NTNU | Norwegian University of Science and Technology

POMDPS

An overview

Tiago Veiga

March 24, 2022

Outline

- Introduction
- Offline
- Online planning
- Reinforcement Learning
- Conclusions

Introduction

- Planning: find a sequence of actions that transform the system state into one of goal states;

Introduction

- Planning: find a sequence of actions that transform the system state into one of goal states;
- POMDPs are a mathematical framework for modelling sequential decision making under uncertainty;

Introduction

- Planning: find a sequence of actions that transform the system state into one of goal states;
- POMDPs are a mathematical framework for modelling sequential decision making under uncertainty;
- Uncertainty:

Introduction

- Planning: find a sequence of actions that transform the system state into one of goal states;
- POMDPs are a mathematical framework for modelling sequential decision making under uncertainty;
- Uncertainty:
  - Actions: don’t know where the agent may end up;

Introduction

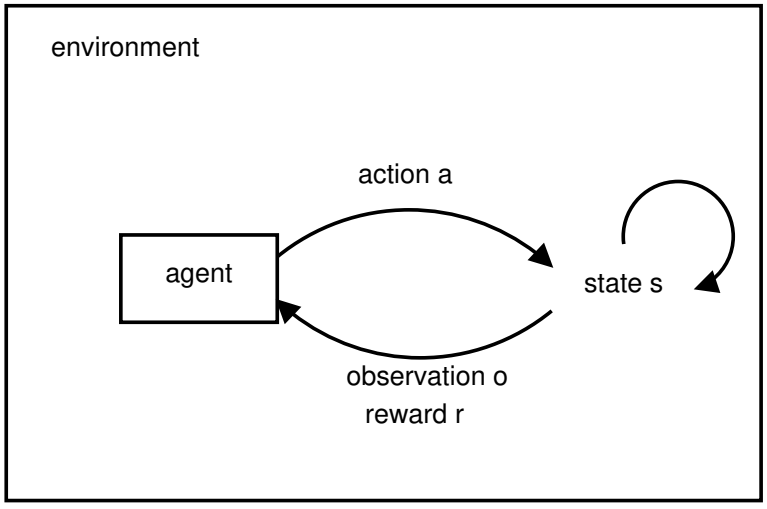
- Planning: find a sequence of actions that transform the system state into one of goal states;
- POMDPs are a mathematical framework for modelling sequential decision making under uncertainty;
- Uncertainty:
  - Actions: don’t know where the agent may end up;
  - Observations: no sure where the agent is at each moment.

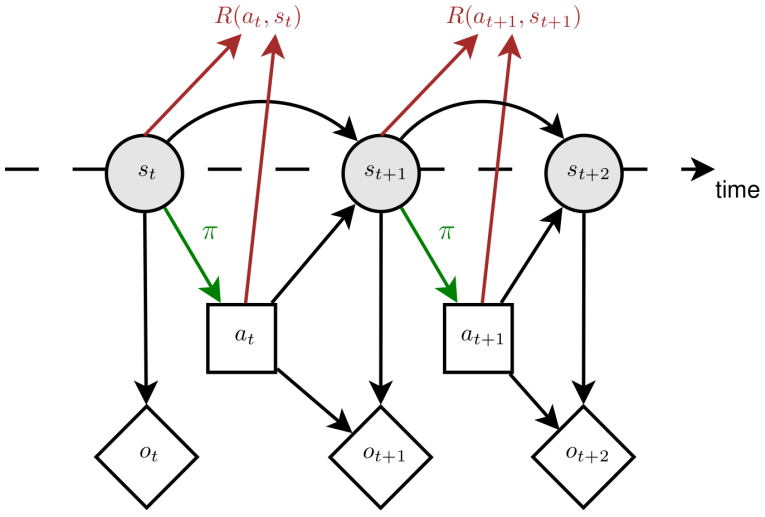
Markov model family

- Does this anything to do with other Markov models?
- In common: Markov property.

Markov Models		**Do we have control over the state transitons?**	
		NO	**YES**
Are the states completely observable?	**YES**	Markov Chain	MDP Markov Decision Process
	NO	HMM Hidden Markov Model	POMDP Partially Observable Markov Decision Process

