Decision Making and Reinforcement Learning

Module 5: Partially Observable MDPs

Tony Dear, Ph.D.
Department of Computer Science
School of Engineering and Applied Sciences

Topics

- Partial observability and POMDP framework
- Belief state definition and computation
- Representation of POMDPs as belief MDPs
- Online planning in MDPs and POMDPs

Learning Objectives

- Model a partially observable decision problem as a POMDP
- Represent a POMDP as a belief MDP
- Visualize belief MDPs using backup diagrams
- Implement online planning agents for MDPs and POMDPs

Partial Observability

- MDP formalism assumes *full observability* of environment

- Decision-making agent always knows the state that it is in
- Policies can thus be defined as functions of individual states

Partial Observability

- MDP formalism assumes *full observability* of environment

- Decision-making agent always knows the state that it is in
- Policies can thus be defined as functions of individual states

- But for many real-world problems, we only have *partial observability*

- Agent does *not* know its true state, so cannot execute an action $\pi(s)$!
- Instead, we may have a *degree of belief* of being in different states

Partially Observable MDPs

- We need to modify the MDP model to deal with partial observability
- Define the **partially observable MDP (POMDP)**:

Partially Observable MDPs

- We need to modify the MDP model to deal with partial observability
- Define the **partially observable MDP (POMDP)**:

- Same elements from MDPs: States S, actions A
- Transitions $T: S \times A \times S \to [0,1]$; rewards $R: S \times A \times S \to \mathbb{R}$

Partially Observable MDPs

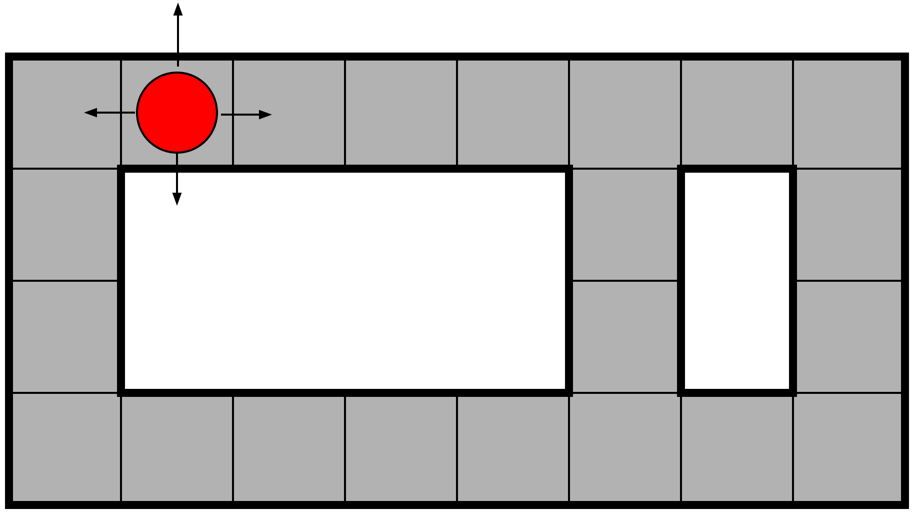
- We need to modify the MDP model to deal with partial observability
- Define the **partially observable MDP (POMDP)**:

- Same elements from MDPs: States S, actions A
- Transitions $T: S \times A \times S \to [0,1]$; rewards $R: S \times A \times S \to \mathbb{R}$

- New components: Set of possible *observations* or *evidence* Z
- **Sensor model** $O: S \times Z \to [0,1]$; $O(s,z) = \Pr(z|s)$

POMDP Example

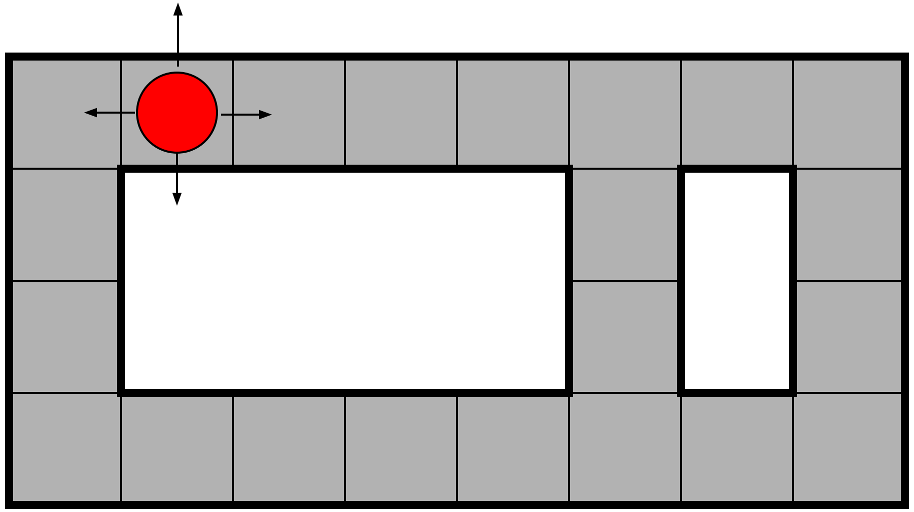
- Consider a robot roaming around a hallway
- States: Locations in the hallway
- Actions: Move in a certain direction



Example from Michael Pfeiffer

POMDP Example

- Consider a robot roaming around a hallway
- States: Locations in the hallway
- Actions: Move in a certain direction
- Robot may have a *sensor* (e.g., lidar) described by a sensor model
- Describes likelihood of seeing a wall adjacent to the robot



Example from Michael Pfeiffer

Belief States

- Policies that map from states to actions are no longer sufficient

Belief States

- Policies that map from states to actions are no longer sufficient

- Idea: Define a **belief state** $b: S \rightarrow [0,1]$ as a *probability distribution*
- $b(s)$ is the probability that the agent is in state s
- This is a *continuous* state space—infinitely many distributions!

