## 15-887

## Planning, Execution and Learning

Planning under Uncertainty: Partially Observable Markov Decision Processes (POMDP)

Maxim Likhachev<br>Robotics Institute Carnegie Mellon University

## Graph vs. MDP vs. POMDP

- Consider a path planning example



## Graph vs. MDP vs. POMDP

- Consider a path planning example
- Assume perfect action execution and full knowledge of the state (i.e., perfect localization)



## Graph vs. MDP vs. POMDP

- Consider a path planning example
- Assume perfect action execution and full knowledge of the state (i.e., perfect localization)



## Graph:

Implicitly defined as $\{S, A, C\}$, where $S$ - set of states, $A$ - set of actions, $C$ - costs of all $(s, a)$ pairs.

## Graph vs. MDP vs. POMDP

- Consider a path planning example
- Assume perfect action execution and full knowledge of the state (i.e., perfect localization)



## Graph:

Implicitly defined as $\{S, A, C\}$,
where $S$ - set of states, $A$ - set of actions, $C$ - costs of all $(s, a)$ pairs.

## Graph vs. MDP vs. POMDP

- Consider a path planning example
- Assume imperfect action execution and full knowledge of the state (i.e., perfect localization)


Let's assume $50 \%$ chance of ending up on the left and $50 \%$ ending up on the right

## Graph vs. MDP vs. POMDP

- Consider a path planning example
- Assume imperfect action execution and full knowledge of the state (i.e., perfect localization)



## MDP:

Defined as $\{S, A, T, C\}$, where $S$ - set of states, $A$ - set of actions, $\boldsymbol{T}\left(\mathbf{s}, \boldsymbol{a}, \mathbf{s}^{\prime}\right)-\operatorname{Prob}\left(\mathbf{s}^{\prime} \mid \mathbf{s}, \boldsymbol{a}\right), C$ - costs of all ( $(,, a)$ pairs

## Graph vs. MDP vs. POMDP

- Consider a path $\mathrm{p}^{1} \quad$ What is an optimal policy here?
- Assume imperfect action execution and full knowledge of the state (i.e., perfect localization)



## MDP:

Defined as $\{S, A, T, C\}$, where $S$ - set of states, $A$ - set of actions, $T\left(s, a, s^{\prime}\right)-\operatorname{Prob}\left(s^{\prime} \mid s, a\right), C-$ costs of all ( $\left.s, a\right)$ pairs

## Graph vs. MDP vs. POMDP

- Consider a path $\mathrm{p}^{1} \quad$ What is an optimal policy here?
- Assume imperfect action execution and full knowledge of the state (i.e., perfect localization)



## MDP:

Defined as $\{S, A, T, C\}$, where $S$ - set of states, $A$ - set of actions, $T\left(s, a, s^{\prime}\right)-\operatorname{Prob}\left(s^{\prime} \mid s, a\right), C-$ costs of all ( $\left.s, a\right)$ pairs

## Graph vs. MDP vs. POMDP

- Consider a path planning example
- Assume imperfect action execution and full knowledge of the state (i.e., perfect localization)



## MDP (rewards version):

Defined as $\{S, A, T, R\}$, where $S$ - set of states, $A$ - set of actions, $T\left(s, a, s^{\prime}\right)-\operatorname{Prob}\left(s^{\prime} \mid s, a\right), R$ - rewards for all $(s, a)$ pairs

## Graph vs. MDP vs. POMDP

- Consider a path planning example
- Assume imperfect action execution and partial observability of the state (i.e., imperfect localization)


Let's assume UAV initially knows it is at $S_{0}$ During execution: it can only sense adjacent obstacles and being at goal

After taking this action, UAV doesn't know whether it is at state $S_{1}$ or $S_{2}$

## POMDP:

## Graph vs. MDP vs. POMDP

- Consider a path p . What is an optimal policy here?
- Assume imperfect action execution and partial observability of the state (i.e., imperfect localization)



## Graph vs. MDP vs. POMDP

- Consider a path planning example
- Assume imperfect action execution and partial observability of the state (i.e., imperfect localization)


Let's assume UAV initially knows it is at $S_{0}$ During execution: it can only sense adjacent obstacles and being at goal

After taking this action, UAV doesn't know whether it is at state $S_{1}$ or $S_{2}$

POMDP: $\{S, A, T, R, \Omega, O\}$, where $S, A, T\left(s, a, s{ }^{\prime}\right), R(s, a)$ - all as in $M D P, \Omega$ - set of all possible observation vectors o, $\boldsymbol{O}\left(\boldsymbol{s}^{\prime}, \boldsymbol{a}, \boldsymbol{o}\right)-\boldsymbol{P r o b}\left(\boldsymbol{o} \mid \mathbf{s}^{\prime}, \boldsymbol{a}\right)$ probability of seeing o after executing action a and ending up at state $s$ '

## Graph vs. MDP vs. POMDP

- Consider a path planning example
- Assume imperfect action execution and partial observability of the state (i.e., imperfect localization)


Causal relationship


POMDP: $\{S, A, T, R, \Omega, O\}$, where $S, A, T(s, a, s), R(s, a)-$ all as in $M D P, \Omega$ - set of all possible observation vectors o, $\boldsymbol{O}\left(\boldsymbol{s}^{\prime}, \boldsymbol{a}, \boldsymbol{o}\right)-\boldsymbol{P r o b}\left(\boldsymbol{o} \mid \mathbf{s}^{\prime}, \boldsymbol{a}\right)$ probability of seeing o after executing action a and ending up at state s'

## Graph vs. MDP vs. POMDP

Example of POMDP problems
where the robot knows its own pose perfectly
(perfect localization)?