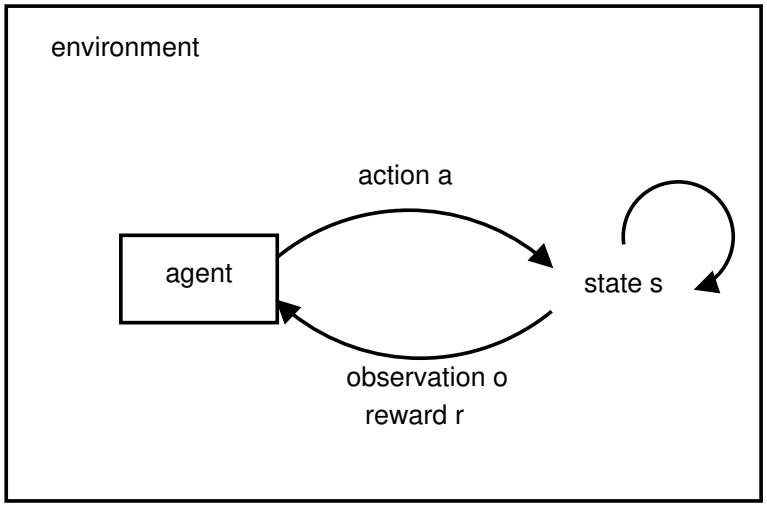
POMDPs

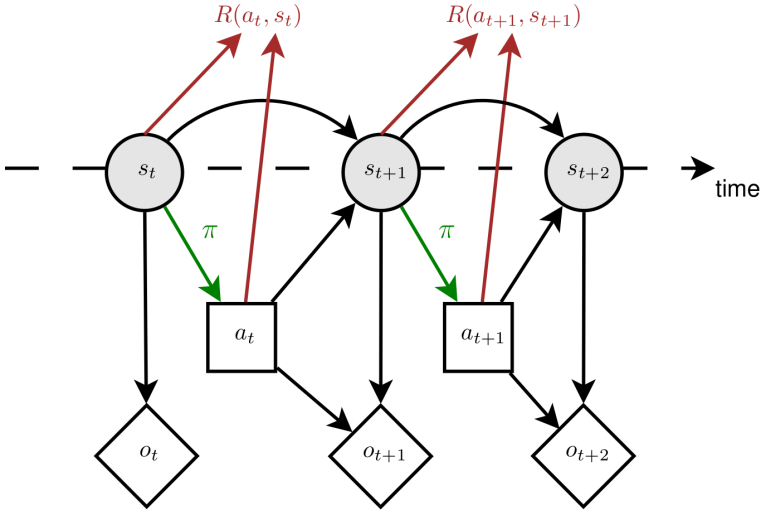




- POMDPs are defined by $(S, A, Z, T, O, R, \gamma)$:
  - Sets of states S, observations Z and actions A.
  - Transition model $T = p(s'|s, a)$: models effects of actions.
  - Observation model $O = p(o|s', a)$: relates observations to states.
  - Task defined by reward $R(s, a)$.
  - A discount rate $0 \leq \gamma < 1$.

POMDPs





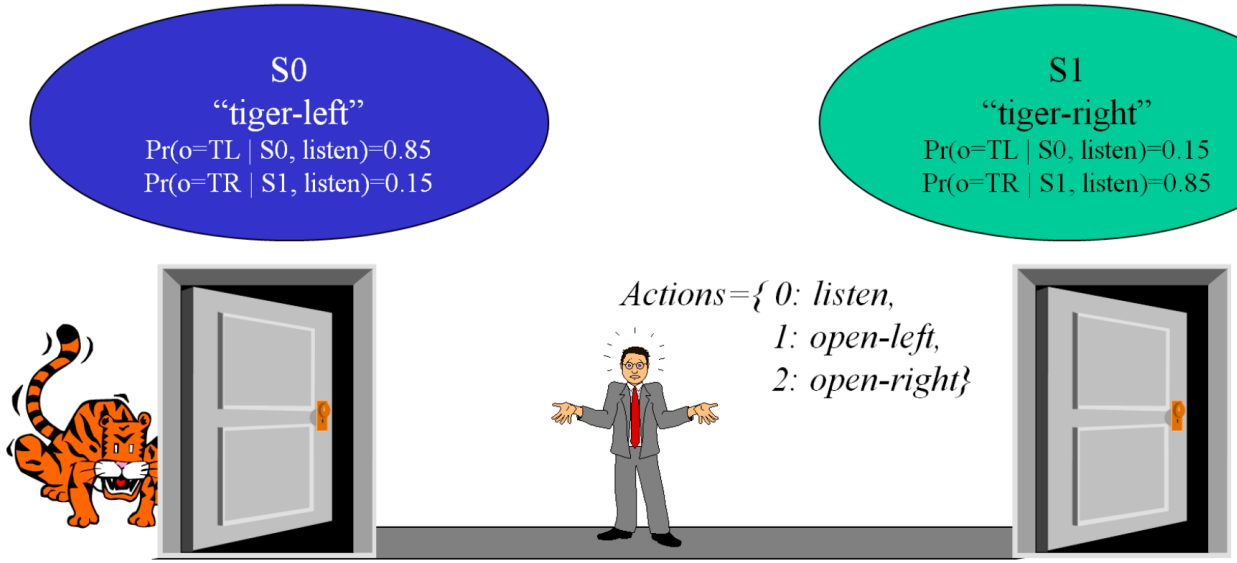
- Beliefs: probability distribution over states;
- Belief is updated using Bayes rule:

$$b_a^o(s') = p(s'|o, a, b) = \frac{p(o|s', a)}{p(o|b, a)} \sum_{s \in S} p(s'|s, a) b(s)$$

- It is a sufficient statistic of the history.

POMDPs

Tiger problem



Reward Function

- *Penalty for wrong opening: -100*
- *Reward for correct opening: +10*
- *Cost for listening action: -1*

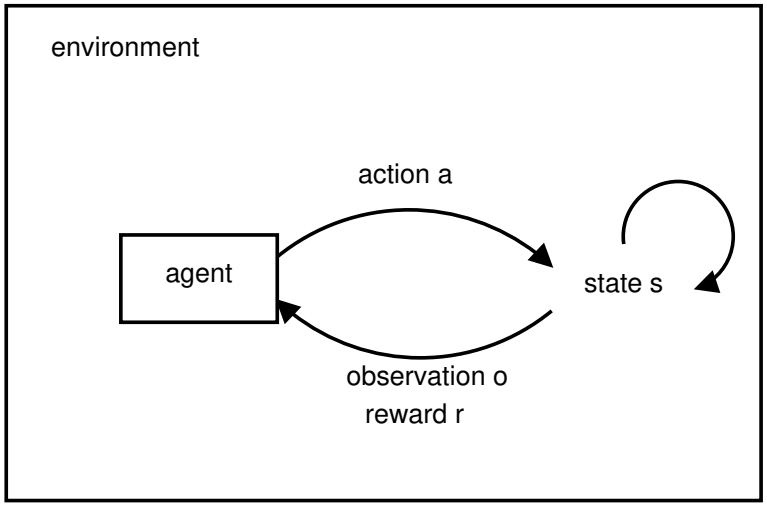
Observations

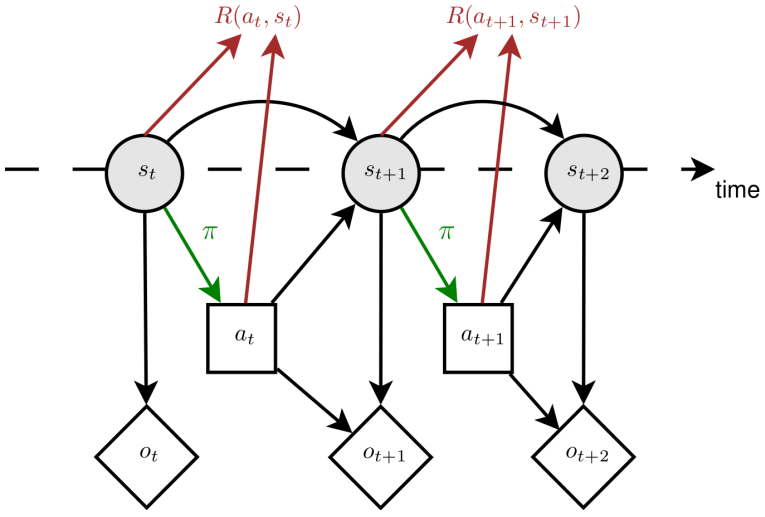
- *to hear the tiger on the left (TL)*
- *to hear the tiger on the right(TR)*

