Reinforcement Learning for Staghunt

Adrian Manhey
amanhey.edu
Occidental College

Figure 1.

1 Introduction and Problem Context

The goal of this project is to use Reinforcement Learning to train a reinforcement learning model to play stag-hunt. This project is part of a larger research study led by Professor Rabkina at Occidental College, where researchers are using Analogical Theory of Mind (AToM) to develop virtual agents with social reasoning abilities similar to humans [**Rabking2019**]. Stag-hunt is a stochastic prisoner's dilemma-style game where the goal is to accumulate the highest number of points by capturing high-value mobile targets (stags) or low-value static targets (hares). Agents, whether virtually created or human players, control hunters which can hunt stags or hares, but stags require two or more agents to capture them. All moves happen at the same time and agents are allowed to occupy the same spaces.

Figure **??** is an example stag-hunt scenario, taken from scenario (d) of Figure 1 of Rabkina et al. (2019), where a pair of agents (A and B) cooperate to capture a stag and another agent captures a hare individually. The circles represent hunters position at a given time-stamp and the dotted lines represent movement. A key element of this game is an agent inferring other agents' intent to collaborate with them to pursue the high reward. This process involves social reasoning called Theory of Mind [**Ruhl2020**], where an entity makes inferences about another's mental state (knowledge, beliefs, preferences, desires, goals, and intentions). The AToM model posits that a combination of analogical processes and feedback lead to the development of theory of mind reasoning without an underlying model of cooperation. However, to compare the results of the AToM model to other implementations, specifically a Bayesian model [**Shum2019**], there needs to be data collected around human perception of cooperation of the agents. Shum et al. 2019 proposes a Bayesian model, which does not require training but does require an explicit representation of team hierarchy and was found to have strong correlations with human predictions. However, the AToM model is more accurate than the Bayesian model at most time stamps and does not require a predefined underlying hierarchy of team cooperation. Furthermore, the ability to make predictions about the future actions of agents is an additional effect of reasoning about an agent's cooperation with others. To make accurate comparisons, different models can be implemented and then a study can be conducted to determine how people inference the cooperativeness of the agents.

For this problem I am approaching it with reinforcement learning due to its presence in similar problems and academic interest. The initial approach is to develop a Q-learning algorithm for stag-hunt where one agent utilizes Q-learning and the other agent uses a more basic graph search algorithm, such as A*. As the project develops, changes may be made to adjust the difficulty of the implementation.

2 Technical Background

OpenAi's Gym environment designed for "developing and comparing reinforcement learning algorithms." [**Gym**] The environment has the versatility to provide many example environments but also has the flexibility to allow designers to create their own. In terms of Q-learning, with the environment created the algorithm can be designed ourselves, in the most likely case resulting in a table of state and action pairs (s, a) with an associated Q-value, representing their "quality". Q-values are initialized to an arbitrary value and as the agent explores the environment they are updated. The method for updating a Q-value consists of two parameters: $\alpha \in (0, 1]$ is the learning rate and $\gamma \in [0, 1]$ is the discount factor, which makes short-term or long-term rewards more valuable. The equation for this relationship can be written as

$$Q(s, a) = (1 - \alpha)Q(s, a) + \alpha(r + \gamma \text{max}Q_a(s', A)),$$

where r is the reward of an action, s' is the next state, and A is all the possible actions that can be taken. Additionally the algorithm uses ϵ, which can be thought of the exploration parameter. This tells the algorithm how often to pick a random action instead of the current most beneficial one from the Q-table. In sum, we are learning the proper action by taking the old Q-value and adding the learned value of the immediate reward and possible future rewards.

This algorithm also has the potential to be optimized by modifying the values of the parameters either before or during training. The learning rate can be decreased as the

knowledge base increases and the exploration parameter can be decreased as the number of trials increase.

3 Prior Work

This section describes of related and/or existing work. This could be scientific or scholarly, but may also be a survey of existing products/games. The goal of this section is to put your project in the context of what has already been done.

4 Methods and Evaluation Metrics

First idea would be to implement a Q-learning algorithm like the taxi tutorial.

5 Ethical Considerations

Think about what the research is and the goals of the project, particularly the social goals. If the goals are still dealing with people, there will be various cultural and racial considerations to be made. This is built on the assumption that the project is focused on robot-human interaction. However, if this is built on robot-robot cooperation then it doesn't matter. You can also make a simplified model of a person, defined as a sequence of events that can be skipped through. Also need to consider how that agent is going to perform with the desired player given the testing player, if it trains with a RL agent can it play with a human agent.

6 Timeline

The first step is going to be doing more research around reinforcement learning, especially in stag-hunt. So far there have been quite a few resources for both single and mult-agent stag-hunt using reinforcement learning so finding which algorithm will work best for this project will be key. After the Occidental College URC program begins, I plan to begin doing this research in tandem to my other role on the parent project. Most of the summer will mainly consist of this research so that in the fall the project has a clear direction. In May I plan to learn more about the existing models of the project and how the new implementation may account for some gap in functionality or theory. June will likely then be reserved for researching models I would want to build myself, which involves loosely designing the model I choose in terms of high-level implementation. This phase of the project will be building the prior work that has been done in terms of stag-hunt, reinforcement learning, or multi-agent games. In July I plan to continue designing the model and developing the kind of technical information I may need. Additionally, this time period of high- to mid-level design would provide a good opportunity to start putting together a list of resources needed by the project, i.e. are there any computational tools I might need to request in the fall? The semester starts in August so I would like to finish the design of the model in order to have it ready to implement in September. This phase revolves around creating the specific methods I'm using in my implementation. In September I plan to finish implementing the model and start trying to improve it, e.g. optimizing the parameters. October will consist of evaluating the model and gaining insight into the performance of the model and how well it meets the goals of my comprehensive project and the parent project. The last two months will involve any needed refactoring and preparing for presentations.