2) = 1$ ($\mathcal{J}_2$ is the all-ones 2×2 matrix). Following the non-correlated policy, both agents are reinforced to trace *a cyclic trajectory that never converges to the equilibrium* [49, 71].

4 Graph Probabilistic Recursive Reasoning

As a remedy for the weakness exhibited by a non-correlated joint policy, we herein extend [71] to *graph probabilistic recursive reasoning* (**GrPR2**), which applies to GDec-POMDPs. Specifically, unlike previous works, each agent i takes an iterative best response to other agents' policies, over all possible interactions induced by the interaction graph. Thereby, we restate the joint policy at time t as:

$$\pi_\theta(u^i_t, u^{-i}_t, \mathcal{A}_t|s_t) = \pi^i_{\theta^i}(u^i_t|\omega^i_t)\pi^{-i}_{\theta^{-i}}(u^{-i}_t|\omega^{-i}_t, u^i_t)\rho_\varphi(\mathcal{A}_t|s_t) \quad (3)$$

where $\pi^{-i}_{\theta^{-i}}(u^{-i}_t|\omega^{-i}_t, u^i_t)$ represents other agents' consideration of agent i's action $u^i_t \sim \pi^i_{\theta^i}(\cdot|\omega^i_t)$, provided their own observations ω^{-i}_t induced by the interaction graph $\mathcal{A}_t \sim \rho_\varphi(\cdot|s_t)$. A *level-1* recursive scheme is formed, from both agent i's and other agents' perspectives. Full knowledge regarding the actual conditional policy $\pi^{-i}_{\theta^{-i}}$ is impractical, which can be approximated by a best-fit model $\psi^{-i}_{\phi^{-i}}(u^{-i}_t|s_t, \mathcal{A}_t, u^i_t)$ with parameters ϕ^{-i}. Hence, (3) is estimated by $\hat{\pi}^i_\theta$, after substituting $\pi^{-i}_{\theta^{-i}}$ with $\psi^{-i}_{\phi^{-i}}$. The learning task can thus be re-formulated as maximizing (1) with respect to θ^i, φ and ϕ^{-i}.

Under the GrPR2 settings, we provide a new graphical multi-agent policy gradient theorem (Subsection 4.1). By variational inference, we then capture the uncertainties of other agents' conditional policies for the sake of their approximation (Subsection 4.2).

4.1 GrPR2 – Policy Gradients

In Lemma 4.1, we establish the GrPR2 policy gradient for updating θ^i and φ, resp. (**See Appendix C for a detailed proof [10]**).

Lemma 4.1. $\nabla_{\theta^i}\eta^i \approx \mathbb{E}[\log \rho_\varphi(\mathcal{A}|s)\nabla_{\theta^i}\log \pi^i_{\theta^i}(u^i|o^i(s,\mathcal{A}))\cdot Q^i_{\psi^{-i}_{\phi^{-i}}}(s,u^i)]$ *and* $\nabla_\varphi\eta^i \approx \mathbb{E}_{s\sim p,\mathcal{A}\sim\rho_\varphi}[\nabla_\varphi \log \rho_\varphi(\mathcal{A}|s)\bar{Q}^i_{\psi^{-i}_{\phi^{-i}}}(s)]$, *where* $Q^i_{\psi^{-i}_{\phi^{-i}}}(s,u^i) := \int_{u^{-i}} \psi^{-i}_{\phi^{-i}}(u^{-i}|s,\mathcal{A},u^i)Q^i(s,u^i,u^{-i})\mathrm{d}u^{-i}$, $\bar{Q}^i_{\psi^{-i}_{\phi^{-i}}}(s) := \int_{u^i} Q^i_{\psi^{-i}_{\phi^{-i}}}(s,u^i)\mathrm{d}u^i$, *and* $s \sim p, u^i \sim \pi^i_{\theta^i}, \mathcal{A} \sim \rho_\varphi$.

Opposed to [12, 71], we note that the direction of the policy updates are guided by terms which shape the reward after considering the affect upon agent i's neighbors over all possible interactions.

4.2 Variational Inference of Agents' Policies

We infer $\psi^{-i}_{\phi^{-i}}$ via variational inference (VI) [32]. Let $\tau = (s_t, \mathcal{A}_t, u_t)_{t=1}^T$ be the trajectory of length $T > 0$. Assuming that the agent cannot influence the environment transition probability, each agent i approximates the true distribution $p(\tau)$ of τ being generated via:

$$\hat{p}_{\pi_{\theta^i}}(\tau) = p(s_1)\prod_{t=1}^{T} p(s_t|s_t, u^i_t, u^{-i}_t, \mathcal{A}_t)\hat{\pi}^i_\theta(u^i_t, u^{-i}_t, \mathcal{A}_t|s_t) \quad (4)$$

Accordingly, each agent i aims to find the best approximation of $\hat{\pi}^i_\theta$, which can be achieved via minimizing the KL-divergence, given as follows (See Appendix D.1 for details [10]):

$$D_{KL}(\hat{p}^i(\tau)||p(\tau)) = -\sum_{t=0}^{T}\mathbb{E}[r^i(s_t,u_t)+\mathbb{H}(\hat{\pi}^i_\theta(u^i_t,u^{-i}_t,\mathcal{A}_t|s_t))] \quad (5)$$

where $s_t, u_t, \mathcal{A}_t \sim \hat{p}^i$. The entropy term $\mathbb{H}(\cdot)$, conditioned on both the state and agent i's action, promotes the explorations for agent i's best response, other agents' estimated policy and interaction graphs. Minimizing (5) yields the following theorem.

Theorem 4.2. *Agent i's optimal Q-function, which minimizes* (5), *is* $Q^i_{\pi_{\theta^i}}(s,u^i) = \log\int_{u^{-i}}\exp(Q^i_{\pi_\theta}(s,u^i,u^{-i}))\mathrm{d}u^{-i}$, *and the respective optimal conditional policies of other agents is* $\psi^{-i}_{\phi^{-i}}(u^{-i}|s,\mathcal{A},u^i) = \frac{1}{Z}\exp(Q^i_{\pi_\theta}(s,u^i,u^{-i}) - Q^i_{\pi_{\theta^i}}(s,u^i))$ ***(See Appendix D.2 [10])***.

Theorem 4.2 yields that $\psi^{-i}_{\phi^{-i}}$ can be learned by minimizing the KL-divergence between it and the *advantage* function $\exp(Q^i_{\pi_\theta} - Q^i_{\pi_{\theta^i}})$. We define the GrPR2 soft Bellman evaluation operator by $\mathcal{T}Q^i_{\pi_\theta}(s,u^i,u^{-i}) = r^i(s,u^i,u^{-i}) + \gamma\mathbb{E}[Q^i_{\pi_{\theta^i}}(s',(u^i)')]$, where $s', (u^i)' \sim p, \pi_{\theta^i}$. In the following theorem, we prove the convergence under self-play when there is one equilibrium, which leads to a fixed-point iteration that resembles value iteration.

Theorem 4.3. *In a symmetric game with only one equilibrium and with a Q-function $Q_\star$ and a policy $\pi_\star$, $\mathcal{T}$ is a contraction mapping if the equilibrium is either of the following: (1) the global optimum, i.e., $\mathbb{E}_{\pi_\star}[Q^i] \geq \mathbb{E}_\pi[Q^i]$; (2) a saddle point, i.e., $\mathbb{E}_{\pi_\star}[Q^i] \geq \mathbb{E}_{\pi^i}\mathbb{E}_{\pi^{-i}_\star}[Q^i]$ or $\mathbb{E}_{\pi_\star}[Q^i] \geq \mathbb{E}_{\pi^i_\star}\mathbb{E}_{\pi^{-i}}[Q^i]$* ***(See Appendix D.3 [10])***.

