Online POMDP Algorithms for Very Large Observation Spaces

Wee Sun Lee
NATIONAL UNIVERSITY OF SINGAPORE

06/06/2017
Final Report

DISTRIBUTION A: Distribution approved for public release.
Online POMDP Algorithms for Very Large Observation Spaces

Wee Sun Lee

NATIONAL UNIVERSITY OF SINGAPORE
21 LOWER KENT RIDGE ROAD
SINGAPORE, 119077 SG

AOARD
UNIT 45002
APO AP 96338-5002

A DISTRIBUTION UNLIMITED: PB Public Release

PARTIALLY OBSERVABLE MARKOV DECISION PROCESS (POMDP) PROVIDES A MATHEMATICALLY ELEGANT MODELING TOOL FOR PLANNING AND CONTROL UNDER UNCERTAINTY. SUBSTANTIAL PROGRESS HAS BEEN ACHIEVED IN THE PAST DECADE, ALLOWING SOME LARGE-SCALE PROBLEMS TO BE SOLVED USING POMDPs. HOWEVER, VERY LARGE OBSERVATION SPACES STILL POSE SUBSTANTIAL DIFFICULTIES FOR EFFECTIVE PLANNING. IN THIS PROJECT, TWO ASPECTS OF THESE DIFFICULTIES ARE STUDIED. ONE CHALLENGE POSED BY VERY LARGE OBSERVATION SPACES IS THAT MONTE-CARLO METHODS USED FOR SCALING UP THE SOLVERS TO SOLVE VERY LARGE PROBLEMS MAY FAIL TO SAMPLE RARE BUT CRITICAL EVENTS THAT ARE IMPORTANT FOR PLANNING. THE PI'S TEAM DEVELOPED METHODS FOR HANDLING THESE DIFFICULTIES BY USING IMPORTANCE SAMPLING TO FOCUS ON SAMPLING THESE EVENTS. THEY SHOW THAT OUR ONLINE PLANNING METHOD RETAINS GOOD THEORETICAL PROPERTIES WHEN IMPORTANCE SAMPLING IS USED AND PROPOSE A METHOD FOR LEARNING THE IMPORTANCE SAMPLING DISTRIBUTION. EXPERIMENTALLY, THE METHOD WORKS WELL IN SIMULATION AND ON REALISTIC DATA. ANOTHER ISSUE WITH VERY LARGE OBSERVATION SPACES IS THE HIGH COMPUTATIONAL COMPLEXITY OF HANDLING THE VERY LARGE SPACE. THE TEAM STUDIED THE APPROACH OF USING MAXIMUM LIKELIHOOD DETERMINATION, WHERE ONLY THE MOST LIKELY OBSERVATIONS ARE USED DURING THE SEARCH FOR SOLUTION. THEY SHOWED THAT SOLUTIONS TO SOME SUBCLASSES OF POMDP PROBLEMS CAN BE WELL APPROXIMATED IN POLYNOMIAL TIME USING THIS APPROACH.

DATA MINING

Name of Principal Investigators (PI and Co-PIs):
- e-mail address: leews@comp.nus.edu.sg
- Institution: National University of Singapore
- Mailing Address: Department of Computer Science, Computing 1, 13 Computing Drive, Singapore 117417, Republic of Singapore
- Phone: +65 65164526
- Fax: +65 67794580

Period of Performance: 05/13/2015 – 05/13/2017

Abstract: Partially Observable Markov Decision Process (POMDP) provides a mathematically elegant modeling tool for planning and control under uncertainty. Substantial progress has been achieved in the past decade, allowing some large-scale problems to be solved using POMDPs. However, very large observation spaces still pose substantial difficulties for effective planning. In this project, we study two aspects of these difficulties. One challenge posed by very large observation spaces is that Monte-Carlo methods used for scaling up the solvers to solve very large problems may fail to sample rare but critical events that are important for planning. We develop methods for handling these difficulties by using importance sampling to focus on sampling these events. We show that our online planning method retains good theoretical properties when importance sampling is used and propose a method for learning the importance sampling distribution. Experimentally, the method works well in simulation and on realistic data. Another issue with very large observation spaces is the high computational complexity of handling the very large space. We study the approach of using maximum likelihood determinization, where only the most likely observations are used during the search for solution. We showed that solutions to some subclasses of POMDP problems can be well approximated in polynomial time using this approach.

