Decision-Making Among Bounded Rational Agents

Junhong Xu[1], Durgakant Pushp[1], Kai Yin[2], and Lantao Liu[1]

Indiana University, Bloomington[1], IN 47408, USA
Expedia Group[2]
{xu14, dpushp, lantao}@iu.edu, yinkai1000@gmail.com

Abstract. When robots share the same workspace with other intelligent agents (e.g., other robots or humans), they must be able to reason about the behaviors of their neighboring agents while accomplishing the designated tasks. In practice, frequently, agents do not exhibit absolutely rational behavior due to their limited computational resources. Thus, predicting the optimal agent behaviors is undesirable (because it demands prohibitive computational resources) and undesirable (because the prediction may be wrong). Motivated by this observation, we remove the assumption of perfectly rational agents and propose incorporating the concept of bounded rationality from an information-theoretic view into the game-theoretic framework. This allows the robots to reason other agents' sub-optimal behaviors and act accordingly under their computational constraints. Specifically, bounded rationality directly models the agent's information processing ability, which is represented as the KL-divergence between nominal and optimized stochastic policies, and the solution to the bounded-optimal policy can be obtained by an efficient importance sampling approach. Using both simulated and real-world experiments in multi-robot navigation tasks, we demonstrate that the resulting framework allows the robots to reason about different levels of rational behaviors of other agents and compute a reasonable strategy under its computational constraint. [1]

Keywords: Bounded Rationality, Game Theory, Multi-Robot System

1 Introduction

We consider the problem of generating reasonable decisions for robots in multi-agent environments. This decision-making problem is complex because each robot's motion trajectory depends on and affects the trajectories of others. Thus

[1] A preliminary version of this work appeared as a poster in *2021 NeurIPS Workshop on Learning and Decision-Making with Strategic Feedback*.
The video of the real-world experiments can be found at `https://youtu.be/hzCitSSuWiI`.
We gratefully acknowledge the support of NSF with grant No. 2006886 and 2047169.

they need to anticipate how others will respond to their decisions. The game-theoretic framework provides an appealing model choice to describe this complex decision-making problem among agents [16] and has been applied to various robotics applications, e.g., drone racing [24] and swarm coordination [1]. In an ideal situation, where all the agents are perfectly rational (i.e., they have unlimited computational resources), they can select the motion trajectories that reach the Nash equilibrium (if exists). However, since these trajectories live in a continuous space, agents need to evaluate infinitely many trajectories and the interaction among them, which is intractable. As a remedy, most of the works consider constraining the search space of the multi-robot problem via sampling [25] or locally perturbing the solution [24] to find a good trajectory within a reasonable amount of time.

Most game-theoretic planners mentioned above do not explicitly consider the agents' computational limitations in their modeling process, i.e., they assume each agent is perfectly rational. They consider these limitations only externally, e.g., by truncating the number of iterations during optimization. In contrast, we propose a novel and more principled treatment for modeling agents with limited computational resources by directly modeling the agents being only *bounded-rational* [6] in the game-theoretic framework. Bounded Rationality (BR) has been developed in economics [22] and cognitive science [7] to describe behaviors of humans (or other intelligent agents), who have limited computational resources and partial information about the world but still need to make decisions from an enormous number of choices, and has been applied to analyze the robustness of controllers in the single-agent setting [17]. In this work, we use the information-theoretic view of BR [6], which states that the agent optimizes its strategy under an information-theoretic constraint (e.g., KL-divergence), describing the cost of transforming its a-prior strategy into an optimized one. This problem can be solved efficiently by evaluating a finite number of trajectory choices from its prior trajectory distribution. Since BR explicitly considers the computational constraints during the modeling process, the resulting solution provides an explainable way to trade-off computational efficiency and performance. Furthermore, by incorporating BR into the game-theoretic framework, robots can naturally reason about other agents' sub-optimal behaviors and use these predictions to take advantage of (or avoid) other agents with less (or higher) computational resources.

2 Related Work

Our work is related to motion planning in multi-agent systems and game-theoretic frameworks. Here we provide a brief overview of these topics. The game-theoretic approach in robotics has gained increasing popularity recently. For example, the authors in [23, 24] combine Model-Predictive Control (MPC) and Iterated Best Response (IBR) to approximate the Nash Equilibrium solution of a two-player vehicle racing game. Recently, a game-theoretic iterative linear quadratic regulator (iLQR) has been proposed to solve a general-sum stochastic game [21].