5 Optimality in Cooperative MARL

In this section, we initially devise a variational lower bound on the likelihood of each agent's optimality in the *cooperative* MARL (CMARL) version of the GrPR2 settings (Subsection 5.1). Then, we discuss the manner in which policies are learned for its optimization. Finally, we propose an exact tabular Q-iteration method that holds convergence guarantees (Subsection 5.2).

5.1 The Variational Lower Bound of Optimality

Extending [68], we supply a variational lower bound on the likelihood of each agent's optimality in the *cooperative* MARL (CMARL) version of the GDec-POMDP settings. Formally, an *optimal* behavior in such settings stands for agent i *best responding* to other agents'

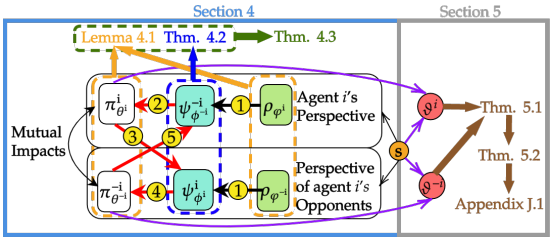


Figure 1: At state s, GrPR2 (3) decomposes the correlated interactions between agents. ①: agent i considers each possible adjacency matrix, as sampled from its graph reasoning policy ρ_{φ^i}, inducing a probability distribution over latent adjacency matrices (See Appendix B [10]). i's opponents perform the same procedure w.r.t. $\rho_{\varphi^{-i}}$. ②: agent i best responds, after accounting *all* possible topologies and potential impacts on actions executed by its opponents given its own action u^i. ③: agent i's behavior is the prior of other agents for them to learn their own impact on agent i. ④-⑤: as in ②-③, from the perspective of agent i's opponents. Iterating over those steps yields the recursive reasoning procedure compound by *GrPR2*, while affecting agent i's optimality ϑ^i. Under the setups of Sections 4–5, the lemmas and theorems specify how the optimal conditional policies are learned.

policy $\pi^{-i}_{\theta^{-i}}$. Letting $V^i(s;\pi^i,\pi^{-i}) := V^i_\pi(s)$ for any valid joint policy π, it can be defined as the policy $\pi^i_\star$ with $V^i(s;\pi^i_\star,\pi^{-i}_{\theta^{-i}}) \geq V^i(s;\pi^i_{\theta^i},\pi_{\theta^{-i}})$ for all valid $\pi^i_{\theta^i}$. Clearly, if all agents act in best response to others, the game reaches a Nash equilibrium [51].

Since GRP2 is a probabilistic model, we introduce a binary random variable $\vartheta^i_t \in \{0,1\}$, standing for the optimality of agent i's policy at time t. Opposed to previous work on recursive reasoning [68, 72], ϑ^i_t is not only dependent on the joint actions, but also on the interaction graph. Accordingly, we assume that given that other players act optimally, the higher the reward agent i receives, the higher the probability that agent i's current policy is optimal, i.e., $\mathbb{P}(\vartheta^i_t = 1|\vartheta^{-i}_t = 1, s_t, u^i_t, u^{-i}_t, \mathcal{A}^i_t) \propto \exp(r^i(s_t,u^i_t,u^{-i}_t))$. In the CMARL, if all agents play optimally, then agents receive the maximum reward that is also the Nash equilibrium (NE). Hence, from agent i's perspective, it aims to maximize $\mathbb{P}(\vartheta^i_{1:T}=1|\vartheta^{-i}_{1:T}=1)$, which is the probability of obtaining its maximum cumulative reward (i.e., best response) towards the NE along a trajectory $\tau^i := [(s_t,u^i_t,u^{-i}_t,\mathcal{A}^i_t)]^T_{t=1}$. Hereafter, we omit ϑ_t's value for brevity. As we assume no knowledge of the optimal policies and the model of the environment, we treat them as latent variables. Thus, we apply VI [32] with an auxiliary distribution $q^i(\tau^i|\vartheta^i_{1:T},\vartheta^{-i}_{1:T})$. q^i captures agent i's conditional policy on the current state, its reasoning on the interaction graph, other agents' actions and the beliefs regarding their actions. Accordingly, agent i will learn the optimal policy, while modelling its neighbors' actions. We regard the variational form $q^i(\tau^i|\vartheta^i_{1:T},\vartheta^{-i}_{1:T}) = \mathbb{P}(s_1)\prod^T_{t=2}\mathbb{P}(s_{t+1}|s_t,u_t,\mathcal{A}^i_t)\pi_\theta(u^i_t,u^{-i}_t,\mathcal{A}_t|s_t)$, where we refer to an alternate decomposition of the joint policy: $\pi_\theta(u^i_t,u^{-i}_t,\mathcal{A}_t|s_t) = \pi^i_{\theta^i}(u^i_t|\omega^i_t,u^{-i}_t)\psi^{-i}_{\phi^{-i}}(u^{-i}_t|s_t,\mathcal{A}_t)\rho_{\varphi^i}(\mathcal{A}_t|s_t)$. Intuitively, this decoupling dictates that agent i will *best respond* to all potential actions of its neighbors. Thereby, agent i's modelling $\psi^{-i}_{\phi^{-i}}$ of other agents' policies will approach its optimum ahead of agent i's own policy, and thus help agents to establish a mutual trust. We derive the following objective of agent i (where $s_t,u_t,\mathcal{A}_t \sim \hat{p}^i$):

$$\mathcal{J}^i(\hat{\pi}^i_\theta) := \sum_{t=0}^{T}\mathbb{E}\big[\mathbb{E}_{u_t}[r^i(s_t,u^i_t,u^{-i}_t) + \mathbb{H}(\pi^i_{\theta^i}(u^i_t|\omega^i_t,u^{-i}_t))] \tag{6}$$

$$-\mathbb{E}_{u^{-i}_t}[D_{KL}(\rho_{\varphi^i}(\mathcal{A}_t|s_t)||\mathbb{P}(\mathcal{A}_t|s_t))+ \tag{7}$$

$$+D_{KL}(\psi^{-i}_{\phi^{-i}}(u^{-i}_t|s_t,\mathcal{A}_t)||\mathbb{P}(u^{-i}_t|s_t,\mathcal{A}_t,\vartheta^{-i}_t))]\big] \tag{8}$$

which is a lower bound on the likelihood of $\log\mathbb{P}(\vartheta^i_{1:T}|\vartheta^{-i}_{1:T})$ (The derivation is deferred to Appendix F [10]). As we are solely focused on the case where $\vartheta^i_t = \vartheta^{-i}_t = 1$, they are omitted hereafter.

We make the observation that (6)-(8) resemble the maximum entropy objective in single-agent RL [26, 33, 69]. In the context of CMARL, we note that trajectories generated by policies updated by optimizing the entropy term in (6) also optimizes both $\psi^{-i}_{\phi^{-i}}$, $Q^i_{\pi_\theta}$ and ρ_{φ^i}. Indeed, without the regularizers (7)-(8), at iteration d, we: (1) fix $\pi^{i,d}_{\theta^i}$ to learn $\psi^{-i,d+1}_{\phi^{-i}}$ according to the interaction graphs generated by $\rho^d_{\varphi^i}$; (2) learn $\pi^{i,d+1}_{\theta^i}$ by the trajectories generated by $\psi^{-i,d+1}_{\phi^{-i}}$ and $\rho^d_{\varphi^i}$; and (3) fix $\pi^{i,d+1}_{\theta^i}$ and $\psi^{-i,d+1}_{\phi^{-i}}$ for learning $\rho^{d+1}_{\varphi^i}$. An EM-like scheme is induced, whose E-step ((1)-(2)) converges to the optimal conditional policies as proved in Theorem 5.1. We prove the convergence of the EM-like algorithm in Appendix H [10].

