Optimal policy for partially Observable MDP with r blind actions

Asif Shaikh, Ronil Mandaviya, Vansh Kapoor

December 2022

1 Basic Formulation

1.1 Problem Statement

A Partially Observable MDP is a generalization of the MDP setting where the system dynamics are determined by an MDP, but the agent cannot directly observe the underlying state. In this setting, 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. We must find an efficient algorithm to evaluate the optimal policy for the agent in this setting.

1.2 Notation

- $\mathbf{S} = \{\mathbf{s_1}, \mathbf{s_2}, \ldots, \mathbf{s_n}\}$- Set of states of the MDP
- $\mathbf{A} = \{\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_m}\}$- Set of actions available to the agent
- $\mathbf{n}$ - number of states
- $\mathbf{m}$ - number of actions
- $\mathbf{P_{ij}(a)}$ - 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
- $\mathbf{K}$ - the cost of querying present state
- $\mathbf{r}$ - maximum number of consecutive blind actions that the agent can take.
- $\pi\mathbf{(i, t)}$ - 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. $0 \leq T_\pi(i) \leq r$
- $\mathbf{B}_\pi\mathbf{(i, t, j)}$ - 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.

2 Theory

2.1 Size of policy space

2.2 Bellman Equation

The Bellman equation defined below can be used to evaluate the value function for any policy (π, T_π)

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

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

It is easy to show that the above is a contraction mapping.

Definition: A mapping T : B(l) $\rightarrow$ B(l) is said to be a contraction mapping if

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

for some $\beta < 1$, where $||u|| = \max u$ (taking L_∞ norm)

Proof: Let the bellman operator defined above be T

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

2.3 Bellman Optimality Operator

$$\begin{aligned}
T^*u(i) - T^*v(i) = \min_\pi \{C(i,\pi(i,0)) + \sum_{t=1}^{T_\pi(i)} \gamma^t B_\pi(i,t,\cdot) C(\cdot,\pi(i,t)) + \gamma^{T_\pi(i)+1}\big(K + \sum_{j=1}^{n} B_\pi(i, T_\pi(i)+1, j) u(j)\big)\} - \\
\min_\pi \{C(i,\pi(i,0)) + \sum_{t=1}^{T_\pi(i)} \gamma^t B_\pi(i,t,\cdot) C(\cdot,\pi(i,t)) + \gamma^{T_\pi(i)+1}\big(K + \sum_{j=1}^{n} B_\pi(i, T_\pi(i)+1, j) v(j)\big)\}
\end{aligned}$$

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

$$\Rightarrow T^*u(i) - T^*v(i) \leq C(i,\pi^*(i,0)) + \sum_{t=1}^{T_{\pi^*}(i)} \gamma^t B_{\pi^*}(i,t,\cdot)C(\cdot,\pi^*(i,t)) + \gamma^{T_{\pi^*}(i)+1}\big(K + \sum_{j=1}^{n} B_{\pi^*}(i,T_{\pi^*}(i)+1,j)u(j)\big) -$$
$$C(i,\pi^*(i,0)) + \sum_{t=1}^{T_{\pi^*}(i)} \gamma^t B_{\pi^*}(i,t,\cdot)C(\cdot,\pi^*(i,t)) + \gamma^{T_{\pi^*}(i)+1}\big(K + \sum_{j=1}^{n} B_{\pi^*}(i,T_{\pi^*}(i)+1,j)v(j)\big)$$
$$\leq \gamma^{T_{\pi^*}(i)+1}\Big(\sum_{j=1}^{n} B_{\pi^*}(i,T_{\pi^*}(i)+1,j)(u_j - v_j)\Big)$$
$$\Rightarrow T^*u(i) - T^*v(i) \leq \gamma^{T_{\pi^*}(i)+1}||u-v||$$
$$\Rightarrow ||T^*u - T^*v|| \leq \gamma^*||u-v||$$

Sensing Cost=0.01, gamma=0.5
transition 0 R 0 0.06593861175481464 0.2781420941052756
transition 0 R 1 0.06593861175481464 0.7218579058947244
transition 0 B 0 0.28856606907502125 0.31379133602017495
transition 0 B 1 0.28856606907502125 0.686208663979825
transition 1 R 0 0.5018966865305635 0.9347459600535482
transition 1 R 1 0.5018966865305635 0.06525403994645185
transition 1 B 0 0.4114868983999753 0.481229557314791
transition 1 B 1 0.4114868983999753 0.518770442685209

R B
0.3583201285214027 0.6720146979977837
R BRR
0.35801600010782264 0.6712892561762362

3 MDP modelling of POMDP by state space expansion

The POMDP can be modelled as an MDP by expanding the state space into belief states. After taking a sequence $\mathbf{s}$ of blind actions starting from state i, the resultant belief state of the agent can be represented as $\mathbf{s_i}$. From this state, the possible actions that the agent can take are a, $a \in A$ and $a', a \in A$, where a' represents an action followed by sensing. The cost incurred with such an action is $C(s,a') = C(s,a) + \gamma K$. On taking blind action $\mathbf{a}$ from state $\mathbf{s}$, the resultant state is $\mathbf{s} + \mathbf{a}$. The set of all belief states is represented by $\mathbf{S'}$. $\mathbf{B(s)}$ represent the distribution over the state space of the agent in belief state s.

4 Sufficient conditions for always sensing to be the optimal policy

In the expanded state space MDP model for a two-state ($S = \{0,1\}$), two-action ($A = \{R,B\}$) POMDP, with $W = \infty$ let π be a policy where $\pi(s) = a', a \in A, \forall s \in S$ and $\forall s \in S'$, i.e., the policy for each state has a sensing action. Additionally, $\pi(0,1) = \pi^*_{K=0,W=0}(0,1)$. If at each state, π cannot be improved by replacing the action a' at s with a, i.e., π cannot be improved by Not Sensing at any state. For a POMDP satisfying the above, the π constructed here is the optimal policy.

Here we are considering a model in which we expand the state space rather than the action space when we go blind. Here the state space is of the form 'i' + string where $i \in \{0,1\}$ and $string \in \{R,B\}^n$ where n is the blind window size. We maintain a belief vector over all possible states.

$\mathbf{Q}_\pi(\mathbf{s},\mathbf{a})$ - Action value function of state s after taking action 'a' in it and then following policy π.

$$Q_\pi(s,a) = C(s,a) + \gamma \sum_{s'} V_\pi(s')$$

Here we want the optimal action in any state to be **a + sensing** instead of just **a**.
Let $\pi(s) = a_1 + \text{sensing}$, $\pi(s+a_1) = a_2 + \text{sensing}$ ($a_1, a_2 \in \{\text{R,B}\}$). For the dominant policy condition to hold true,

$$\mathbf{Q}_\pi(\mathbf{s},\mathbf{a_1}) > \mathbf{V}_\pi(\mathbf{s}) \tag{2}$$

. Thus,

$$\mathbf{C}(\mathbf{s},\mathbf{a_1}) + \gamma \mathbf{V}_\pi(\mathbf{s}+\mathbf{a_1}) > \mathbf{V}_\pi(\mathbf{s}) \tag{3}$$

$$\mathbf{V}_\pi(\mathbf{s}) = \mathbf{B}(\mathbf{s})\mathbf{C}(\pi(\mathbf{s})) + \gamma\mathbf{K} + \gamma(\mathbf{B}(\mathbf{s})\mathbf{T}_{\pi(\mathbf{s})}\mathbf{V}_\pi(\mathbf{0},\mathbf{1})) \tag{4}$$

Here,

- $\mathbf{B}(\mathbf{s})$ - 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**.
- $\mathbf{T}_{\pi(\mathbf{s})}$ - Transition probability matrix denoting probabilities of transition of being in state 0(1) and taking action according to policy π.
- $\mathbf{V}_{\pi(\mathbf{0},\mathbf{1})}$ - Column vector containing value functions of state 0 and 1 after following policy π.

Substituting (4), in (3), we get

$$B(s)C(a_1) + \gamma\Big(B(s+a_1)C(\pi(s+a_1)) + \gamma K + \gamma\big(B(s+a_1)T_{\pi(s+a_1)}V_\pi(0,1)\big)\Big) > B(s)C(\pi(s)) + \gamma K + \gamma\big(B(s)T_{\pi(s)}V_\pi(0,1)\big)$$

$$B(s)C(a_1) + \gamma\Big(B(s)T_{a_1}C(a_2) + \gamma K + \gamma\big(B(s)T_{a_1}T_{a_2}V_\pi(0,1)\big)\Big) > B(s)C(a_1) + \gamma K + \gamma\big(B(s)T_{a_1}V_\pi(0,1)\big)$$

Now $V_\pi(0,1)$. Hence $V_\pi(0,1)$ depends only on the actions taken at states 0 and 1 under the assumed conditions. There are two possible policies, $\pi(0) = R, \pi(1) = B$ or $\pi(0) = B, \pi(1) = R$. The sensing is implicit here because of our initial assumption. But $V_\pi(0,1)$ itself depends on K. This relation is given as $V_\pi(0,1) = V_{\pi,K=0}(0,1) + \frac{\gamma}{1-\gamma}KE$, where $E = [1\ 1]^T$.

$$B(s)T_{a_1}C(a_2) + \gamma K + \gamma\big(B(s)T_{a_1}T_{a_2}(V_{\pi,K=0}(0,1) + \frac{\gamma}{1-\gamma}KE)\big) > K + \big(B(s)T_{a_2}(V_{\pi,K=0}(0,1) + \frac{\gamma}{1-\gamma}KE)\big)$$

$$B(s)\big(T_{a_1}C(a_2) + \gamma T_{a_1}T_{a_2}V_{\pi,K=0}(0,1) - T_{a_1}V_{\pi,K=0}(0,1)\big) > (1-\gamma)K + KB(s)E\frac{\gamma}{1-\gamma}(1-\gamma)$$

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

$$\begin{aligned} B(s)\big(T_{a_1}C(a_2) + \gamma T_{a_1}T_{a_2}V_{\pi,K=0}(0,1) - T_{a_1}V_{\pi,K=0}(0,1)\big) &> K \\ B(s)T_{a_1}\Big(C(a_2) - \big(I - \gamma T_{a_2}\big)V^*_{K=0}(0,1)\Big) &> K \end{aligned}$$

$$K < \min_{a_1,a_2,B(s)} B(s)T_{a_1}\Big(C(a_2) - \big(I - \gamma T_{a_2}\big)V^*_{K=0}(0,1)\Big)$$

$$K < \min_{a_1,a_2,B(s)} B(s)T_{a_1}\Big(C(a_2) + \gamma T_{a_2}V^*_{K=0}(0,1) - V^*_{K=0}(0,1)\Big)$$

$$K < \min_{a_1,a_2,B(s)} B(s)T_{a_1}\Big(Q^{\pi^*}_{K=0}(a_2)(0,1) - V^*_{K=0}(0,1)\Big)$$