EE691: RnD Project
POMDP With State Sensing Cost

Vansh Kapoor, Ronil Mandavia
Guide: Prof. Jayakrishnan Nair

Department of Electrical Engineering
IIT Bombay

Outline

1. Introduction
2. Bellman Operator
3. What Next?
4. Conjectures
5. Sufficient Condition for Always Sensing
6. Timeline
7. Future Work

Problem Statement

- A Partially Observable MDP (POMDP) is a generalization of the MDP setting where an MDP determines the system dynamics, but the agent cannot directly observe the underlying state.
- We are working on a POMDP where the agent is aware of its initial state, but is unaware of their exact state after taking any action, except when it pays a fixed cost K, and the state is revealed to the agent.
- An action where the agent does not query its state is called a blind action, and the agent is said to be in a blind state.
- We apply an additional constraint that the agent can take at most r blind actions, i.e., after r consecutive blind actions, the agent is forced to pay price K and query its state.

Notation

- $\mathbf{S} = \{\mathbf{s_1}, \mathbf{s_2}, \dots, \mathbf{s_n}\}$- Set of states of the MDP and $\mathbf{A} = \{\mathbf{a_1}, \mathbf{a_2}, \dots, \mathbf{a_m}\}$- Set of actions available to the agent
- $\mathbf{p_{ij}(a)}$ - Transition probability from state s_i to s_j on taking action a
- $\mathbf{C(i, a)}$ - Cost incurred on choosing action a in state s_i.
- γ - discount factor
- **K** - the cost of querying present state
- **r** - maximum number of consecutive blind actions that the agent can take.

Notation

- $\pi(\mathbf{i},\mathbf{t})$ - policy specifying the action to be taken when the last state that the agent sensed is s_i, and it has already taken taking t-1 consecutive blind actions. $1 \leq t \leq T_\pi(i)+1$
- $\mathbf{T}_\pi(\mathbf{i})$ - Number of blind actions taken from state i under policy π. $0 \leq T_\pi(i) \leq r$
- $\mathbf{B}_\pi(\mathbf{i},\mathbf{t},\mathbf{j})$ - Probability of the agent being in state j after beginning in state i, taking t blind actions according to policy π.
- $\mathbf{V}_\pi(\mathbf{i})$ - Value function of state i after following policy π.

Bellman Equation

The Bellman equation defined below can be used to evaluate the value function for any policy

$$V_\pi(i) = C(i, \pi(i,0)) + \sum_{t=1}^{T_\pi(i)} \gamma^t B_\pi(i,t,\cdot) C(\cdot, \pi(i,t)) + \gamma^{T_\pi(i)+1} (K + \sum_{j=1}^{n} B_\pi(i, T_\pi(i)+1, j) V_\pi(j))$$

Here $B_\pi(i,t,\cdot)$ is a row vector and $C(\cdot, \pi(i,t))$ is a column vector

Contraction Mapping

Definition

A mapping T : B(I) → B(I) is said to be a contraction mapping if for some $\beta < 1$

$$||Tu - Tv|| \leq \beta ||u - v||$$

where

$$||u|| = \max_i u_i$$

(taking L_∞ norm)

Bellman Operator is a Contraction Mapping

Proof.

Let the Bellman operator defined above be T

$$
\begin{aligned}
||Tu - Tv||_\infty &= ||\gamma^{T_\pi(i)+1}\Big(\sum_{j=1}^{n} B_\pi(i, T_\pi(i)+1, j)(u_j - v_j)\Big)||_\infty \\
&= \max_i \gamma^{T_\pi(i)+1}\Big(\sum_{j=1}^{n} B_\pi(i, T_\pi(i)+1, j)|u_j - v_j|\Big) \\
&\leq \max_i \gamma^{T_\pi(i)+1}\Big(\sum_{j=1}^{n} B_\pi(i, T_\pi(i)+1, j)\max_j |u_j - v_j|\Big) \\
&\leq \max_i \gamma^{T_\pi(i)+1}\max_j |u_j - v_j|\sum_{j=1}^{n} B_\pi(i, T_\pi(i)+1, j)
\end{aligned}
$$

$$
||Tu - Tv||_\infty \leq \max_i \gamma^{T_\pi(i)+1}||u - v||_\infty
$$

□

Bellman Optimality Operator

Proof.

