6.4110 Recitation: POMDPs

Partially Observable MDPs (POMDPs)

A POMDP is defined by:

$$\langle \mathcal{S}, \mathcal{A}, T, R, \mathcal{O}, O, \gamma \rangle$$

- $T(s, a, s') = P(s'|s, a)$
- $O(s', a, o) = P(o|s', a)$

Goal:

$$\max \mathbb{E}\left[\sum_t \gamma^t r_t\right]$$

Key challenge: the state is not observable.

Belief State

Belief is a probability distribution over states:

$$b(s) = P(s), \quad \sum_s b(s) = 1$$

$$b_t(s) = P(s_t \mid a_0, o_0, \ldots, a_{t-1}, o_{t-1})$$

Key property: b_t is a sufficient statistic.

Belief Update (Bayes Filter)

$$b'(s') = \eta\, O(s', a, o) \sum_s T(s, a, s')\, b(s)$$

$$\eta = \frac{1}{P(o|b, a)}$$

$$P(o|b, a) = \sum_{s'} O(s', a, o) \sum_s T(s, a, s') b(s)$$

Interpretation:

- Prediction: $\sum_s T(s, a, s') b(s)$
- Update: weight by $O(s', a, o)$

Belief MDP

- State: b
- Action: a
- Reward:
$$R(b,a) = \sum_s b(s)R(s,a)$$
- Transition:
$$b' = bf(b,a,o), \quad \text{with prob } P(o|b,a)$$

Key idea: planning occurs in belief space.

Expectimax in POMDPs

$$V(b) = \max_a \sum_o P(o|b,a)\, V(b')$$

Structure:

- Max over actions
- Expectation over observations

Policy Trees

A policy maps observations to actions.

$$a \to \begin{cases} o_1 : a_1 \\ o_2 : a_2 \end{cases}$$

Value of a Policy Tree

Value starting from state s:

$$V(s) = R(s,a) + \sum_o P(o|s,a)\, V_{\text{subtree}}(s)$$

At belief:

$$V(b) = \sum_s b(s)V(s)$$

Key Insight: Value of Information

Some actions reduce uncertainty and improve future decisions.
Optimal policies balance:

- immediate reward
- information gain

1 Prospecting

Disclaimer*: this example is completely made-up and probably ridiculous from the petroleum engineering standpoint.*

Your job is to automate the decision-making process for oil exploration and drilling. In a particular site, your goal is to strike pumpable oil, if possible.

- At that site, there might be shallow oil, deep oil, or no oil.
- You can: drill a test well and take a sample (but it won't work to pump oil from a test well), drill a shallow well, drill a deep well, or give up on this site.
- If you drill a test well, you get one of two observations: "oil" or "no oil". The probability of "no oil" when there is no oil is 1. The probability of "no oil" when there is deep oil is 0.3. The probability of "no oil" when there is shallow oil is 0.1.
- If you drill a test well, and there is shallow oil, then there is a 0.2 chance that the floor of the reservoir will rupture and the oil will become deep, otherwise it will stay shallow. Drilling a test well has no other possible effects.
- It costs -10 to drill a test well, -50 to drill a shallow well, and -200 to drill a deep well. A shallow well is guaranteed to hit oil if there is shallow oil. A deep well is guaranteed to hit oil if there is shallow or deep oil. If you hit oil with a shallow or a deep well, it is worth +1000. Giving up is worth 0. There is no discounting.

(a) Write down the state space, action space, and the reward function for this problem.

(b) If your initial belief $b = (1/3, 1/3, 1/3)$, what is the belief state resulting from drilling a test well and observing "oil"? Assume that the observation depends on the starting state (before the test well has its potential effect on the state).

(c) If your initial belief $b = (1/3, 1/3, 1/3)$, what is the belief state resulting from drilling a test well and observing "no oil"? Again, assume that the observation depends on the starting state (before the test well has its potential effect on the state).

(d) What is the value of the following policy tree at each of the three possible world states? Drill a test well; if the observation is "oil" then drill a deep well, otherwise give up. Show your math work.

(e) The following is a picture of the belief space, partitioned into regions in which different policy trees dominate. The vertices of the belief simplex are labeled with coordinates: the first coordinate is *shallow oil*, the second is *deep oil*, and the third is *no oil*.

For each of the following policies, indicate which color region it corresponds to. Note that there is a tiny yellow triangle at the top.

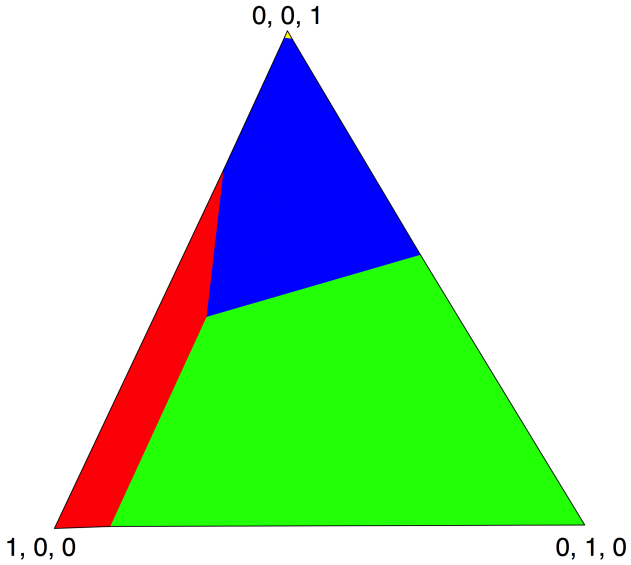
Policy A : Always drill shallow.

Policy B : Always drill deep.

Policy C : Always give up.

Policy D : Follow the policy in part c.

Note small yellow triangle at top



Yellow:	◯ Policy A	◯ Policy B	◯ Policy C	◯ Policy D
Blue:	◯ Policy A	◯ Policy B	◯ Policy C	◯ Policy D
Red:	◯ Policy A	◯ Policy B	◯ Policy C	◯ Policy D
Green:	◯ Policy A	◯ Policy B	◯ Policy C	◯ Policy D