Belief States

- Policies that map from states to actions are no longer sufficient

- Idea: Define a **belief state** $b: S \rightarrow [0,1]$ as a *probability distribution*
- $b(s)$ is the probability that the agent is in state s
- This is a *continuous* state space—infinitely many distributions!

- Ex: Mini-gridworld
  - Completely certain that we're in state A
  - No idea which state we're in (uniform)

1	0	0

Belief State Update

- How does a belief state b change, given action a and observation z?
- We obtain *conditional probabilities* of being in different successor states s'

Belief State Update

- How does a belief state b change, given action a and observation z?
- We obtain *conditional probabilities* of being in different successor states s'

- Take action a: Sum over possible starting states s, weighted by belief of being in each

$$\Pr(s'|b,a) = \sum_{s} \Pr(s'|s,a)\, b(s)$$

Belief State Update

- How does a belief state b change, given action a and observation z?
- We obtain *conditional probabilities* of being in different successor states s'

- Take action a: Sum over possible starting states s, weighted by belief of being in each

$$\Pr(s'|b,a) = \sum_{s} \Pr(s'|s,a)\, b(s)$$

- Make observation z: Apply *Bayes' rule* using sensor model

$$\Pr(s'|b,a,z) \propto \Pr(z|s')\, \Pr(s'|b,a) = \Pr(z|s') \sum_{s} \Pr(s'|s,a)\, b(s)$$

Belief State Update

- How does a belief state b change, given action a and observation z?
- We obtain *conditional probabilities* of being in different successor states s'

- Take action a: Sum over possible starting states s, weighted by belief of being in each

$$\Pr(s'|b,a) = \sum_{s} \Pr(s'|s,a)\, b(s)$$

- Make observation z: Apply *Bayes' rule* using sensor model

$$\Pr(s'|b,a,z) \propto \Pr(z|s') \Pr(s'|b,a) = \Pr(z|s') \sum_{s} \Pr(s'|s,a)\, b(s)$$

- We apply the above update to each state s' and normalize to obtain $b'(s'|b,a,z)$

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: “Intended direction” prob 0.8

A	B	C

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: "Intended direction" prob 0.8

A	B	C

- At any state, suppose that the agent can observe whether there is a wall (**1**) or no wall (**0**) in either the left or right direction, chosen randomly

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: "Intended direction" prob 0.8

A	B	C

- At any state, suppose that the agent can observe whether there is a wall (**1**) or no wall (**0**) in either the left or right direction, chosen randomly

- Sensor model: $\Pr(z = 1 | s = A) = 0.5$ $\quad \Pr(z = 0 | s = A) = 0.5$

$$\Pr(z = 1 | s = B) = 0 \qquad \Pr(z = 0 | s = B) = 1$$

$$\Pr(z = 1 | s = C) = 0.5 \qquad \Pr(z = 0 | s = C) = 0.5$$

Example: Mini-Gridworld

A	B	C

- States: A, B, C; actions: L, R
- Transition model: "Intended direction" prob 0.8

- Suppose current belief state is $\left(b(A), b(B), b(C)\right) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- Agent takes action L and observe "wall":

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: “Intended direction” prob 0.8

A	B	C

- Suppose current belief state is $\left(b(A), b(B), b(C)\right) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- Agent takes action L and observe “wall”:

$$b'(A):\ \ \Pr(1|\mathrm{A}) \times \sum_s \Pr(s' = A|s, a = L)b(s)$$

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: "Intended direction" prob 0.8

A	B	C

- Suppose current belief state is $\big(b(A), b(B), b(C)\big) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- Agent takes action L and observe "wall":

$$b'(A):\ \ \Pr(1|\mathrm{A}) \times \sum_s \Pr(s' = A|s, a = L)b(s)$$

$$b'(A):\ \ 0.5 \times \left(0.8 \times \frac{1}{3} + 0.8 \times \frac{1}{3} + 0 \times \frac{1}{3}\right) = \frac{4}{15}$$

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: “Intended direction” prob 0.8

A	B	C

- Suppose current belief state is $\big(b(A), b(B), b(C)\big) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- Agent takes action L and observe “wall”:

$$b'(B):\;\; \cancel{\Pr(1|B)} \times \sum_s \Pr(s' = A|s, a = L)b(s)$$

0

Example: Mini-Gridworld

A	B	C

- States: A, B, C; actions: L, R
- Transition model: “Intended direction” prob 0.8