Yet, it is unrealistic to train such models since other agents have no access to agent i's policy, and thus their learning of agent i's policy might substantially differ from its true conditional policy $\pi^i_{\theta^i}$. Hence, best responding to policies that are far from the actual ones (from either agent i's or other agents' perspective) can lead to poor performance. Fortunately, (7)-(8) prevent agent i from attaining an unrealistic modelling of other agents' policies. (8) concerns the prior $\mathbb{P}(u^{-i}_t|s_t,\mathcal{A}_t)$ of optimal conditional policies of agents other than agent i. Via setting it to be the observed empirical distribution of other agents' actions given states, the KL-divergence penalizes for deviating from the empirical distribution. Similarly, the same applies to (7) that regards $\mathbb{P}(\mathcal{A}_t|s_t)$, which is the prior of adjacency matrices. We remark that this indicates the vulnerability of (5) opposed to (6)-(8), which is minimized for merely estimating other agents' policies without accounting for any sort of optimality. **See Figure 1 for the established reasoning structure w.r.t.** (3).

5.2 Graphical Multi-Agent Soft Q-Learning

As (6) resembles Soft Q-learning [26], we herein derive a GDec-POMDP version. For this sake, we initially add a weighting factor of α for the entropy term in (6), where (6) can be restored via $\alpha = 1$. Thus, we define the graphical soft (GS) Q-function and state-value function (resp.) as follows (where $[(s_{t+\ell},u^i_{t+\ell},u^{-i}_{t+\ell},\mathcal{A}^i_{t+\ell})]^\infty_{\ell=1} \sim q^i$):

$$Q^i_{GS}(s_t,u^i_t,u^{-i}_t) = r^i_t + \mathbb{E}\big[\sum_{\ell=1}^{\infty}\gamma^\ell(r_{t+\ell} + \alpha\mathbb{H}(\pi^i_{\theta^i}(u^i_{t+\ell}|\omega^i_{t+\ell},u^{-i}_{t+\ell})) \tag{9}$$

$$-D_{KL}(\psi^{-i}_{\phi^{-i}}(u^{-i}_{t+\ell}|s_{t+\ell},\mathcal{A}_{t+\ell})||\mathbb{P}(u^{-i}_{t+\ell}|s_{t+\ell},\mathcal{A}_{t+\ell})) \tag{10}$$

$$-D_{KL}(\rho_{\varphi^i}(\mathcal{A}_{t+\ell}|s_{t+\ell})||\mathbb{P}(\mathcal{A}_{t+\ell}|s_{t+\ell}))\big] \tag{11}$$

$$V^i_{GS}(s) = \log\sum_{u^{-i},u^i,\mathcal{A}}\mathbb{P}(u^{-i}|s,\mathcal{A})\mathbb{P}(\mathcal{A}|s)(\exp(\frac{1}{\alpha}Q^i_{GS}(s,u^i,u^{-i})))^\alpha \tag{12}$$

In the following theorem, we thus infer the optimal conditional policies of agent i and all complementary agents of i, from agent i's perspective (**See Appendix G.2 for a detailed proof [10]**).

Theorem 5.1. *From agent i's perspective, for* (6)*, the optimal conditional policy of agent i is $\pi^i_\star(u^i|\omega^i,u^{-i}) \propto \exp(\frac{1}{\alpha}Q^i_{GS}(s_t,u^i,u^{-i}))$ and the optimal conditional policies for all complementary agents of i are $\psi^i_\star(u^{-i}|s,\mathcal{A},u^i) \propto \mathbb{P}(u^{-i}|s,\mathcal{A})\mathbb{P}(\mathcal{A}|s)(\Sigma_{u^i}\exp(\frac{1}{\alpha}Q^i_{GS}(s_t,u^i,u^{-i})))^\alpha$.*

For learning the GS Q-function, we observe it satisfies the *graphical multi-agent soft Bellman equation* given by $Q^i_{GS}(s_t,u^i_t,u^{-i}_t) = r^i_t + \gamma\mathbb{E}_{s_{t+1}}[V^i_{GS}(s_{t+1})]$. Subsequently, we prove that the induced fixed-point iteration converges to the optimal GS action- and state-value functions under certain circumstances as in Theorem 4.3.

Theorem 5.2. *In a symmetric game with only one equilibrium (meeting one of the conditions in Theorem 4.3), and with a Q-function $Q_\star$, a state-value function $V_\star$ and a policy $\pi_\star$, assume $Q^i_{GS}, V^i_{GS}, Q_\star, V_\star < \infty$. Then, the fixed-point iteration $Q^i_{GS}(s_t,u^i_t,u^{-i}_t) = r^i_t + \gamma\mathbb{E}_{s_{t+1}}[V^i_{GS}(s_{t+1})]$, where $V^i_{GS}(s_{t+1})$ is as in* (12)*, converges to $Q_\star$ and $V_\star$, respectively.*

Proof. For learning Q^i_{GS}, we observe that it satisfies the *graphical multi-agent soft Bellman equation* given by $Q^{\pi^i,\psi^{-i},\rho^i}_{GS}(s_t,u^i_t,u^{-i}_t) = r^i_t + \gamma\mathbb{E}_{s_{t+1}}[V^{\pi^i,\psi^{-i},\rho^i}_{GS}(s_{t+1})]$. We define the graphical soft Bellman evaluation operator as follows: $\mathcal{T}_{GS}Q^i(s,u^i,u^{-i}) = r^i(s,u^i,u^{-i}) + \gamma\mathbb{E}_{s'}[\log\sum_{\tilde{u}^{-i},\tilde{u}^i,\mathcal{A}}\mathbb{P}(\tilde{u}^{-i}|s,\mathcal{A})\cdot\mathbb{P}(\mathcal{A}|s)(\exp(\frac{1}{\alpha}Q^i(s',\tilde{u}^i,\tilde{u}^{-i})))^\alpha]$. The proof follows from arguments similar to Theorem 4.3. □

Remark 1. *The EM-like scheme in Subsection 5.1 and Theorems 5.1-5.2 guarantee monotonic increase in the **probability** that ψ^{-i} is optimal. By acting optimally to the converged opponent model, we recover agent i's optimal policy, but not the optimum in the game.*

6 Hierarchical Graph Recursive Reasoning

GrPR2 operates as a level-1 recursive reasoning, without accounting for different *hierarchical* levels of rationality. Thus, following [72], we propose **GrPR2-L** which deepens the recursion to level-k ($k \geq 2$), and extends the incorporation of *Level-k models* into games with incomplete information [31] to the context of *graphical games* [34]. Formally, agent i at level k assumes that other agents are at level $k-1$ and then best responds by integrating over all possible interactions induced by the interaction graph and best responses from lower-level agents to agent i of level $k-2$:

$$\pi^i_k(u^i_k|\omega^i) \propto \int_{\mathcal{A}}\rho_{\varphi^i}(\mathcal{A}|s)\int_{u^{-i}_{k-1}}\pi^i_k(u^i_k|\omega^i,u^{-i}_{k-1})\cdot \int_{u^i_{k-2}}\psi^{-i}_{k-1}(u^{-i}_{k-1}|s,\mathcal{A},u^i_{k-2})\pi^i_{k-2}(u^i_{k-2}|\omega^i)\mathrm{d}u^i_{k-2}\mathrm{d}u^i_{k-1}\mathrm{d}\mathcal{A} \tag{13}$$

where the subscript stands for the level of thinking and π^i_0 is uniformly distributed. Agent i perceives that others will best respond to its own fictitious action u^i_{k-2} at level $k-2$ via $\psi^{-i}_{k-1}(u^{-i}_{k-1}|s,\mathcal{A}) = \int_{u^i_{k-2}}\psi^{-i}_{k-1}(u^{-i}_{k-1}|s,\mathcal{A},u^i_{k-2})\pi^i_{k-2}(u^i_{k-2}|\omega^i)\mathrm{d}u^i_{k-2}$.