In addition, the game-theoretic framework is also used in self-driving vehicles for trajectory prediction [5, 20] and motion planning among pedestrians [4, 12]. The above works are all based on the concept of rational agents who are assumed to be able to maximize their utility. In contrast, the proposed bounded rational framework explicitly considers the information-processing constraints of intelligent agents.

Although bounded rational solutions are used in almost all robotic systems in practice, e.g., anytime planners terminate the computation if the time runs out [8], only a few works attempt to model this bounded rational assumption explicitly. Recently, authors in [17] analyze the single-agent robust control performance using the information-theoretic bounded rationality [6, 15]. Another closely related but independently developed literature is KL-Control [3, 9]. When computing the optimal policy, it includes an additional information-theoretic cost measured by the KL-divergence between a prior policy distribution and a posterior after optimization. This is similar to the effect of the information-theoretic constraints of bounded rationality, where the final optimal control distribution can be sampled from the exponential family using approximate Bayesian inference [10, 26]. Although these methods have similar traits, they generally only focus on single-agent problems. In contrast, our work integrates the bounded rationality idea into the game-theoretic framework and provides a method to compute agents' strategies under computational limits.

3 Problem Formulation

In this section, we define multi-agent decision-making using the formulation of Multi-Agent Markov Decision Processes (MMDPs) or Markov Game (MGs) [11], and provide the resulting Nash Equilibrium (NE) solution concept under the perfect rationality assumption. In the subsequent sections, we show how this assumption can be mitigated using the BR concept and derive a sampling-based method to find a Nash Equilibrium strategy profile under bounded rationality.

3.1 Multi-Agent Markov Decision Process

In an MMDP with N agents, each agent i has its own state space $s^i \in \mathcal{S}^i$ and action space $a^i \in \mathcal{A}^i$, where a^i and s^i denote the state and action of agent i; $\mathcal{S}^i$ and $\mathcal{A}^i$ denote the corresponding spaces. We denote the joint states and actions of all the agents as $S = [s^1, ..., s^N]$ and $A = [a^1, ..., a^N]$. The agents progress in the environment as follows. At every timestep t, all the agents simultaneously execute their actions A_t to advance to the next states S_{t+1} according to their stochastic transition function $s^i_{t+1} \sim p^i(s^i_{t+1}|S_t, A_t)$. At the same time, they receive rewards $R_t = [r^1_t, ..., r^N_t]$, where $r^i_t = f^i_t(S_t, A_t)$ is the reward function for agent i. Each agent's stochastic transition and reward functions depend on all agents' states and actions in the system. Under the perfectly rational assumption, the goal for agent i is to find a strategy that maximizes an expected utility function

$$\pi^{i,*}_t = \arg\max_{\pi^i_t} U^i(S_t, \Pi_t), \tag{1}$$

where $U^i(S_t, \Pi_t) = \mathbb{E}\Big[\sum_{k=t}^{H+t} r_t^i(S_k, A_k)\Big]$ is agent i's utility function at t; H is planning horizon, and $\Pi_t = [\pi_t^1, ..., \pi_t^N]$ denotes the strategy profile for every agent. In this work, we assume that the agents' strategies take a specific form: a distribution over the action sequence $\mathbf{a}_t^i \sim \pi_t^i(\mathbf{a}_t^i | S_t, \Pi_t^{-i})$, where $\mathbf{a}_t^i = [a_t^i, ..., a_{t+H}^i]$ is the action sequence up to horizon H and Π_t^{-i} is the strategy profile without agent i. This policy form is well-suited for most robotics problems because the trajectory induced by the action sequence can be tracked by a low-level controller.

3.2 Iterative Best Response for MMDPs

To solve the problem defined in Eq. (1), each agent needs to predict how other agents will behave and respond to each other. For brevity, we will write $\pi^i(a_t^i | S_t)$ as agent i's strategy and omit the dependency on Π^{-i}. One possible and common way to predict other agents' behaviors is by assuming all other agents are perfectly rational, and thus the strategy profile of all the agents reaches the Nash Equilibrium (NE) (if exists) [16], which satisfies the following relationship:

$$U^i(S_t, \Pi_t^{-i,*}, \pi_t^{i,*}) \geq U^i(S_t, \Pi_t^{-i,*}, \pi_t^i), \forall i \in \{1, ..., N\}, \text{for any } \pi_t^i, \tag{2}$$

Intuitively, if the agents satisfy the NE, no agent can improve its utility by unilaterally changing its strategy. To compute NE, we can apply a numerical procedure called Iterative Best Response (IBR) [19]. Starting from an initial guess of the strategy profiles of all the agents, we update each agent's strategy to the best response to the current strategies of all other agents. The above procedure is applied iteratively for each agent until the strategy profile does not change. If every agent is perfectly rational, the robot can use NE strategy profile to predict other agents' behaviors and act correspondingly.[2]. However, there is a gap between this perfect rational assumption and the real world, as most existing methods can only search the strategy profile in a neighborhood of the initial guess [23, 24] due to computational limits. In the following section, we fill this gap by explicitly formulating this bounded rational solution.

