A Receding-Horizon Markov-Game Approach for Human-Robot Coordination

Shih-Yun Lo[1], Benito Fernandez[1], and Peter Stone[2]

***Abstract*— There is yet no formal mathematical model to analyze the real-time coordination between humans and robots, which commonly happens when they co-exist in a shared workspace: human-robot teaming, and their resource allocation to resolve conflicts. Mis-coordination deteriorates task efficiency of both parties, as well as human trust and patience on robotic agents. In this work, we model human-robot coordination from a game-theoretic perspective, to capture how human's 1) perceived robot capability, 2) personal preferences, 3) self-interested level, and 4) belief propagation rate affect their behaviors, and the associated game outcomes in a finite-horizon setting.**

In this work, we provide 1) human-robot path crossing as an instantiation, and 2) approximate game-theoretic models for online human decision prediction, and 3) efficient pruning algorithm for robot strategy planning in a receding-horizon fashion. We would also show the effectiveness of our human behavior model by reproducing real-world observed human-robot crossing interactions, along with their verbal responses after the experiments to support our human-behavior hypotheses.

I. Preliminaries

A. Game Definition

Consider a game G with k players, where each player $i \in \{1...k\}$ has a finite action set A^i. The set of action profiles is denoted as $S = A^1 \times A^2 \times ... \times A^k$. The utility of an agent i is a function, denoted as $f^i : s \rightarrow \mathbb{R}$, evaluated at $s \in S$. The utility functions can also be transferred into positive real values $r^i : s \rightarrow \in \mathbb{R}^+$ without loss of generality.

We model human-robot interaction as a game, because each agent's utility not only depends on his or her own action, but also on others' actions. Let $x \in X$ be the joint state space of all agents, and $a^H \in A^H$ and $a^R \in A^R$ be the action space of humans and robots, respectively. Consider a two-player game with agent R and H, an action profile $s^* = (a^{R*}, a^{H*})$ is a *Nash Equilibrium*, if no agent benefits from unilateral deviation from his or her current action:

$$\forall i \in \{H, R\}, s^* \in S, f^i(a^{i*}, a^{-i*}) \geq f^i(a^i, a^{-i*}). \quad (1)$$

Here a^{-i} refers to actions from all agents but i, e.g. $a^{-H} = a^R$.

B. Cooperative Games v.s. Non-cooperative Games

One common game example is social navigation [1]: each agent optimizes his or her own path to reach a final goal, while other agents' choices of trajectories may intersect and then cause delays. Agents therefore need to plan based on predictions of other agents' actions, to efficiently resolve such resource conflicts. Another example is the table carrying task [2], where agents share the same objective to collaboratively carry out a large piece of furniture.

Games with shared collective payoffs, $f^i = f^{-i}$, are categorized as cooperative games, which are often discussed separately from those with individual outcomes, especially those with profit conflicts: the *non-cooperative games*. For cooperative games, there is only one strategy profile that maximizes the collective payoff; where as in non-cooperative games, multiple Nash equilibria may exist. In those games, cooperative behaviors may also be observed, strategies such as *early commitments* and *threats* are commonly used during the bargaining process to maximize self-interest.

In either, player *trust, reciprocity, and perceived capability* affect how he or she predicts the other agents' behaviors; whereas *self-interest level and personal preferences* affect his or her own utility function to parametrize game outcomes. Details will be discussed in Sec. on human behaviors and interactions with robots.

C. Real-time Game Formulation

Therefore, to analyze how human factors affect the dynamics of human-robot interaction and the game convergence over time, we propose a real-time game setting defined by the following:

1) Game start criteria: In a scenario where agents' actions affect the outcome of each other, we define as it meets the criteria of *potential interaction*, $\mathcal{C}_{PI}$; when agents become aware of each other and start acting according to the perceived strategy of the other agents, it meets the criteria of *mutual awareness*, $\mathcal{C}_{MA}$. Once the two criteria hold, game begins. (A counterexample is when agents are aware of each other but their actions do not affect each other's outcome. In such scenarios, such as pedestrians walking on different sides of a traffic intersection, no potential interaction is concerned; another example is when only partial party is aware of the other agents. In that case, merely one-way adaptation among agents happens, which we do not treat as interactions.

