Multi-Agent Learning for Coordinated Robotic Weed Killing

Wyatt McAllister1

[1]University of Illinois at Urbana-Champaign

January 23, 2018

Industrial Application

Robotic Weeding

- Recent work on robotic weeding has focused both on the design of weeding robots and on the challenge of plant recognition [Bakker et al., 2010, Griepentrog et al., 2004, Gai et al., 2015, Lund and Søgaard, 2004].
- Our work aims to leverage strategies for multi-robot coordination in order to extend past work to create a scalable weeding solution.

Problem in Application

Weeding Under a Crop Canopy

- For many crops, including 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.

Further References: Coordinated Robotics

- Cooperative mobile robotics: Antecedents and directions [Cao et al., 1997]
- Reinforcement Learning in the Multi-Robot Domain [Matarić, 1997]
- The impact of diversity on performance in multi-robot foraging [Balch, 1999]
- Learning in behavior-based multi-robot systems: Policies, models, and other agents [Mataric, 2001]
- Multi-Agent Reinforcement Learning (MARL): a critical survey [Shoham et al., 2003]

MDP Formulation

Environmental Parameters

- N_{dim} is the number of squares in a row (85).
- Y_{dim} is the length of each row (209 feet).
- $N_W(x)$ is the number of weeds in each row.
- $R_W(x, y)$ is the reward for each weed at each location (x, y) (the reward increases from a baseline value as the weed grows).
- v_i is the velocity of agent i.
- T_{kill} is the time to kill a weed.

Station, Action, Reward

- State: Current row for each agent.

$$x_i(t) \in S \quad \forall i \in I, \quad \begin{matrix} S \equiv \{1, ..., N_{\text{dim}}\} \\ I \equiv \{1, ..., N_{\text{agents}}\} \end{matrix} \tag{1}$$

- Action: Target row for each agent.

$$a_i(t) = x_i(t+1) \in A \equiv S \tag{2}$$

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

$$R_i(a_i(t)) = \sum_{y=0}^{N_{\text{dim}}} R_W(a_i(t), y) \tag{3}$$

Planned Operation time

- The planned operation time is the sum of the time it takes to move to the proposed row $T_{\text{move to row}}$, the time it takes to move down it $T_{\text{move down row}}$, and the time it takes to weed all the squares in the row $T_{\text{weed row}}$.

$$T_i\left(a_i\left(t\right)\right) = T_{\text{move to row}} + T_{\text{move down row}} + T_{\text{weed row}} \tag{4}$$

$$T_{\text{move to row}} = \frac{\left(x_i\left(t+1\right) - x_i\left(t\right)\right)}{v_i} \tag{5}$$

$$T_{\text{move down row}} = \frac{Y_{\text{dim}}}{v_i} \tag{6}$$

$$T_{\text{weed row}} = T_{\text{kill}} \cdot N_W\left(x_i\left(t+1\right)\right) \tag{7}$$

Value Function

- For this problem, we want to maximize the overall value function, which is the sum over all agents of the planned reward, time discounted by the planed operation time.
- This structure has long been common in multi-robot foraging problems [Matarić, 1997].

$$V(t+1) = \sum_{i \in I} \gamma^{T_{\textbf{operation},\mathrm{i}}(a_i(t))} R_{\mathrm{i}}(a_i(t)) \tag{8}$$

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.

$$\{x_i(t), a_i(t), R_i(a_i(t))\} \Rightarrow \text{Known} \quad \forall \text{i} \tag{9}$$

Single Agent Selection

- In our environment, we assume a field where robots can only cross rows at the edges of the field, and two robots cannot move side by side.
- Under this assumption, it is necessary to select one agent per row.

$$a_i(t) : a_i(t) \neq a_j(t) \quad \forall i \neq j \tag{10}$$

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

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.

$$\begin{aligned} & x_i(t) = x_j(t) \\ & \Rightarrow R_i(a_i(t)) = R_j(a_j(t)) \quad \forall (i,j) \in I \end{aligned} \tag{11}$$

- 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 feasible 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.

$$R_W(x, y) \Rightarrow \text{Known} \, \forall (x, y) \in S \tag{12}$$

- 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.

$$R_W(x, y) \Rightarrow \text{Known} \, \forall \{(x, y)_{\text{visited}} \pm r_{\text{obs.}}\} \tag{13}$$

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

Assumptions for MDP Approach

MDP Model Vs. POMDP Model Choice

- Our choice to utilize an MDP model over a POMDP model in this case is motivated by the fact that even though the entire environment may not be observed, the reward in the observed space is known with high fidelity.
- This allows us to simplify the problem to an MDP problem in the observed space, without having to compute the probability of states given observations.