POMDPs

Tiger problem - Belief example

POMDPs





- Goal: compute policy that maximizes long term reward;

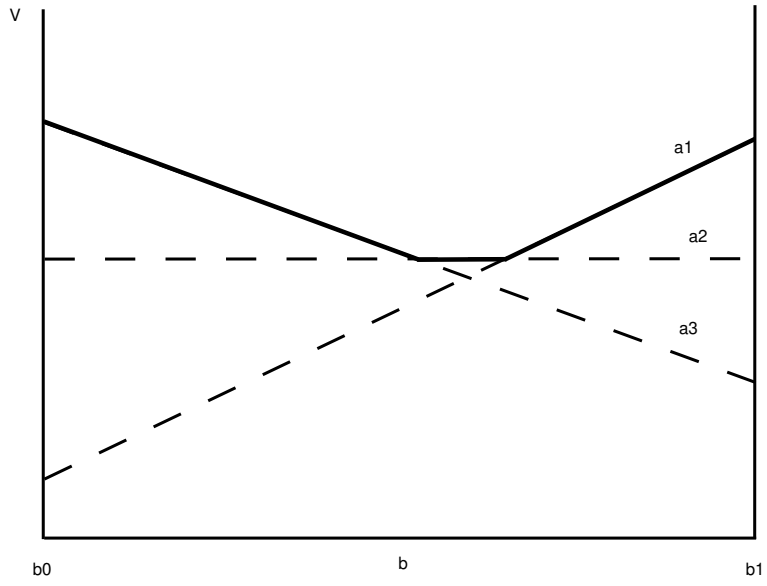
$$V^{\pi}(b) = E_{\pi}\left[\sum_{t=0}^{h} \gamma^t r(b_t, \pi(b_t)) \middle| b_0 = b\right]$$

$$\pi^*(b) = \underset{\pi}{\operatorname{argmax}}\, V^{\pi}(b)$$

- MDP: $\pi : S \rightarrow A$; POMDP: $\pi : B \rightarrow A$ or $\pi : h \rightarrow A$.

POMDPs

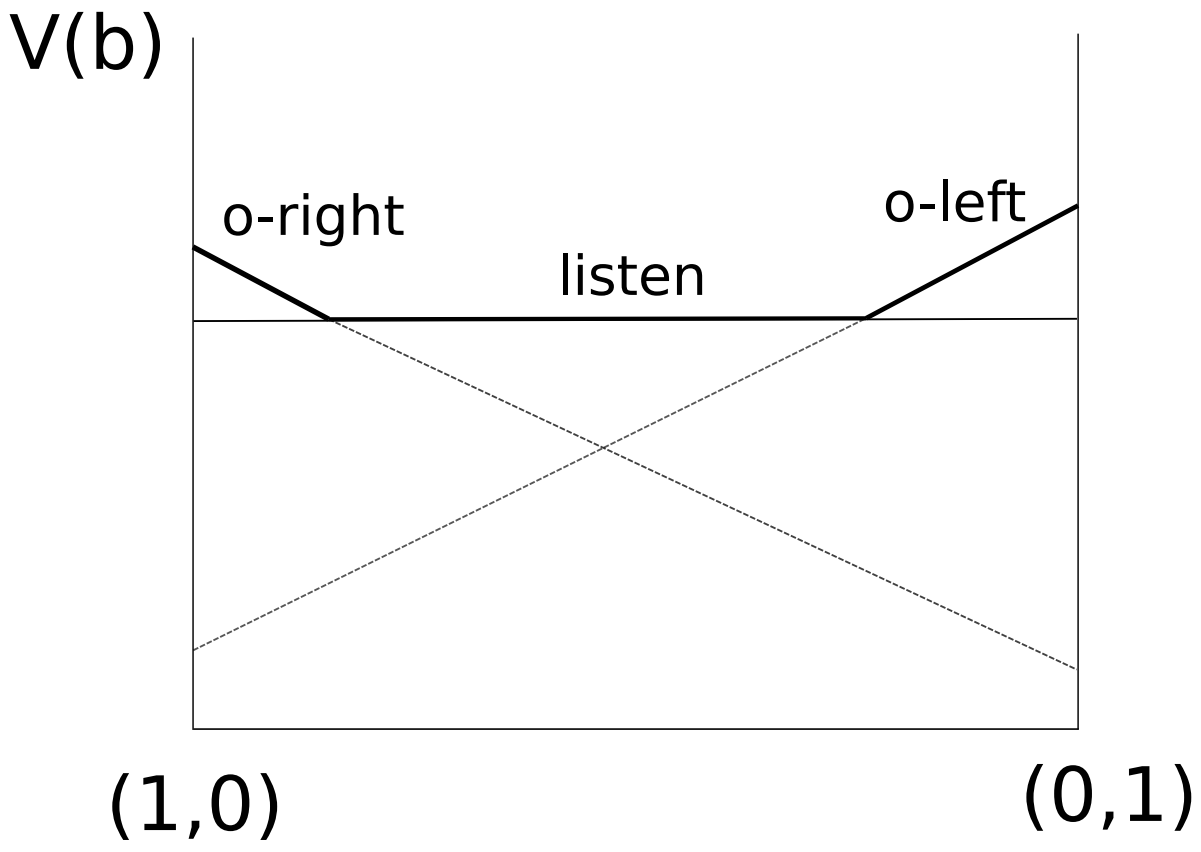
Value Function



- Value function is PWLC;
- Can be represented by a finite set of $|S|$-dimensional vectors;
- Value at a belief point is given by maximizing vector.

