

## 1 Introduction

Many scenarios in real life occur where one must make a sequence of decisions under uncertainty. Quite often in these scenarios two unknowns exist: the result of taking an action at a given state and the true state of the agent at any point in time. This is similar to that of a hidden Markov model (HMM), only in that one must make a sequence of actions instead of a single action; thus the problem falls no longer in supervised learning but reinforcement learning. The classic example is a robot which tries to navigate a discrete environment using a GPS, and it performs action that lead to different states with various probabilities. The transition probabilities are unknown, and because of the GPS, there is also inaccuracy in what the underlying state is.

One can formalize the problem, under Markovian assumptions, as a partially observable Markov decision process (POMDP). In this project we provide a software library for implementing and solving them, which is modular and flexible enough for further development and user-specified agents/environments. We encode a variety of basic tasks and solve them using a combination of value iteration and several standard (PO)MDP solvers—one is a variant of Thompson sampling [2], which is a Bayesian approach following a Dirichlet-multinomial posterior over each state-action pair.

The repository can be found at <https://github.com/dustinvtran/bayesrl>. To install from pip, run

```
1 pip install -e "git+https://github.com/dustinvtran/bayesrl.git#egg=bayesrl"
```

## 2 Technical Background

### 2.1 POMDP

POMDP is a generalization of a Markov decision process (MDP). In a MDP, for each possible state of the process, a decision has to be made regarding which action should be executed in that state. The chosen action and given state affects the costs (or rewards) incurred. The goal is to learn the optimal *policy*, which is a choice of actions that in expectation leads to the optima reward in a pre-defined number of steps (in the case of finite horizon; or in infinite time). In a POMDP model, the agent does not fully observe the underlying states but a (non-sufficient) statistic of it. Thus POMDPs also maintain a probability distribution over the set of possible states, based on a set of observations and observation probabilities, and the underlying MDP.

More formally, a POMDP is a collection  $(S, A, T, R, \Omega, O, \gamma)$ , where

- $S$  is a set of states
- $A$  is a set of actions
- $T$  is a set of transition probabilities between states. If the agent is currently in state  $s \in S$ , and it takes action  $a \in A$ , the agent will transition to a new state  $s'$  with probability  $T(s' | s, a)$ .
- $R : S \times A \rightarrow \mathbb{R}$  is a reward function that assigns a numeric reward (or cost if the value is negative) for each state and action.
- $\Omega$  is a set of observations

- $O$  is a set of conditional observation probabilities. If the agent is now in state  $s$ , it receives an observation  $o$  according to  $O(o | s)$
- $\gamma \in [0, 1]$  is a discount factor that determines how much rewards should be discounted over time

We use our testing environment `GridWorld` as an example. `GridWorld` represents a 2D maze where the agent can be in discrete locations. Certain locations are impossible for the agent to move to, representing “walls”. Every action can move the agent between two adjacent grid locations, or fail, and cause the agent to take a uniformly random action instead according to some user-specified probability. The goal is to reach the goal location in least amount of time, as each move is -1 reward, hitting a wall is -1 reward, and reaching the end positions incurs a reward of +50.

For `GridWorld`,  $S$  consists of all the possible (row, column) location tuples inside the maze.  $A$  contains the four possible actions the agent can take: up, down, left, right.  $T$  describes a transition model that allows the agent to move without hitting the wall. We define  $R$  as the above. To make this a partially observable problem, we implemented two observation models.

- The easier model gives the agent more information about the environment. The agent knows which of its four neighbors are walls, giving rise to 16 total observations.
- In the second observation model, the agent can only observe how many of its four neighbors are walls, giving rise to 5 possible observations.  $O$  is such that  $Pr(true\_observation | state) = true\_observation\_prob$ , where  $true\_observation\_prob$  can be adjusted, and

$$Pr(other\_observation | state) = \frac{1 - true\_observation\_prob}{total\_num\_of\_observations - 1} \quad (1)$$

Since we cannot work with the underlying states directly in POMDP, we also need  $B$  which is the set of belief states, or the probabilities the agent is at all possible states.