- Assume imperfect action execution and partial observability of the state (i.e., imperfect localization)


Causal relationship


POMDP: $\{S, A, T, R, \Omega, O\}$, where $S, A, T\left(s, a, s^{\prime}\right), R(s, a)-$ all as in MDP, $\Omega$ - set of all possible observation vectors o, $\boldsymbol{O}\left(\boldsymbol{s}^{\prime}, \boldsymbol{a}, \boldsymbol{o}\right)-\boldsymbol{P r o b}\left(\boldsymbol{o} \mid \mathbf{s}^{\prime}, \boldsymbol{a}\right)$ probability of seeing o after executing action a and ending up at state $s$ '

## Belief State Space

- Belief state $\boldsymbol{b}$ : Probability distribution over the states the robot believes it is currently in


Causal relationship


POMDP: $\{S, A, T, R, \Omega, O\}$, where $T\left(s, a, s^{\prime}\right)=P\left(s^{\prime} \mid s, a\right), R(s, a), O\left(s^{\prime}, a, o\right)=\operatorname{Prob}\left(o \mid s^{\prime}, a\right)$

## Belief State Space

- Belief state $\boldsymbol{b}$ : Probability distribution over the states the robot believes it is currently in

$$
\begin{aligned}
& b-a \text { vector of size } N(\# \text { of states in } S \text { ) } \\
& \Sigma^{N} b_{i}=1 \text {, and } b_{i} \geq 0 \text { for all } i
\end{aligned}
$$

Initially, the robot knows it is in $s_{0}$.
Thus, initial $b=[1,0,0,0,0,0,0,0]^{\top}$. That is, $P\left(s_{0}\right)=1$


Causal relationship


POMDP: $\{S, A, T, R, \Omega, O\}$, where $T\left(s, a, s^{\prime}\right)=P\left(s^{\prime} \mid s, a\right), R(s, a), O\left(s^{\prime}, a, o\right)=\operatorname{Prob}\left(o \mid s^{\prime}, a\right)$

## Belief State Space

- Belief state $\boldsymbol{b}$ : Probability distribution over the states the robot believes it is currently in

$$
\begin{aligned}
& b-a \text { vector of size } N \text { (\# of states in } S \text { ) } \\
& \qquad \Sigma^{N} b_{i}=1 \text {, and } b_{i} \geq 0 \text { for all } i
\end{aligned}
$$

Initially, the robot knows it is in $s_{0}$.
Thus, initial $b=[1,0,0,0,0,0,0,0]^{\top}$. That is, $P\left(s_{0}\right)=1$


POMDP: $\{S, A, T, R, \Omega, O\}$, where $T\left(s, a, s^{\prime}\right)=P\left(s^{\prime} \mid s, a\right), R(s, a), O\left(s^{\prime}, a, o\right)=\operatorname{Prob}\left(o \mid s^{\prime}, a\right)$

## Belief State Space

- Belief state $\boldsymbol{b}$ : Probability distribution over the states the robot believes it is currently in

Belief State Space
(for K actions, M possible observations)


POMDP: $\{S, A, T, R, \Omega, O\}$, where $T\left(s, a, s^{\prime}\right)=P\left(s^{\prime} \mid s, a\right), R(s, a), O\left(s^{\prime}, a, o\right)=\operatorname{Prob}\left(o \mid s^{\prime}, a\right)$

## Belief State Space

- Belief state $\boldsymbol{b}$ : Probability distribution over the states the robot believes it is currently in

$$
\begin{aligned}
& b^{\prime}: P\left(s^{\prime} \mid b, a, o\right) \text { for every } s^{\prime} \text { in } S \text {; } \\
& b^{\prime}\left(s^{\prime}\right)=P\left(s^{\prime} \mid b, a, o\right)=\frac{\left.o\left(s^{\prime}, a, o\right) \sum_{s} s T\left(s, a, s^{\prime}\right) * b(s)\right\}}{P(o \mid b, a)}
\end{aligned}
$$

Here how outcome beliefs
are computed

## Belief State Space

(for K actions, M possible observations)


POMDP:
$\{S, A, T, R, \Omega, O\}$, where $T\left(s, a, s^{\prime}\right)=P\left(s^{\prime} \mid s, a\right), R(s, a), O\left(s^{\prime}, a, o\right)=\operatorname{Prob}\left(o \mid s^{\prime}, a\right)$

## Belief State Space

- Belief state $\boldsymbol{b}$ : Probability distribution over the states the robot believes it is currently in
$b^{\prime}: P\left(s^{\prime} \mid b, a, o\right)$ for every s'in $S$;
$b^{\prime}\left(s^{\prime}\right)=P\left(s^{\prime} \mid b, a, o\right)=\frac{O\left(s^{\prime}, a, o\right) \sum_{s}\left\{T\left(s, a, s^{\prime}\right) * b(s)\right\}}{P(o \mid b, a)}$


Derivation:

$$
P\left(s^{\prime} \mid b, a, o\right)=\frac{P\left(o \mid b, a, s^{\prime}\right) P\left(s^{\prime} \mid b, a\right)}{P(o \mid b, a)}=\frac{P\left(o \mid s^{\prime}, a\right) \sum_{s}\left\{P\left(s^{\prime} \mid s, a\right) * P(s)\right\}}{P(o \mid b, a)}
$$

ations)


POMDP: $\{S, A, T, R, \Omega, O\}$, where $T\left(s, a, s^{\prime}\right)=P\left(s^{\prime} \mid s, a\right), R(s, a), O\left(s^{\prime}, a, o\right)=\operatorname{Prob}\left(o \mid s^{\prime}, a\right)$

## Belief State Space

- Belief state $\boldsymbol{b}$ : Probability distribution over the states the robot believes it is currently in


## What is Belief State Space?

It is MDP!
We just need to compute transition probabilities $\tau\left(b, a, b^{\prime}\right)=P\left(b^{\prime} \mid b, a\right)$ and reward function $\rho(b, a)$

Belief State Space
(for K actions, M possible observations)


POMDP:
$\{S, A, T, R, \Omega, O\}$, where $T\left(s, a, s^{\prime}\right)=P\left(s^{\prime} \mid s, a\right), R(s, a), O\left(s^{\prime}, a, o\right)=\operatorname{Prob}\left(o \mid s^{\prime}, a\right)$

## Belief State Space

- Belief state $\boldsymbol{b}$ : Probability distribution over the states the robot believes it is currently in

$$
\tau\left(b, a, b^{\prime}\right)=P\left(b^{\prime} \mid b, a\right)=\sum_{o \text { leading to } b^{\prime}} P(o \mid b, a)=\sum_{o \text { leading to } b^{\prime}} \sum_{s^{\prime}} P\left(o \mid s^{\prime}, a\right) \sum_{s} P\left(s^{\prime} \mid s, a\right) b(s)
$$

Belief State Space (for K actions, M possible observations)


## Belief State Space

- Belief state $\boldsymbol{b}$ : Probability distribution over the states the robot believes it is currently in

$$
\tau\left(b, a, b^{\prime}\right)=P\left(b^{\prime} \mid b, a\right)=\sum_{o \text { leading to } b^{\prime}} P(o \mid b, a)=\sum_{o \text { leading to } b^{\prime}} \sum_{s^{\prime}} P\left(o \mid s^{\prime}, a\right) \sum_{s} P\left(s^{\prime} \mid s, a\right) b(s)
$$

$$
\rho(b, a)=\sum_{s} R(s, a) b(s)
$$

## Belief State Space

 (for K actions, M possible observations)

## Belief State Space

- Belief state $\boldsymbol{b}$ : Probability distribution over the states the robot believes it is currently in

$$
\tau\left(b, a, b^{\prime}\right)=P\left(b^{\prime} \mid b, a\right)=\sum_{o \text { leading to } b^{\prime}} P(o \mid b, a)=\sum_{o \text { leading to } b^{\prime}} \sum_{s^{\prime}} P\left(o \mid s^{\prime}, a\right) \sum_{s} P\left(s^{\prime} \mid s, a\right) b(s)
$$

$$
\rho(b, a)=\sum_{s} R(s, a) b(s)
$$

## Belief State Space

 (for K actions, M possible observations)

The size of Belief MDP is infinite :


So, finding an optimal policy for POMDP = finding an optimal policy for Belief MDP ©)

## Belief State Space

- Belief state $\boldsymbol{b}$ : Probability distribution over the states the robot believes it is currently in
- Popular techniques for solving POMDPs
- by discretizing belief statespace into a finite \# of states [Lovejoy, '91]
- by taking advantage of the geometric nature of value function [Kaelbing, Littman \& Cassandra, '98]
- by sampling-based approximations [Pineau, Gordon \& Thrun, '03]

Belief State Space (for K actions, M possible observations)


POMDP:
$\{S, A, T, R, \Omega, O\}$, where $T\left(s, a, s^{\prime}\right)=P\left(s^{\prime} \mid s, a\right), R(s, a), O\left(s^{\prime}, a, o\right)=\operatorname{Prob}\left(o \mid s^{\prime}, a\right)$

## Summary

- MDP generalizes Graph representation
- POMDP generalizes MDP representation
- POMDP - representation of problems where the state of relevant variables is NOT fully known
- Solving POMDP can be represented as solving a Belief MDP (whose size is infinite though)
- Approximation techniques exist but intractability is still a huge issue for using POMDP planning in real world

