Targeted Information Gain for Real-Time Reactive Policy Learning for Robotic Weed Killing

Wyatt McAllister1

[1]University of Illinois at Urbana-Champaign

November 17, 2017

Motivation

Distributed Autonomous Systems

- Increase the capability of autonomous devices for distributed applications in dynamic and uncertain environments.
- Guarantee these systems have high performance, but also have the capability to improve their performance by interaction with users and the environment.
- Use such systems to augment human capability by both improving quality of life through improved infrastructure, and compensating users in goods and services for their contribution to teaching this infrastructure.

Key Directions to Achieve These Goal

Critical Goals

- Demonstrate the capability of autonomous decision making for novel industrial applications, and enhance performance in such applications.
- Develop a scalable multi-tiered task communication framework for specifying clear instructions to different systems in different human languages.
- Create a hierarchical learning and decision framework that dynamically integrates multiple task specific learners and planners into a system which has the ability to plan abstractly to complete high level objectives.

Key Directions to Achieve These Goal

Step 1

- **Demonstrate the capability of autonomous decision making for novel industrial applications, and enhance performance in such applications.**
- Develop a scalable multi-tiered task communication framework for specifying clear instructions to different systems in different human languages.
- Create a hierarchical learning and decision framework that dynamically integrates multiple task specific learners and planners into a system which has the ability to plan abstractly to complete high level objectives.

Industrial Application

High Level Learning and Planning for the TERRA Project

- Design distributed autonomous systems for scalable applications in industrial agriculture.
- Use many small low cost robots to provide for small agriculture applications in developing nations as well as large farms.

Problem in Application

Weeding Under a Crop Canopy

- For many crops, such as corn, weeding must be done under a canopy, and therefore under partially observable conditions.
- Robots should have the ability to coordinate their weeding under this partial observability, to optimize yield using minimal time and resources.
- In these circumstances, robots can only classify weeds within a local radius, and thus only information about the neighboring rows is known.

Relevant Sources

Robot Foraging

- Foraging has long been considered a key problem in multi-agent robotics [Cao et al., 1997].
- Recent work has solved this problem under partial observability, solving a search and rescue problem with ground robots and UAVs [Liu et al., 2017].
- However, this work assumed the capability of the UAVs to localize the victims. We aim to design a real-time system for the weeding task using only the information collected from the surroundings of the ground robots.

MDP Formulation

Station, Action, Reward

- State: Current row for each agent.

$$S\left(t\right)=\left\{x_i\left(t\right),i\in I\right\},I\equiv\left\{1,...,N_{\mathrm{agents}}\right\}$$

- Action: Target row for each agent.

$$A\left(t\right)=\left\{a_i\left(t\right)=x_i\left(t+1\right),i\in I\right\}$$

- Reward: We want to simultaneously maximize the number of weeds killed and minimize the operation time. Therefore, the reward is the sum over all agents of number of weeds killed, time discounted by the operation time. This structure is similar to that used in past work [Matarić, 1997].

$$R\left(t\right)=\sum_{i\in I}\gamma^{T_{\textsf{operation},\mathrm{i}}\left(a_i\left(t\right)\right)}R_{\textsf{weeds killed},\mathrm{i}}\left(a_i\left(t\right)\right)$$

Explanation of Reward

- For each agent, the planned reward is the reward of the proposed row.

$$R_i\left(a_i\left(t\right)\right) = \sum_{y=0}^{N_{\mathrm{dim}}} R_W\left(a_i\left(t\right), y\right)$$

- For each agent, the planned operation time is the operation time of the proposed row.

$$T_i\left(a_i\left(t\right)\right) = \frac{\left(x_i\left(t+1\right) - x_i\left(t\right)\right)}{v_i} + \frac{Y_{\mathrm{dim}}}{v_i} + T_{\mathrm{kill}} \cdot N_W\left(x_i\left(t+1\right)\right)$$

- N_{dim} is the number of squares in a row (85), Y_{dim} is the length of each row (209 feet), $N_W\left(x\right)$ is the number of weeds in each row, and $R_W\left(x, y\right)$ is the reward for each weed at each location $\left(x, y\right)$ (the reward increases from a baseline value as the weed grows).

Assumptions for MDP Approach

Full Communication - Centralized Approach

- For this project, we start with the assumption of full communication between all the agents about their current location, the action they have selected, and the total reward they have collected from the environment.

$$\{S\left(t\right), A\left(t\right), R\left(t\right)\} \Rightarrow \text{Known}$$

- The reason for this is that each acre of crop land is only 209 feet by 209 feet, so each agent will be at most 295 feet away, well within the range of shortwave communication.