2) Game horizon and termination criteria: When interactions start at time $t = 0$, the game repeats *finitely* until a termination criteria $\beta : X \times S \rightarrow \{True, False\}$ is satisfied. It may be a predefined criteria, $t > T$, where T is a pre-specified time frame of mandatory co-working. Termination criteria may also take in a dynamic format, by ending the game whenever $\mathcal{C}_{PI}$ no longer holds. An example would be crowd interactions: when pedestrians went past the route intersections with one another, game terminates; as no further intervention is expected.

*This work was supported by

3) Game outcome and discount factor: The game outcome V^i of each player i is defined as the discounted cumulative reward till game termination:

$$V^i = \Sigma_{t=0}^{\beta_t=True} \gamma^t r^i(s_t), \tag{2}$$

where β_t is short for $\beta(x_t, s_t)$ and $0 < \gamma \leq 1$ is the discount rate.

In the repeated game setting, the strategy profile $\int_{t:t+h}$ is composed by the action profiles $s_{t:t+h}$ over certain horizon h: $\mathcal{S} = S_0 \times S^1 \times ... \times A^h$.

D. Early termination

A game may early terminate when a Nash Equilibrium is reached before termination and no agent conveys intent to deviate.

———(Recap finitely repeated game convergence,

E. Bayesian games

Games where agents receive outcomes depending on their types $\theta \in \Theta$ are defined as Bayesian games. The individual utility function becomes: $f^i : S \times \Theta \to \mathbb{R}$, and the type variable s θ^i are not directly observable by other agents $-i$. As a result, agents observe game outcomes and update their beliefs of the other agent's type $b_t^i(\theta^{-i})$.———— policy π^i for action execution————— $\Pi : X \times S \to U$

II. Human Decision-Making Model

When planning for long-horizon purposes, or repeated games, agents ideally want to optimize their outcome in a full-horizon manner:

$$a_{0:T}^{i*} = \underset{a_{0:T}^i}{\operatorname{argmax}} \mathbb{E}_{a_{0:T}^{-i}} \left[\Sigma_{t=0}^{\beta_t=True} \gamma^t r^i(s_t) \right]. \tag{3}$$

Here game terminates at $t = T$, and we assume agents have knowledge of such timing even in the dynamic termination setting β_t.

However, due to computational resource limitation as well as the demand of fast re-planning in dynamic environments, agents may only be available to plan for finite-step lookahead. Agents then instead try to optimize:

$$a_{0:H}^{i*} = \underset{a_{0:H}^i}{\operatorname{argmax}} \mathbb{E}_{a_{0:H}^{-i}} \left[\Sigma_{t=0}^{\beta_t=True} \gamma^t r^i(s_t) + \hat{V}_{H+1}^i(s_H) \right], \tag{4}$$

where $H < T$ is a fixed horizon for lookahead, and $\hat{V}_{H+1}(s_H)$ is the cumulative reward estimate for following s_H from $t = H + 1$ to T.

When the interaction continues, $t = h$, where $0 < h < T$, agents may re-consider their strategy for $t = h : H'$, where $H' = max(h+H, T)$, based on update of the estimate of the other agents' strategies $p(a_{h:H'}^{-i} | s_{0:h-1})$, given past observations. This online replanning strategy, also called receding-horizon planning, can be formulated as the following:

$$a_{h:H'}^{i*} = \underset{a_{h:H'}^i}{\operatorname{argmax}} \mathbb{E}_{a_{h:h+H}^{-i} | s_{0:h-1}} \left[\Sigma_{t=h}^{\beta_t=True} \gamma^t r^i(s_t) + \hat{V}_{H'+1}^i(s_H) \right], \tag{5}$$

which is to plan for certain $H' \leq H$ horizon, execute action and observe other agents' behaviors, update the belief of agents' strategies, and plan again at the next time step.

A. Finite-step lookahead

Given domain information complexity, dynamic variations, and familiarity with the environments, agents may plan at different lookahead horizons.

1) 0-step: Agents assume the current game is the termination game. If all agents plan for 0-step lookahead, players play one mixed strategy equilibrium. If there is only one equilibrium, the game converges immediately.