## 2.2 Thompson Sampling

Thompson Sampling is used to learn the transition distribution  $T$ , where one first specifies a prior according to one’s knowledge about the transitions before trying any actions. After a constant set of time steps, the transition probabilities are then recalculated using a posterior update following a Dirichlet-multinomial distribution. In MDP, since the agent observes the states directly, the posterior transition probabilities are updated directly using the transition counts. In POMDP, the posterior probabilities are iteratively updated following an update on belief state transition probabilities.

Note that we’ve also implemented a variety of other standard MDP solvers for benchmarking: Q-Learning, SARSA, and R-MAX.

## 2.3 Value Iteration

Given estimated transition probabilities, we then solve for the underlying policy with value iteration. Value  $V$ , is the expected total reward given a policy  $\pi$ , where a policy decides which action to take given the belief state.  $a = \pi(b)$ . The expected reward for policy  $\pi$  starting from belief  $b_0$  is defined as

$$V^\pi(b_0) = \sum_{t=0}^{\infty} \gamma^t r(b_t, a_t)$$

where  $r(b_t, a_t) = \sum_{s \in S} b_t(s) R(s, a_t)$ . The optimal policy should maximize the long term reward

$$\pi = \operatorname{argmax}_{\pi} V^\pi(b_0)$$

At each time step, we update the belief states based on the observation, and then update the values based on the updated belief states. The action that gives the largest expected reward over the belief states is selected for the next time step. The values gradually improve until convergence. By improving the values, the policy is implicitly improved.

### 3 Implementation

#### 3.1 Agent Environment Paradigm

For MDP, we follow the paradigm set forth in Sutton and Barto (Figure 3.1, [3]). It suggests that RL agents

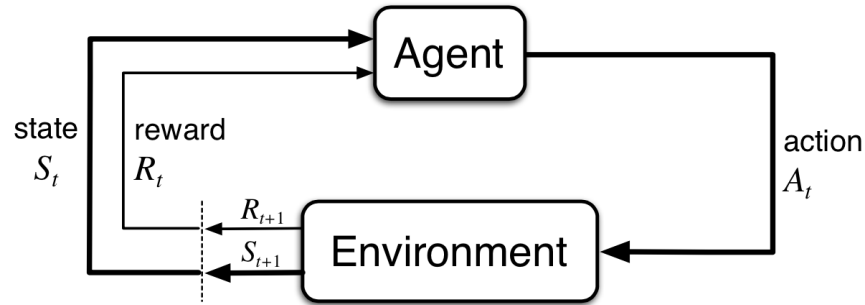


Figure 3.1: The agent–environment interaction in reinforcement learning.

need only output a suggested action after previous history and a given state and reward. Overall by following the paradigm for the software design, we make the learning process explicit and intuitive.

##### 3.1.1 Agent

We implement a base class `Agent` which is a collection of objects and functions to be used for all other agents. Agents differ primarily in their `interact()` function, which determines the next action to perform given a state and reward from the environment.

The model-based algorithms R-MAX and Thompson sampling inherit from `ModelBasedAgent`, which is a class that itself inherits from `Agent`; `ModelBasedAgent` adds subroutines specific to model-based approaches such as value iteration.

In order to reduce the most redundant code, we also could have used an additional class that inherits from `Agent` for temporal difference method agents; this would be used by both `SARSAgent` and `QLearningAgent`, as they differ only in their `value_table` assignment. However, the efficiency gain in such an abstraction is not worth the loss of readability in our opinion.

##### 3.1.2 Environment

An `Environment` object is initialized at some state, with an arbitrarily defined state and action space. Actions are performed on an `Environment` object under the subroutine `perform_action()`, and the output is a new state and its reward.

### 3.1.3 Trial and Plot

As for trials, we implement a class `Trial` which contains all information for running multiple trials, i.e., independent collections of episodes to learn and act upon. We also add a `Plot` class which is a wrapper containing all `Trial` objects; this is convenient for generating plots on collections of trials coming from possibly many different agents.