$\mathrm{T}^* u(i) - T^* v(i) =$

$$\min_\pi \{ C(i,\pi(i,0)) + \sum_{t=1}^{T_\pi(i)} \gamma^t B_\pi(i,t,\cdot) C(\cdot,\pi(i,t)) + \gamma^{T_\pi(i)+1} (K + \sum_{j=1}^{n} B_\pi(i, T_\pi(i)+1, j) u(j)) \} -$$

$$\min_\pi \{ C(i,\pi(i,0)) + \sum_{t=1}^{T_\pi(i)} \gamma^t B_\pi(i,t,\cdot) C(\cdot,\pi(i,t)) + \gamma^{T_\pi(i)+1} (K + \sum_{j=1}^{n} B_\pi(i, T_\pi(i)+1, j) v(j)) \}$$

Let $\pi^* =$

$$argmin_\pi \{ C(i,\pi(i,0)) + \sum_{t=1}^{T_\pi(i)} \gamma^t B_\pi(i,t,\cdot) C(\cdot,\pi(i,t)) + \gamma^{T_\pi(i)+1} (K + \sum_{j=1}^{n} B_\pi(i, T_\pi(i)+1, j) v(j)) \}$$

□

Bellman Optimality Operator

Proof.

$\Rightarrow T^*u(i) - T^*v(i)$

$\leq C(i,\pi^*(i,0)) + \sum_{t=1}^{T_{\pi^*}(i)} \gamma^t B_{\pi^*}(i,t,\cdot)C(\cdot,\pi^*(i,t)) + \gamma^{T_{\pi^*}(i)+1}(K + \sum_{j=1}^{n} B_{\pi^*}(i, T_{\pi^*}(i)+1,j)u(j)) -$

$\text{C(i,}\pi^*(i,0)) + \sum_{t=1}^{T_{\pi^*}(i)} \gamma^t B_{\pi^*}(i,t,\cdot)C(\cdot,\pi^*(i,t)) + \gamma^{T_{\pi^*}(i)+1}(K + \sum_{j=1}^{n} B_{\pi^*}(i, T_{\pi^*}(i)+1,j)v(j))$

$\leq \gamma^{T_{\pi^*}(i)+1}\left(\sum_{j=1}^{n} B_{\pi^*}(i, T_{\pi^*}(i)+1,j)(u_j - v_j)\right)$

$\Rightarrow T^*u(i) - T^*v(i) \leq \gamma^{T_{\pi^*}(i)+1}||u-v||_\infty$

$\Rightarrow ||T^*u - T^*v||_\infty \leq \gamma^*||u-v||_\infty$

□

Application of Result

- The optimal value function is the unique fixed point of the Bellman optimality operator
- Thus value iteration and policy iteration can be used to find the optimal value function

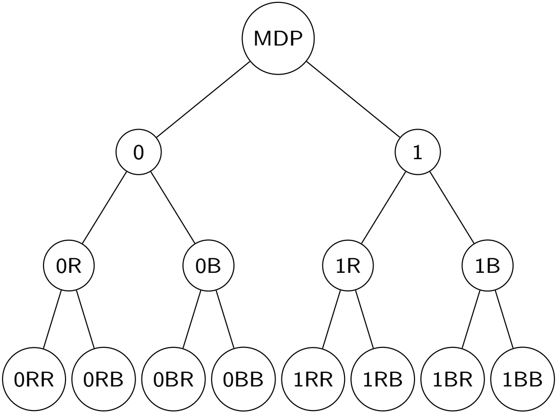
State Space Expansion

- Even if we limit the state space and action space to two states and actions say Red and Blue, the complexity of searching the optimal policy grows exponentially with max number of allowed blind steps.
- Moving forward we restrict our state space $\mathrm{S} = \{0,1\}$ and action space $\mathrm{A} = \{\mathrm{Red},\mathrm{Blue}\}$ and work on this MDP to establish subsequent results

State Space Expansion

- This POMDP can be modeled as an MDP by expanding the state space into set of states based on the sequence of actions.
- For example, if we reach the state 0R if we take action Red and decide to go blind on the subsequent state (defined as Red_blind). Similarly, we reach 0RR if we once again take Red_blind action at state 0R.
- Thus we can take 4 possible actions from every state(Red_blind, Red_see, Blue_blind, Blue_see)
- The set of all belief states is represented by $\mathbf{S}'$ and $\mathbf{B}(\mathbf{s})$ represents the distribution over the state space of the agent in a belief state s.