———(proof)assuming their value estimates are consistent with the true value

2) threats and early commitments: When agents perceive the game has more than one shot, early commitments and threats are commonly seen to arrive at an equilibrium with higher self-interests. In the real-time game setting, oftentimes one agent senses the interaction and acts first, which easily allows early commitments to play the role and influence other agents' strategies. Such strategy is commonly seen in dense crowd navigation (cite..), where agents look at each other (to initialize the game through signaling awareness), act, and then quickly look away without waiting for responses. Agents then continue the same action through out the interaction and benefit from ruthless road-taking.

3) T.5-step: In real-time coordination, intentions need time to convey and to comprehend. The mismatch of reaction time among agents introduces a source of mis-coordination, and result in potential situations where it too late for an agent to respond after comprehension. We refer this as T.5-step game termination.

B.

To predict about other agent's actions a_t^{-H}, we assume humans maintain their beliefs b_t^H of their *reaction* given past interactions:

$$a_t^H \sim \pi^H(x_t, b_t^H(a_t^{-H} | s_{0:t})). \tag{6}$$

To anticipate the other agents' reaction $a_t^{-H} | s_{0:t}$, one may further hypothesize how the others anticipate his or her reaction to the past interactions,

$$b_t^H(a_t^{-H} | s_{0:t}) \sim b_t^H(\pi^{-H}(x_t, , b_t^{-H}(a_t^H | s_{0:t}))), \tag{7}$$

and further, they may anticipate N^{-H} steps of future reactions,

$$b_t^H(a_t^{-H} | s_{0:t}) \sim b_t^H(\pi^{-H}(x_t, , b_t^{-H}(\pi^H(..., \tag{8}$$

$$\pi^H(x_t, , b_t^{-H}(a_t^H | s_{0:t})))))), \tag{9}$$

1) Bounded memory for belief propagation: ——- $b_t^H(\pi^{-H})$ takes into account finite-step of history interactions

$$a_t^H \sim \pi^H(x_t, b_t^H(a_t^{-H} | s_{t-n:t})). \tag{10}$$

2) Bounded computation for N-step anticipations: ————

$$b_t^H(a_t^{-H} | s_{0:t}) \sim b_t^H(\pi^{-H}(x_t, , b_t^{-H}(a_t^H | s_{0:t}))), \tag{11}$$

III. Variations in human-robot interaction

A. Perceived robot capability

When deploying robots into shared workspaces with humans, people may not have prior experience working with those agents. Intention confusion, incapability in predicting future behaviors, and mistrust are commonly described in the literature in human-robot interaction. The large gap between true robot capability and human perceived robot capability, has shown to deteriorate either joint or individual work efficiency in different task domains.

Here, we characterize human perceived robot capability for human-robot interaction into two categories: 1) functional capability and 2) social inference capability:

1) functional capability: This includes the belief of whether the robot is able to *identify interaction*. Before engaging interactions, agents need to ensure that $\mathcal{C}_{PI}$ and $\mathcal{C}_{MA}$ are met by all parties, or confusions may arise. Due to lack of tools for social signals, such as gazes, humans may be uncertain if the robot is aware of upcoming interactions. This may result in distantly-avoiding behaviors through out the interaction, since people are not sure whether to engage (cite..). This also includes the knowledge of robot action set A^R, and the belief of whether the robot is able to *succeed in its target actions*, especially when complex domains are considered (cite..).

2) social inference capability: This involves whether the robot is able to play the game,to interact in a socially-aware and socially informative manner. It includes the ability to *identify human's intended high-level actions* based the inference from past movements or trajectory, and to *express its own intent* in a clear, context-aware way, or, legibly (cite..).

———mathematical formulation of social inference: to respond within proper response time————- The above two capabilities are prerequisites to enable smooth human-robot interactions.

B. Social trust and reciprocity

——In a social environment, where public goods are di

C. personal preferences

D. self-interest level

IV. Convergence Criteria

ACKNOWLEDGMENT

References

[1] Christoforos I Mavrogiannis and Ross A Knepper. Decentralized multi-agent navigation planning with braids. In *Proceedings of the Workshop on the Algorithmic Foundations of Robotics. San Francisco, USA*, 2016.

[2] Stefanos Nikolaidis, Anton Kuznetsov, David Hsu, and Siddharta Srinivasa. Formalizing human-robot mutual adaptation: A bounded memory model. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pages 75–82. IEEE Press, 2016.