REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
22-05-2021	Final Report	3-Sep-2015 - 31-Aug-2018

4. TITLE AND SUBTITLE
Final Report: Probably Approximately Correct Protocols for Reactive Control and Learning

5a. CONTRACT NUMBER
W911NF-15-1-0592

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Texas at Austin
101 East 27th Street
Suite 5.300
Austin, TX 78712 -1532

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)
67811-CS.5

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is 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

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Ufuk Topcu
UU	UU	UU	UU		19b. TELEPHONE NUMBER 949-202-6227

RPPR Final Report
as of 28-May-2021

Agency Code: 21XD

Proposal Number: 67811CS **Agreement Number: W911NF-15-1-0592**
INVESTIGATOR(S):

Name: Ufuk Topcu
Email: utopcu@austin.utexas.edu
Phone Number: 9492026227
Principal: Y

Organization: **University of Texas at Austin**
Address: 101 East 27th Street, Austin, TX 787121532
Country: USA
DUNS Number: 170230239 EIN: 746000203
Report Date: 30-Nov-2018 Date Received: 22-May-2021
Final Report for Period Beginning 03-Sep-2015 and Ending 31-Aug-2018
Title: Probably Approximately Correct Protocols for Reactive Control and Learning
Begin Performance Period: 03-Sep-2015 **End Performance Period:** 31-Aug-2018
Report Term: 0-Other
Submitted By: Ufuk Topcu Email: utopcu@austin.utexas.edu
Phone: (949) 202-6227
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 1 **STEM Participants:**

Major Goals: The objective of this project is to develop decision-making algorithms for autonomous and intelligent systems that jointly learn and react in environments with stochastic as well as adversarial uncertainties. The algorithms will be not only efficient in learning in terms of their use of samples, time, and space (i.e., in the traditional probably approximate correctness “PAC” sense) but also provably correct (by synthesis) with respect to rich temporal logic mission specifications.

The effort investigates a collection of questions including the following: What is an appropriate adaptation of the probably approximate correctness notions in the presence of temporal logic constraints on the evolution of the underlying systems? How can such adaptation be extended to environments with both stochastic uncertainties and adversarial opponents as well as applications in which safety is critical while approximately optimal performance is acceptable? How can we cope with the possibly high computational cost in protocol synthesis for joint control and learning? How can we interpret the traditional exploration vs. exploitation trade-offs under temporal logic specifications? How can we develop protocols that not only execute correctly in their nominal domains but also whose performance and correctness are invariant to or degrade gracefully under domain variations? How can we incorporate performance criteria that account for the effects of lack of knowledge about environment and/or opponents in decision-making?

Accomplishments: We now summarize the results of the main studies enabled by this project.

Safety-Constrained Reinforcement Learning for MDPs

We considered controller synthesis for stochastic and partially unknown environments in which safety is essential. Specifically, we abstracted the problem as a Markov decision process in which the expected performance is measured using a cost function that is unknown prior to run-time exploration of the state space. Standard learning approaches synthesize cost-optimal strategies without guaranteeing safety properties. To remedy this, we first computed safe, permissive strategies. Then, exploration is constrained to these strategies and thereby meets the imposed safety requirements. Exploiting an iterative learning procedure, the resulting policy is safety-constrained and optimal. We showed correctness and completeness of the method and investigated the use of several heuristics to increase its scalability.

Correct-by-synthesis reinforcement learning with temporal logic constraints

We considered a problem on the synthesis of reactive controllers that optimize some a priori unknown performance

criterion while interacting with an uncontrolled environment such that the system satisfies a given temporal logic specification. We decoupled the problem into two subproblems. First, we extracted a (maximally) permissive strategy for the system, which encodes multiple (possibly all) ways in which the system can react to the adversarial environment and satisfy the specifications. Then, we quantified the a priori unknown performance criterion as a (still unknown) reward function and compute an optimal strategy for the system within the operating envelope allowed by the permissive strategy by using the so-called maximin-Q learning algorithm. We established both correctness (with respect to the temporal logic specifications) and optimality (with respect to the a priori unknown performance criterion) of this two-step technique for a fragment of temporal logic specifications. For specifications beyond this fragment, correctness can still be preserved, but the learned strategy may be sub-optimal. We presented an algorithm to the overall problem, and demonstrated its use and computational requirements on a set of robot motion planning examples.