Assumptions for MDP Approach

Single Agent Selection

- In our environment, we assumed robots can only cross rows at the edges of the field, and two robots cannot always move side by side. Therefore, selecting one agent per row is sensible in order to maximize efficiency, so no two agents pick the same row.

$$A = \{a_i(t) : a_i(t) \neq a_j(t) \quad \forall i \neq j\}$$

- We therefore select the agent with the maximum value for each row, and break ties based on increasing order of agent indices.

Assumptions for MDP Approach

One Step Look Ahead

- In this set of experiments, all agents are homogeneous, meaning they have identical capabilities. Therefore, they each have identical value for the same rows under identical initial conditions.

$$x_i\left(t\right) = x_j\left(t\right) \Rightarrow R_i\left(a_i\left(t\right)\right) = R_j\left(a_j\left(t\right)\right) \quad \forall\left(i, j\right) \in I$$

- This implies that the value of the one-step policy for different time instances will not change depending on the states of the agents, and thus one step learning is sensible for this environment.

Assumptions for MDP Approach

The Case of Full Observability

- As a baseline, we consider the case of full observability, where we assume the number and heights of weeds in every row is known.

$$R_W(x) \Rightarrow \text{Known} \quad \forall x$$

- In this case, we can easily benchmark the performance of the learning approach.

The Case of Partial Observability

- In the partially observable case, we assume that robots can classify weeds within a certain radius.

$$R_W(x) \Rightarrow \text{Known} \forall \{x : \{x_{visited}\} \pm r_{obs.}\}$$

- We collect information on the neighboring rows, and update the global environmental model with this information.

RP Approach

Multi-Agent Reactive Policy Learning with One Step Look Ahead

- We take a factored MDP approach to this problem, which allows us to break up the problem into a set of factored MDPs for each agent.
- This decentralized approach is guaranteed to find an optimum solution, since factored MDPs which are observation, transition, and reward independent, may optimize the total additive reward by optimizing each agents local reward [Amato et al., 2013].

RP Approach

Multi-Agent Reactive Policy Learning with One Step Look Ahead

- For the Reactive Policy (RP), we plan simultaneously across all the agents, evaluating the expected return for a transition from the agent's current state to any other new state.

$$Q_{t+1}^{i}\left(x_i\left(t\right), a_i\left(t\right)\right) += \alpha\left(\gamma^{T_i\left(a_i\left(t\right)\right)} \cdot R_i\left(a_i\left(t\right)\right) - Q_t^{i}\left(x_i\left(t\right), a_i\left(t\right)\right)\right)$$

- By simultaneously optimizing over the value function for each agent, we plan a coordinated policy sending each agent to the row with maximum value.

$$A\left(t+1\right) = \left\{ \underset{a_i(t)}{\arg\max}\, Q_{t+1}^{i}\left(x_i\left(t\right), a_i\left(t\right)\right) \quad \forall i \in I \right\}$$

Factored MDP Model

Observation, Transition, and Reward Independence [Amato et al., 2013]

- The total observation space is constructed via the observations of each agent, and the factored model is thus observation independent.
- The transition of each agent depends only on its current state and action, not on the other agents, and the factored model is thus transition independent.
- The reward of each agent depends only on its current state and action, not on the other agents, and the factored model is thus reward independent.

Simulation Environment

JavaScript Weed World

- Implemented the Weed World Environment in collaboration with Denis Osipychev, as a grid world of 85 rows of 2.5 foot squares, totaling 4400 square feet or one acre.
- Included weed growth, with existing weeds growing at a fixed rate, and new weeds seeded by weeds above a certain height.

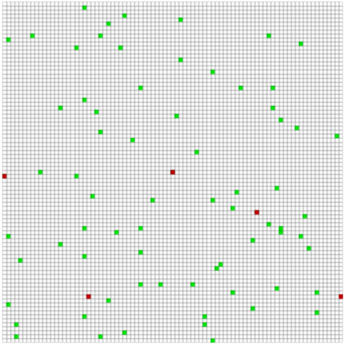

Figure 1: Simulation Environment

Experiment 1

- To establish a baseline, we first assume full observability, giving the planner full knowledge of the locations of all the weeds.
- We then assume that the environment is partially observable, giving the planner knowledge of the weeds adjacent to squares the robots have passed before as the simulation runs.
- We compare the performance of RP approach under full observability to that of partial observability.