POMDPs

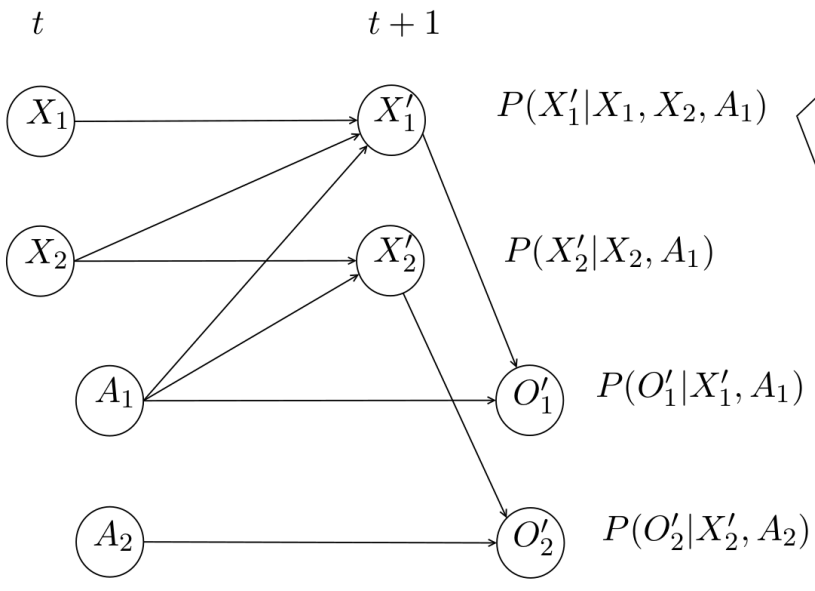
Value Function - Tiger



POMDPs

Factored models

- Cross-product of variables:
  - $S = X_1 \times \ldots \times X_n$
  - $O = Z_1 \times \ldots \times Z_n$
  - $A = A_1 \times \ldots \times A_n$



X_1	X_2	A_1	x_1'	$\bar{x}_1'$
x_1	x_2	a_1	0.4	0.6
x_1	x_2	$\bar{a}_1$	0.5	0.5
x_1	$\bar{x}_2$	a_1	0.4	0.6
x_1	$\bar{x}_2$	$\bar{a}_1$	0.3	0.7
$\bar{x}_1$	x_2	a_1	0.5	0.5
$\bar{x}_1$	x_2	$\bar{a}_1$	0.6	0.4
$\bar{x}_1$	$\bar{x}_2$	a_1	0.7	0.3
$\bar{x}_1$	$\bar{x}_2$	$\bar{a}_1$	0.4	0.6

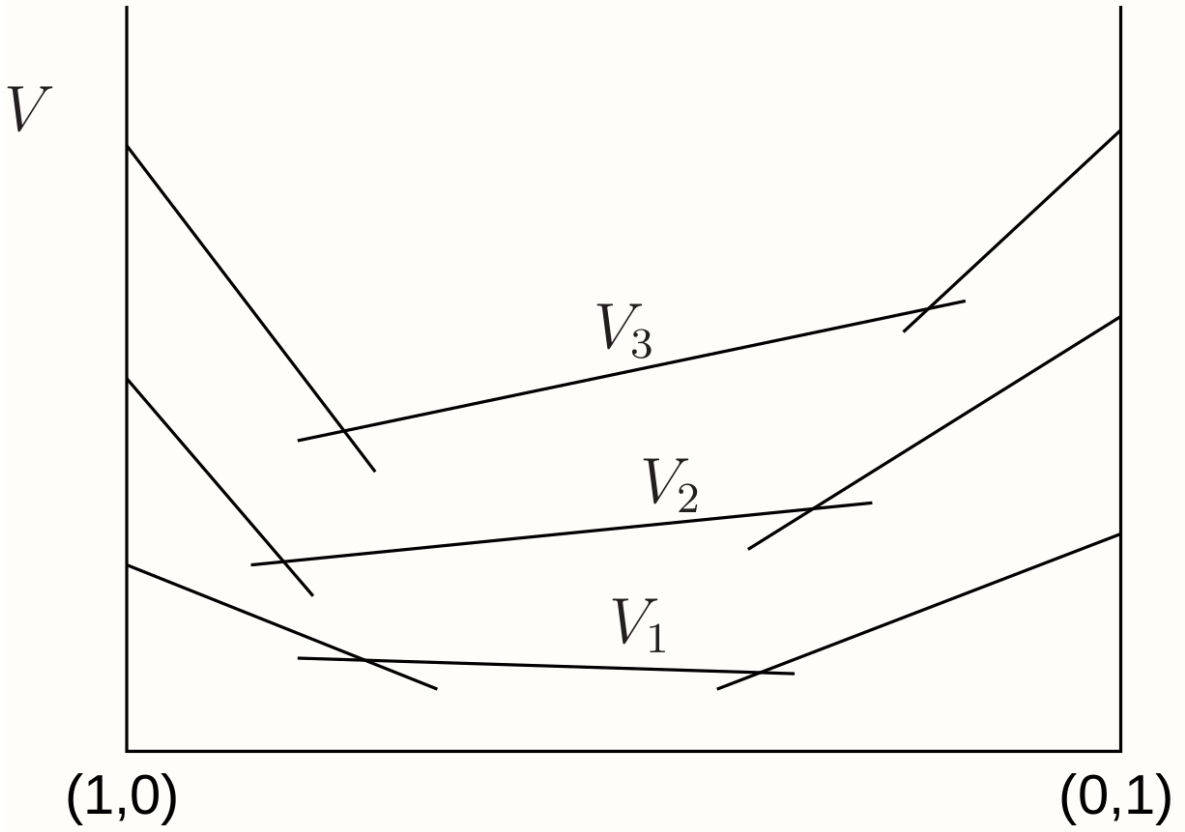
Challenges

- Curse of dimensionality:
  - Dimension of each vector is $|S|$;
- Curse of history:
  - Number of vectors grows exponentially with the planning horizon and observations.
- Non-stationary environments:
  - Model may change.