- Suppose current belief state is $\left(b(A), b(B), b(C)\right) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- Agent takes action L and observe “wall”:

$$b'(C):\ \ \Pr(1|C) \times \sum_s \Pr(s' = C|s, a = L)b(s)$$

$$b'(C):\ \ 0.5 \times \left(0 \times \frac{1}{3} + 0.2 \times \frac{1}{3} + 0.2 \times \frac{1}{3}\right) = \frac{1}{15}$$

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: "Intended direction" prob 0.8

A	B	C

- Suppose current belief state is $\left(b(A), b(B), b(C)\right) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- Agent takes action L and observe "wall":

$$\begin{pmatrix} b'(A) \\ b'(B) \\ b'(C) \end{pmatrix} \propto \begin{pmatrix} 4/15 \\ 0 \\ 1/15 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} b'(A) \\ b'(B) \\ b'(C) \end{pmatrix} = \begin{pmatrix} 0.8 \\ 0 \\ 0.2 \end{pmatrix}$$

Belief Transition Model

- We can find an equivalent *observable* MDP to represent a POMDP
- State space is simply the (continuous) set of belief states

Belief Transition Model

- We can find an equivalent *observable* MDP to represent a POMDP
- State space is simply the (continuous) set of belief states

- Transition model: probability of moving between different belief states
- Given a fixed observation z, we have seen how to compute the new belief state b' given the old belief state b

Belief Transition Model

- We can find an equivalent *observable* MDP to represent a POMDP
- State space is simply the (continuous) set of belief states

- Transition model: probability of moving between different belief states
- Given a fixed observation z, we have seen how to compute the new belief state b' given the old belief state b

- The *transition probability* $\Pr(b'|b, a)$ is equal to the probability of observing z!

Belief Transition Model

- Suppose z is the observation that helps us to deterministically update the belief state from b to b'

$$\Pr(b'|b,a) = \Pr(z|b,a)$$

Belief Transition Model

- Suppose z is the observation that helps us to deterministically update the belief state from b to b'

$$\Pr(b'|b,a) = \Pr(z|b,a)$$

- What is the probability of observing z given belief state b and action a?

$$\Pr(z|b,a) = \sum_{s'} \Pr(z|s') \Pr(s'|b,a) = \sum_{s'} \Pr(z|s') \sum_{s} \Pr(s'|s,a)\, b(s)$$

Belief Transition Model

- Suppose z is the observation that helps us to deterministically update the belief state from b to b'

$$\Pr(b'|b,a) = \Pr(z|b,a)$$

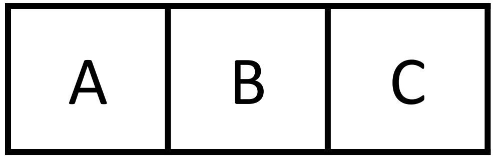
- What is the probability of observing z given belief state b and action a?

$$\Pr(z|b,a) = \sum_{s'} \Pr(z|s') \Pr(s'|b,a) = \sum_{s'} \Pr(z|s') \sum_{s} \Pr(s'|s,a)\, b(s)$$

- The above sum is simply the sum of the "unnormalized" likelihoods of b'!

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: “Intended direction” prob 0.8



Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: “Intended direction” prob 0.8

A	B	C

- Recall that $b' = (0.8,0,0.2)$ is the successor belief state to $b = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ upon taking action L and observing “wall”

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: “Intended direction” prob 0.8

A	B	C

- Recall that $b' = (0.8,0,0.2)$ is the successor belief state to $b = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ upon taking action L and observing “wall”

- The *transition probability* $\Pr(b'|b,a)$ is equal to $\Pr(z=1|b,a)$:

$$\Pr(z|b,a) = \sum_{s'} \Pr(z|s') \Pr(s'|b,a) = 0.5\left(\frac{4}{15}\right) + 0(0) + 0.5\left(\frac{1}{15}\right) = \frac{1}{3}$$

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: “Intended direction” prob 0.8

A	B	C

- What is the successor belief state b'' if we observe “no wall”?

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: “Intended direction” prob 0.8

A	B	C

- What is the successor belief state b'' if we observe “no wall”?

$$\begin{pmatrix} b''(A) \\ b''(B) \\ b''(C) \end{pmatrix} \propto \begin{pmatrix} 0.5 \times (0.8 \times 1/3 + 0.8 \times 1/3 + 0 \times 1/3) \\ 1 \times (0.2 \times 1/3 + 0 \times 1/3 + 0.8 \times 1/3) \\ 0.5 \times (0 \times 1/3 + 0.2 \times 1/3 + 0.2 \times 1/3) \end{pmatrix} = \begin{pmatrix} 4/15 \\ 1/3 \\ 1/15 \end{pmatrix}$$

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: "Intended direction" prob 0.8

A	B	C

- What is the successor belief state b'' if we observe "no wall"?

$$\begin{pmatrix} b''(A) \\ b''(B) \\ b''(C) \end{pmatrix} \propto \begin{pmatrix} 0.5 \times (0.8 \times 1/3 + 0.8 \times 1/3 + 0 \times 1/3) \\ 1 \times (0.2 \times 1/3 + 0 \times 1/3 + 0.8 \times 1/3) \\ 0.5 \times (0 \times 1/3 + 0.2 \times 1/3 + 0.2 \times 1/3) \end{pmatrix} = \begin{pmatrix} 4/15 \\ 1/3 \\ 1/15 \end{pmatrix}$$