4 Methodology

This section first provides details on the formulation of bounded rationality and integrates it into the game-theoretic framework. Then, we propose a sampling-based approach to generate bounded-rational stochastic policies.

4.1 Bounded Rational Agents in Game-Theoretic Framework

In the standard game-theoretic frameworks, agents are rational, i.e., they optimize their decisions via evaluating an infinite number of action sequences without considering the computational resources. In contrast, we model each agent

[2] We make a simplifying assumption that there is only one NE in the game.

as bounded rational – it makes decisions that maximize its utility subject to a certain computational constraint. Following the work in information-theoretic bounded rationality [6, 14], this constraint is explicitly defined by the neighborhood of a default policy q^i. Intuitively, this default policy describes the nominal behavior of the agent. For example, in a driving scenario, the default policy of an aggressive driver may be more likely to drive at a high speed. The agent's goal is to search for an optimized posterior policy bounded within the neighborhood of q^i. This size of the neighborhood may reflect the limited computational resources or other practical considerations. In the following, we omit the time subscript t for clarity. We use KL-divergence to characterize this neighborhood

$$\pi^{i,*} = \arg\max_{\pi^i} U^i(S, \mathit{\Pi}), \text{ s. t. } KL(\pi^i||q^i) \leq K_i. \tag{3}$$

K_i is a constant denoting the amount of computation (measured in bits) agent i can deviate from the default policy. Using Lagrange multipliers, we can rewrite the constrained optimization problem in Eq. (3) as an unconstrained one $\pi^{i,*} = \arg\max_{\pi^i} U^i(S, \mathit{\Pi}) - \frac{1}{\beta_i} KL(\pi^i||q^i)$, where $\beta_i > 0$ indicates the *rationality level*. To see how this bounded-optimal stochastic policy can be computed, we can first observe that the unconstrained problem can be written as

$$\begin{aligned} & U^i(S, \mathit{\Pi}) - \frac{1}{\beta_i} KL(\pi^i||q^i) \\ &= -\frac{1}{\beta_i}\Big(KL(\pi^i||q) - \beta U^i(S, \mathit{\Pi})\Big) \\ &= -\frac{1}{\beta}\int \pi^i(\mathbf{a}^i|S)\Big(\log \frac{\pi^i(\mathbf{a}^i|S)}{q^i(\mathbf{a}^i)e^{\beta U^i(S,\mathit{\Pi})}}\Big)d\mathbf{a}^i \\ &= -\frac{1}{\beta} KL(\pi^i||\psi^i), \end{aligned} \tag{4}$$

where $\psi^i(\mathbf{a}^i|\mathbf{s}) \propto q^i(\mathbf{a}^i)e^{\beta U(\mathbf{s},\mathbf{a})}$. Since KL-divergence is non-negative, the maximum of $-\frac{1}{\beta} KL(\pi^i||\psi^i)$ is obtained only when $KL(\pi^i||\psi^i) = 0$, which means $\pi^i = \psi^i$. Therefore, the optimal action sequence distribution of agent i (while keeping other agents' strategies fixed) under the bounded rationality constraint is

$$\pi^{i,*}(\mathbf{a}^i|S, \mathit{\Pi}^{-i}) = \frac{1}{Z} q^i(\mathbf{a}^i)e^{\beta \cdot U^i(S,\mathit{\Pi})}, \tag{5}$$

where $Z = \int q^i(\mathbf{a}^i)e^{\beta \cdot U^i(S,\mathit{\Pi})}d\mathbf{a}^i$ is a normalization constant. This bounded-optimal strategy provides an intuitive and explainable way to trade-off between computation and performance. When β increases from 0 to infinity, the agent becomes more rational and requires more time to compute the optimal behavior. When $\beta_i = 0$, the bounded-rational policy becomes the prior, meaning agent i has no computational resources to leverage. On the other hand, when $\beta_i \to \infty$, the agent becomes entirely rational and selects the optimal action sequence deterministically. The rationality parameter β allows us to model agents with different amounts of computational resources.