Probably Approximately Correct Learning in Stochastic Games with Temporal Logic Specifications

We considered a controller synthesis problem in turn- based stochastic games with both a qualitative linear temporal logic (LTL) constraint and a quantitative discounted-sum objective. For each case in which the LTL specification is realizable and can be equivalently transformed into a deterministic Buchi automaton, we showed that there always exists a memoryless almost-sure winning strategy that is "-optimal with respect to the discounted-sum objective for any arbitrary positive ". Building on the idea of the R-MAX algorithm, we propose a probably approximately correct (PAC) learning algorithm that can learn such a strategy efficiently in an online manner with a-priori unknown reward functions and unknown transition distributions. To the best of our knowledge, this is the first result on PAC learning in stochastic games with independent quantitative and qualitative objectives.

Environment-Independent Task Specifications via GLTL

We proposed a new task-specification language for Markov decision processes that is designed to be an improvement over reward functions by be- ing environment independent. The language is a variant of Linear Temporal Logic (LTL) that is extended to probabilistic specifications in a way that permits approximations to be learned in finite time. We provided several small environments that demonstrate the advantages of our geometric LTL (GLTL) language and illustrate how it can be used to specify standard reinforcement-learning tasks straightforwardly.

Safe Reinforcement Learning via Shielding

Reinforcement learning algorithms discover policies that maximize reward, but do not necessarily guarantee safety during learning or execution phases. We introduced a new approach to learn optimal policies while enforcing properties expressed in temporal logic. To this end, given the temporal logic specification that is to be obeyed by the learning system, we proposed to synthesize a reactive system called a shield. The shield monitors the actions from the learner and corrects them only if the chosen action causes a violation of the specification. We investigated which requirements a shield must meet to preserve the convergence guarantees of the learner. Finally, we demonstrated the versatility of our approach on several challenging reinforcement learning scenarios.

Training Opportunities: The project has enabled the training of the following researchers.

Min Wen (Ph.D. student)
Nils Jansen (postdoctoral scholar)
Jie Fu (postdoctoral scholar
Mohammed AlShiekh (research scientist)

Results Dissemination: The results have been disseminated through conference (e.g., AAAI, IJCAI, TACAS, IROS) and journal (e.g., IEEE TAC), research seminars and modules in the graduate course taught by the PI.

Honors and Awards: The results of the project contributed in the preliminary results that gave rise to the PI's NSF CAREER award.

Protocol Activity Status:

Technology Transfer: Nothing to Report

RPPR Final Report
as of 28-May-2021

PARTICIPANTS:

Participant Type: PD/PI
Participant: Ufuk Topcu
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Jie Fu
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Nils Jansen
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Min Wen
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: International Joint Conference on Artificial Intelligence
Date Received: 20-Jul-2016 Conference Date: 09-Jul-2016 Date Published: 09-Jul-2016
Conference Location: New York, NY
Paper Title: Probably Approximately Correct Learning in Stochastic Games with Temporal Logic Specifications: Technical Report
Authors: Min Wen, Ufuk Topcu
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 28-May-2021

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Formal Methods Europe
Date Received: 25-Aug-2017 Conference Date: 08-Nov-2016 Date Published: 08-Nov-2016
Conference Location: Cyprus
Paper Title: Synthesis of Shared Control Protocols with Provable Safety and Performance Guarantees
Authors: Nils Jansen, Ufuk Topcu
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Twenty-Sixth International Joint Conference on Artificial Intelligence
Date Received: 25-Aug-2017 Conference Date: 19-Aug-2017 Date Published:
Conference Location: Melbourne, Australia
Paper Title: Reduction Techniques for Model Checking and Learning in MDPs
Authors: Suda Bharadwaj, Stephane Le Roux, Guillermo Perez, Ufuk Topcu
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Twenty-Sixth International Joint Conference on Artificial Intelligence
Date Received: 25-Aug-2017 Conference Date: 19-Aug-2017 Date Published:
Conference Location: Melbourne, Australia
Paper Title: Learning from Demonstrations with High-Level Side Information
Authors: Min Wen, Ivan Papusha, Ufuk Topcu
Acknowledged Federal Support: **Y**

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Ufuk Topcu
Signature Date: 5/22/21 3:10PM

This file includes the text already entered in the text boxes.

Major goals

The objective of this project is to develop decision-making algorithms for autonomous and intelligent systems that jointly learn and react in environments with stochastic as well as adversarial uncertainties. The algorithms will be not only efficient in learning in terms of their use of samples, time, and space (i.e., in the traditional probably approximate correctness "PAC" sense) but also provably correct (by synthesis) with respect to rich temporal logic mission specifications.