- Normalize: $b'' = (0.4,0.5,0.1)$; $\Pr(b''|b,a) = \frac{4}{15} + \frac{1}{3} + \frac{1}{15} = \frac{2}{3}$
- Note that $\Pr(b'|b,a) + \Pr(b''|b,a) = 1$

Belief MDPs

- Now we have a state space and transition function
- We can define a reward function using *expected rewards on actual states*:

$$R(b, a, b') = \sum_{s} b(s) \sum_{s'} \Pr(s'|s, a)\, R(s, a, s')$$

Belief MDPs

- Now we have a state space and transition function
- We can define a reward function using *expected rewards on actual states*:

$$R(b, a, b') = \sum_{s} b(s) \sum_{s'} \Pr(s'|s, a) R(s, a, s')$$

- Original rewards are weighted by a) belief probabilities of starting in states s, and b) transition probabilities of ending up in successor states s'
- As before, we sum over all starting states s and successor states s'

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: “Intended direction” prob 0.8

A	B	C

- “Observable” rewards: $R(s, a, A) = +3, R(s, a, B) = -2, R(s, a, C) = +1$
- $b = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$; agent takes action L

Example: Mini-Gridworld

- States: A, B, C; actions: L, R
- Transition model: "Intended direction" prob 0.8

A	B	C

- "Observable" rewards: $R(s,a,A) = +3, R(s,a,B) = -2, R(s,a,C) = +1$
- $b = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$; agent takes action L

$$R(b, L, b') = \left(\frac{1}{3} \times 0.8 + \frac{1}{3} \times 0.8 + \frac{1}{3} \times 0\right) \times 3 + \left(\frac{1}{3} \times 0.2 + \frac{1}{3} \times 0 + \frac{1}{3} \times 0.8\right) \times -2$$

$$+ \left(\frac{1}{3} \times 0 + \frac{1}{3} \times 0.2 + \frac{1}{3} \times 0.2\right) \times 1 = 1.067$$

Policies and Value Functions

- A policy $\pi: S \rightarrow A$ for a belief MDP maps from belief states to actions
- Value function for π can be written as

$$V^{\pi}(b) = \sum_{b\prime} T(b, \pi(b), b')[R(b, \pi(b), b') + \gamma V^{\pi}(b')]$$

Policies and Value Functions

- A policy $\pi: S \rightarrow A$ for a belief MDP maps from belief states to actions
- Value function for π can be written as

$$V^{\pi}(b) = \sum_{b'} T(b, \pi(b), b')[R(b, \pi(b), b') + \gamma V^{\pi}(b')]$$

- Bellman equations for optimal value function and policy:

$$V^{*}(b) = \max_{a} \sum_{b'} T(b, a, b')[R(b, a, b') + \gamma V^{*}(b')]$$

$$\pi^{*}(b) = \operatorname*{argmax}_{a} \sum_{b'} T(b, a, b')[R(b, a, b') + \gamma V^{*}(b')]$$

Value Iteration

- Theoretically, value iteration still works for finding optimal values V^*

$$V_{i+1}(b) \leftarrow \max_a \sum_{b'} T(b,a,b')[R(b,a,b') + \gamma V_i(b')]$$

- In each iteration, take each belief state b and compute the following:

Value Iteration

- Theoretically, value iteration still works for finding optimal values V^*

$$V_{i+1}(b) \leftarrow \max_a \sum_{b\prime} T(b,a,b')[R(b,a,b') + \gamma V_i(b')]$$

- In each iteration, take each belief state b and compute the following:

- The set of successor states b'
- Transition probabilities to b' given b an a
- Expected rewards of the (b,a,b') transition

Value Iteration

- Theoretically, value iteration still works for finding optimal values V^*

$$V_{i+1}(b) \leftarrow \max_a \sum_{b'} T(b,a,b')[R(b,a,b') + \gamma V_i(b')]$$

- What's the problem? Potentially infinitely many belief states!

Value Iteration

- Theoretically, value iteration still works for finding optimal values V^*

$$V_{i+1}(b) \leftarrow \max_a \sum_{b'} T(b,a,b')[R(b,a,b') + \gamma V_i(b')]$$

- What's the problem? Potentially infinitely many belief states!

- Even when starting with a deterministic belief state, number of successors can grow exponentially over time

- Finding optimal policies for POMDPs is computationally intractable

Online Planning

- Instead of solving for optimal value functions and policies offline, consider planning *online* using *look-ahead search*

Online Planning

- Instead of solving for optimal value functions and policies offline, consider planning *online* using *look-ahead search*

- Idea: Make decisions while interacting with environment
- Look ahead a finite number of steps from current belief state

