Planning and Acting in Partially Observable Stochastic Domains

Jarkko Salojärvi

Structure

- MDP's
- POMDP's
- Value Iteration for POMDP's

Markov Decision Process

MDP is described by a tuple $\langle \mathcal{S}, \mathcal{A}, T, R \rangle$, where

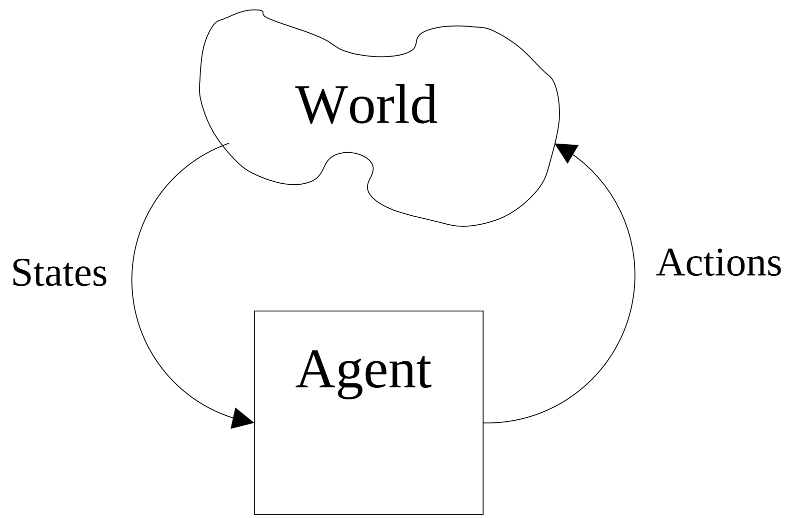
$\mathcal{S}$ is a finite set of states of the world

$\mathcal{A}$ is a finite set of actions

$T : \mathcal{S} \times \mathcal{A} \to \Pi(\mathcal{S})$ is the state transition function.

$R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function.

Markov Decision Process (2)



- Uncertainty about the effects of an agent's actions
- Currect state is always known.
- The next state and the expected reward depend only on the previous state and the action taken.

Solving MDP: Value iteration

$V_1(s) = 0;\ \forall s$

$t = 1$

loop

    $t = t + 1$

    loop $\forall s \in \mathcal{S}$

        loop $\forall a \in \mathcal{A}$

            $Q_t^a(s) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} T(s,a,s') V_{t-1}(s')$

        end loop

    $V_t(s) = \max_a Q_t^a(s)$

    end loop

until $|V_t(s) - V_{t-1}(s)| < \epsilon\ \forall s \in \mathcal{S}$

Partially Observable MDP

- Uncertainty about the currect state.
- \+ *Observations* on the state of the world.

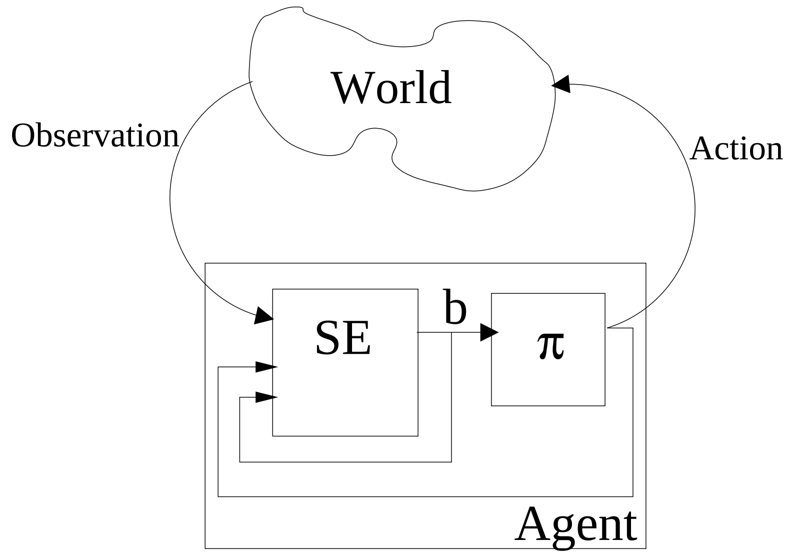
Probability distribution b over possible states.

