Multi-Agent Specialization and Coordination in a Gridworld Task

Chase McDonald, Thuy Ngoc Nguyen, Cleotilde Gonzalez
Dynamic Decision Making Laboratory
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213

Abstract

Coordination of individual behavior is essential for success in any goal-directed team task. In particular, the ability to coordinate in the absence of explicit communication among team members will depend on the task structure and features of the environment. This work investigates how environmental variables—agent positioning, reward distribution, and inter-agent visibility—impact the ability of agents to learn coordinated behavior in a goal-directed multi-agent task. We also relate the learnt coordination to learning behavioral specialization of agents and discuss the relationship between coordination and specialization. Our agents are cognitive models built from Instance-Based Learning Theory, a theory of human decisions from experience. The results and insights from our simulation reveal how environmental factors can facilitate or inhibit coordinated behavior for successful performance of a team task.

Introduction

Environmental factors play a critical role in how and what we learn through experience. Indeed, distinct environmental attributes can give rise to disparate behavior, particularly in the case of social or multi-agent settings. This research investigates how environment variables impact agents' ability to learn to coordinate without communication in a multi-agent task. We rely on Instance-Based Learning Theory (IBLT) (Gonzalez, Lerch, and Lebiere 2003) to instantiate a cognitive model, which is used to simulate individual agents. IBLT is employed given a cognitively plausible set of decision mechanisms that often account for human decisions in risky tasks (Gonzalez and Dutt 2011) with delayed feedback (Nguyen and Gonzalez 2020b). Through a simulation experiment, we explore three environmental variables and how they determine the specialization and coordination among agentsin a multi-agent target-seeking task: (1) the distribution of rewards; (2) the starting point of the agents (i.e., spawn location); and (3) the inter-agent visibility. Reward schemes are fundamental in reinforcement and varying rewards can facilitate coordination among (Tampuu et al. 2017; Grzes 2017). Similarly, levels of observability are key in the development of environments, as is reflected in the key distinction of Markov Decision Processes (MDPs) and Partially Observable MDPs (Gronauer and Diepold 2021).

We test the following hypotheses. First, we expect that the distribution of the rewards will influence agents' coordination; for example, if there is only one target of high value and other targets of low value, it is expected that all agents will aim for the high value target (i.e. low coordination). Second, we expect that the relative spawn locations will also influence the agents' coordination. If the starting point is different for all members of a team, coordinating may be easier compared to when all team members have the same starting location due to potential collisions with other agents. Finally, we also expect that a larger "field of view", FoV (i.e., range of inter-agent visibility surrounding an agent), would result in better coordination among the agents, as they will be able to observe other team members and condition their actions on the observations of others.

Multi-Agent Gridworld Environment

We use a gridworld with an 11×11 grid that contains walls and four colored targets. An example is depicted in Figure 1. Each target has an associated reward that ranges from 0 to 1, which an agent receives upon reaching the target. Gridworld simulations are comprised of episodes, and each episode entails a maximum of 31 timesteps, in each of which agents make a decision. Each decision entails selecting one of the four cardinal directions to move in.

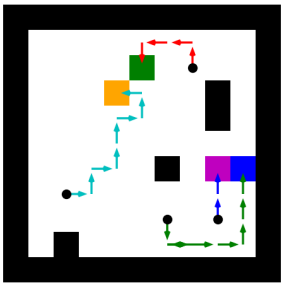

Figure 1: A Gridworld configuration. Black squares represent walls and colored squares are targets. Four agents (black circles) start in different spawn locations.

A gridworld is formalized as a Markov Decision Process

(MDP). Each MDP $\mathcal{M}$ has a state space $\mathcal{S}$, and, in its most simple form, each $(x; y)$-coordinate in the grid represents a state $S \in \mathcal{S}$. At each within-episode time step $t \in \{1; :::; T\}$, an agent j observes their state $S_{j;t}$, then takes an action $A_{j;t}$ from a common action space $\mathcal{A}$ (up/down/left/right) to move into state $S_{j;t+1}$ and observes the reward (or cost) $R_{j;t}$. By executing a policy $_j$ in the environment $\mathcal{M}$, an agent creates a trajectory denoted by $\mathcal{T}_j = \{(S_{j;t}; A_{j;t})\}_{t=1}^{T}$. Each environment is a 11×11 grid that contains obstacles (black cells) and four colored targets, an example of which is depicted in Figure 1. Each target has an associated reward that ranges from 0 to 1 which an agent receives upon reaching the target. See also: (Nguyen and Gonzalez 2020a,b).

The multi-agent gridworld simulations include four agents, each acting in agreement with the IBL algorithm (described below). Each agent is penalized for each step taken in the environment (-0.01) and for running into walls or other agents (-0.05). If two agents attempt to move to the same position, one is randomly prioritized and moves successfully, while the other receives the collision penalty (-0.05) and their position is unchanged. At the beginning of each episode, agents spawn at their respective location and each agent simultaneously takes an action until all agents have reached a target or 31 steps have elapsed, forming a trajectory. Importantly, no two agents may reach the same target; once a target has been reached by one agent, they occupy that position and stop interacting with the environment until the episode terminates.