The effort investigates a collection of questions including the following: What is an appropriate adaptation of the probably approximate correctness notions in the presence of temporal logic constraints on the evolution of the underlying systems? How can such adaptation be extended to environments with both stochastic uncertainties and adversarial opponents as well as applications in which safety is critical while approximately optimal performance is acceptable? How can we cope with the possibly high computational cost in protocol synthesis for joint control and learning? How can we interpret the traditional exploration vs. exploitation trade-offs under temporal logic specifications? How can we develop protocols that not only execute correctly in their nominal domains but also whose performance and correctness are invariant to or degrade gracefully under domain variations? How can we incorporate performance criteria that account for the effects of lack of knowledge about environment and/or opponents in decision-making?

Accomplishments

We now summarize the results of the main studies enabled by this project.

Safety-Constrained Reinforcement Learning for MDPs

We considered controller synthesis for stochastic and partially unknown environments in which safety is essential. Specifically, we abstracted the problem as a Markov decision process in which the expected performance is measured using a cost function that is unknown prior to run-time exploration of the state space. Standard learning approaches synthesize cost-optimal strategies without guaranteeing safety properties. To remedy this, we first computed safe, permissive strategies. Then, exploration is constrained to these strategies and thereby meets the imposed safety requirements. Exploiting an iterative learning procedure, the resulting policy is safety-constrained and optimal. We showed correctness and completeness of the method and investigated the use of several heuristics to increase its scalability.

Correct-by-synthesis reinforcement learning with temporal logic constraints

We considered a problem on the synthesis of reactive controllers that optimize some a priori unknown performance criterion while interacting with an uncontrolled environment

such that the system satisfies a given temporal logic specification. We decoupled the problem into two subproblems. First, we extracted a (maximally) permissive strategy for the system, which encodes multiple (possibly all) ways in which the system can react to the adversarial environment and satisfy the specifications. Then, we quantified the a priori unknown performance criterion as a (still unknown) reward function and compute an optimal strategy for the system within the operating envelope allowed by the permissive strategy by using the so-called maximin-Q learning algorithm. We established both correctness (with respect to the temporal logic specifications) and optimality (with respect to the a priori unknown performance criterion) of this two-step technique for a fragment of temporal logic specifications. For specifications beyond this fragment, correctness can still be preserved, but the learned strategy may be sub-optimal. We presented an algorithm to the overall problem, and demonstrated its use and computational requirements on a set of robot motion planning examples.

Probably Approximately Correct Learning in Stochastic Games with Temporal Logic Specifications

We considered a controller synthesis problem in turn- based stochastic games with both a qualitative linear temporal logic (LTL) constraint and a quantitative discounted-sum objective. For each case in which the LTL specification is realizable and can be equivalently transformed into a deterministic Buchi automaton, we showed that there always exists a memoryless almost-sure winning strategy that is "-optimal with respect to the discounted-sum objective for any arbitrary positive ". Building on the idea of the R-MAX algorithm, we propose a probably approximately correct (PAC) learning algorithm that can learn such a strategy efficiently in an online manner with a-priori unknown reward functions and unknown transition distributions. To the best of our knowledge, this is the first result on PAC learning in stochastic games with independent quantitative and qualitative objectives.

Environment-Independent Task Specifications via GLTL

We proposed a new task-specification language for Markov decision processes that is designed to be an improvement over reward functions by be- ing environment independent. The language is a variant of Linear Temporal Logic (LTL) that is extended to probabilistic specifications in a way that permits approximations to be learned in finite time. We provided several small environments that demonstrate the advantages of our geometric LTL (GLTL) language and illustrate how it can be used to specify standard reinforcement-learning tasks straightforwardly.

Safe Reinforcement Learning via Shielding

Reinforcement learning algorithms discover policies that maximize reward, but do not necessarily guarantee safety during learning or execution phases. We introduced a new approach to learn optimal policies while enforcing properties expressed in temporal logic. To this end, given the temporal logic specification that is to be obeyed by the learning system, we proposed to synthesize a reactive system called a shield. The

shield monitors the actions from the learner and corrects them only if the chosen action causes a violation of the specification. We investigated which requirements a shield must meet to preserve the convergence guarantees of the learner. Finally, we demonstrated the versatility of our approach on several challenging reinforcement learning scenarios.

Training Opportunities

The project has enabled the training of the following researchers.

* Min Wen (Ph.D. student)
* Nils Jansen (postdoctoral scholar)
* Jie Fu (postdoctoral scholar
* Mohammed AlShiekh (research scientist)

Results Dissemination

The results have been disseminated through conference (e.g., AAAI, IJCAI, TACAS, IROS) and journal (e.g., IEEE TAC), research seminars and modules in the graduate course taught by the PI.

Honors

The results of the project contributed in the preliminary results that gave rise to the PI's NSF CAREER award.