Following the *cognitive hierarchy* (CH) model [5, 72], we propose *Cognitive Hierarchy GrPR2* (**GrPR2-CH**), which lets each level-k player best respond to a *mixture* of strictly lower levels in the hierarchy, induced by truncation up to level $k-1$ from the underlying level distribution. Formally, let $f = (f_h)_{h\geq 0}$ be a distribution over $\mathbb{N}$ which represents the hierarchy of levels. The probability that a k-level player assigns independently for each of the other players to belong to the h-level is $g_k(h) = f_h/(\sum_{m=0}^k f_m)$. For a k-level player, we are now capable of mixing all k levels of thinking $\{\hat{\pi}^i_m\}^k_{m=0}$ into its belief about other agents at lower levels by $\bar{\pi}^i_k(u^i_{0:k}|\omega^i) = \sum_{h=0}^k g_k(h)\hat{\pi}^i_h(u^i_h|\omega^i,u^{-i}_{0:h-1})$, where $\hat{\pi}^i_0(u^i_0|\omega^i) := \hat{\pi}^i_h(u^i_h|\omega^i,u^{-i}_{0:-1})$. As in [72], we deduce **GrPR2-M** by choosing $f_h = \frac{e^{-\lambda}\lambda^h}{h!}$ to be a Poisson distribution, with a mean of λ. As in TD-λ [67], λ acts as a hyperparameter.

6.1 Optimal Hierarchical Level of Reasoning

In this section, we prove that level-3 reasoning is the optimal hierarchical level, maximizing each agent's expected return. **See Appendix I for extensive details [10]**. From the perspective of each k_1-level agent i and a lower level $2 \leq k_2 < k_1$, we seek to find an upper bound $C \geq 0$ on the discrepancy between the expected return $\eta^i(\bar{\pi}^i_{k_1},\bar{\psi}^{-i}_{k_1-1})$, incurred by executing agent i's k_1-level reasoning, and the expected return $\eta^i(\bar{\pi}^i_{k_2},\bar{\psi}^{-i}_{k_2-1})$ resulting from executing agent i's k_2-level reasoning, i.e., $|\eta^i(\bar{\pi}^i_{k_1},\bar{\psi}^{-i}_{k_1-1}) - \eta^i(\bar{\pi}^i_{k_2},\bar{\psi}^{-i}_{k_2-1})| \leq C$. If the model is improved by at least C, we can guarantee improvement for varying cognitive levels. Thus, the following lemma depicts a general discrepancy bound on the expected returns using policies of different cognitive levels of reasoning by agent i.

Lemma 6.1. *Assume that the agent i's reward is bounded by $r^i_{\max} = \max_{s,u^i,u^{-i}} r^i(s,u^i,u^{-i})$, and let the expected transition distribution be bounded by $M \geq 0$. Then, the discrepancy bound on the expected returns is expressed by $r^i_{\max}M\sum_{h_1=k_2+1}^{k_1}\sum_{h_2=k_2}^{k_1-1}|g_{k_1}(h_1)g_{k_1-1}(h_2) - g_{k_2}(h_1)g_{k_2-1}(h_2)| =: r^i_{\max}M\cdot C(k_1,k_2,\lambda)$ (**See Appendix I.2 [10]**).*

Lemma 6.1 indicates that the key influence on the monotonic improvement in expected returns between an agent's cognitive levels is *the difference between the beliefs they induce regarding the distribution of other agents over all lower level in the hierarchy*. Under GrPR2-CH, in the following theorem we deal with maximizing the *discrepancy bounds ratio* $\hat{C}(k,k-1,\lambda) := \frac{C(k+1,k,\lambda)}{C(k,k-1,\lambda)}$ w.r.t. λ ($k \geq 2$), as it will illustrate the best monotonic improvement that can be guaranteed under a specific choice of parameters.

Theorem 6.2. *$k = 3, 4$ maximize $\hat{C}(k,k-1,\lambda)$ w.r.t. λ, i.e., achieve the **maximal** discrepancy bounds ratios (**See Appendix I.3 [10]**).*

A natural question is: *which one is the optimal level - 3 or 4?* Intuitively, for $k = 4$, the approximated conditional policy of other agents is used *twice*, thus amplifying the variance of each agent's k-level policy. Further, as illustrated by the following theorem, the mixing of hierarchical policies may introduce a bias into the learned policy that depends on λ and k. Formally, the bias is induced by the difference between the mixed policy and the (potentially locally) optimal policy at convergence (**See Appendix I.3.2 [10]**).

Theorem 6.3. *Let $\bar{\pi}^i_{k,(d)}$ and $\{\hat{\pi}^i_{m,(d)}\}^k_{m=0}$ be the d^{th} updated k-level conditional policies of agent i under GrPR2-M and GrPR2-L (resp.), where $\bar{\pi}^i_{k,\star}$ and $\{\hat{\pi}^i_{m,\star}\}^k_{m=0}$ denote the optimal policies at convergence (resp.). Let $\hat{\pi}^i_{0,(d)} \equiv \hat{\pi}^i_0$ be uniformly distributed. Letting $D_{TV}(\cdot,\cdot)$ denote the total variational distance between two probability measures (i.e., policies), agent i's k-level policy bias ($k \geq 2$),*

given by $D_{TV}(\bar{\pi}^i_{k,(d)}, \pi^i_\star)$*, is bounded as by:* $D_{TV}(\bar{\pi}^i_{k,(d)}, \bar{\pi}^i_{k,\star}) \geq D_{TV}(\hat{\pi}^i_0, \bar{\pi}^i_{k,\star}) - \sum_{h=1}^{k} g_k(h) D_{TV}(\hat{\pi}^i_0, \hat{\pi}^i_{h,(d)})$*, and:* $\forall \varepsilon > 0 \exists \tilde{d} \in \mathbb{N} : D_{TV}(\bar{\pi}^i_{k,(d)}, \bar{\pi}^i_{k,\star}) < \varepsilon \sum_{h=0}^{k} g_k(h) \forall d \geq \tilde{d}$*. Then, the mixture introduces a bias which increases with* λ *and* $D_{TV}(\hat{\pi}^i_0, \bar{\pi}^i_{k,\star})$*.*

As observed in the proof of Theorem 6.2, a higher value of λ is required for attaining the maximal value of $\hat{C}(k, k-1, \lambda)$ as k increases, and thus we summarize that **level**-3 **reasoning is the optimal hierarchical level**. We remark that this result is empirically supported: in Appendix F of [72], we observe that convergence to an equilibrium in *Keynes Beauty Contest* [35] occurs when the level of reasoning k ranges between 1 to 3, yet for $k = 4$ it **fails**.