Offline

Exact solution

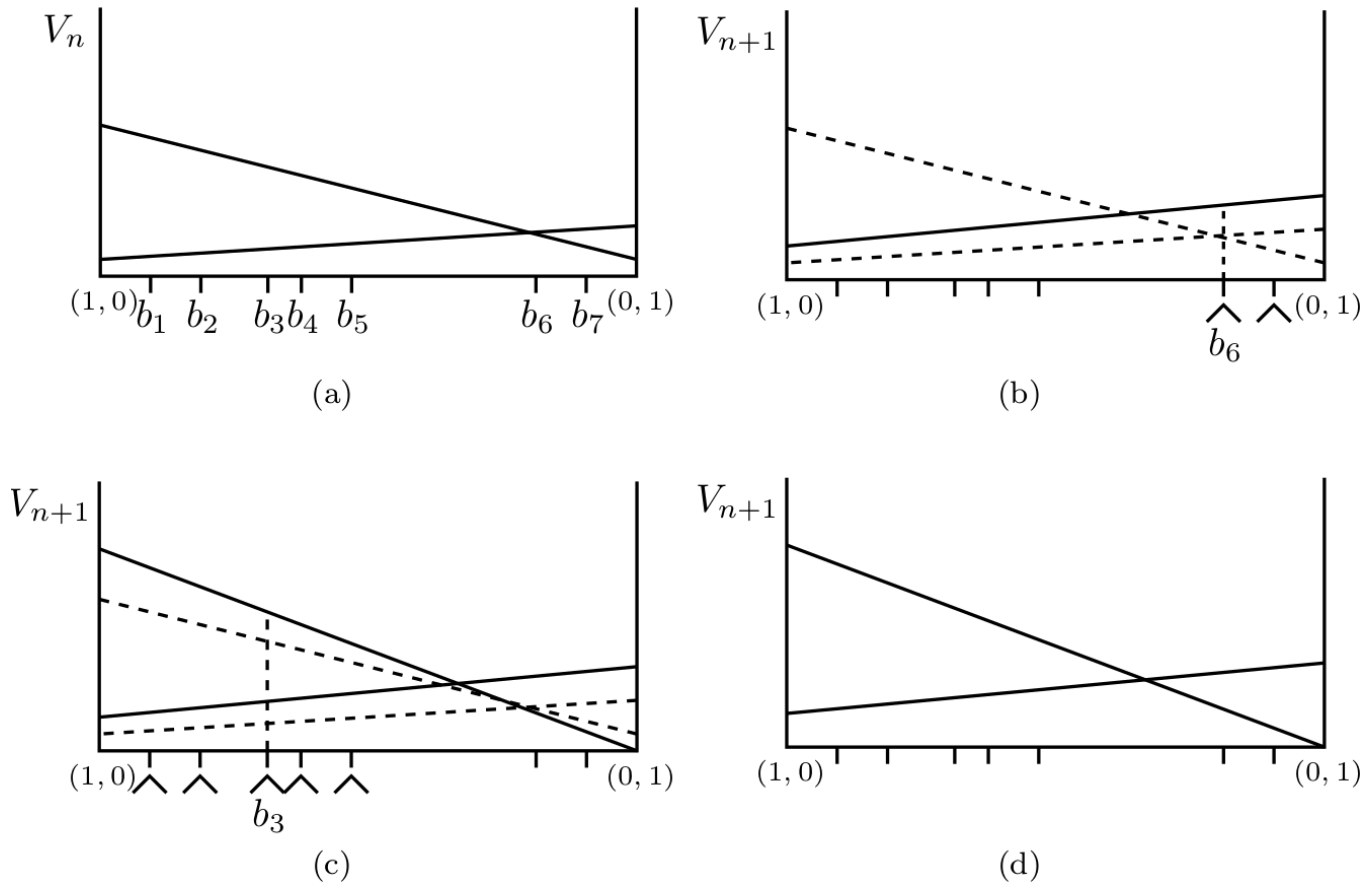
- Find all possible vectors, then prune useless ones;
- Iterate until convergence.



Offline

Point-based

- Find best vectors for a sampled subset of belief points.



Offline

Point-based

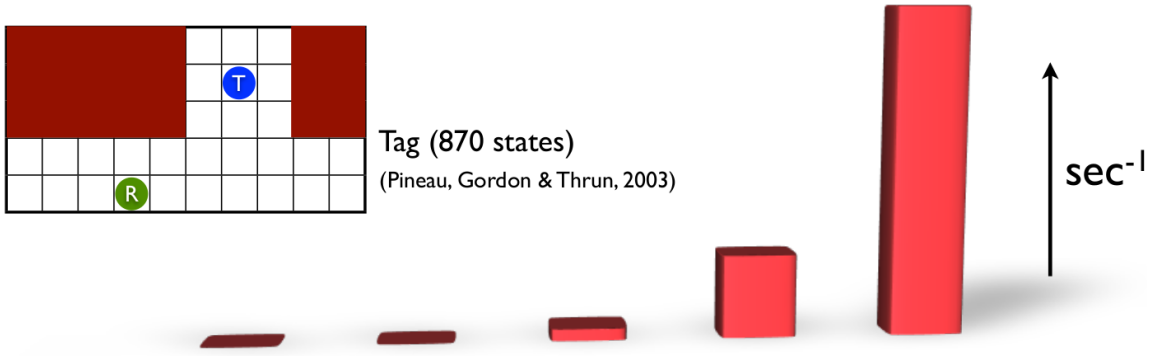
Algorithm	Collect	Update
PBVI	L_1-norm	full backup
Perseus	Random	asynchronous backup
HSVI	Bound uncertainty	newest points backup
GapMin	Bound uncertainty	full backup
PEMA	error minimization	full backup
FSVI	MDP heuristic	newest points backup

Outline of several point-based algorithms.