Similar to the rational case, our goal is to find the Nash Equilibrium strategy profile for a group of bounded-rational agents whose bounded-optimal policies are defined in Eq. (5). This can be done using the IBR procedure analogous to Section 3. Instead of optimizing the policy in Eq. (1), each bounded-rational agent finds the optimal strategy distribution defined in Eq. (5) while keeping other agents' strategies fixed. This procedure is carried out for each agent for a fixed number of iterations or until no agent's stochastic policy changes.

4.2 Importance Sampling for Computing Bounded-Rational Strategies

The previous section describes the bounded rationality concept, its integration with the game-theoretic framework, and uses the IBR numerical method to solve for a bounded rational Nash Equilibrium strategy profile. To actually compute the bounded-rational strategies, we need an efficient way to query samples from the distribution in Eq. (5) for each agent. Since it is relatively easier to sample the actions from the default $q^i(\mathbf{a}^i)$, we can use importance sampling to estimate the expectation of the optimal action sequence as the best response for the agent i while keeping others' actions fixed [2]

$$\begin{aligned}\mathbb{E}_{\mathbf{a}_t^{i,*}\sim\pi_t^{i,*}}[\mathbf{a}_t^{i,*}|S_t, \Pi_t^{-i}] &= \frac{1}{Z}\int \mathbf{a}_t^i q_t^i(\mathbf{a}_t^i) e^{\beta U^i(S_t,\Pi_t)} d\mathbf{a}_t^i \\ &= \frac{1}{Z}\mathbb{E}_{\mathbf{a}_t^i\sim q_t^i(\mathbf{a_t^i})}[w(\mathbf{a}_t^i)\mathbf{a}_t^i] \\ &\approx \frac{1}{Z}\frac{1}{K}\sum_{k=1}^{K} w(\mathbf{a}_{t,k}^i)\mathbf{a}_{t,k}^i,\end{aligned} \tag{6}$$

where $w(\mathbf{a}_t^i) = \exp\{\beta \sum_{k=t}^{H+t} r_t^i(S_k, A_k)\}$ and $\mathbf{a}_{t,k}^i$ denotes the k^{th} sample from the default policy with N samples in total. Similarly, the normalization constant can also be approximated as

$$\begin{aligned}Z &= \int q^i(\mathbf{a}_t^i) e^{\beta U^i(S_t,\Pi_t)} d\mathbf{a}_t^i \\ &= \mathbb{E}_{\mathbf{a}_t^i\sim q^i(\mathbf{a}_t^i)}[w(\mathbf{a}_t^i)] \\ &\approx \frac{1}{K}\sum_{k=1}^{K} w(\mathbf{a}_{t,k}^i).\end{aligned} \tag{7}$$

At a high level, the importance sampling procedure proceeds as follows. The agents first propose action sequence samples from their default policies $q_t^i(\mathbf{a}^i)$ and then assign each sequence a weight $w(\mathbf{a_t}^i)$ indicating its value based on the agents' utilities and rationality levels. Finally, by combining Eq. (6) and Eq. (7), we can use the weighted average to compute the expected optimal action sequence

$$\mathbb{E}_{\mathbf{a}_t^{i,*}\sim\pi_t^{i,*}}[\mathbf{a}_t^{i,*}|S_t, \Pi_t^{-i}] \approx \frac{\sum_{k=1}^{N} w(\mathbf{a}_{t,k}^i)\mathbf{a}_{t,k}^i}{\sum_{k=1}^{N} w(\mathbf{a}_{t,k}^i)}. \tag{8}$$

To find the bounded-rational Nash Equilibrium strategy profile, we replace the optimization procedure in IBR Eq. (2) using the above importance sampling. One important observation is that the shape of the prior distribution q^i, the number of samples for evaluation N, and the rationality level β play essential roles in the final performance. Their relationships will be explored in the experimental section.

5 Simulated Experiments

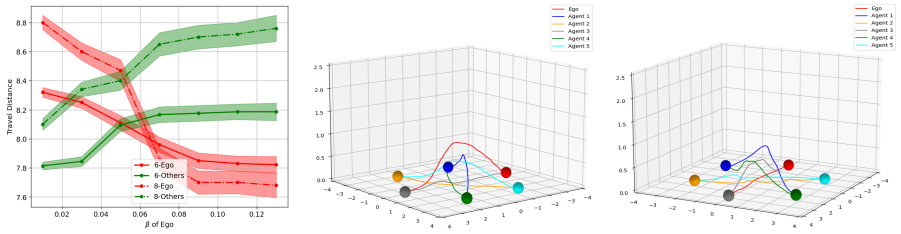
We conduct extensive simulation experiments to demonstrate that integrating bounded rationality with the game theory framework (1) allows each agent to reason about other agents' rationality levels to exploit (or avoid) others with less (or higher) computational capabilities and (2) explicitly trades-off between the performance and computation by varying the rationality level β and the number of sampled trajectories. We also show qualitatively that our method can generate group behaviors that approximately reach a bounded rational Nash Equilibrium strategy profile for a varying number of agents.