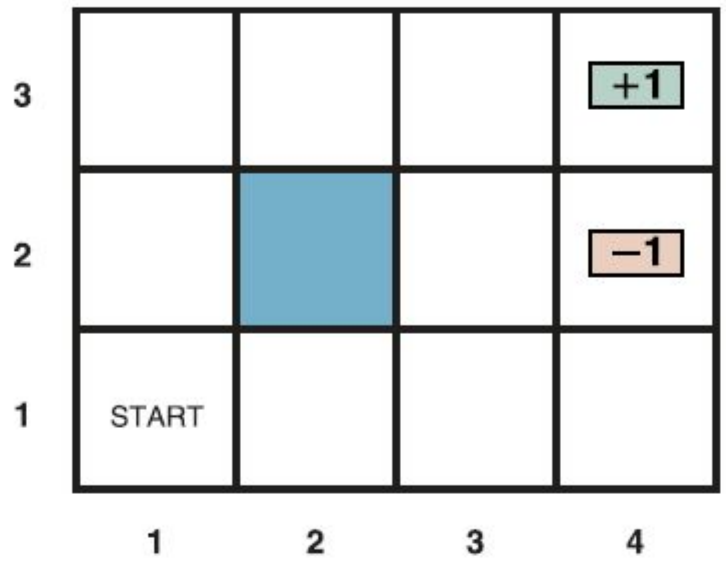
Online Planning

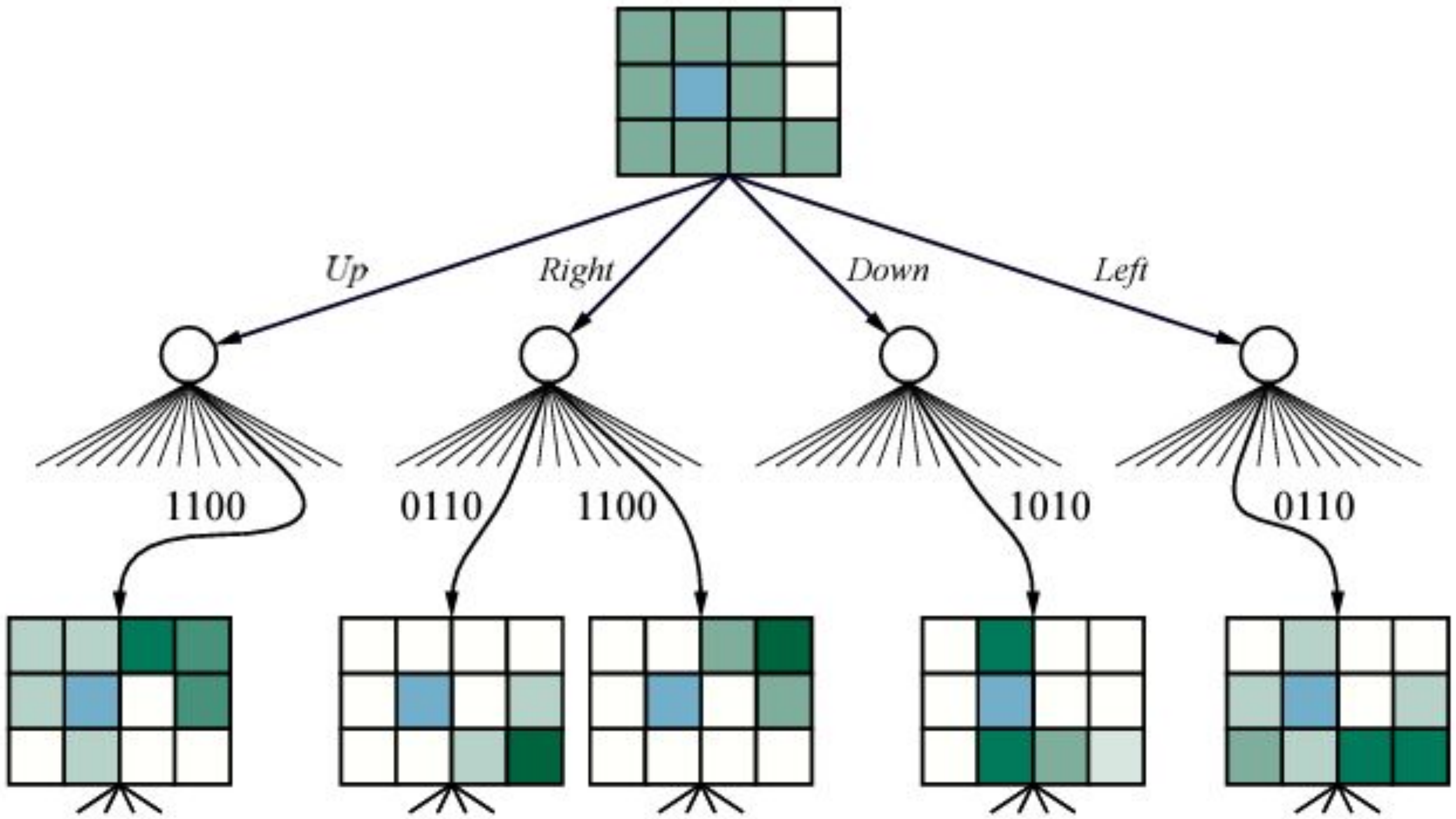
- Instead of solving for optimal value functions and policies offline, consider planning *online* using *look-ahead search*

- Idea: Make decisions while interacting with environment
- Look ahead a finite number of steps from current belief state

- Only consider locally reachable successor states
- Solve a finite-horizon MDP at any given time
- Use estimates of state utilities instead of exact values

POMDP Backup Diagram





Complexity of Exhaustive Search

- Suppose we search to a depth d in backup tree diagram from current state
- In each layer, consider $|A|$ actions and $|Z|$ percepts for each action

Complexity of Exhaustive Search

- Suppose we search to a depth d in backup tree diagram from current state
- In each layer, consider $|A|$ actions and $|Z|$ percepts for each action

- One step lookahead: $O(|A||Z|)$ time to compute all successors
- d-step lookahead: $O(|A|^d|Z|^d)$ time to compute all successors

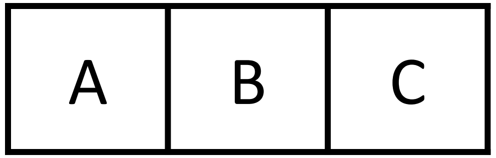
Complexity of Exhaustive Search

- Suppose we search to a depth d in backup tree diagram from current state
- In each layer, consider $|A|$ actions and $|Z|$ percepts for each action

- One step lookahead: $O(|A||Z|)$ time to compute all successors
- d-step lookahead: $O(|A|^d|Z|^d)$ time to compute all successors

- In practice, we can get away with small d if discount factor is also small
- Sampling methods can be used to cut down on branching factor

Example: Mini-Gridworld



- States: A, B, C; actions: L, R
- Transition model: “Intended direction” prob 0.8
- Current belief state: $b = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- Look ahead one step:
  - Action L, observation 1: $b_1' = (0.8,0,0.2)$ with prob $1/3$
  - Action L, observation 0: $b_2' = (0.4,0.5,0.1)$ with prob $2/3$

Example: Mini-Gridworld

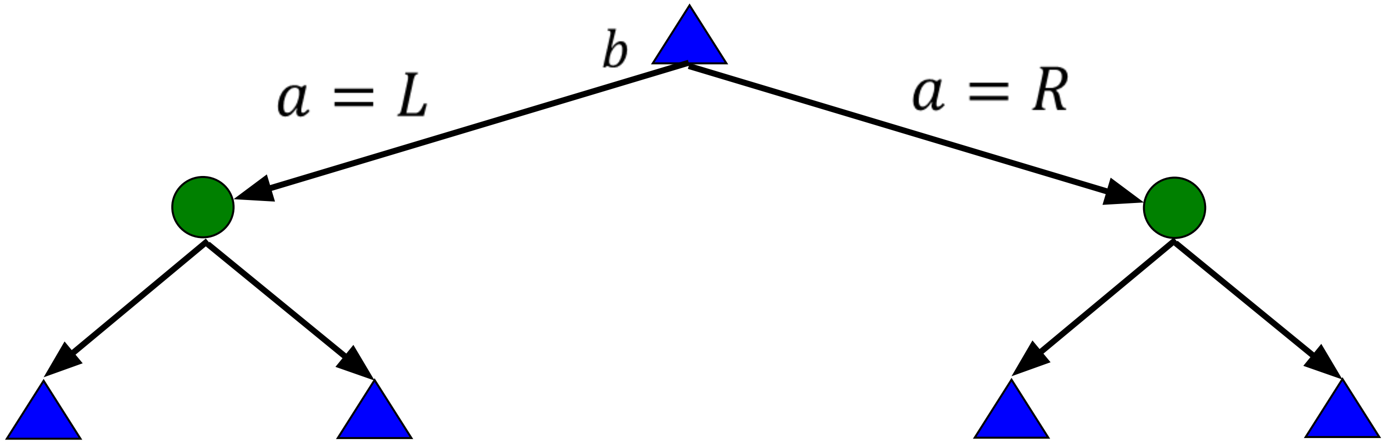
- States: A, B, C; actions: L, R
- Transition model: "Intended direction" prob 0.8
- Current belief state: $b = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$

A	B	C

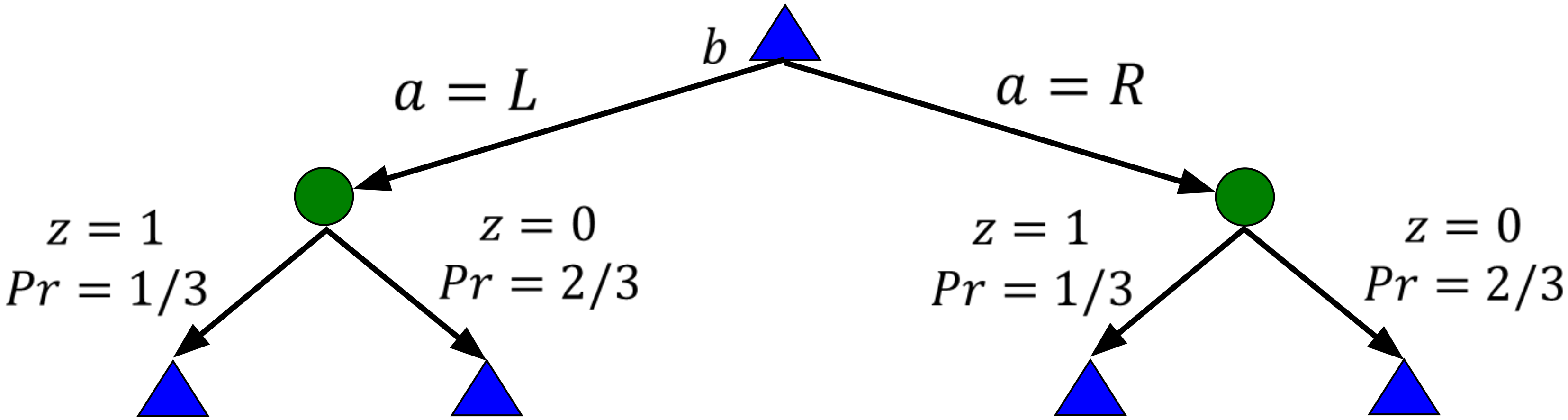
- Look ahead one step:
  - Action L, observation 1: $b_1' = (0.8, 0, 0.2)$ with prob $1/3$
  - Action L, observation 0: $b_2' = (0.4, 0.5, 0.1)$ with prob $2/3$

  - Action R, observation 1: $b_3' = (0.2, 0, 0.8)$ with prob $1/3$
  - Action R, observation 0: $b_4' = (0.1, 0.5, 0.4)$ with prob $2/3$

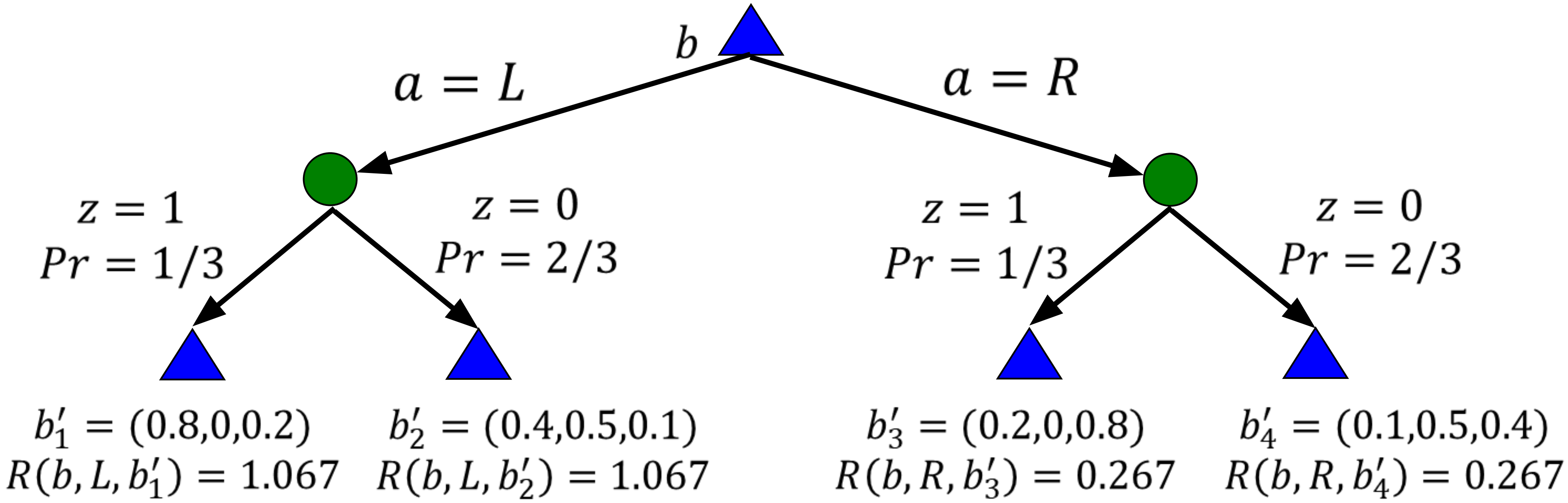
Example: Mini-Gridworld



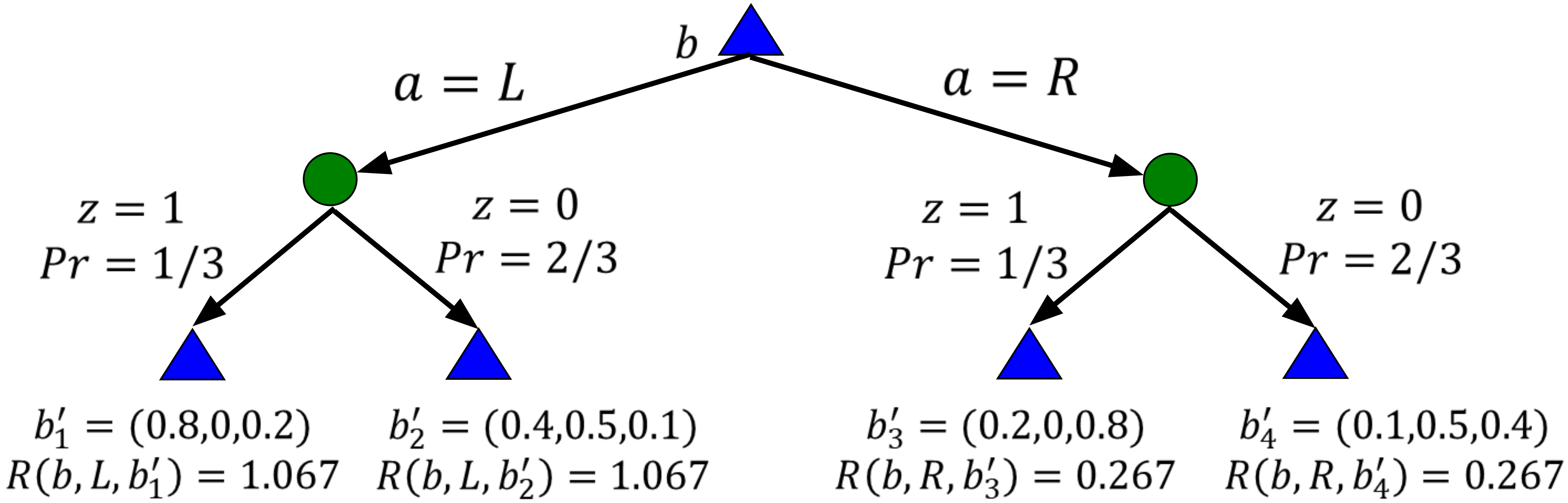
Example: Mini-Gridworld



Example: Mini-Gridworld

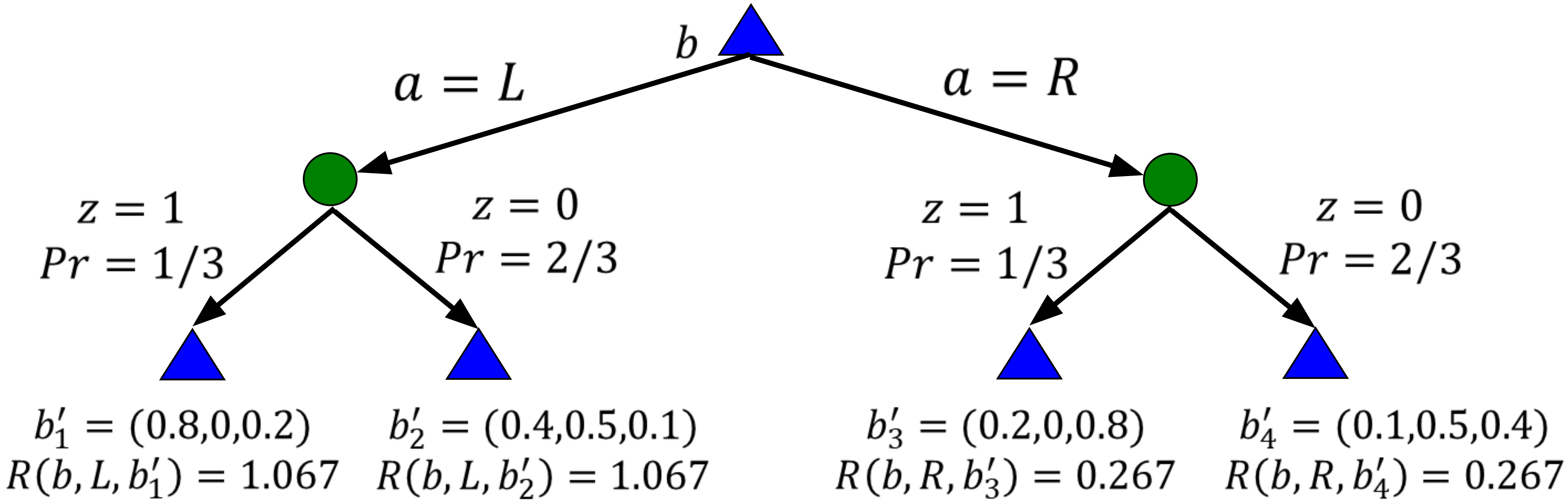


Example: Mini-Gridworld



- If we wanted to look ahead further, we can continue building up the tree
- Once we decide to stop, we can compute or estimate values $V(b)$
- Decision at the root (current belief state) is the action that gives the higher value

Example: Mini-Gridworld



$$V(b) = \max_{L,R} \begin{cases} \frac{1}{3}\left(1.067 + \gamma V(b'_1)\right) + \frac{2}{3}\left(1.067 + \gamma V(b'_2)\right) \\ \frac{1}{3}\left(0.267 + \gamma V(b'_3)\right) + \frac{2}{3}\left(0.267 + \gamma V(b'_4)\right) \end{cases}$$

Finer Points

- Many tricks for improving performance of lookahead search

Finer Points

- Many tricks for improving performance of lookahead search
- *Reuse information*
  - Subtrees can be carried over from one step to the next as agent progresses
  - Value approximations may be computed offline (e.g., exact MDP)

Finer Points

- Many tricks for improving performance of lookahead search

- *Reuse information*
  - Subtrees can be carried over from one step to the next as agent progresses
  - Value approximations may be computed offline (e.g., exact MDP)

- *Search wisely*
  - Tree branches can be pruned based on lower/upper bound information
  - Ordering of expanded nodes can be informed by heuristics
  - Instead of computing evidence probabilities, estimate them by sampling

Summary

- Many real problems exhibit partial observability
- A POMDP incorporates observations and a sensor model

- POMDPs can be represented as belief MDPs
- States are belief states; transitions and rewards all functions of beliefs

- Exact solutions of belief MDPs cannot be found efficiently
- Online planning can provide approximate solutions and decision making