State Space Expansion



Conjectures

Based on the observations from the simulations we ran these were the conjectures we had:

- If we increase the number of blind steps and the optimal policy does not change once, it will remain the same thereafter
- The first action for an MDP and a POMDP are similar

Example of an MDP which follows Conjecture 1

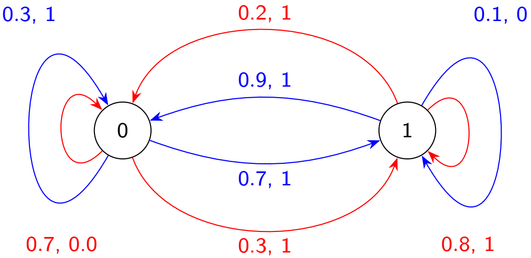


Figure: Two-state two-action MDP with action 0 in blue and action 1 in red with sensing cost 0.5

Example of an MDP which follows Conjecture 1

No. of blind steps	Optimal Policy for state 0	Value Function for state 0	Optimal Policy for state 1	Value Function For state 1
1	R	0.5	B	0.5
2	RR	0.352	BR	0.254
3	RR	0.349	BRR	0.237
4	RR	0.349	BRR	0.237
5	RR	0.349	BRR	0.237
6	RR	0.349	BRR	0.237
7	RR	0.349	BRR	0.237
8	RR	0.349	BRR	0.237
9	RR	0.349	BRR	0.237
10	RR	0.349	BRR	0.237

Table: Change in optimal policy and value function with the number of blind steps

Counter-example for conjecture 1

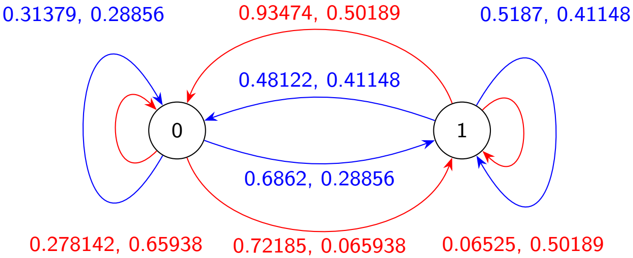


Figure: Two-state two-action MDP with action 0 in blue and action 1 in red with sensing cost 0.01.

Establishing a Sufficient Condition

We now move towards establishing a sufficient condition for sensing to be the optimal action even if there is no bound on the maximum number of blind steps

Notation

- $\mathbf{B}(\mathbf{s})$ - Row vector containing beliefs of being in state 0 and state 1.
- $\mathbf{C}(\mathbf{a})$ - Column vector containing costs of being in state 0 and 1 and taking action **a**
- $\mathbf{P}_{\pi(\mathbf{s})}$ - Transition probability matrix denoting probabilities of transition when in state 0/1 in the actual MDP and taking action according to policy π
- $\tilde{\mathbf{V}}$ - Column vector containing value functions of state 0 and 1 after following optimal policy π
- $\mathbf{Q}_\pi(\mathbf{s}, \mathbf{a})$ - Action value function of state s after taking action 'a' in it and then following optimal policy π with sensing cost K

Notation

- $\mathbf{V}^{*}_{\mathbf{K=0}}$ - Column vector containing optimal value functions of state 0 and 1 with no sensing cost
- $\mathbf{Q}^{*}_{\mathbf{K=0}}(\mathbf{a})$ - Column vector of Action value function of states 0 and 1 after taking action 'a' in it and then following the optimal policy with no sensing cost

Theorem

Theorem

If the Sensing Cost K of the MDP is below a certain threshold given by the expression

$$\min_{a_1, a_2, B(s)} B(s) P_{a_1} \left(Q^*_{K=0}(a_2) - V^*_{K=0} \right)$$

then sensing is optimal at every state

Proof

Proof.

Here we want the optimal action in any state to be **a + sensing** instead of just **a** (action a and going blind)

We claim that

$$Q_\pi(s, a_1) > Q_\pi(s, a_1 + sensing)$$

$\forall\ a_1 \in \{\text{R,B}\}$ and s in expanded state space is a sufficient condition for sensing to be optimal at every state

$$\Rightarrow B(s)C(a_1) + \gamma\Big(B(s)T_{a_1}C(a_2) + \gamma K + \gamma\big(B(s)P_{a_1}P_{a_2}\tilde{V}\big)\Big) > B(s)C(a_1) + \gamma K + \gamma\big(B(s)P_{a_1}\tilde{V}\big)$$

□

Proof

Proof.