Reactive Policy (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 factored 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].

Multi-Agent 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.

$$V_{t+1}^{i}\left(x_i\left(t\right), a_i\left(t\right)\right) = \\ \alpha\left(\gamma^{T_i(a_i(t))} \cdot R_i\left(a_i\left(t\right)\right) - V_t^i\left(x_i\left(t\right), a_i\left(t\right)\right)\right) \tag{14}$$

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

$$a_i\left(t\right) = \underset{a_i(t)}{\arg\max}\, V_{t+1}^{i}\left(x_i\left(t\right), a_i\left(t\right)\right) \qquad \forall i \in I \tag{15}$$

Targeted Observation (TO)

- The RP approach above uses Sequential Observation (SO) when there is no prior information, going to the next available adjacent unexplored row.
- We would like to explore the case of Targeted Observation (TO), inspired by Frontier Based Exploration in [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 Observation (TO)

- We denote the exploration value for each unexplored row by $E^i_{t+1}\left(x_i\left(t\right), a_i\left(t\right)\right)$, which is calculated for each agent as the sum of the values of adjacent explored rows.

$$\begin{aligned} &E^i_{t+1}\left(x_i\left(t\right), a_i\left(t\right)\right) = \\ &V^i_{t+1}\left(x_i\left(t\right)+1, a_i\left(t\right)\right) + V^i_{t+1}\left(x_i\left(t\right)-1, a_i\left(t\right)\right) \end{aligned} \tag{16}$$

- We then explore rows with exploration value greater than or equal to the maximum value for explored rows.

$$\begin{aligned} &\arg\max_{a_i(t)} E^i_{t+1}\left(x_i\left(t\right), a_i\left(t\right)\right) \geq \arg\max_{a_i(t)} V^i_{t+1}\left(x_i\left(t\right), a_i\left(t\right)\right) \\ &\Rightarrow a_i\left(t\right) = \arg\max_{a_i(t)} E^i_{t+1}\left(x_i\left(t\right), a_i\left(t\right)\right) \end{aligned} \tag{17}$$

- If no rows have been explored, then we simply follow Sequential Observation (SO).

Targeted Observation (TO)

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 some rows have been explored, and the maximum exploration value for the unexplored rows is greater than or equal to the maximum value for the explored rows, go 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 Observation (SO).

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 dynamic weed-growth model, real-time visualization, and portable visualization for on-line deployment.

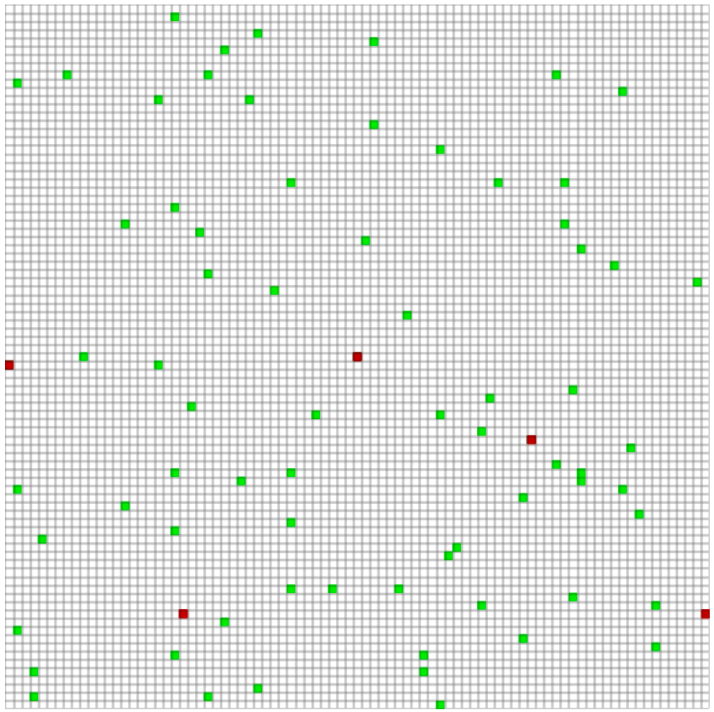


Figure 1: Simulation Environment

Dynamic Weed Growth Model

- Weeds are initially distributed uniformly at random with an arbitrary reward of 0.001 corresponding to a heigh of one inch.
- Existing weeds growing at a fixed rate until they are killed or reach a height of 8 inches, at which point they may seed empty squares within a local radius dependent on their height.
- Weeds may spawn in empty squares with a probability of 10^{-6}, corresponding to the rate of growth from the underlying seed bank.

