Lecture 10 - *Partially Observable Markov Decision Processes*

OBJECTIVE: POMDPs are a natural extension of MDPs to worlds where the states are only partially observed. POMDPs are powerful and realistic models, but are extremely hard to optimize. In this lecture, we introduce POMDPs and illustrate them with several examples.

◇ DEFINITION

In the POMDP setting, the agent has a belief $b = p(\mathbf{s}_t = \mathbf{s}|\mathbf{y}_{1:t}, \mathbf{a}_{1:t-1})$ over the hidden states given everything it has seen up to time t. By taking an action $\mathbf{a}_t = \mathbf{a}$, the agent will transition to a new state $\mathbf{s}'$ according to the transition model $p(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ and make an observation $\mathbf{y}_{t+1} = \mathbf{y}$ according to the observation model $p(\mathbf{y}|\mathbf{s}', \mathbf{a})$. The agent receives a reward $r(\mathbf{s}, \mathbf{s}', \mathbf{a}, \mathbf{y})$. Of course, a priori, the agent does not know

its exact state, the future state and the future observations.

The following decision graph illustrates the elements and sequence of operations in a POMDP:

⋆

◇ POMDP EXAMPLES

POMDPs are of fundamental importance to AI, Games, control, robotics, planning, scheduling, HCI, sequential experimental design and sequential decision making in general. The examples provided here are by no means exhaustive, but a small taste of what is possible:

1. **Boutilier and Poole's coffee delivery robot:**
   (a) **States:** David has coffee,David wants coffee, raining, wet robot, robot carries umbrella.
   (b) **Observations:** David's answer.
   (c) **Control actions:** Get coffee, check if David wants coffee.

2. **Human learning and instruction:**
   (a) **States:** The student's knowledge of the subject.
   (b) **Observations:** Response to questions or exams.
   (c) **Control actions:** Choose alternative presentation, decide whether to proceed or revise the material.
3. **Computers that ask questions:**
   (a) **States:** Status of human operator (e.g. knowledge level, irritation level, business, and so on).
   (b) **Observations:** Answers to questions.
   (c) **Control actions:** A sequence of questions aimed at maximizing knowledge about the status of the human operator (preference elicitation). Here one also must minimize the irritation factor!

4. **Intelligent image/music/speech labelling:**
   (a) **States:** Parameters of a recognition model (e.g. a tiger detector).
   (b) **Observations:** Label indicating whether the chosen image is a tiger or not.
   (c) **Control actions:** Choose among all the images retrieved by the Google query: tiger.

5. **Sequential image processing:**
   (a) **States:** Labels for image features (image explanation in text).
   (b) **Observations:** Image features.
   (c) **Control actions:** Whether to use a more expensive detector and, if so, where in the image it should be used.

6. **Diagnosis:**
   (a) **States:** Internal status of machine (e.g. Mars rover).
   (b) **Observations:** Internal and external sensors.
   (c) **Control actions:** Diagnose and fix faults automatically, before they happen.

7. **Medical prognosis and drug design:**
   (a) **States:** Physiological status of patient.
   (b) **Observations:** Response to treatment.
   (c) **Control actions:** Select types of drug and treatment.

8. **Sensor networks:**

   (a) **States:** Properties of phenomenon being observed.

   (b) **Observations:** Physical measurements with network of sensors.

   (c) **Control actions:** Where to place the sensors, which sensor should be switch on?

9. **Search for a moving/changing target:**

   (a) **States:** Target position, velocity and other properties.

   (b) **Observations:** Result of search (e.g., found,not found, found a friend).

   (c) **Control actions:** Where to search and resource expenditure.

10. **Spoken dialog systems:**

    (a) **States:** What the human user wants (e.g. pay a phone bill).

    (b) **Observations:** Speech utterances.

    (c) **Control actions:** Greet, query customer (propose options), confirm statements, end the conversation.

11. **Task assistance for the elderly and people with Dementia:**

    (a) **States:** Environmental variables (hands location), task variables, patient's mental state and so on. See the wonderful PhD theses of Joelle Pineau and Pascal Poupart for more details.

    (b) **Observations:** Video of person involved in the task.

    (c) **Control actions:** When and what to say to assist person.