Offline

Point-based

- Sampling strategy matters:

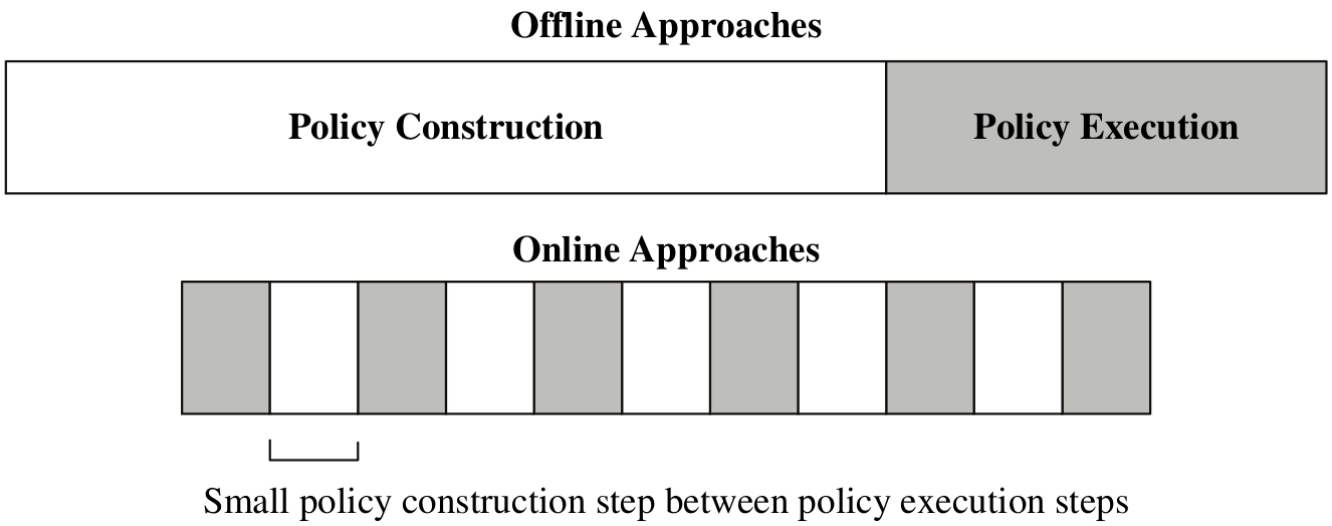


	PBVI 2003 Pineau, Gordon & Thrun	HSVI 2003 Smith & Simmons	Perseus 2005 Spaan & Vlassis	HSVI2 2005 Smith & Simmons	SARSOP 2008 Kurniawati, Hsu & Lee	....
Time (sec.)	180,880	10,113	1,670	24	6	....
Reward	-9.18	-6.17	-6.17	-6.36	-6.13	

- Good survey: *A survey of point-based POMDP solvers*, Shani et al, 2012

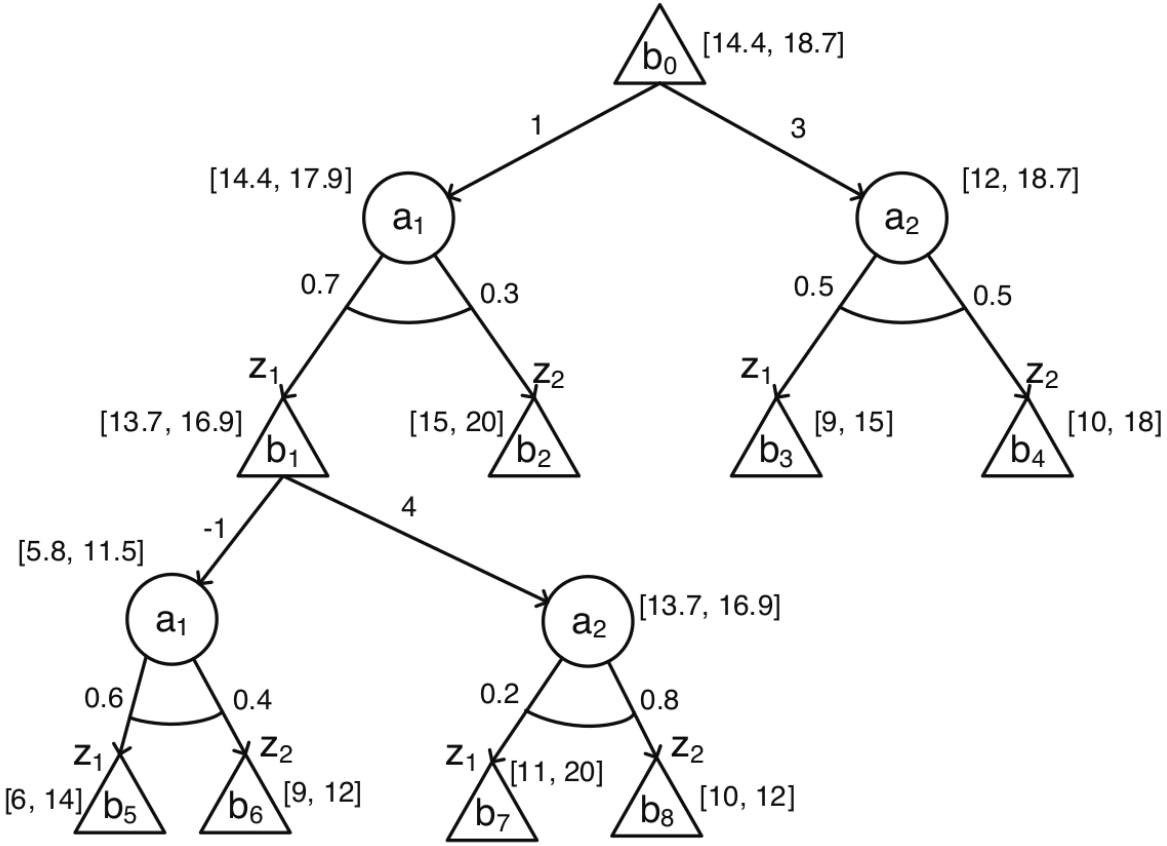
Online

- Offline: good global policy;
- Online: good local policy.



Online

- Forward search in belief state space.



Online

- Good survey: *Online Planning Algorithms for POMDPs*, Ross et al (2008);
- Most methods fit into three major categories:
  - Branch and bound pruning;
  - Monte Carlo sampling;
  - Heuristic search.