The relation between the Value functions is given by $\tilde{V} = V^*_{K=0} + \frac{\gamma}{1-\gamma}KE$, where $E = [1\ 1]^T$

$$\Rightarrow B(s)P_{a_1}C(a_2) + \gamma K + \gamma\big(B(s)P_{a_1}P_{a_2}(V^*_{K=0} + \frac{\gamma}{1-\gamma}KE)\big) > K + \big(B(s)P_{a_2}(V^*_{K=0} + \frac{\gamma}{1-\gamma}KE)\big)$$

$$\Rightarrow B(s)\big(P_{a_1}C(a_2) + \gamma P_{a_1}P_{a_2}V^*_{K=0} - P_{a_1}V^*_{K=0}(0,1)\big) > (1-\gamma)K + KB(s)E\frac{\gamma}{1-\gamma}(1-\gamma)$$

□

Proof

Proof.

Now, it is easy to see that $B(s)E = 1$.

$$\begin{aligned} B(s)\left(P_{a_1}C(a_2) + \gamma P_{a_1}P_{a_2}V^*_{K=0} - P_{a_1}V^*_{K=0}\right) &> K \\ B(s)P_{a_1}\left(C(a_2) - \left(I - \gamma P_{a_2}\right)V^*_{K=0}\right) &> K \end{aligned}$$

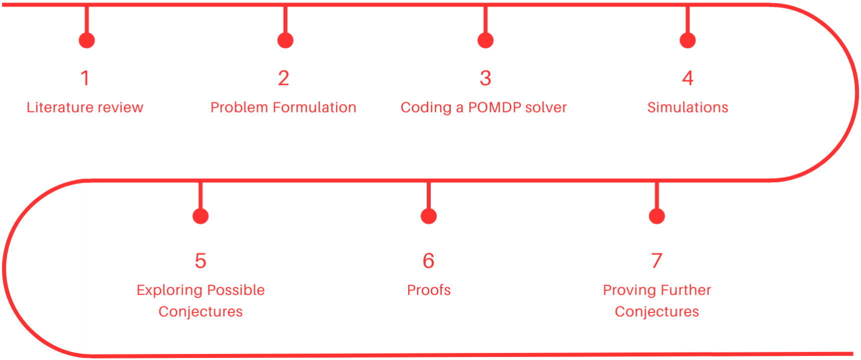
$$K < \min_{a_1,a_2,B(s)} B(s)P_{a_1}\left(C(a_2) - \left(I - \gamma P_{a_2}\right)V^*_{K=0}\right)$$

$$K < \min_{a_1,a_2,B(s)} B(s)P_{a_1}\left(C(a_2) + \gamma P_{a_2}V^*_{K=0} - V^*_{K=0}\right)$$

$$K < \min_{a_1,a_2,B(s)} B(s)P_{a_1}\left(Q^*_{K=0}(a_2) - V^*_{K=0}\right)$$

□

Timeline



1
Literature review

2
Problem Formulation

3
Coding a POMDP solver

4
Simulations

5
Exploring Possible Conjectures

6
Proofs

7
Proving Further Conjectures

Future Work

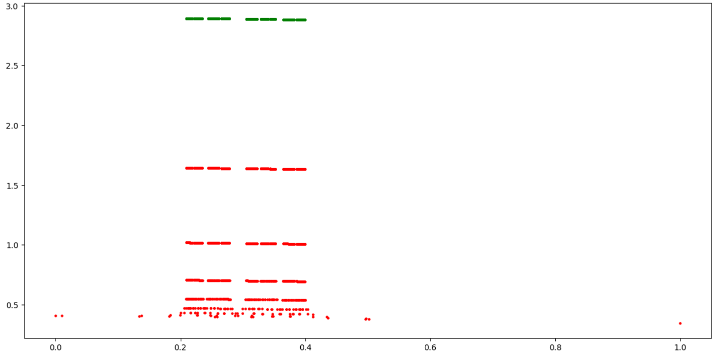


Figure: The x-axis denotes the probability that any belief state is state 1 and the y-axis denotes the value function of the belief states. Action red implies taking action red and action green implies taking action red + sensing

Future Work

As we can see that the value functions form clusters and the origin of these clusters will be explored by us in the upcoming weeks

References

The code repository can be found here