5.1 Simulation Setup

In this experiment, we consider the task of navigating a group of aerial vehicles in a 3D space. Each agent's goal is to swap its position with the diametrically opposite one while avoiding collisions with each other. The distance between each agent and its goal is $6m$. This environment is neither fully cooperative as each agent has a different goal location nor fully competitive because their objectives are not totally reversed (zero-sum). Thus, the agents need to collaborate with each other to avoid blockage in the center and at the same time compete with each other to follow their shortest paths to the goals. The agents are homogeneous in terms of their sizes and physical capabilities. Each agent has a size of $0.125m$ in x, y, z directions (similar to the dimension of Crazyfly drones used in the next section). Similar to [24], we set their transition functions using a deterministic single integrator model with the minimum and maximum speeds as $a_{min} = 0m/s$ and $a_{max} = 1m/s$. We set a uniform prior $q^i(\mathbf{a}) = Uniform(a_{min}, a_{max})$ as the default policy for all the agents. The number of IBR iterations is set to 10 as we found that under varying parameters, IBR usually converges to a bounded rational NE strategy at 10 iterations. In all the simulations, we assume that the rationality levels of all the agents are known to each other. The one-step reward function is the same for each agent and set to penalize collisions and large distances to the goal. We run 50 times for each simulation with $T = 80$ timesteps.

5.2 Results

We first show that using the proposed framework agents can naturally reason about the behaviors of others with the same physical capabilities but different

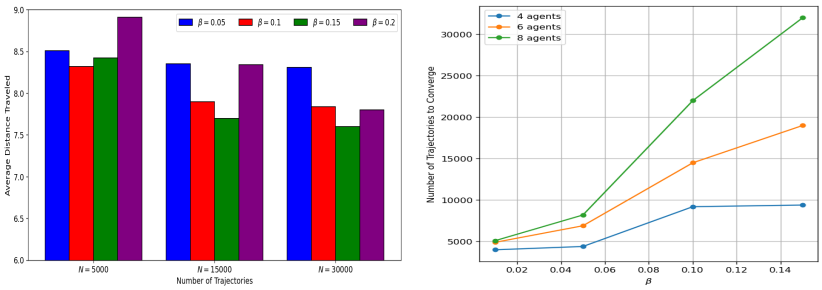


(a) (b) Ego's $\beta = 0.01$ (c) Ego's $\beta = 0.13$

Fig. 1. (a) Comparison of the performance by increasing the ego's rationality level from $\beta = 0.01$ to $\beta = 0.13$ while keeping other agents' rationality levels fixed in six (solid lines) and eight (dashed lines) agent environments. The x-axis and y-axis indicate the ego's β values and traveled distances. The green lines are the average travel distance of other agents and the red lines indicate the ego's travel distance. (b)(c) Show the agents' trajectories in six-agent environment with $\beta = 0.01$ and $\beta = 0.13$, respectively. The ego's trajectories are depicted in red.

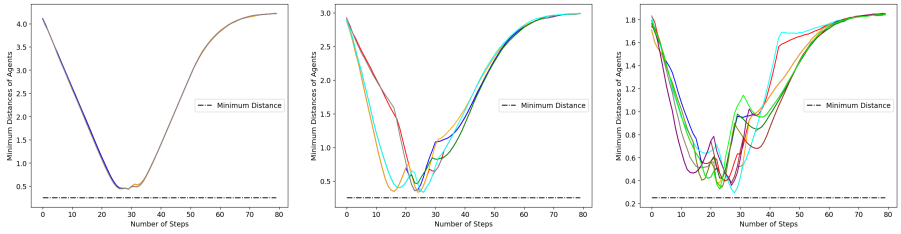
rationality levels β. Since we want to examine how varying β affects the performance, we need to ensure that the policy converges to the "optimal" one under the computational constraints. Thus, we sample a large number of trajectories, 5×10^5, for policy computation for each agent. In this simulation, we fix other agents' $\beta = 0.05$ and vary one agent's (called *ego* agent) β from 0.01 to 0.13 and compare the travel distances between the robot and other agents (the distance of other agents is averaged). Fig. 1(a) shows the performance comparison in six and eight agent environments. In general, when the ego has a lower rationality level β than other agents, it avoids them by taking a longer path. When all the agents have the same β, the ego and other agents have a similar performance. As β increases, the ego begins to exploit its advantage in computing more rational decisions and taking shorter paths. We also notice that the ego generally performs better with a large β when there are more agents in the scene. The trajectories of the six-agent environment for $\beta = 0.01$ and $\beta = 0.13$ are plotted in Fig 1(b) and Fig 1(c), respectively. When the ego's $\beta = 0.01$ (in the red trajectory), it takes a detour by elevating to a higher altitude to avoid other agents. In contrast, when its $\beta = 0.13$, it pushes other agents aside and takes an almost straight path. These results are aligned with Fig. 1(a). We omit the trajectories of eight agent environments to avoid clutter. The readers are encouraged to watch the videos at `https://youtu.be/hzCitSSuWiI`

