# REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188



| 1. REPORT DATE (DD-MM-YYYY) | 2. REPORT TYPE | 3. DATES COVERED (From - To) |
|---|---|---|
| 30-09-2016 | Final Report | 1-May-2011 - 30-Apr-2015 |

**4. TITLE AND SUBTITLE**
Final Report: Concurrent Learning of Control in Multi-agent Sequential Decision Tasks

**5a. CONTRACT NUMBER**
W911NF-11-1-0124

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**
611102

**6. AUTHORS**
Bikramjit Banerjee

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAMES AND ADDRESSES**
University of Southern Mississippi
118 College Drive #5157

Hattiesburg, MS 39406 -0001

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)**
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

**10. SPONSOR/MONITOR'S ACRONYM(S)**
ARO

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**
57785-NS.19

**12. DISTRIBUTION AVAILIBILITY STATEMENT**
Approved for Public Release; Distribution Unlimited

**13. SUPPLEMENTARY NOTES**
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

**14. ABSTRACT**
The overall objective of this project was to develop multi-agent reinforcement learning (MARL) approaches for intelligent agents to autonomously learn distributed control policies in decentralized partially observable Markov decision processes (Dec-POMDPs), without prior knowledge of the model parameters. During this project, we developed algorithms for decentralized learning of policies in Dec-POMDPs,
established performance bounds, evaluated these algorithms both theoretically and empirically,

**15. SUBJECT TERMS**
decentralized POMDPs, multi-agent learning

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 15. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Bikramjit Banerjee |
| UU | UU | UU | UU | | 19b. TELEPHONE NUMBER 601-266-6287 |

# RPPR

as of 17-Apr-2018

Agency Code:

Proposal Number: **Agreement Number:**

Organization:
Address: , ,
Country:
DUNS Number: EIN:
**Report Date:** Date Received:
for Period Beginning and Ending
**Title:**
**Begin Performance Period:** **End Performance Period:**
**Report Term:** -
Submitted By: Email:
Phone:

**Distribution Statement:** -

**STEM Degrees:** **STEM Participants:**

**Major Goals:**

**Accomplishments:**

**Training Opportunities:**

**Results Dissemination:**

**Plans Next Period:**

**Honors and Awards:**

**Protocol Activity Status:**

**Technology Transfer:**

# Final Report

## Project Title: Concurrent Learning of Control in Multi-agent Sequential Decision Tasks

Proposal # 57785-NS
Agreement # W911NF-11-1-0124

**Project period: May-01-2011 to April-30-2015**

PI: Bikramjit Banerjee
The University of Southern Mississippi
118 College Dr. # 5106
Hattiesburg, MS 39406-0001

## Project Objective

The overall objective of this project was to develop multi-agent reinforcement learning (MARL) approaches for intelligent agents to autonomously learn distributed control policies in decentralized partially observable Markov decision processes (Dec-POMDPs), without prior knowledge of the model parameters.

## Problem Model

We can define a Dec-POMDP as a tuple $\langle n, S, A, P, R, \Omega, O \rangle$, where:

- $n$ is the number of agents in the multi-agent system.
- $S$ is a finite set of (unobservable) environment states.
- $A = \times_i A_i$ is a set of joint actions, where $A_i$ is the set of individual actions that agent $i$ can perform.
- $P(s'|s, \vec{a})$ gives the probability of transitioning to state $s' \in S$ when joint action $\vec{a} \in A$ is taken in state $s \in S$.
- $R(s, \vec{a})$ gives the immediate reward the agents receive upon executing action $\vec{a} \in A$ in state $s \in S$.
- $\Omega = \times_i \Omega_i$ is the set of joint observations, where $\Omega_i$ is the finite set of individual observations that agent $i$ can receive from the environment.
- $O(\vec{\omega}|s', \vec{a})$ gives the probability of the agents jointly observing $\vec{\omega} \in \Omega$ if the current state is $s' \in S$ and the previous joint action was $\vec{a} \in A$.

Learning agents attempt to learn coordinated *policies* that achieve highest expected reward, without prior knowledge of the above model parameters (i.e., functions $P, R, O$). For finite horizon problems, a policy is a mapping from the history of an agent's own actions and observations, to its own action; i.e., for agent $i$ and step $t$, the policy is $\pi_i^t : (A_i \times \Omega_i)^{t-1} \mapsto A_i$. In infinite horizon problems, each agent learns a finite state controller.