6.1.1 ***On the Intuition behind Theorem 6.3.*** By Theorem 6.3, each agent's explorable region of the state space grows with the decrease in λ, k (and vice versa). Thus, a higher value of both λ, k constrains policy search near the optimal policy more heavily. Further, the difference between $\hat{\pi}^i_0$ and the optimal k-level policy at convergence $\bar{\pi}^i_{k,\star}$ ($D_{TV}(\hat{\pi}^i_0, \bar{\pi}^i_{k,\star})$) *may* bias the final policy, depending on λ, k. This implies the **weak spot** of the classical Level-k and CH models: **as** $\hat{\pi}^i_0$ **is uniformly distributed, it *may* be far from the optimal policy, thus biasing the final policy, depending on the explorable region, which grows as** λ, k **decrease**. We conclude that the choice of the underlying 0-level reasoning $\hat{\pi}^i_0$ plays a critical role in the learning of the optimal policy. That is, the bias incurred by the 0-level reasoning is reduced as $\hat{\pi}^i_0$ is closer to the optimal policy. Additionally, if the optimal trajectory resides within the explorable region, then the corresponding optimal policy can be learned. Otherwise, the policy will remain suboptimal. Hence, ***smaller*** λ ***and*** k ***(and thus a larger explorable region) will increase the possibility of reaching the optimal policy.***

7 Graphical Multi-Agent Soft Actor-Critic

For practical implementation in complex domains, we propose the *GrPR2 Actor-Critic* (**GrPR2-AC**), consisting a model-free approximation of the tabular algorithm proposed earlier in Subsection 5.2 that follows the learning scheme presented in Subsection 5 and [72]. **See Appendix J for extensive details and pseudo-codes [10]**. As function estimators, we use neural networks (NNs). As in [12], ρ_{φ^i} is represented by a normalizing flow (See Appendix B [10]). For policy evaluation, each agent rolls both its graph reasoning policy and approximated conditional policies recursively up to level k following Section 6. *Soft learning* [27] is then applied to maximize $\mathcal{J}^i(\hat{\pi}^i_\theta)$ in (6): ξ^i is updated via minimizing the soft Bellman residual $\mathcal{J}_{Q^i}(\xi^i) = \mathbb{E}_{\mathcal{D}^i}[\frac{1}{2}(Q^i_{\xi^i}(s_t, u^i_t, u^{-i}_t) - r^i(s_t, u^i_t, u^{-i}_t) - \gamma \mathbb{E}_{s_{t+1}}[\bar{V}^i(s_{t+1})])^2]$, where $\mathcal{D}^i$ is the replay buffer. Noting that the entropy and KL-divergence terms in (6)-(8) are the expansion of $\mathbb{H}(\hat{\pi}^i_\theta(u^i_t, u^{-i}_t, \mathcal{A}_t|s_t))$, we infer $\mathcal{J}^i(\hat{\pi}^i_\theta) = -D_{KL}(\hat{p}^i(\tau)||p(\tau))$ by (5). Thus, letting $Q^i_{\xi^i}(s, u^i) = \log \int_{\mathcal{A}} \rho^i(\mathcal{A}|s) \int_{u^{-i}} \psi^{-i}_{\phi^{-i}}(u^{-i}|s, u^{-i}) \cdot \exp(Q^i_{\xi^i}(s, u^i, u^{-i}))\mathrm{d}u^{-i}\mathrm{d}\mathcal{A}$, which marginalizes the joint Q-function via the estimated opponent model, the value function of the k-level policy π^i_k is $\bar{V}^i(s_{t+1}) = \mathbb{E}_{u^i_k \sim \pi^i_k}[Q^i_{\xi^i}(s_t, u^i_t) - \log \pi^i_k(u^i_k|\omega^i)]$. Further, the optimal opponent model $\psi^{-i}_{\phi^{-i}}$ follows Theorem 4.2, which yields that ϕ^{-i} can be updated by minimizing $\mathcal{J}_{\psi^{-i}}(\phi^{-i}) := \mathbb{E}_{s,u^i \sim \mathcal{D}^i}[D_{KL}(\psi^{-i}_{\phi^{-i}}(\cdot|s, u^i)|| \exp(Q^i_{\xi^i}(s, u^i, \cdot) - Q^i_{\xi^i}(s_t, u^i_t)))]$. Estimation of $Q^i_{\xi^i}(s_t, u^i_t, u^{-i})$ and $Q^i_{\xi^i}(s_t, u^i_t)$ are maintained separately for robust training, and ϕ^{-i}'s gradient is computed by SVGD [46]. Thereby, θ^i can be learned by improving towards the current soft Q-function $Q^i_{\xi^i}(s_t, u^i_t)$ by minimizing the KL-divergence $\mathcal{J}_{\pi^i_k}(\theta^i) = \mathbb{E}\left[D_{KL}\left(\pi^i_k(\cdot|\omega^i)\middle\| \frac{\exp(Q^i_{\xi^i}(s, \cdot, u^{-i}))}{\sum_{\tilde{u}^i} Q^i_{\xi^i}(s_t, \tilde{u}^i, u^{-i}))}\right)\right]$, with $s_t, \mathcal{A}_t \sim \mathcal{D}^i, \hat{u}^{-i}_{t+1} \sim \psi^{-i}$. The optimal $\pi^i_{\theta^i}$ and $\psi^{-i}_{\phi^{-i}}$ are thus recovered as in [27], avoiding the intractable inferences in (48)-(49). By the reparameterization trick $u^i_t = f_{\theta^i}(\epsilon^i; \omega^i)$ ($\epsilon \sim \mathcal{N}(0, \mathbb{I})$), $\mathcal{J}_{\pi^i_k}(\theta^i) = \mathbb{E}_{s,\mathcal{A},u^i_k,\epsilon}[\log \pi^i_{\theta^i,k}(f_{\theta^i}(\epsilon^i; \omega^i)|\omega^i) - Q^i_{\xi^i}(s, f_{\theta^i}(\epsilon^i; \omega^i))]$. Hence, agent i learns the best response policy by considering all possible actions of opponents at lower levels over all possible interactions, as compound by $Q^i_{\xi^i}(s_t, u^i_t)$, and $\partial \mathcal{J}_{\pi^i_k}/\partial \theta^i$ thus propagates from all higher levels during training. To afford the implementation, we follow the compromises made by [72], as detailed in Appendix J.2 [10]. For reducing the variance of the stochastic gradient in Lemma 4.1 (used for improving ρ_{φ^i}) and its inconsistency w.r.t other policies' update, it is estimated by off-policy importance sampling [12].

8 Experiments

In this section, we evaluate GrPR2 on the high-dimensional task of *Cooperative Navigation* in the Particle World environment [47], with n agents of size 0.05 and n landmarks. In this task, n agents must cooperate through physical actions to reach a set of n landmarks. Agents observe the relative positions of nearest agents and landmarks, and are collectively rewarded based on the proximity of any agent to each landmark, i.e., the agents have to "cover" all of the landmarks. Further, the agents occupy significant physical space and are penalized when colliding with each other. Our agents learn to infer the landmark they must cover, and move there while avoiding other agents. Though the environment holds a continuous state space, agents' actions space is discrete, and given by all possible directions of movement for each agent {up, down, left, right, stay}. Given an interaction graph, we augment this task for enabling local information sharing between neighbors, as outlined subsequently.