Next, we evaluate the performance of a group of agents with the same β to show that the trade-off between the performance and computation can be directly controlled by the rationality level. Since most of the computation occurs when evaluating a large number of sampled trajectories in the proposed importance sampling-based method, we can use the number of evaluated trajectories as a proxy to measure the amount of computation an agent possesses. In Fig. 2(c), we analyze the relationship between β and the computation constraint in the six-agent environment. We can observe that when the computation is limited, i.e., only a small number of action sequences can be evaluated, a larger β (more



(a) (b)

Fig. 2. (a) Compares the performance (average traveled distance of the group) of different β values using a different number of trajectories in the six-agent environment. E The x and y axes are the number of trajectories and the average traveled distance of the group. (b) Shows the number of trajectories required to converge for different β in four, six, and eight agent environments.



(a) (b) (c)

Fig. 3. Minimum distances of each agent to other agents at every timestep in (a) four-agent (b) six-agent (c) ten-agent environments. The x-axis and y-axis are the timesteps and agents' minimum distances, respectively. The colored solid lines represent the statistics of each agent, which is the same as their trajectory color in Fig. 4. The dashed grey line shows the minimum safety distance ($0.25m$) that needs to be maintained to avoid collisions.

rational) actually hurts the performance. When the more rational agents have the resources to evaluate more trajectories, they travel less distance on average than the less rational groups. This result demonstrates that by controlling the rationality parameter the bounded rational framework can effectively trade-off between optimality and limited computation. In Fig. 2(b), we also evaluate the number of trajectories that need to be sampled for the method to converge at different rationality levels in four, six, and eight agent environments. The result aligns with the previous observation – in general, when β is larger, more trajectories need to be evaluated to converge to "optimal" under the bounded rational constraints. Furthermore, when more agents are present in the environment, the method requires more trajectories to converge.

Finally, we show qualitative trajectory results of the group behaviors under bounded rationality using a fixed $\beta = 0.1$ and number of samples $N = 5 \times 10^5$

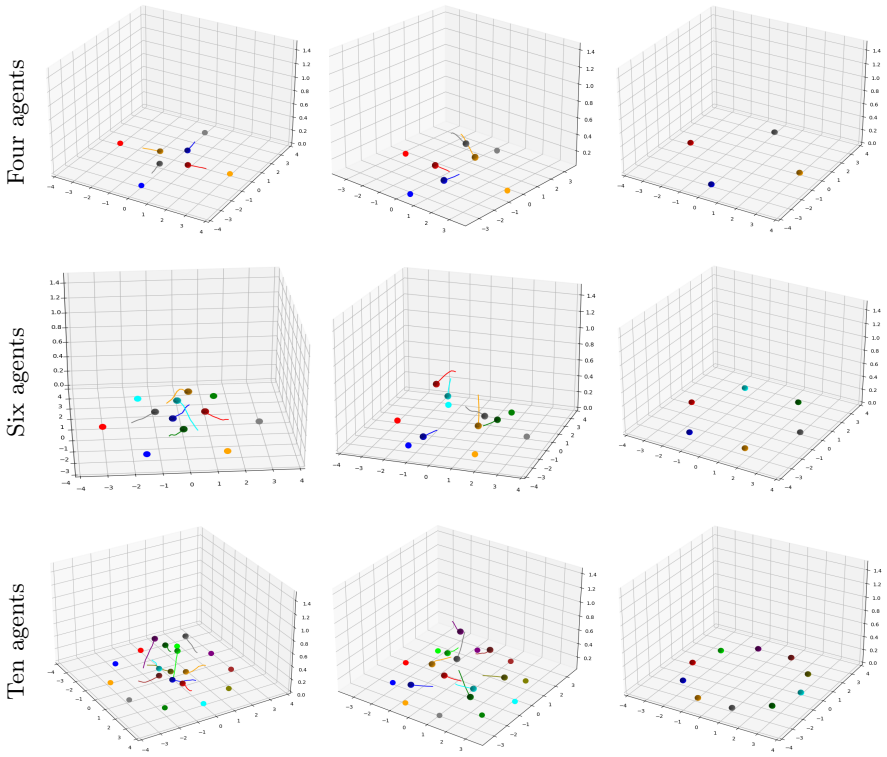
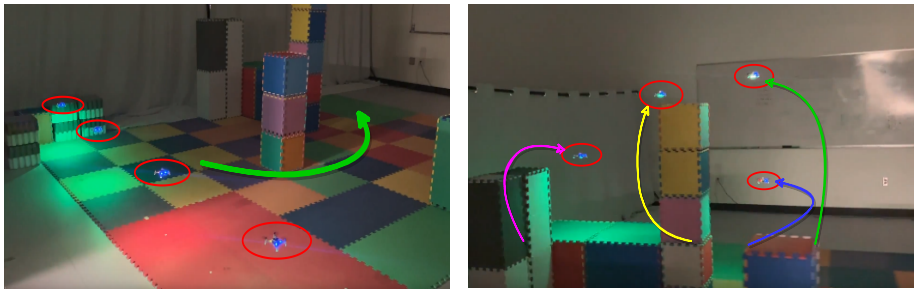