Online

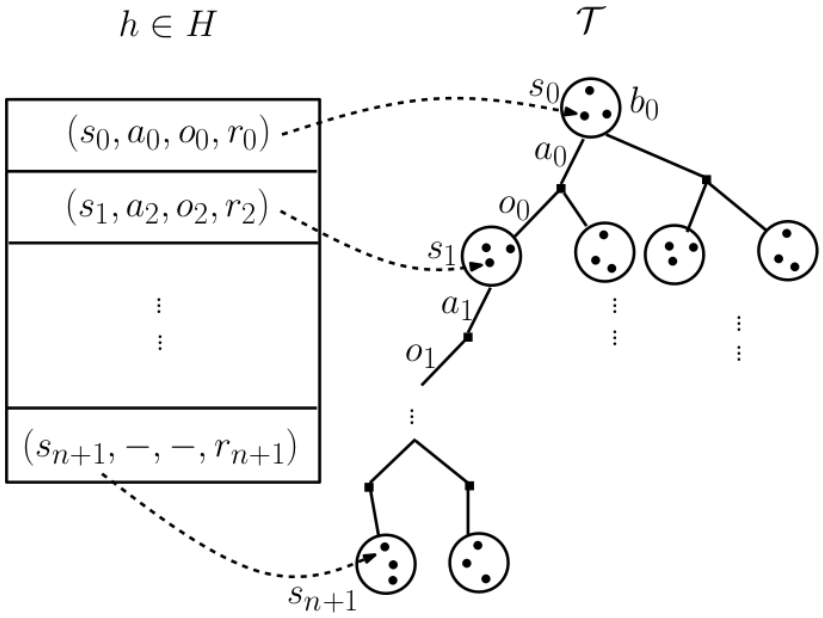
Branch and Bound Pruning

- Prune nodes that are suboptimal compared to others that have already been expanded;
- Maintain a lower and upper bound for each node;
  - $L_T(b) = R_B(b,a) + \gamma \sum_{o \in O} p(o|b,a) L_T(\tau(b,a,o))$
  - $U_T(b) = R_B(b,a) + \gamma \sum_{o \in O} p(o|b,a) U_T(\tau(b,a,o))$
- Bounds at fringe nodes are obtained offline:
  - LB: Blind policy; PBVI;
  - UB: MDP, QMDP, FIB.
- RTBSS (Paquet et al, 2005), FSBS (Ballesteros et al, 2013).

Online

Monte Carlo sampling

- Approximate probabilities by sampling from generative model;
- Leadings online methods:
  - POMCP (Silver and Veness, 2010);
  - ABT (Kurniawati and Yadav, 2013);
  - DESPOT (Somani et al, 2013).



Online

Heuristic Search

- Guide the search to the most relevant branch nodes;
- Chooses node to expand, according to heuristic $H(b)$;
- AEMS (Ross et al, 2007), FHHOP (Zhang et al, 2012).

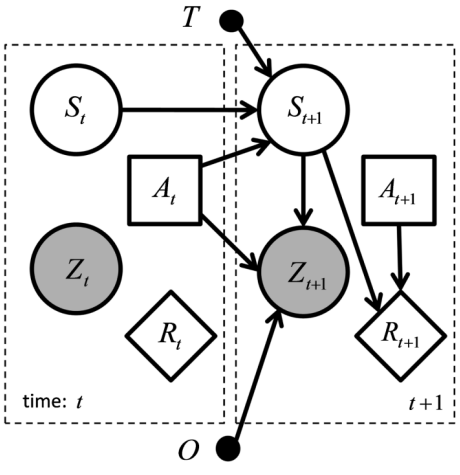
$$
\begin{aligned}
H_T^*(b) &= \begin{cases} H_T(b) & \text{if } b \in \mathcal{F}(T) \\ \max_{a \in A} H_T(b,a) H_T^*(b,a) & \text{otherwise} \end{cases} \\
H_T^*(b,a) &= \max_{z \in Z} H_T(b,a,z) H_T^*(\tau(b,a,z))
\end{aligned}
$$

$$
\begin{aligned}
b_T^*(b) &= \begin{cases} b & \text{if } b \in \mathcal{F}(T) \\ b_T^*(b, a_b^T) & \text{otherwise} \end{cases} \\
b_T^*(b,a) &= b_T^*(\tau(b,a,z_{b,a}^T)) \\
a_b^T &= \operatorname{argmax}_{a \in A} H_T(b,a) H_T^*(b,a) \\
z_{b,a}^T &= \operatorname{argmax}_{z \in Z} H_T(b,a,z) H_T^*(\tau(b,a,z))
\end{aligned}
$$

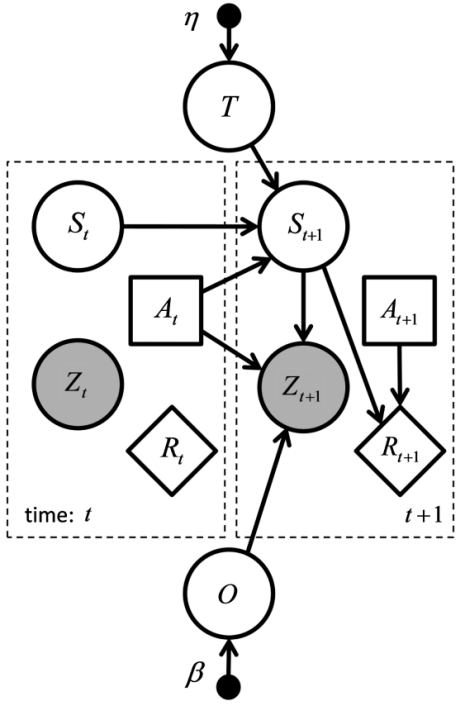
RL with POMDPs

- Model-based:
  - Bayes-Adaptive POMDP (Ross et al, 2011; Katt et all, 2017);
  - T and O are unknown and parameters of the model.

POMDP



BAPOMDP



RL with POMDPs

- Model-based:
  - Bayes-Adaptive POMDP (Ross et al, 2011; Katt et all, 2017);
  - T and O are unknown and parameters of the model.
- Model-free
  - Memoryless techniques: $\pi : O \rightarrow A$;
  - Memory-based: store (partial) history of a_t and o_t;
  - New trends with deep nets.