8.1 Experimental Setup

GrPR2-CH *[9].* Each agent queries its neighbors about their observations, and incorporates that information into its choice of actions and estimated critic. Such basic paradigm was chosen for depicting that our methods are effective in combining local observations into the learning process in environments where the global state may not be available, without any additional information.

Attentive GrPR2 (GrPR2-A) *[8].* We further implement an extension of *FlowComm* [12], supporting level-1 recursive reasoning *with communication.* FlowComm embeds ρ_φ into *MAAC* [29], which trains decentralized policies in multiagent settings, using *centrally* computed critics that share an attention mechanism, selecting relevant information for each agent. At state s, letting $\mathcal{A} \sim \rho_\varphi(\cdot|s)$ and v^i be the *encoding* of agent i's observations, $m^i = \sum_{j=1}^{n} \mathcal{A}_{ij} v^j$ is the *message* i can receive and $\omega^i = m^i \cup v^i$. The above methods can easily be extended to include additional information.

Table 1: Maximum return at the end of the training phase.

n	**GrPR2-CH ($k=2$)**	GrPR2-CH ($k=3$)	GrPR2-L ($k=2$)
4	**-47.803**	-60.737	-58.77
8	**4046.613**	4035.438	4033.093

n	GrPR2-L ($k=3$)	DDPG	OM	ToM	MADDPG
4	-57.421	-61.404	-57.631	-58.649	-56.75412
8	4030.009	4031.2	4034.01	4037.029	4038.256

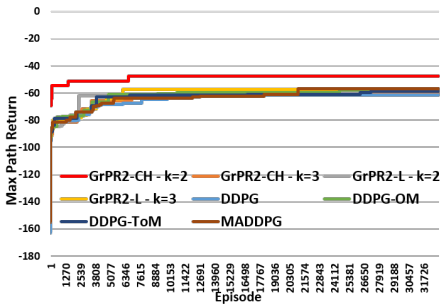
Hyperparameter Settings. Under ***GrPR2-CH***, let k be the highest level of reasoning. As it was previously mentioned, under their GR2-L framework, Wen et al. [72] empirically observe in their ablation study that convergence to an equilibrium occurs when the level of reasoning k ranges between 1 to 3, yet for $k = 4$ it fails. Thus, we adopt $k \in \{2, 3\}$. Further, they demonstrate that a Poisson mean of $\lambda = 1.5$ leads to the best performance in the cooperative navigation task, and we thereby follow their empirical results. The Q-values are updated using an Adam optimizer with learning rate 10^{-4}. The DDPG policy also utilizes an Adam optimizer with a learning rate of 10^{-4}. The methods use a replay pool of size $100k$. Training does not start until the replay buffer has at least $1k$ samples. A batch size of 64 is used. All the policies and Q-functions are modeled by an MLP with 2 hidden layers followed by *ReLU* activation. Through the actor-critic scheme GrPR2-AC, we set the exploration noise to 0.1 in the first $1k$ steps. The annealing parameter is decayed in linear scheme with training step grows to balance the exploration. Deterministic policies are implemented with an additional OU Noise to improve exploration with parameters $\theta = 0.15$ and $\sigma = 0.3$. We update the target parameters softly by setting the target smoothing coefficient to 0.001. We train with 6 random seeds for all environments. For the cooperative navigation task, all the models are trained up to $300k$ steps with a maximum episode length of 30. All hidden layers have 100 hidden units. The update interval is 4.

Under ***GrPR2-A***, agents' local observations are first encoded, and then concatenated with the received message. An LSTM layer encodes the message and observations before feeding them into two fully-connected NNs, producing policies. Critics use three-layered fully-connected NNs. All hidden layers have 64 hidden units. Unlike GrPR2-CH, we use a replay buffer size of 10^6. The learning rate for actor and critic are 10^{-3}. Other hyperparameters are as above.

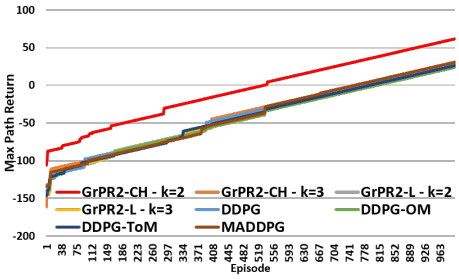
For the graph reasoning policy, we follow [12] by choosing L in Eq.(7) of Appendix B [10] to be a logical gate for reversible additions and subtractions, where g_{φ^i} is implemented by three fully connected layers followed by a tanh activation. Further, we use *four* coupling layers to make sure the elements in $\mathcal{A}$ are dependent on each other and we allow parameter sharing across coupling layers.

8.2 Results – GrPR2-CH

We compare GrPR2 with DDPG independent learner [45], MADDPG [47], which we regard as level-0 reasoning. For fair comparison, we also include the level-0 model of opponent modeling [28] by augmenting DDPG with an opponent module (DDPG-OM) that predicts opponents' behaviors in future states, and a level-1 Theory-of-Mind model [56] that captures the dependency of an agent's policy on



(a) Max Path Return – $n = 4$



(b) Max Path Return – $n = 8$

Figure 2: Figures 2a and 2b report maximum return attained up to each episode for $n = 4$ and 8 (respectively).

opponents' mental states (DDPG-ToM). Unless stated otherwise, each method conducts *centralized training with decentralized execution* (CTDE) [39, 52], with a *centralized* graph reasoning policy, but *decentralized* policies and critics. GrPR2-CH can employ a *decentralized* ρ_{φ^i}, yet this choice was made for fair comparison.

Figure 2a illustrates the maximum return attained by each method along the training phase for a set of $n = 4$ agents. It can be clearly observed that GrPR2-CH with a hierarchical level of $k = 2$ substantially outperforms all other methods. Specifically, during the very first episode, it already reaches a remarkably higher reward of -69.462 compared to the other baselines, whose rewards range from -162.969 to -141.898. Further, not only that GrPR2-CH with $k = 2$ converges faster than all other methods, it also reaches the highest maximum return at the end of the training phase. As further emphasized by Table 1, it is worthy of noting that GrPR2-CH with $k = 2$ exhibits a higher maximum reward across the learning phase than GrPR2-CH with $k = 3$. This can be theoretically justified by the intuition behind Theorem 6.3 provided in Subsection 6.1.1: *a smaller value of k will induce a larger explorable region, thus increasing the possibility of reaching the optimal policy.* Further, one may argue that this empirical result inconsistent with Theorem 6.2. Yet, recall that the expected transition distribution is assumed to be bounded by Lemma 6.1, whose implication is Theorem 6.2. Such an assumption is not necessarily satisfied in high-dimensional domains such as the Particle World, which thus explains the mentioned empirical result. However, aligned with Theorem 6.2, we note that GrPR2-L with $k = 3$ overtakes all other models while being competitive with DDPG-OM, where GrPR2-CH with $k = 2$ and MADDPG are the exceptions. GrPR2-CH and GrPR2-L with

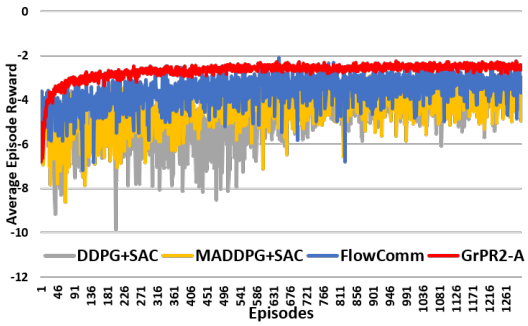
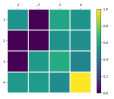


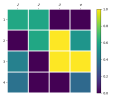
Figure 3: The average episode reward along the training phase with $n = 8$ and heterogeneous graphs.