Fig. 4. Agents' trajectories in environments with four, six, and ten agents. The three columns show the snapshots at $t = 20, 60, 80$, respectively. Each agent, its trajectory, and its goal are assigned the same unique color. We only show the last 10 steps of trajectories to avoid clutter.

for each agent in Fig. 4. Additionally, we provide the minimum distances of each agent to other agents in Fig. 3. The result shows that in each environment, the agent can maintain a safe distance $> 0.25m$ to avoid collisions.

6 Physical Experiments

We use the Crazyflie 2.1 nano-drones under a motion capture system to validate our method in the real world. For this hardware experiment, we consider two types of tasks with a varying number of agents. The first task is to navigate a group of drones to a designated goal region while avoiding static obstacles and inter-drone collisions. The second task is position swapping similar to the previous section. The size of the workspace considered in all the experiments is $4.2m \times 5.4m \times 2m$ in the x, y, and z axes, and the size of the drone is $92mm \times 92mm \times 29mm$. To mitigate the downwash effect and control inaccuracy, we buffer the drone's collision size to be $0.5m \times 0.5m \times 0.5m$. We use



(a) (b)

Fig. 5. (a) Shows the experimental setup with 4 Crazyflie drones. Green arrow points to the goal position. All the agents are navigating to the same goal point. (b) Snapshot of the experiment. Shows the path followed by the agents to avoid the obstacles.

(a) (b)

Fig. 6. Shows the experimental setup with 6 Crazyflie drones. The drones are divided into two groups - red and green. (a) Shows the initial position of the drones. The task is to swap the positions. Black lines show assigned swapping tasks among agents. (b) Snapshot of the experiment during swapping. It shows the path followed by the agents to avoid collision with each other.

Crazyswarm [18] platform to control the drones. For each run, we generate trajectories using $N = 3 \times 10^5$ and $\beta = 0.1$ for all the agents using the proposed method. These trajectories contain a sequence of (x, y, z) coordinates. We use minimum snap trajectory optimization and control strategy [13] to track the trajectories generated by the proposed planner.

We show two representative scenarios for each task. For complete experimental videos, please refer to `https://youtu.be/hzCitSSuWiI` Fig. 5 shows that a group of four drones have to go from the red zone to the green zone while avoiding four obstacles of various sizes distributed around the center of the workspace. Note that one of the drones opted to go over the obstacle of height 1.5m which shows that it finds a path through the space as narrow as the size of the drone $(0.5m)$ in the z-axis. This event is captured in the snapshot shown in Fig. 5(b). Fig. 6 shows the position swapping scenario. We use the same dynamics model as the previous section to generate the trajectories. We observe that the out-

comes of the physical experiments are consistent with the results obtained in the simulation.