Instance-Based Learning Model in the Multi-Agent Gridworld task

Each agent acts according to an algorithm defined in the Instance-Based Learning Theory (IBLT) (Gonzalez, Lerch, and Lebiere 2003). Each agent in the team stores and manages its own memory of instances. In IBLT an instance is a memory unit stored when a decision is evaluated or experienced, and it is comprised of three distinct parts: a situation (or state) S (the attributes describing the context of the decision), a decision (or action) A, and a utility (or reward) R. The IBL agent of the gridworld stores instances that contain the location $(x; y)$, the decision is the possible movement of the agent, and the outcome that the agent receives as a result.

To make choices, IBL models use the choice mechanism *Blending* (Gonzalez and Dutt 2011), defined as an expected value for each alternative, where the outcomes of past instances are weighted by the probability of retrieving such instances from memory. The IBL model selects the alternative with the highest blended value.

Let an option $k = (S; A)$ be defined by taking an action A in state S. At time t, assume that there are $n_{k;t}$ different generated instances $(k; x_{i;k;t})$ for $i = 1; :::; n_{k;t}$, which correspond to selecting k and achieving outcome $x_{i;k;t}$. Then, IBLT associates each instance i in memory with an activation value (Eqn. 1), representing how readily available that information is in memory (Anderson and Lebiere 2014):

$$Act_{i;k;t} = \ln \sum_{t' \in T_{i,k,t}} (t - t')^{-d} + \ln \frac{1 - \ _{i;k;t}}{\ _{i;k;t}} \quad (1)$$

where d and are decay and noise parameters, respectively; $T_{i;k;t} \subset \{0; :::; t-1\}$ is the set of previous steps in which the instance i was observed, and $_{i;k;t} \sim Uniform(0; 1)$ is a randomly generated value. The activation is used to determine the probability of retrieval (Eqn. 2) of an instance from memory:

$$p_{i;k;t} = \frac{e^{Act_{i,k,t}/\tau}}{\sum_{j=1}^{n_{k,t}} e^{Act_{j,k,t}/\tau}} \quad (2)$$

This, in turn, is used to calculate the expected utility, or blended value (Eqn. 3, of an option k, based on a blending mechanism designed for choice tasks (Lebiere 1999; Lejarraga, Dutt, and Gonzalez 2012; Gonzalez and Dutt 2011):

$$V_{k;t} = \sum_{i=1}^{n_{k;t}} p_{i;k;t} x_{i;k;t} \quad (3)$$

where is the Boltzmann constant that we define as $= \frac{1}{\sqrt{2}}$.

Simulation Experiment

We ran a 2 (Distribution of rewards: Dirichlet or Uniform) $\times$ 2 (Spawn location: common or distinct)$\times$ 2 (Field of View (FoV): 1 or 3) simulation experiment.

Reward distribution: target rewards are distributed according to a Dirichlet distribution, where one target is valued significantly higher than the rest, or uniformly, where reaching any target will return a reward of 1.0.

Spawn location: the Spawn location is either common to all agents (with the common spawn point changing in each episode) or distinct (agents rotate through these distinct positions between episodes).

Inter-agent visibility: the FoV refers to the observability of the agent's surroundings. FoV1 indicates the the agent "sees" nothing around it, where FoV3 indicates that the agent observes whether another agent is in a space of size 3×3 around its current $x - y$ coordinate. This changes the instance representation of agent j from $S_j = (x_j; y_j)$ to $S_j = (x_j; y_j; \{x_k; y_k\}_{k=1, k \neq j}^{\#\text{agents}})$ where $x_k; y_k$ are null values if agent k is not within the 3×3 field of view.

For the dependent measures, we considered the following metrics:

(1) Efficiency: the ratio of their positive rewards to the movement cost, formally measured for each agent's trajectory $\mathcal{T}_j$:

$$\text{Efficiency}(\mathcal{T}_j) = \left(\ |\mathcal{T}_j| \right)^{-1} \sum_{i=1}^{|\mathcal{T}_j|} \max(0; R_{j;t})$$

where $= 0.01$ is the step penalty.

(2) Coordination: the proportion of times that all agents reach a target in a single episode. In our task, we define coordinated behavior to be behavior that results in successful completion of the task—reaching all of the unique targets in an episode. In order for this to occur, the agents must each navigate to distinct targets, and avoid competition or strategic mismatches that result in collisions.