## 3.2 POMDP

For POMDP, we assume that the agent is given an observation model of the environment it is acting in, in the form of a conditional probability distribution  $P(\text{observation} \mid \text{state})$ . Aström has shown that a properly updated probability distribution over the state space  $S$  is sufficient to summarize all the observable history of a POMDP agent without loss of optimality [1]. Thus we add a step of updating the belief state to the MDP paradigm. And the belief state is used to update the transition model instead of the underlying state.

## 3.3 Organization of Code

We follow the directory structure specified in the problem set, with two exceptions:

- `documentation/` does not exist. Instead, documentation is written in the `README.md` inside the current working directory. Any additional documentation not purely necessary for the problem set submission is in the Github wiki (which is a subset of anything in this writeup).
- `source/` is named `bayesrl/` in order to follow Python convention for installing modules.

## 4 Analysis

### 4.1 Runtime

All solvers we implement are guaranteed to converge in polynomial time for MDPs, although unfortunately we have not seen stricter theoretical results on the upper bounds than this. This certainly makes sense as it is true for all environments regardless of the pathological scenario. However, it would certainly be interesting to examine bounds under stricter assumptions where we fix the environment and perhaps certain parameter settings to simplify the analysis.

### 4.2 Memory

We store several arrays for the computation. The transition probabilities table is of dimension  $|S|^2|A|$ . The value table is of dimension  $|S||A|$ . The transition observation table is of dimension  $|S|^2|A|$ . Thus the space requirement for the algorithm is  $O(|S|^2|A|)$ .

### 4.3 Limitations

If the number of states is large, the algorithm quickly becomes intractable. We are quite interested in examining function approximations, which allow one to essentially apply a supervised learning algorithm to predict the optimal action given state characteristics, rather than to hardcode them manually. Note however that this makes the runtime even worse as there is an additional error accumulating as a result of the predictions.

## 5 Experiments

The implementation of Thompson Sampling for MDPs ran very successfully: as time progressed, the agent was able to get to the goal much faster, and get a high reward. It did not work, however, for POMDPs. As time progressed, the agent seemed to take about the same amount of time to reach the goal on every execution, not improving in performance. We hypothesize that the reason is due to a lack of a good prior on the transition model: Thompson sampling relies a lot on being able to count the number of transitions between states, given an action.