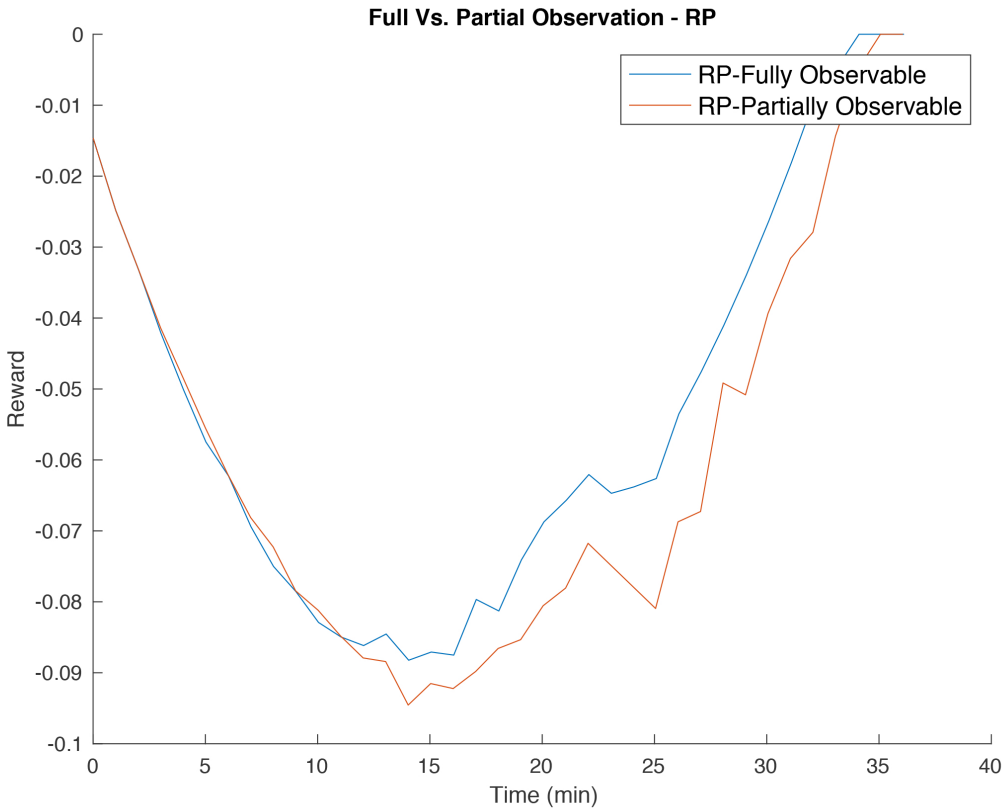
$$P_{spawn} = 10^{-6} \tag{18}$$

$$R_W(x, y) = 0.001 + \frac{1}{24 \cdot 60 \cdot 60} \frac{inch}{day} \tag{19}$$

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.

Full Vs. Partial Observability - RP Approach



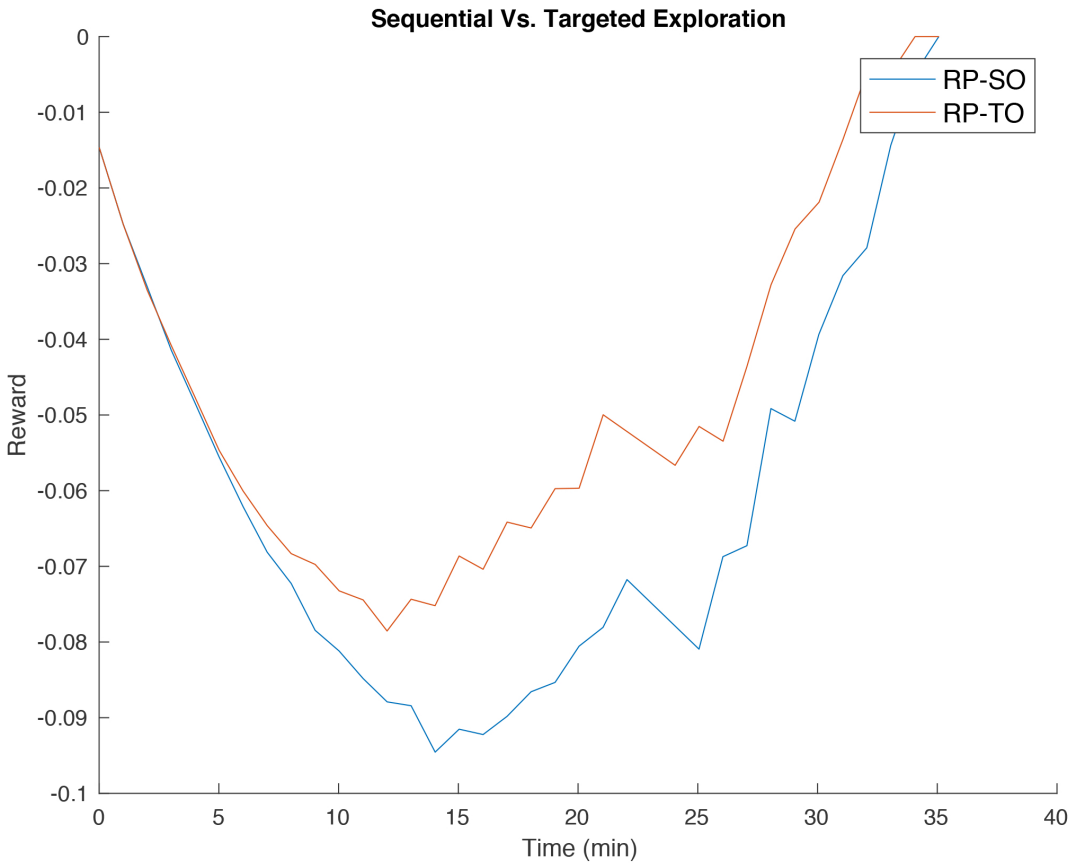
- We see that the performance of the RP approach does drop for the partially observable case, as expected.

Figure 2: Full Vs. Partial Observability - RP

Experiment 2

- We now compare the RP approach with Sequential Observation (SO) to that with Targeted Observation (TO) for the partially observable case.

Sequential Vs. Targeted Observation



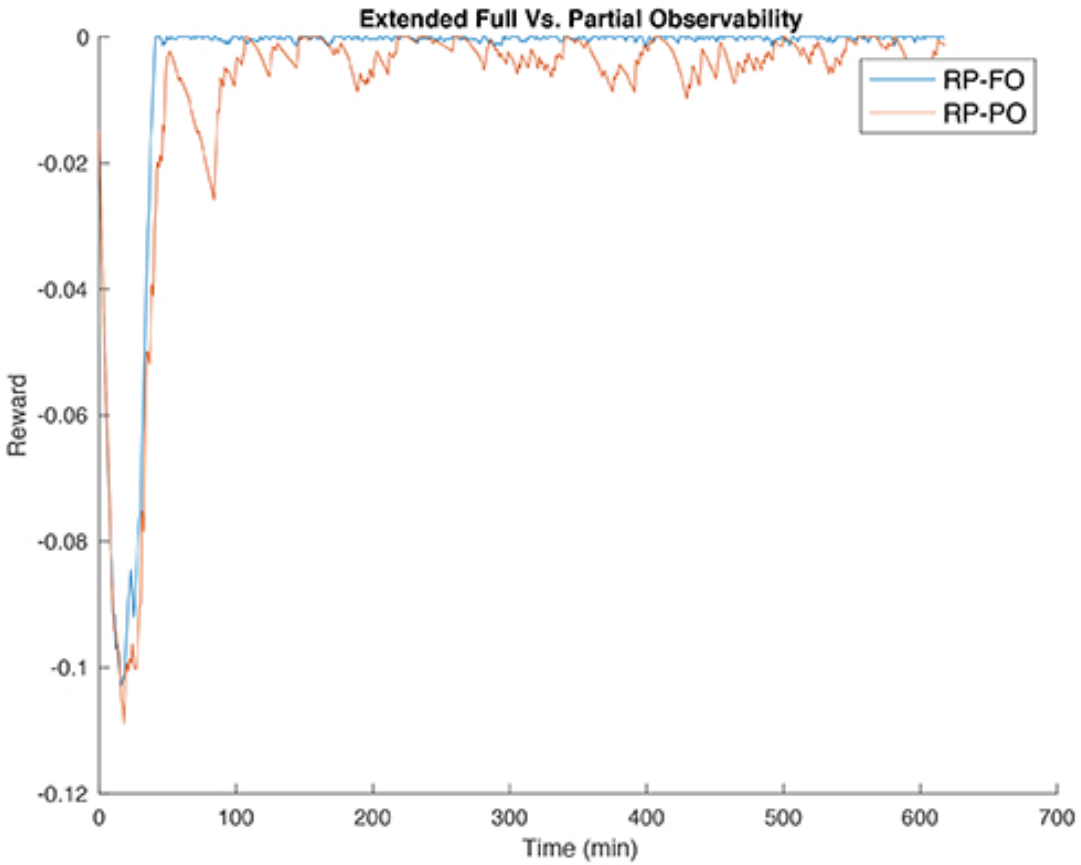
- We see significant improvement with Targeted Observation, as desired.