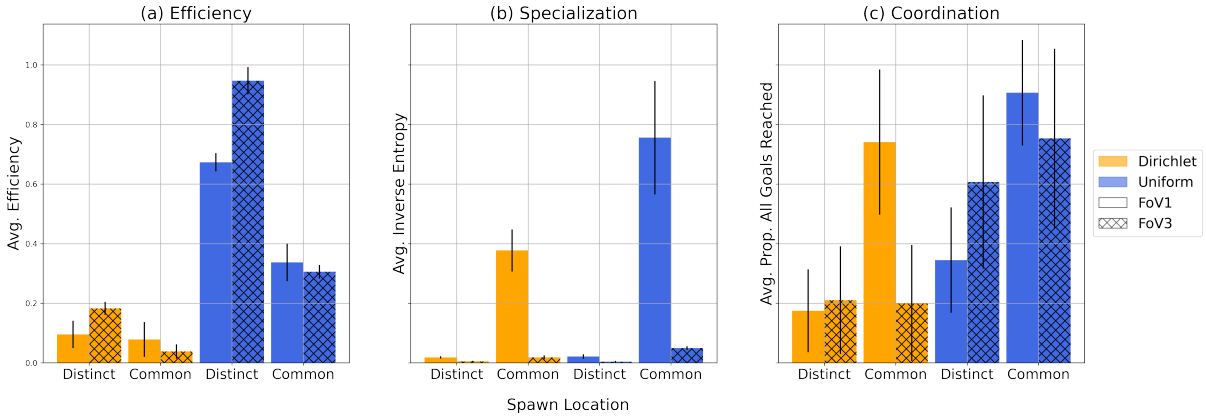

Figure 2: Results for the simulations in terms of (a) efficiency, (b) specialization, and (c) coordination.

(3) Specialization: for each agent we use the inverse entropy of the distribution over outcomes. Formally, this is defined by

$$\text{Entropy(j)}^{-1} = -\sum_{o} p(o_j)\log p(o_j)^{-1}$$

where $= 10^{-4}$ and $p(o_j)$ is given by the proportion of times that agent j reached outcome o in a given gridworld. This metric captures the extent to which agents successfully reach a single target across episodes, as opposed to varied targets. Such behavior would correspond to varying learned strategies, such as "always go to the green target" as opposed to "go to the nearest target."

Simulations consist of 11 different gridworld configurations (different location of goals and obstacles and specific reward values in the Dirichlet case). Agents learn over 1,000 episodes, where the environment and agent position are reset at the initialization of each episode. These runs of 1,000 episodes occur independently for each gridworld configuration. We then average over the results of these independent runs. We perform min-max normalization for both the efficiency and specialization metrics.

Results

Results for each condition are shown in Figure 2. **Efficiency** (Fig. 2a) is higher with uniform than with Dirichlet rewards. We also observe that common spawn points result in lower levels of efficiency compared to distinct spawns. The increased field of view (FoV3) results in higher efficiency in both reward distributions but only when spawn positions are distinct. Essentially, there is higher efficiency when agents do not need to compete for a high-value goal, spawning from different locations reduces such competition making the agents more efficient, particularly with a larger FoV.

We observe increased **specialization** (Fig. 2b) when agents spawn from a common location and have a FoV1 rather than FoV3. With a FoV3 an agent "sees" other agents getting a target, enabling them to alter their trajectory towards a different target.

When the spawn locations are identical, we observe high specialization. Figure 3 shows the target reached, or lack thereof, in a sample trial on a single grid configuration for the distinct and common spawn conditions, both with FoV1 and Dirichlet Rewards. In the Distinct condition, agents reach a variety of different targets, and often fail to reach one at all. In contrast, in the common spawn condition, agents learn to specialize and attempt to reach a particular target—resulting in higher coordination (when all targets are reached) and relatively fewer occurrences of any agent not reaching a target at all.

These results offer a potential explanation for a lower efficiency in the common spawn conditions: agents attempt to reach a specific target, even if it is less efficient than reaching a goal near the spawn location; a distinct spawn condition may lead agents to seek out the goal that they believe is closest to them.

Finally, we observe better **coordination** (Fig. 2c) in Uniform compared to Dirichlet rewards. Coordinated behavior is also learned *faster* in the uniform reward case, and particularly so with FoV1 and common spawn. Figure 4 depicts the level of coordination over time for each condition, showing that Uniform reward conditions tend to reach higher levels of coordination both overall and earlier in the trials. It similarly shows common spawn conditions outperform their distinct spawn counterpart conditions. The results for improved coordination in Uniform conditions may be due to agents being able to "settle" for distinct outcomes that are equally valued, as opposed to lower value outcomes in the Dirichlet case. FoV1 results in better coordination compared to FoV3 in the common spawn conditions; particularly in the case of a Dirichlet reward distribution. This may be attributed to the increased difficulty of strategy formation: agents have the ability to alter their paths based on the location of other agents, and doing so is more complex as they must both navigate to a goal and select actions such that collisions are avoided. We also note that the highest specialization corresponds to the highest coordination. High specialization corresponds to simple strategies, and it is easier to coordinate when the behavior is less complex.