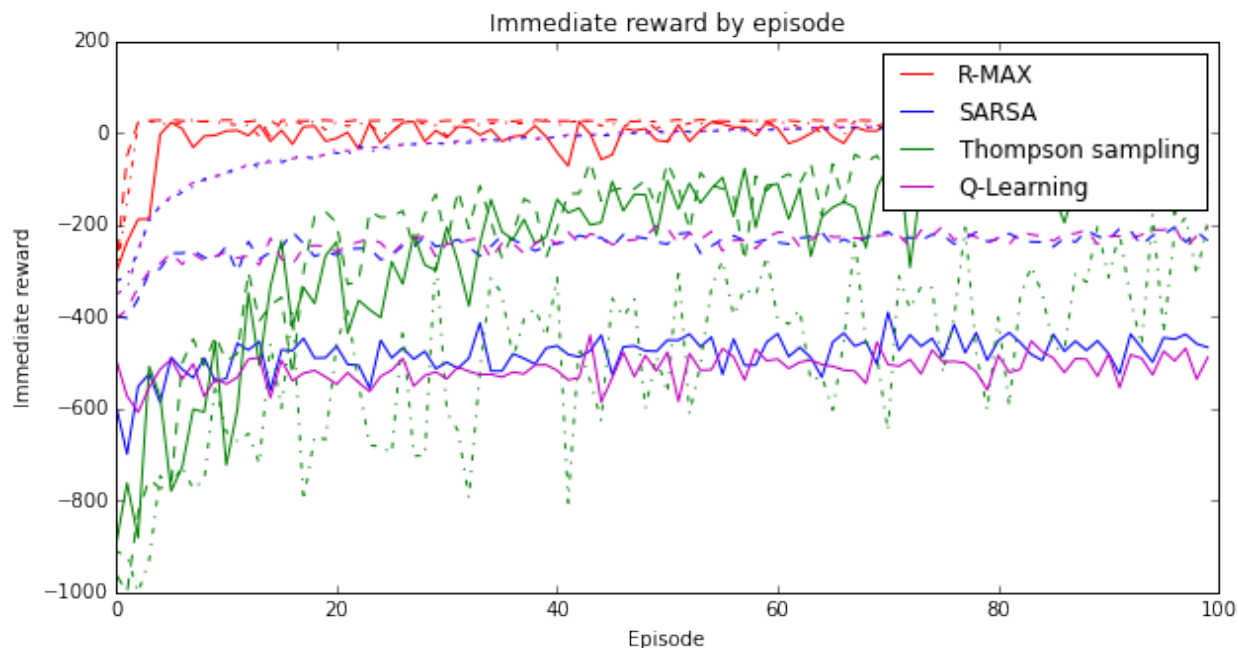


Figure 1: Immediate reward for MDP solvers in a variety of parameter settings on the Gridworld example. Each of the settings for Thompson sampling vary how strong the misspecification of the prior as a uniform distribution, and yet in all three scenarios it indicates convergence. We also see that R-MAX performs better than the temporal difference methods.

This is very hard in a POMDP with a weak prior on its transition model, as it may have a completely wrong idea of where it ends up at each step. The observation model is supposed to improve its accuracy, but it was not sufficient in this case. In fact, we can see this empirically by essentially hardcoding the optimal path by putting high probability on where it should go, and indeed it converges to the optimal quite quickly.

To have an idea of how differently successful the same approach was on MDPs vs POMDPs, we show you our results:

These graphs show the cumulative reward that the agent got as time progressed. In the MDP case, the reward started low, and kept getting lower, as the agent explored. It, however, started getting higher as the agent started acting in a more deliberate manner, to maximize reward. The POMDP agent, however, seems to be in a state where its reward just gets increasingly negative. It is unclear whether it never leaves the exploration phase, or if it simply learns a completely wrong transition model, and thus computes a very wrong policy.

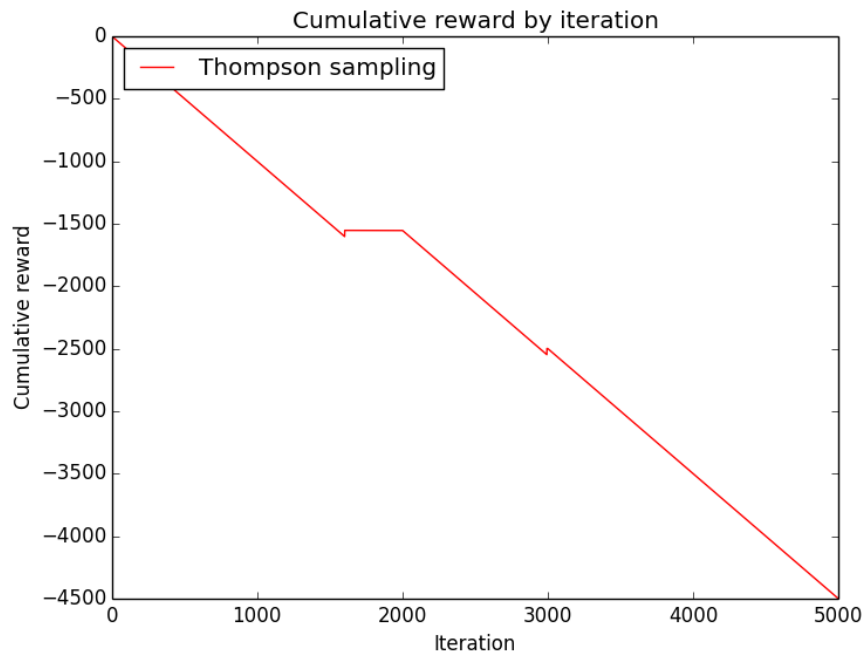


Figure 2: Cumulative Reward for POMDP

## References

- [1] Karl Johan Aström. Optimal control of markov decision processes with incomplete state estimation. *Journal of Mathematical Analysis and Applications*, 10:174205, 1965.
- [2] Malcolm Strens. A bayesian framework for reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, 2000.
- [3] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*, volume 116. Cambridge Univ Press, 1998.