Introduction: Partially observable Markov Decision Processes (POMDP) is a mathematically elegant modeling tool that has been shown to be useful in various problems of planning and control under uncertainty, including dialog systems, assistive technologies, and autonomous vehicle navigation. Substantial progress has been made in the last decade, addressing problems with large state spaces. In particular, as part of an earlier AOARD grant (FA2386-12-1-4031), we have developed an effective online anytime POMDP solver, DESPOT, that is able to scale to very large state spaces [8]. DESPOT uses a set of sampled scenarios in order to construct a relatively small search tree, allowing the search for a good action to be done more efficiently. We have continued work on the DESPOT algorithm within this grant, extending the search algorithm, publishing the algorithm in the Journal of Artificial Intelligence Research [8], and releasing C++ open-source software (https://github.com/AdaCompNUS/despot). DESPOT has also been implemented as open source software in the Julia language by the Stanford Intelligent Systems Laboratory (https://github.com/JuliaPOMDP/DESPOT.jl) [2].

DESPOT is a Monte Carlo algorithm for doing online search. One issue with sampling algorithms is that they may sometimes fail to sample rare but critical events. This problem is particularly bad when the observation space is very large as only very few scenarios will agree with what is observed in most parts of the search tree. For example, we implemented DESPOT for driving autonomously through a crowded environment [1] (Figure 1), and in simulations with measured pedestrian trajectories, we find that DESPOT has a 0.0013
collision rate (after removing cases where the pedestrian walks into the vehicle instead of
the vehicle moving into the pedestrian). While the collision rate in the simulations is
artificially high due to the fact that the pedestrians in the measured trajectories are not
aware of the existence of the vehicle, it does give an indication that more work should be
done for the case of distracted pedestrians (e.g. those distracted while looking at their
mobile phones while walking).

Figure 1: Autonomous vehicle driving through a crowd. The right hand side shows the
starting point and destination of the vehicle.

To handle the problem, we use importance sampling to focus the sampling on the rare but
critical events (in the case of the autonomous vehicle, these would be trajectories that leads
to collisions). We prove that the theoretical guarantees for the DESPOT algorithm can be
modified to give similar guarantees when importance sampling is used. We further give an
algorithm for learning the importance sampling distribution using the model.

With importance sampling, we are able to remove all the collisions in the measured
pedestrian trajectories used in our study. The paper reporting on the importance sampling
algorithm [6] was published in the Workshop on the Algorithmic Foundations of Robotics
where it was a best paper nominee. The paper has also been invited for submission to the
International Journal on Robotics Research special issue representing some of the best
papers appearing at the conference. We are also using the new algorithm to participate in
the IEEE RAS–SIGHT Humanitarian Robotics & Automation Technology Challenge 2017 on
landmine detection and we are one of three teams that have qualified for the final challenge
to be held at the end of May 2017.

Another issue with very large observation spaces is the computational complexity of dealing
with the large number of observations. We investigated the approach of determinizing the
observation to use only the most likely observation. We call this approach maximum
likelihood determinization. With this approach, the observation branching is eliminated and
we have a deterministic search problem. Furthermore, for some subclasses of POMDPs, the
determinized problem can be approximated in polynomial time. As part of an earlier AOARD
grant (FA2386-12-1-4031), we have developed a polynomial time approximation algorithm
for adaptive informative path planning using this approach [4]. In adaptive informative path
planning, the aim is to minimize the travel cost for adaptively finding a path to gather
information in order to identify a target hypothesis. In this project, we have extended this
result to more general objective functions that are point-wise submodular, allowing the
method to be more widely used. This result was published in NIPS 2015 [3]. We have also
investigated the application of maximum likelihood determinization to a Bayesian version of
the Canadian Traveller Problem. In the Canadian Traveller Problem, a traveller is travelling
from a start location to a target location on a road network. Some of the roads may be
blocked due to snow without the traveller’s knowledge, and the aim of the traveller is to
minimize the expected travelling time to reach the target. In the Bayesian version of the

DISTRIBUTION A. Approved for public release: distribution unlimited.
problem, the road blockages are correlated, and observation on one road provides information on the states of the other roads. Like the adaptive informative path planning problem, the agent has to do information gathering, but in the Bayesian Canadian Traveller Problem, the agent is doing information gathering in order to identify road conditions so that it reach the goal quickly. This gives rise to the exploitation vs exploration problem, where the agent has to balance exploration to gather information about the road network with exploitation, which uses the knowledge already gained to plan the shortest path to the goal. Using determinization, we are able to give a polynomial time algorithm with guaranteed approximation for this problem. The result has been submitted to UAI 2017 [5].

In the following section, we describe selected experiments done as part of the project and the results obtained. For details of the algorithms, including theoretical properties and proofs, and other experiments, we refer the reader to the publications [3,5,6,7].