Experiment 1

Full Vs. Partial Observability - RP Approach

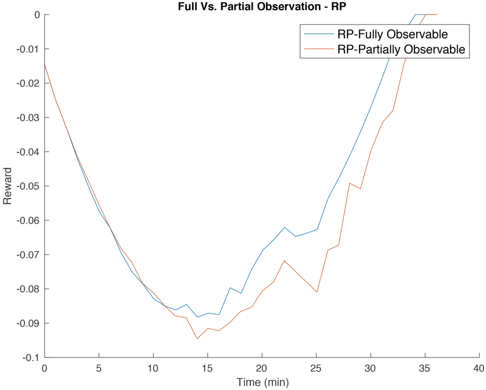

Figure 2: Full Vs. Partial Observability - RP

- We see that the performance of the RP approach does drop for the partially observable case, as expected.

Next Steps

- In these experiments, the RP approach uses Sequential Information Gain (SIG) when there is no prior information in the partially observable case, simply going to the next available adjacent unexplored row.
- We would like to explore the case of targeted Information Gain (TIG), inspired by [Yamauchi, 1997], to see if this improves performance.
- We aim to give exploration preference to rows near those known to have high reward, and characterize the change in performance for the RP approach.

Targeted Information Gain (TIG)

Input: $x_i(t)$: state of agents
Input: $R(x)$: reward for each row
Input: $N_w(x)$: number of weeds in each row
Output: $a_i(t)$: action for each agent

- Step 1: Rank all unexplored rows adjacent to those previously explored by the value of the adjacent row. Call this ranking the exploration value of the row.
- Step 2: If an agent is not assigned to a row and some rows have been explored, send the agent to the row with the highest exploration value.
- Step 3: If an agent is not assigned to a row and no rows have been explored, follow Sequential Information Gain (SIG).

Experiment 2

- We now compare the RP approach with Sequential Information Gain (SIG) to that with Targeted Information Gain (TIG) for the partially observable case.

Experiment 2

Sequential Vs. Targeted Information Gain

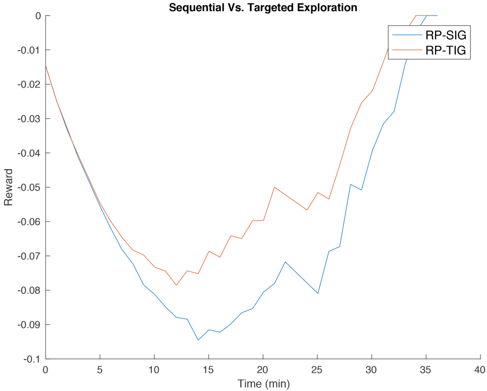

Figure 3: Sequential Vs. Targeted Information Gain RP Approach

- We see significant improvement with targeted information gain, as desired.

Discussion

Performance of Reactive Policy

- We have demonstrated here that we are able to learn an optimum reactive policy for our Weed World environment.
- Performance may be improved if the entire policy is learned in real-time (if the optimal sequence of rows for each agent is learned at every step based on available information).

Challenges and Future Work

Hybridized Targeted Information Gain and Neural Network Approach

- Due to the size of the state space, this approach will require a Neural Network framework, and will thus not be guaranteed to find an optimal solution.
- We hypothesize that the highest performing algorithm will utilize targeted information gain while gathering information and training the neural network, using the trained policy when there is sufficient information.
- We will next extend our simulation to utilize neural networks in order to test this hypothesis.

Links

- Literature Review
- Weed World Implementation

References I

Amato, C., Chowdhary, G., Geramifard, A., Ure, N. K., and Kochenderfer, M. J. (2013).
Decentralized control of partially observable markov decision processes.
In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 2398–2405. IEEE.

Cao, Y. U., Fukunaga, A. S., and Kahng, A. (1997).
Cooperative mobile robotics: Antecedents and directions.
Autonomous robots, 4(1):7–27.

Liu, M., Sivakumar, K., Omidshafiei, S., Amato, C., and How, J. P. (2017).
Learning for multi-robot cooperation in partially observable stochastic environments with macro-actions.
CoRR, abs/1707.07399.

Matarić, M. J. (1997).
Reinforcement learning in the multi-robot domain.
Autonomous Robots, 4(1):73–83.

Yamauchi, B. (1997).
A frontier-based approach for autonomous exploration.
In *Computational Intelligence in Robotics and Automation, 1997. CIRA'97., Proceedings., 1997 IEEE International Symposium on*, pages 146–151. IEEE.