Figure 3: Sequential Vs. Targeted Observation RP Approach

Experiment 3

- We now extend the time horizon to ten hours, to gauge the performance of our algorithm for one solar day in a mid-upper latitude environment.
- This interval was chosen because our robots utilize opticals sensors which whose performance may be affected in low light conditions.
- We compare the RP approach with TO under partial observability to the fully observable case.

Extended Time: Full Vs. Partial Observability



- We see that the RP approach with TO, while not as high performing, is able to track the fully observable case, even in the extended time horizon.

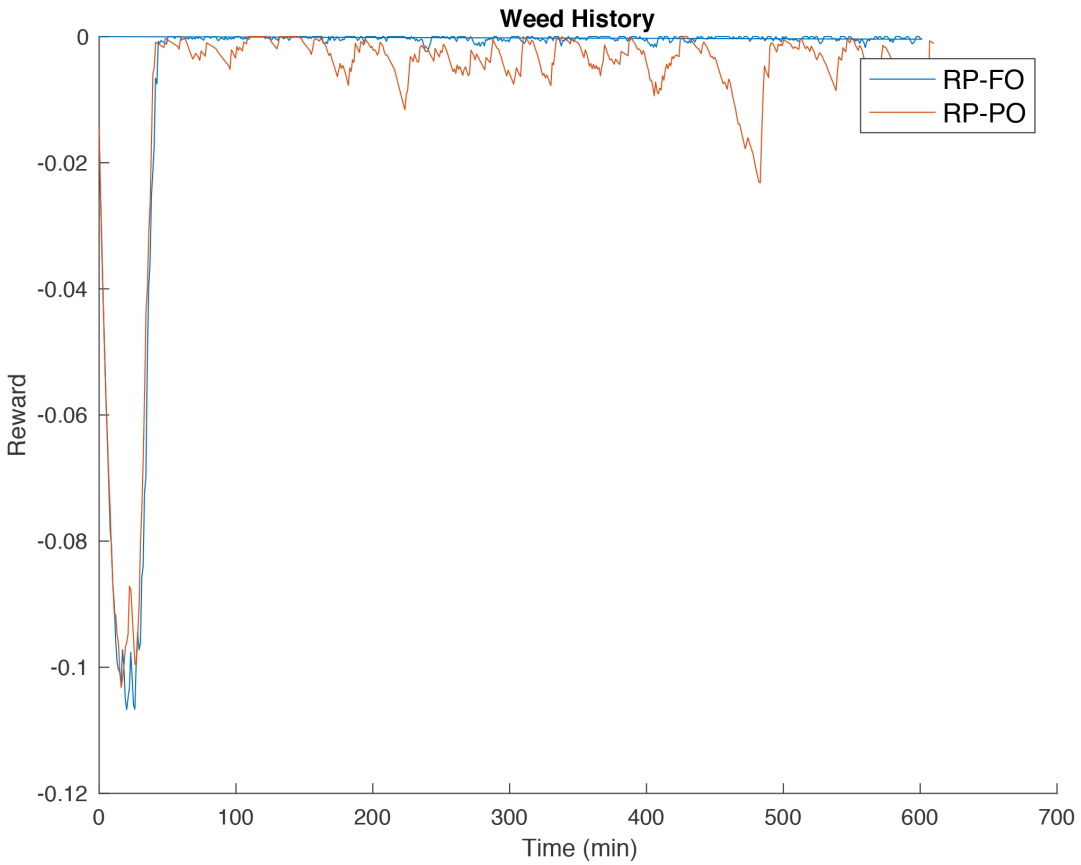
Figure 4: Extended Time: Full Vs. Partial Observability