Deep RL with POMDPs

- Main ideas:
  - Learn directly from the belief (Egorov, 2015);
  - Introduce some kind of memory (RNN,...) (Zhu et al, 2018);
  - Learn generative model of the environment (Igl et al, 2018).
- Deep Q learning:

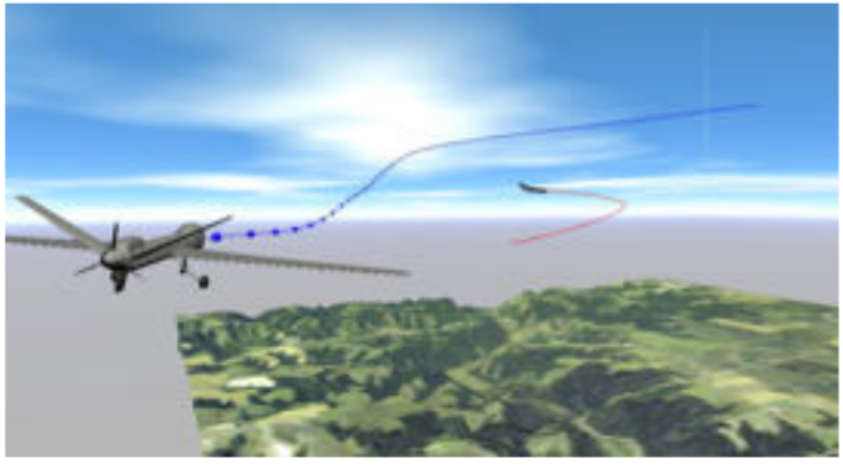
Model	Input	Problem Addressed	Description
DQN	s_t	model-free MDP	full knowledge of the state required
DBQN	b_t	model-based POMDP	updated belief state required
DRQN	$\langle o_1, o_2, \cdots, o_t \rangle$	model-free POMDP	observations as the input
DDRQN	$\langle a_0, a_0, \cdots, a_{t-1} \rangle$ $\langle o_1, o_2, \cdots, o_t \rangle$	model-free POMDP	decoupled actions and observations required
ADRQN	$\langle (a_0, o_1), (a_1, o_2), \cdots, (a_{t-1}, o_t) \rangle$	model-free POMDP	action-observation pairs required as input

Table 1: A comparison among state-of-the-art deep Q-learning approaches

Real World Applications

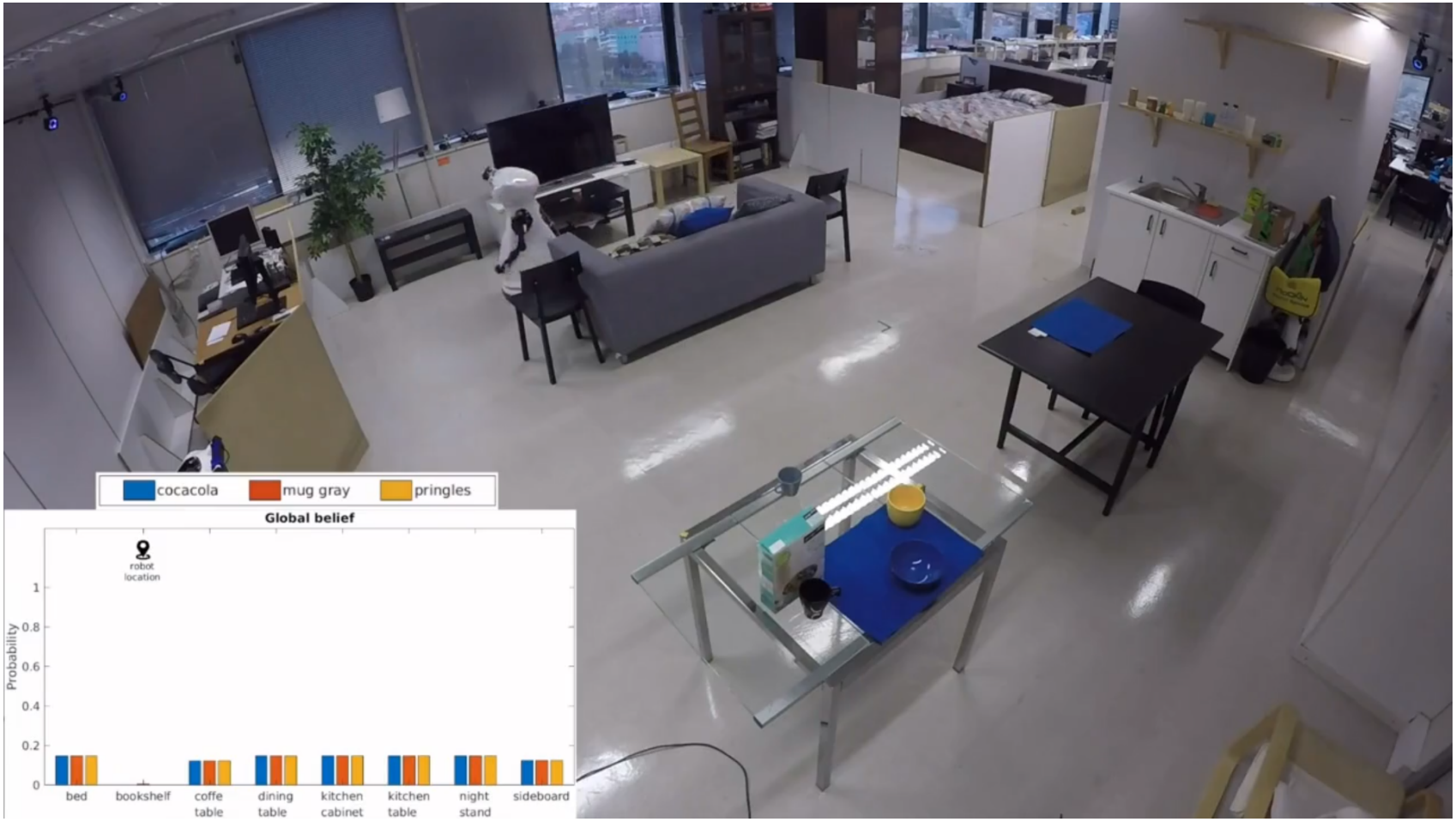
Where POMDPs are used?

- TCAS (Bai et al, 2015);
- Pedestrian avoidance (Bai et al, 2015);
- Home Energy Management (Hansen et al., 2016);
- Cancer screening, surveillance and treatment (Zhang and Denton, 2018).



Applications

Home Scenario



More info

- Begginer tutorial: www.pomdp.org
- Reinforcement Learning, State of the Art (2012)
- Decision-making Under Uncertainty, Mykel Kochenderfer (2015)
- Algorithms for Decision Making, Mykel Kochenderfer (2022)