Experiments and Results:

*Importance Sampling with DESPOT [6]:* In the paper, we develop two versions of importance sampling for use with the DESPOT algorithm: unnormalized importance sampling (UIS-DESPOT) and normalized importance sampling (NIS-DESPOT). Both versions of the algorithm outperform DESPOT when rare critical events are present in the problem. We describe two simulation experiments. In the collision avoidance problem, an aircraft starts in a random position on the one area of the map while another agent randomly moves in another region. The aircraft has to successfully avoid collision when moving past the region of the other agent. The aircraft knows its own position but the observation of the other agent's position is corrupted by noise. At each time step, the aircraft can change direction with a cost of 1, and a penalty of 1000 is applied if there is a collision. As shown in Figure 2, both UIS-DESPOT and NIS-DESPOT again outperform DESPOT in both the total discounted reward as well as in the collision rate.

![Figure 2: Collision avoidance problem. The plot in the middle shows the average total discounted reward for the algorithms as a function of the sample size used, while the plot on the right shows the collision rate as a function of the sample size used.](image)

In the demining experiment, a robot has to detect and report landmines in the field. Each grid point has a probability 0.05 of having a mine. At each time step, the robot can move and observe its adjacent 4 cells with probability 0.9 of accurate observation. If the robot reports a mine correctly, a reward of 10 is provided, otherwise a penalty of 10 is applied. Stepping over a mine causes a high penalty of 1000. As seen in Figure 3, both UIS-DESPOT and NIS-DESPOT again outperform DESPOT.

![Figure 3: Demining experiment results.](image)
Figure 3: Demining problem. The plot in the middle shows the average total discounted reward as a function of the sample size, while the plot on the right shows the explosion rate as a function of the sample size.

Adaptive Informative Path Planning with Submodular Functions [3]: In this paper, we generalized the adaptive informative path planning problem to handle more general objective functions. For the experiments, we use the Gibbs error criterion which is appropriate for distinguishing equivalent classes. We run simulations of a UAV search and rescue problem, where the UAV has to find a survivor in a search and rescue scenario. In this scenario, there is danger zone where the UAV has to identify the exact location of the survivor and a safe zone where the UAV does not have to worry about the exact location of the survivor, giving rise to an equivalent class of positions. As shown in Figure 4, our new method RAC-GE as well as a variant RAC-V outperform competing methods.

Figure 4: Cost as a function of target Gibbs error. For most target Gibbs error, RAC-GE and RAC-V outperform other methods.

Bayesian Canadian Traveller Problem [5]: In the Bayesian Canadian Traveller problem, the agent needs to get to the destination quickly through a road network where some of the roads may be blocked. The road blockages may be correlated and the agent knows the distribution of road blockages. The agent needs to do some exploration to gather information about possible road blockages but also needs to balance the exploration and exploit the gained information to get to the destination quickly. The problem is NP-hard and we give a polynomial time algorithm with guaranteed approximation. We experimented with two problems: a road network with some roads that may be blocked, as well as the reduction from optimal decision tree (ODT) problem that is used to show that the problem is NP-hard. The algorithm, HSPD, outperforms a baseline optimistic algorithm as well as a state of the art algorithm based on the upper confidence tree (UCT) algorithm. The performance gain is substantial when exploration forms an important part for solving the problem, e.g. in
the reduction from ODT. The results are shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>HSPD</th>
<th>Optimistic</th>
<th>UCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODT Reduction</td>
<td>31.5</td>
<td>502</td>
<td>566</td>
</tr>
<tr>
<td>Road Network</td>
<td>38.9</td>
<td>59.1</td>
<td>44.6</td>
</tr>
</tbody>
</table>

Table 1: Average cost for HSPD, Optimistic and UCT on the ODT reduction and road network problems.

Discussion: Large observation spaces create various difficulties for POMDP planning. One issue is the difficulty of sampling rare but critical events when Monte-Carlo methods are used. We have developed an effective importance sampling technique for sampling such events with the online POMDP algorithm DESPOT. Another difficulty is the high computational complexity associated with very large observation spaces. For various subclasses of POMDPs, we are able to develop polynomial time algorithms with approximation guarantees with the use of maximum likelihood determinization. This includes adaptive informative path planning and the Bayesian Canadian Traveller problem.

These techniques have helped make large scale POMDP planning more practical. However, many issues still need to be better handled. One issue is to do effective planning when the action space is very large, such as in the case of multi-agent problems. Another issue is to develop effective learning methods for learning models that are suitable for large scale POMDP planning.

List of Publications and Significant Collaborations that resulted from your AOARD supported project: In standard format showing authors, title, journal, issue, pages, and date, for each category list the following:

a) papers published in peer-reviewed journals,

b) papers published in peer-reviewed conference proceedings,

c) papers published in non-peer-reviewed journals and conference proceedings,
   - Nil

d) conference presentations without papers,
   - Nil

e) manuscripts submitted but not yet published, and
   - Luo, Yuanfu, Haoyu Bai, David Hsu and Wee Sun Lee. "Importance Sampling for


f) provide a list any interactions with industry or with Air Force Research Laboratory scientists or significant collaborations that resulted from this work.

- Nil

References:


