

Math for the human - robot multilevel POMDP

1 Human (PO)MDP

This is a MDP that the human runs in their mind. It can be formalized as a POMDP, but right now we design it to be an MDP with a belief distribution over the robot's distribution of objects. We repeatedly look at what the nested POMDP would look like to model our MDP.

1. $S = \langle \iota, \mathcal{I}, d, H \rangle$
 - (a) Here \mathcal{I} is the set of items the robot can pass.
 - (b) Item ι is the object the human wants, and this is hidden information for the robot.
 - (c) d is the dialogue state - what question was asked previously by the robot.
 - (d) Here $B(i) = P(i == \iota), \forall i \in \mathcal{I}$ is the belief distribution of the robot for human's needed item for all the items in \mathcal{I} . In the POMDP formulation B would be part of the state that is hidden from the human. Instead we model a distribution that tracks B allowing the human to solve an MDP instead of nested POMDPs. We will go back to the nested POMDP model repeatedly to make sure our MDP model is equivalent.
- *2. Human's hunch of the robot's belief $H = P(\hat{B}|\eta)$, where \hat{B} is an estimate of the distribution of B , it is over the set of items \mathcal{I} . η is a set of priors that defines the distribution \hat{B} , hence H is over the space of all possible priors values. We propose to use the Dirichlet distribution to model H . A set of values for possible priors need to be listed. These can be sampled so we have a fixed set of priors instead of infinite possible priors.
3. $A_h = \langle l, g \rangle$, where A_h is the human action set and l and g are language and gesture actions respectively.
4. If this were a POMDP we would need observation functions and a observation set.
 - (a) $\Omega_h = \langle A_r \rangle$, where A_r is the set of robot actions and Ω_h is the set of human observations.

- (b) $O = P(A_r|\iota, \mathcal{I}, d, H, B)$ is the observation function and it is hand coded by us so we know it, since we know the robots response to all the belief states.
 - (c) $T = P(\iota', \mathcal{I}', d', H', B'|\iota, \mathcal{I}, d, H, B, a_h, a_r) = P(\iota', \mathcal{I}', d'|\iota, \mathcal{I}, d, H, B, a_h, a_r) \times P(H'|\iota, \mathcal{I}, d, H, B, a_h, a_r) \times P(B'|\iota, \mathcal{I}, d, H, B, a_h, a_r)$ is the transition function.
- *5. MDP formulation of this problem would not need the observation set or the observation functions, instead H would get updated based on A_r and A_h . This just has a transition function now defined as

$$T = P(\iota', \mathcal{I}', d', H'|\iota, \mathcal{I}, d, H, a_h, a_r) \quad (1)$$

$$= P(\iota', \mathcal{I}', d'|\iota, \mathcal{I}, d, H, a_h, a_r) \times P(H'|\iota, \mathcal{I}, d, H, B, a_h, a_r) \quad (2)$$

. The conditional independence of the human's hunch H from the distribution over the required item, or set of items left over or the last question asked comes from visible robot actions. $P(H'|\iota, \mathcal{I}, d, H, B, a_h, a_r)$ is being designed by us as an approximation and we need to think of data intensive methods of measuring this transition. If a_r is a pick action and i is the object picked:

$$P(\iota', \mathcal{I}', d'|\iota, \mathcal{I}, d, H, a_h, a_r) = \begin{cases} 1/|\mathcal{I}_1| & \text{if } i! = \iota \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

If a_r is an ask question

$$P(\iota', \mathcal{I}', d'|\iota, \mathcal{I}, d, H, a_h, a_r) = \begin{cases} 1 & \text{if } d' = \text{a.ask} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

- 6. The reward for the human subject is not well defined, but we can assume that the net reward is for both human and the robot, since this is a co-operative domain.

2 Robot POMDP

- 1. $S = \langle \iota, \mathcal{I}, d, Mo(H), B \rangle$. Here $Mo(H)$ is the mode of the distribution H which is unknown and B is the known belief over items. The belief state over items is part of the second level state.
- 2. $A_r = \text{wait, pick(object), ask(property), point(property)}$
- 3. $\Omega_r = \langle l, g \rangle$
- *4. $O = P(o|\iota, \mathcal{I}, a_r, a_h, Mo(H))$: this is a unigram based model, if we assume $Mo(H)$ to be a countable set then we can count these observation probabilities, for different a_r, ι and $Mo(H)$. We assume that there is a data intensive way of computing H repeated and calculate its mode $Mo(H)$.

- *5. T The transition is the same as the MDP transition function, except here we need to track only the $Mo(H)$ instead of all of H . This approximation needs to be relaxed for correct modeling.
- 6. R : Reward function can be a combination of regular rewards like large penalties for a wrong pick and low for an ask, point, and wait. Small positive rewards for the correct pick. We have to figure out if a reward hack of $KL(B||Mo(H))$ is worth pursuing. We are not sure this can be called reward shaping as we don't know if the optimal policy will remain the same under this shaping function.