Experiment 4

- We now consider the case in which the initial uniform distribution of weed spawning probabilities for each location increases over time in locations where weeds have spawned before.

$$P_{\text{spawn}} = H_{xy} \cdot 10^{-6} \quad H_{xy} = \sum_{t=0}^{t_{current}} N_{\text{weeds}}(x, y) \tag{20}$$

Extended Time - Weed History: Full Vs. Partial Observability



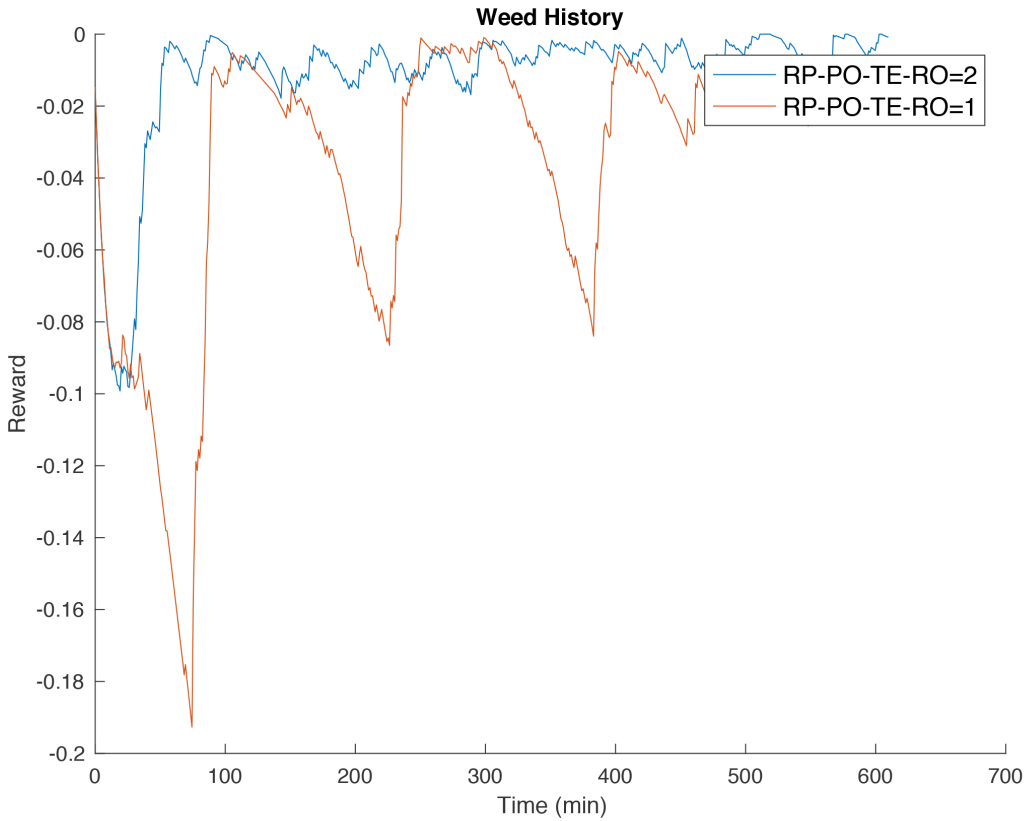
- We see that the RP approach with TO still tracks the fully observable case with the time varying spawning probability distribution.

Figure 5: Extended Time - Weed History: Full Vs. Partial Observability

Experiment 5

- The previous experiments used an observation radius of two.
- The current system is able to collect data about the adjacent two rows, but only the data from the first row has high fidelity.
- We now decrease the observation radius to one and observe the results.