Table 2: Mean rewards with their standard deviations.

GrPR2-A	FlowComm	DDPG+SAC	MADDPG+SAC
-2.698 ± 0.412	-3.601 ± 0.653	-4.747 ± 1.002	-4.419 ± 0.795



(a) Attentive GrPR2



(b) FlowComm

Figure 4: Heatmaps of the average adjacency matrix generated across $3k$ episodes via (4a) GrPR2-A and (4b) FlowComm.

$k = 3$ differ due to Theorem 6.3, as the mixture of polices under GrPR2-CH *may* introduce policy bias, whereas GrPR2-L does not. For the comparison of GrPR2-L with $k = 3$ relative to DDPG-OM and MADDPG, both algorithms account for the opponents, and can thus converge to some local optima, which is not necessarily optimal as exhibited by GrPR2-CH with $k = 2$. However, DDPG fails due to its inherent defect as an independent learning approach.

To illustrate the scalabiltiy of our methods with the growth in the number of agents, Figure 2b provides an additional comparison for $n = 8$. Note that even after the number of agents in *doubled*, our methods preserve their performance attributes. That is, GrPR2-CH with $k = 2$ still significantly surpasses all other baselines, and attains the highest reward of -106.156 during the learning phase's inception, which is followed with the fastest convergence rate to the best maximum return as seen in Table 1.

8.3 Results – Attentive GrPR2

We compare GrPR2-A with FlowComm [12], as well as DDPG and MADDPG with a Soft Actor-Critic (SAC) for fair comparison. Clearly, they are all regarded as level-0 reasoning. For amplifying the superiority of GrPR2-A, we evaluate these methods on [12]'s extension of the cooperative navigation task to a *heterogeneous* communication task. That is, for $n = 8$, the third agent has a larger field of vision compared to the other agents. This allows us to simulate heterogeneity in hierarchical level of thinking, as agent 3 becomes more sophisticated compared to other agents.

Figure 3 compares the average episode reward incurred by each model during the training phase. We stress the remarkable performance of GrPR2-A opposed to other baselines, in terms of attaining a better average reward. Additionally, GrPR2-A exhibits the best consistency in the average episode reward. Table 2 emphasizes this result, as GrPR2-A holds both the *highest* mean reward and the *lowest* standard deviation. Theoretically, an agent operating with a hierarchical recursive reasoning process enables it to selectively interact with opponents of varying and less-sophisticated levels of reasoning and best respond to their actions. Thus, collisions are avoided at maximum. Such abilities are of vital importance, especially in *heterogeneous* communication tasks as the one we regard. This also amplifies the superiority of GrPR2-A compared to the other models, which do not employ such an ability. For stressing the above, for $n = 4$, Figures 4a and 4b supply heatmaps of the average adjacency matrix generated across $3k$ episodes via GrPR2-A and FlowComm (resp.). We note the higher sparsity of the GrPR2-A's heatmap compared to FlowComm's heatmap. We compute the sparsity of the adjacency matrix throughout the learning phase by reporting its mean and standard deviation. For GrPR2-A, we have a value of 0.1875 ± 0.3903, compared to a value of 0.375 ± 0.4841 for FlowComm. GrPR2-A thus learns a more selective prioritization of interactions with other agents, which treats agents' need to avoid collisions by adapting their communication targets dynamically due to the constant change of their physical locations.

9 Conclusion and Future Work

Following humans' inborn recursive reasoning ability, we presented a novel perspective on opponent modeling in domains with only local interactions via **GrPR2-CH**, which enables modelling agents with different *hierarchical* levels of thinking. Unlike previous work on recursive reasoning, level-1 agents iteratively best-respond to other agents' policies *over all possible local interactions*. Agents' policies are approximated via variational inference for capturing their uncertainties, and we proved that an induced variant of Q-learning converges under self-play when there exists only one Nash equilibrium. In *cooperative* MARL, we further devised a variational lower bound on the likelihood of each agent's optimality. We observed that optimizing the resulting objective prevents each agent from attaining an unrealistic modelling of others, and yields an exact tabular Q-iteration method that holds convergence guarantees. After deepening the recursion to level-k, we then proved that: (1) *level-3 reasoning is the optimal hierarchical level*, maximizing each agent's expected return; and (2) *the weak spot of the classical CH models is that 0-level is uniformly distributed*, as it *may* introduce policy bias. Finally, we proposed a practical actor-critic scheme, and illustrated it superiority compared to strong MARL baselines. Naturally, one may argue about the existence of a Perfect Bayesian Equilibrium in the dynamic game induced by GrPR2. Unlike GR2 [72], the dynamic graph imposed by our framework poses an additional challenge, for which future work requires further deepening.

Acknowledgments

This research was funded in part by ISF grant 2306/18.

References

[1] Daniel Barkoczi and Mirta Galesic. 2016. Social learning strategies modify the effect of network structure on group performance. *Nature communications* 7, 1 (2016), 1–8.

[2] Wendelin Böhmer, Vitaly Kurin, and Shimon Whiteson. 2020. Deep coordination graphs. (2020), 980–991.

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

[4] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 2 (2008), 156–172.

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

[6] Colin F Camerer, Teck-Hua Ho, and Juin Kuan Chong. 2015. A psychological approach to strategic thinking in games. *Current Opinion in Behavioral Sciences* 3 (2015), 157–162.

[7] Yinlam Chow, Brandon Cui, MoonKyung Ryu, and Mohammad Ghavamzadeh. 2020. Variational model-based policy optimization. *arXiv preprint arXiv:2006.05443* (2020).

[8] Saar Cohen and Noa Agmon. 2021. Optimizing Multi-Agent Coordination via Hierarchical Graph Probabilistic Recursive Reasoning – Implementation of Attentive Graph Probabilistic Reasoning (GrPR2-A). https://github.com/saarcohen30/GrPR2-A.

[9] Saar Cohen and Noa Agmon. 2021. Optimizing Multi-Agent Coordination via Hierarchical Graph Probabilistic Recursive Reasoning – Implementation of the Hierarchical Graph Probabilistic Reasoning Frameworks. https://github.com/saarcohen30/GrPR2-CH.

[10] Saar Cohen and Noa Agmon. 2022. Optimizing Multi-Agent Coordination via Hierarchical Graph Probabilistic Recursive Reasoning – Supplementary Material. https://www.cs.biu.ac.il/~agmon/AAMAS22Sup.pdf.

[11] Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. 2013. How much does it help to know what she knows you know? An agent-based simulation study. *Artificial Intelligence* 199 (2013), 67–92.

[12] Yali Du, Bo Liu, Vincent Moens, Ziqi Liu, Zhicheng Ren, Jun Wang, Xu Chen, and Haifeng Zhang. 2021. Learning Correlated Communication Topology in Multi-Agent Reinforcement learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 456–464.

[13] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. 2016. Benchmarking deep reinforcement learning for continuous control. *International conference on machine learning* (2016), 1329–1338.

[14] Carlos Florensa, Yan Duan, and Pieter Abbeel. 2017. Stochastic neural networks for hierarchical reinforcement learning. *arXiv preprint arXiv:1704.03012* (2017).

[15] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (2018).

[16] Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip HS Torr, Pushmeet Kohli, and Shimon Whiteson. 2017. Stabilising experience replay for deep multi-agent reinforcement learning. 70 (2017), 1146–1155.