◇ INFORMATION STATES AND FILTERING

In a POMDP, the agents policy states what actions to take using the estimate of the state based on the history of actions and observations. This information about the state is summarized by the filtering distribution, also known as the information state or belief:

$$b_y^a(\mathbf{s}') = p(\mathbf{s}_t = \mathbf{s}' | \mathbf{y}_{1:t}, \mathbf{a}_{1:t-1}).$$

The belief can be computed recursively using marginalization (state prediction) and Bayes rule (correction):

$$\begin{aligned} p(\mathbf{s}_t = \mathbf{s}' | \mathbf{y}_{1:t-1}, \mathbf{a}_{1:t-1}) &= \sum_{\mathbf{s}_{t-1}} p(\mathbf{s}' | \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) p(\mathbf{s}_{t-1} | \mathbf{y}_{1:t-1}, \mathbf{a}_{1:t-2}) \\ p(\mathbf{s}_t = \mathbf{s}' | \mathbf{y}_{1:t}, \mathbf{a}_{1:t-1}) &= \frac{p(\mathbf{y}_t | \mathbf{s}', \mathbf{a}_{t-1}) p(\mathbf{s}' | \mathbf{y}_{1:t-1}, \mathbf{a}_{1:t-1})}{p(\mathbf{y}_t | \mathbf{y}_{1:t-1}, \mathbf{a}_{1:t-1})} \end{aligned}$$

In shorthand notation, we will summarize the Bayesian prediction and update equations with a single filtering equation:

$$b_y^a(\mathbf{s}') = \frac{p(\mathbf{y} | \mathbf{s}', \mathbf{a}) \sum_{\mathbf{s}} p(\mathbf{s}' | \mathbf{s}, \mathbf{a}) b(\mathbf{s})}{p(\mathbf{y} | \mathbf{a}, b)}$$

where

$$p(\mathbf{y} | \mathbf{a}, b) = \sum_{\mathbf{s}'} p(\mathbf{y} | \mathbf{s}', \mathbf{a}) \sum_{\mathbf{s}} p(\mathbf{s}' | \mathbf{s}, \mathbf{a}) b(\mathbf{s})$$

◇ VALUE ITERATION FOR POMDPS

The value function of an information state, when starting with belief b_0 and following policy $\boldsymbol{\pi}$, is given by:

$$V^{\pi}(b_0) = \mathbb{E}_{\pi}\left\{\sum_{t=1} \gamma^t r\left(\mathbf{s}_t, \mathbf{s}'_t, \boldsymbol{\pi}(b_t), \mathbf{y}_t\right)\right\}$$

where the expectation is over $(\mathbf{s}, \mathbf{s}', \mathbf{y}, \mathbf{a})$ for all t.

As we did in the MDP setting, we can apply dynamic programming to obtain a recursive backup for the optimal value function:

$$V^{\star}(b) = \max_{\mathbf{a}}\left\{\sum_{\mathbf{s},\mathbf{s}',\mathbf{y}} b(\mathbf{s})p(\mathbf{s}'|\mathbf{s},\mathbf{a})p(\mathbf{y}|\mathbf{s}',\mathbf{a})\left[r(\mathbf{s},\mathbf{s}',\mathbf{a},\mathbf{y}) + \gamma V^{\star}(b^a_y(\mathbf{s}'))\right]\right\}$$

Using expected rewards and the expression for the normalizing term in the filtering equations, the value function backup

simplifies to:

$$V^{\star}(b) = \max_{\mathbf{a}}\left\{\sum_{\mathbf{s}} b(\mathbf{s})r(\mathbf{s},\mathbf{a}) + \sum_{\mathbf{s},\mathbf{s}',\mathbf{y}} b(\mathbf{s})p(\mathbf{s}'|\mathbf{s},\mathbf{a})p(\mathbf{y}|\mathbf{s}',\mathbf{a})\gamma V^{\star}(b^a_y(\mathbf{s}'))\right\}$$

Moreover, f rom the filtering results, we know that:

$$p(\mathbf{y}|\mathbf{a}, b) = \sum_{\mathbf{s}'} p(\mathbf{y}|\mathbf{s}',\mathbf{a}) \sum_{\mathbf{s}} p(\mathbf{s}'|\mathbf{s},\mathbf{a})b(\mathbf{s})$$

Hence, upon substitution of this filtering equation into the value function update, we get:

$$V^{\star}(b) = \max_{\mathbf{a}}\left\{\sum_{\mathbf{s}} b(\mathbf{s})r(\mathbf{s},\mathbf{a}) + \gamma \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{a}, b)V^{\star}(b^a_y(\mathbf{s}'))\right\}$$

◇ POMDPS AS MDPS

If we define the belief transition model as follows:

$$p(b'|a,b) = \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{a},b)\mathbb{I}(b', b_y^a(\mathbf{s}'))$$

where

$$\mathbb{I}(b', b_y^a(\mathbf{s}')) = \begin{cases} 1 & \text{if } b' = b_y^a(\mathbf{s}') \\ 0 & \text{otherwise.} \end{cases}$$

Then the value update becomes:

$$V^\star(b) = \max_{\mathbf{a}} \left\{ \sum_{\mathbf{s}} b(\mathbf{s}) r(\mathbf{s},\mathbf{a}) + \gamma \sum_{b'} p(b'|a,b) V^\star(b') \right\}$$

From which we see that a POMDP is an MDP in belief space.