## Technical Outcome

During this project, we developed algorithms for decentralized learning of policies in Dec-POMDPs, established performance bounds, evaluated these algorithms both theoretically and empirically, and applied some of our approach to the problem of camera surveillance. The key technical contributions have been described in some details in the interim progress reports submitted before, so I shall only list them here with citations to publications supported by this grant.

**Informed Initial Policy:** [1, 9, 5].

**MCQ-Alt:** [4, 8].

**Iterative MCQ-Alt:** [6].

**Rehearsal Based Learning:** [3, 7, 11].

**Regret Minimization:** [10].

**Application to Distributed Surveillance:** [2, 12].

## Student Support

This grant supportted one M.S. student and one Ph.D. student for the full length of their work. Support included stipend and conference travel. The M.S. thesis [12] was completed in 2014. All work for the Ph.D. dissertation was also completed during the project period; however this doctoral student (Landon Kraemer) joined Amazon, and remains ABD at this time. Furthermore, the grant also provided partial support for another Ph.D. student (Rajesh Yellamraju; he left the project under unforeseen circumstances), and one M.S. student who joined the doctoral program in Spring 2016.

## Publications Acknowledging this Grant

### Journal Papers Published/In Press


[1] Landon Kraemer and Bikramjit Banerjee. Reinforcement Learning of Informed Initial Policies for Decentralized Planning. In *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, Vol 9(4), article 18, ACM Press, 2014.

[2] Bikramjit Banerjee and Landon Kraemer. Stackelberg Surveillance. In *Informatica*, Vol 39(4), pages 451–458, Slovenian Societe Informatika, 2015.

[3] Landon Kraemer and Bikramjit Banerjee. Multi-agent Reinforcement Learning as a Rehearsal for Decentralized Planning. (To appear) In *Neurocomputing*, Elsevier, 2016.


### Conference Papers


[4] Bikramjit Banerjee, Jeremy Lyle, Landon Kraemer, and Rajesh Yellamraju. Sample bounded distributed reinforcement learning for decentralized pomdps. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*, pages 1256–1262, Toronto, Canada, July 2012.

[5] Landon Kraemer and Bikramjit Banerjee. Informed initial policies for learning in dec-pomdps. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (Student Abstract, AAAI-12)*, pages 2433–2434, Toronto, Canada, July 2012.

[6] Bikramjit Banerjee. Pruning for monte carlo distributed reinforcement learning in decentralized pomdps. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI-13)*, pages 88–94, Bellevue, WA, July 2013.

[7] Landon Kraemer and Bikramjit Banerjee. Concurrent reinforcement learning as a rehearsal for decentralized planning under uncertainty (extended abstract). In *Proceedings of the 12th International Conference on Autonomous Agents and Multi-agent Systems (AAMAS-13)*, pages 1291–1292, St. Paul, MN, May 2013.

## Workshop Papers

[8] Bikramjit Banerjee, Jeremy Lyle, Landon Kraemer, and Rajesh Yellamraju. Solving finite horizon decentralized pomdps by distributed reinforcement learning. In *Proceedings of the AAMAS-12 Workshop on Multiagent Sequential Decision Making Under Uncertainty (MSDM-12)*, pages 9–16, Valencia, Spain, June 2012.

[9] Landon Kraemer and Bikramjit Banerjee. Informed initial policies for learning in dec-pomdps. In *Proceedings of the AAMAS-12 Workshop on Adaptive Learning Agents (ALA-12)*, pages 135–143, Valencia, Spain, June 2012.

[10] Bikramjit Banerjee and Landon Kraemer. Counterfactual regret minimization for decentralized planning. In *Proceedings of the AAMAS-13 Workshop on Adaptive Learning Agents (ALA-13)*, pages 84–91, St. Paul, MN, May 2013. Also appeared in Proceedings of the 8th Workshop on Multiagent Sequential Decision Making (MSDM-13), pp 32–39.

[11] Landon Kraemer and Bikramjit Banerjee. Rehearsal based multi-agent reinforcement learning of decentralized plans. In *Proceedings of the 8th AAMAS-13 Workshop on Multiagent Sequential Decision Making (MSDM-13)*, pages 24–31, St. Paul, MN, May 2013.

## Thesis

[12] Madhavi Chittireddy. Reinforcement Learning of Distributed Surveillance Plans. Master's thesis, School of Computing, The University of Southern Mississippi, 2014. Available at `http://www.cs.usm.edu/~banerjee/tr/Madhavi_thesis.pdf`.