7 Conclusion

This paper considers the problem of making sequential decisions for agents with finite computational resources, where they need to interact with each other to complete their designated tasks. This problem is challenging because each agent needs to evaluate its infinite number of decisions (e.g., waypoints or actuator commands) and reason how others will respond to its behavior. While the game-theoretic formulation provides an elegant way to describe this problem, it is based on an unrealistic assumption that agents are perfectly rational and have the ability to evaluate the large decision space. We propose a formulation that replaces this rational assumption with the bounded rationality concept and presents a sampling-based approach to computing agents' policies under their computational constraints. As shown in the experiments, by removing the perfect rational assumption, the proposed formulation allows the agents to take advantage of those with less computational power or avoid those who are more computational-capable. Additionally, when all the agents are similarly computational capable, they exhibit behaviors that avoid being taken advantage of by others.

Bibliography

[1] Mohamed Abdelkader, Samet Güler, Hassan Jaleel, and Jeff S Shamma. Aerial swarms: Recent applications and challenges. *Current Robotics Reports*, 2(3):309–320, 2021.

[2] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[3] Matthew Botvinick and Marc Toussaint. Planning as inference. *Trends in cognitive sciences*, 16(10):485–488, 2012.

[4] Yu Fan Chen, Michael Everett, Miao Liu, and Jonathan P How. Socially aware motion planning with deep reinforcement learning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1343–1350. IEEE, 2017.

[5] Jaime F Fisac, Eli Bronstein, Elis Stefansson, Dorsa Sadigh, S Shankar Sastry, and Anca D Dragan. Hierarchical game-theoretic planning for autonomous vehicles. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9590–9596. IEEE, 2019.

[6] Tim Genewein, Felix Leibfried, Jordi Grau-Moya, and Daniel Alexander Braun. Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, 2:27, 2015.

[7] Gerd Gigerenzer and Henry Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science*, 1(1):107–143, 2009.

[8] Félix Ingrand and Malik Ghallab. Deliberation for autonomous robots: A survey. *Artificial Intelligence*, 247:10–44, 2017.

[9] Hilbert J Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.
[10] Alexander Lambert, Adam Fishman, Dieter Fox, Byron Boots, and Fabio Ramos. Stein variational model predictive control. *arXiv preprint arXiv:2011.07641*, 2020.
[11] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.
[12] Björn Lütjens, Michael Everett, and Jonathan P How. Safe reinforcement learning with model uncertainty estimates. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8662–8668. IEEE, 2019.
[13] Daniel Mellinger and Vijay Kumar. Minimum snap trajectory generation and control for quadrotors. In *2011 IEEE International Conference on Robotics and Automation*, pages 2520–2525, 2011.
[14] Pedro A Ortega and Daniel A Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2153):20120683, 2013.
[15] Pedro A Ortega, Daniel A Braun, Justin Dyer, Kee-Eung Kim, and Naftali Tishby. Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789*, 2015.
[16] Martin J Osborne et al. *An introduction to game theory*, volume 3. Oxford university press New York, 2004.
[17] Vincent Pacelli and Anirudha Majumdar. Robust control under uncertainty via bounded rationality and differential privacy. *arXiv preprint arXiv:2109.08262*, 2021.
[18] James A. Preiss*, Wolfgang Hönig*, Gaurav S. Sukhatme, and Nora Ayanian. Crazyswarm: A large nano-quadcopter swarm. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3299–3304. IEEE, 2017. Software available at `https://github.com/USC-ACTLab/crazyswarm`.
[19] Daniel Reeves and Michael P Wellman. Computing best-response strategies in infinite games of incomplete information. *arXiv preprint arXiv:1207.4171*, 2012.
[20] Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:187–210, 2018.
[21] Wilko Schwarting, Alyssa Pierson, Sertac Karaman, and Daniela Rus. Stochastic dynamic games in belief space. *IEEE Transactions on Robotics*, 2021.
[22] Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118, 1955.
[23] Riccardo Spica, Eric Cristofalo, Zijian Wang, Eduardo Montijano, and Mac Schwager. A real-time game theoretic planner for autonomous two-player drone racing. *IEEE Transactions on Robotics*, 36(5):1389–1403, 2020.
[24] Mingyu Wang, Zijian Wang, John Talbot, J Christian Gerdes, and Mac Schwager. Game theoretic planning for self-driving cars in competitive scenarios. In *Robotics: Science and Systems*, 2019.
[25] Grady Williams, Brian Goldfain, Paul Drews, James M Rehg, and Evangelos A Theodorou. Best response model predictive control for agile interactions between autonomous ground vehicles. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2403–2410. IEEE, 2018.
[26] Grady Williams, Nolan Wagener, Brian Goldfain, Paul Drews, James M Rehg, Byron Boots, and Evangelos A Theodorou. Information theoretic mpc for model-based reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1714–1721. IEEE, 2017.