Extended Time - Weed History: Full Vs. Partial Observability



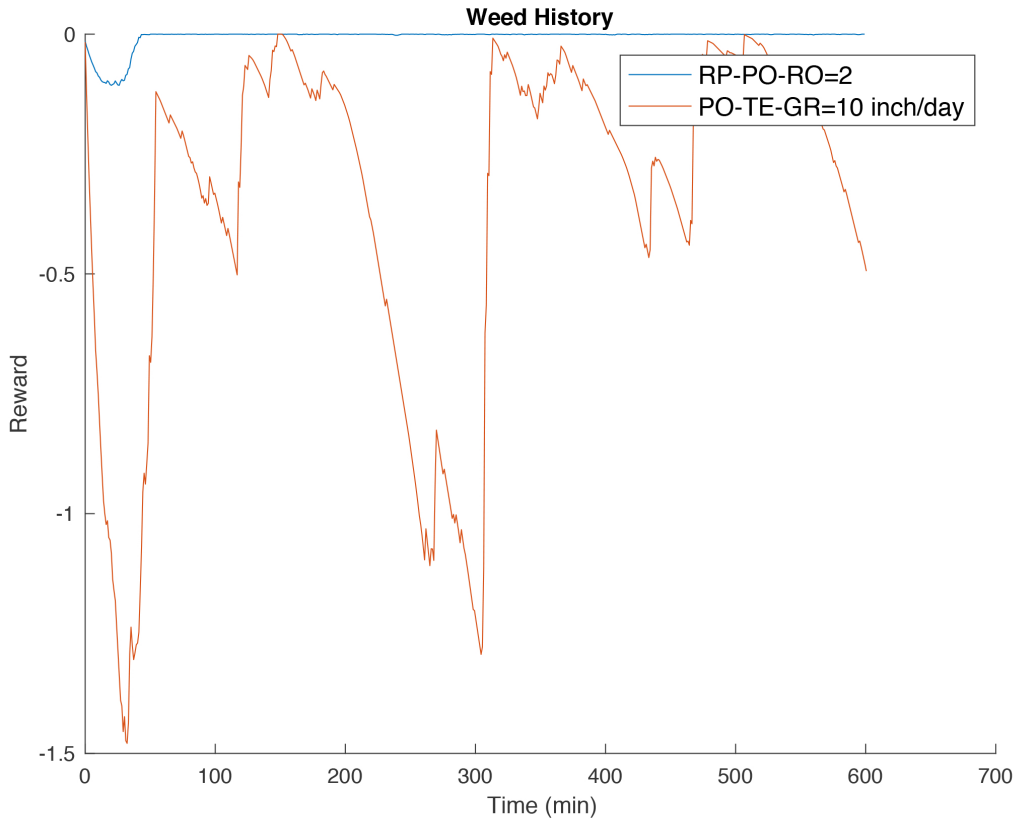
- We see a significant drop in performance for the observation radius of one.
- While the system still weeds the field, it does so more slowly than the case with observation radius of two.

Figure 6: Extended Time - Weed History: Partial Observability, Observation Radius = 1 Vs. Observation Radius = 2

Experiment 6

- The previous experiments utilized a weed growth-rate of one inch per day.
- We will increase this by a factor of ten and repeat the previous experiment with the observation radius equal to one, and with weed history, comparing once more with the fully observable case.

Extended Time - Weed History: Full Vs. Partial Observability



- We see that increasing the growth rate increases the total reward, as expected. However, the system is still able to weed the field in under two hours.

Figure 7: Extended Time, Weed History - Partial Observability, Observation Radius = 1, Weed Growth Rate = 10 inches/day Vs. Full Observability,

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).

Hybridized Targeted Observation 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 Observation 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.


Bakker, T., Asselt, K., Bontsema, J., Müller, J., and Straten, G. (2010).
Systematic design of an autonomous platform for robotic weeding.
Journal of Terramechanics, 47(2):63–73.


Balch, T. (1999).
The impact of diversity on performance in multi-robot foraging.
In *Proceedings of the third annual conference on Autonomous Agents*, pages 92–99. ACM.


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


Gai, J., Tang, L., and Steward, B. (2015).
Plant recognition through the fusion of 2d and 3d images for robotic weeding.
In *2015 ASABE Annual International Meeting*, page 1. American Society of Agricultural and Biological Engineers.


Griepentrog, H.-W., Christensen, S., Søgaard, H. T., Nørremark, M., Lund, I., and Graglia, E. (2004).
Robotic weeding.
In *Proceedings of AgEng*.

References II

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.

Lund, I. and Søgaard, H. T. (2004).
Robotic weeding-plant recognition and micro spray on single weeds.
In *Robotic Weeding-Plant recognition and micro spray on single weeds*.

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

Mataric, M. J. (2001).
Learning in behavior-based multi-robot systems: Policies, models, and other agents.
Cognitive Systems Research, 2(1):81–93.

Shoham, Y., Powers, R., and Grenager, T. (2003).
Multi-agent reinforcement learning: a critical survey.
Web manuscript.

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.