[17] Jakob N Foerster, Yannis M Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with Deep multi-agent reinforcement learning. (2016), 2145–2153.

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

[19] Piotr J Gmytrasiewicz and Edmund H Durfee. 2000. Rational coordination in multi-agent environments. *Autonomous Agents and Multi-Agent Systems* 3, 4 (2000), 319–350.

[20] Alvin I Goldman et al. 2012. Theory of Mind. (2012), 402—-424.

[21] Matthew Gombolay, Anna Bair, Cindy Huang, and Julie Shah. 2017. Computational design of mixed-initiative human–robot teaming that considers human factors: situational awareness, workload, and workflow preferences. *The International journal of robotics research* 36, 5-7 (2017), 597–617.

[22] Matthew Gombolay, Reed Jensen, Jessica Stigile, Toni Golen, Neel Shah, Sung-Hyun Son, and Julie Shah. 2018. Human-machine collaborative optimization via apprenticeship scheduling. *Journal of Artificial Intelligence Research* 63 (2018), 1–49.

[23] Adam S Goodie, Prashant Doshi, and Diana L Young. 2012. Levels of theory-of-mind reasoning in competitive games. *Journal of Behavioral Decision Making* 25, 1 (2012), 95–108.

[24] Alison Gopnik and Henry M Wellman. 1992. Why the Child's Theory of Mind Really Is a Theory. *Mind and Language* 7, 1-2 (1992).

[25] Carlos Guestrin, Daphne Koller, and Ronald Parr. 2001. Multiagent Planning with Factored MDPs. 1 (2001), 1523–1530.

[26] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. 2017. Reinforcement learning with deep energy-based policies. *Proceedings of the 34th International Conference on Machine Learning-Volume 70* 70 (2017), 1352–1361.

[27] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. (2018), 1861–1870.

[28] He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. 2016. Opponent modeling in deep reinforcement learning. (2016), 1804–1813.

[29] Shariq Iqbal and Fei Sha. 2019. Actor-attention-critic for multi-agent reinforcement learning. (2019), 2961–2970.

[30] Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu. 2020. Graph Convolutional Reinforcement Learning. (2020). https://openreview.net/forum?id=HkxdQkSYDB

[31] David Jimenez-Gomez. 2019. False Consensus in Games: Embedding Level-k Models into Games of Incomplete Information. *Available at SSRN 3216040* (2019).

[32] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. 1999. An introduction to variational methods for graphical models. *Machine learning* 37, 2 (1999), 183–233.

[33] HJ Kappen. 2005. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment* 2005, 11 (2005), 11011.

[34] Michael Kearns, Michael L Littman, and Satinder Singh. 2001. Graphical models for game theory. *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence* (2001), 253–260.

[35] John Maynard Keynes and DE Moggridge. 1973. *The Collected Writings of John Maynard Keynes: The General Theory and After. Defence and Development.* Macmillan.

[36] Daewoo Kim, Sangwoo Moon, David Hostallero, Wan Ju Kang, Taeyoung Lee, Kyunghwan Son, and Yung Yi. 2019. Learning to Schedule Communication in Multi-agent Reinforcement Learning. (2019).

[37] Jelle R Kok, Eter Jan Hoen, Bram Bakker, and Nikos Vlassis. 2005. Utile coordination: Learning interdependencies among cooperative agents. (2005), 29–36.

[38] Jelle R Kok and Nikos Vlassis. 2004. Sparse cooperative Q-learning. (2004), 61.

[39] Landon Kraemer and Bikramjit Banerjee. 2016. Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing* 190, C (2016), 82–94.

[40] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences* 40 (2017).

[41] David Lazer and Allan Friedman. 2007. The network structure of exploration and exploitation. *Administrative science quarterly* 52, 4 (2007), 667–694.

[42] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-Agent Reinforcement Learning in Sequential Social Dilemmas. (2017), 464–473.

[43] Sergey Levine and Vladlen Koltun. 2013. Variational policy search via trajectory optimization. *Advances in neural information processing systems* 26 (2013), 207–215.

[44] Sheng Li, Jayesh K Gupta, Peter Morales, Ross Allen, and Mykel J Kochenderfer. 2021. Deep Implicit Coordination Graphs for Multi-agent Reinforcement Learning. (2021), 764–772.

[45] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[46] Qiang Liu and Dilin Wang. 2016. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. *Advances in Neural Information Processing Systems* 29 (2016), 2378–2386.

[47] Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *Advances in Neural Information Processing Systems* 30 (2017), 6379–6390.

[48] Laëtitia Matignon, Laurent Jeanpierre, and Abdel-Illah Mouaddib. 2012. Coordinated multi-robot exploration under communication constraints using decentralized Markov decision processes. *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (2012), 2017–2023.

[49] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. 2017. The numerics of GANs. 30 (2017), 1823–1833.

[50] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

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

[52] Frans A Oliehoek, Matthijs TJ Spaan, and Nikos Vlassis. 2008. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research* 32 (2008), 289–353.

[53] Jan Peters and Stefan Schaal. 2006. Policy gradient methods for robotics. *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2006), 2219–2225.

[54] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.

[55] David V Pynadath and Stacy C Marsella. 2005. PsychSim: Modeling theory of mind with decision-theoretic agents. 5 (2005), 1181–1186.

[56] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. (2018), 4218–4227.

[57] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. (2018), 4295–4304.

[58] Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. 2013. On stochastic optimal control and reinforcement learning by approximate inference. (2013).

[59] Sven Seuken and Shlomo Zilberstein. 2008. Formal models and algorithms for decentralized decision making under uncertainty. *Autonomous Agents and Multi-Agent Systems* 17, 2 (2008), 190–250.

[60] Michael Shum, Max Kleiman-Weiner, Michael L Littman, and Joshua B Tenenbaum. 2019. Theory of minds: Understanding behavior in groups through inverse planning. 33, 01 (2019), 6163–6170.

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

[62] Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. 2018. Learning when to Communicate at Scale in Multiagent Cooperative and Competitive Tasks. (2018).

[63] Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. 2019. Individualized Controlled Continuous Communication Model for Multiagent Cooperative and Competitive Tasks. (2019). https://openreview.net/forum?id=rye7knCqK7

[64] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. 2018. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. (2018), 2085–2087.

[65] Richard S Sutton. 1988. Learning to predict by the methods of temporal differences. *Machine learning* 3, 1 (1988), 9–44.

[66] Richard S Sutton and Andrew G Barto. 1998. Reinforcement learning: an introduction MIT Press. *Cambridge, MA* 22447 (1998).

[67] Gerald Tesauro et al. 1995. Temporal difference learning and TD-Gammon. *Commun. ACM* 38, 3 (1995), 58–68.

[68] Z Tian, Y Wen, Z Gong, F Punakkath, S Zou, and J Wang. 2019. A regularized opponent model with maximum entropy objective. *IJCAI International Joint Conference on Artificial Intelligence* 28 (2019), 602–608.

[69] Emanuel Todorov. 2006. Linearly-solvable Markov decision problems. (2006), 1369–1376.

[70] Marc Toussaint and Amos Storkey. 2006. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. (2006), 945–952.

[71] Ying Wen, Yaodong Yang, Rui Luo, Jun Wang, and W Pan. 2019. Probabilistic recursive reasoning for multi-agent reinforcement learning. (2019).

[72] Ying Wen, Yaodong Yang, and Jun Wang. 2020. Modelling Bounded Rationality in Multi-Agent Interactions by Generalized Recursive Reasoning. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* (2020), 414–421.