POMDP

POMDP is described by a tuple $\langle \mathcal{S}, \mathcal{A}, T, R, \Omega, O \rangle$, where

- $\mathcal{S}, \mathcal{A}, T, R$ describe a Markov decision process.
- Ω is a finite set of observations
- $O : \mathcal{S} \times \mathcal{A} \rightarrow \Pi(\Omega)$ is the observation function.

POMDP



- SE: State estimator
- π: Policy
- b: Internal belief state. A sufficient statistic of the past history and initial belief state.

SE: State estimator

Degree of belief in some state s', $b'(s')$, can be obtained by

$$b'(s') = P(s'|o,a,b) = \frac{P(o|s',a)\sum_s P(s'|a,b,s)P(s|a,b)}{P(o|a,b)}$$

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

Can be constructed simply from a given model.

Finding optimal policy

Optimal policy is the solution to a continuous space belief MDP, defined by

$\mathcal{B}$, the set of belief states

$\mathcal{A}$, the set of actions

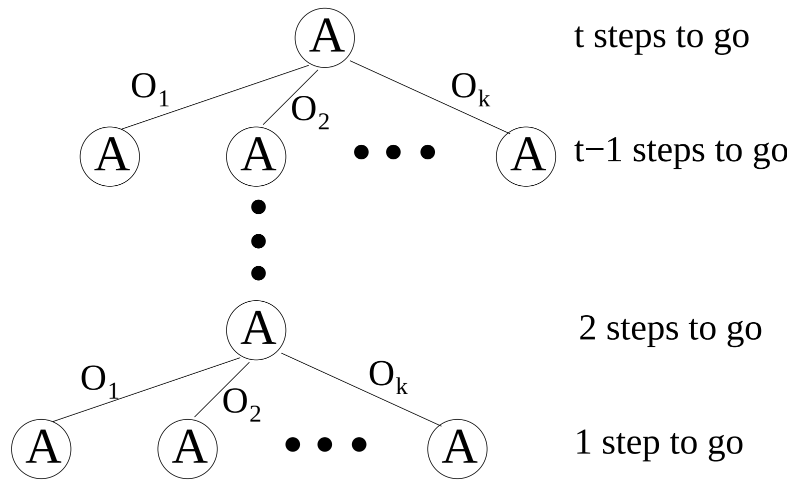
$\tau(b, a, b')$, the state-transition function, defined as $\tau(b, a, b') = P(b'|a, b) = \sum_o P(b'|a, b, o)P(o|a, b)$,

where $P(b'|a, b, o) = 1$ if $SE(b, a, o) = b'$,
(and $P(b'|a, b, o) = 0$ otherwise).
$\rho(b, a)$ is the reward function on belief states,

$$\rho(b, a) = \sum_s b(s)R(s, a)$$

Policy trees

Policy: which action to choose, given the state?
Can be represented as a tree:



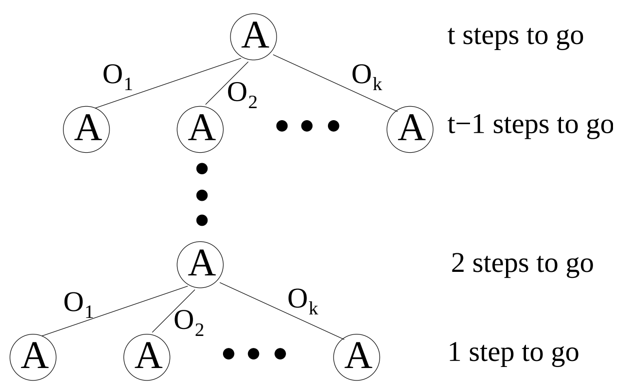
A 1-step policy tree p: $V_p(s) = R(s, a(p))$

. . . policy trees

A t-step policy tree p:

$$V_p(s) = R(s, a(p)) + \gamma \sum_{s'} T(s, a(p), s') \sum_{o_i} O(s', a(p), o_i) V_{o_i(p)}(s'),$$

where $o_i(p)$ is the $t-1$-step policy subtree associated with observation o_i at the top level of a t-step policy tree p.



POMDP: Uncertainty about states

The exact state of the world is not known, only the distribution b over possible states.

The expected value for policy tree p is thus

$$V_p(b) = \sum_s b(s) V_p(s) \ ,$$

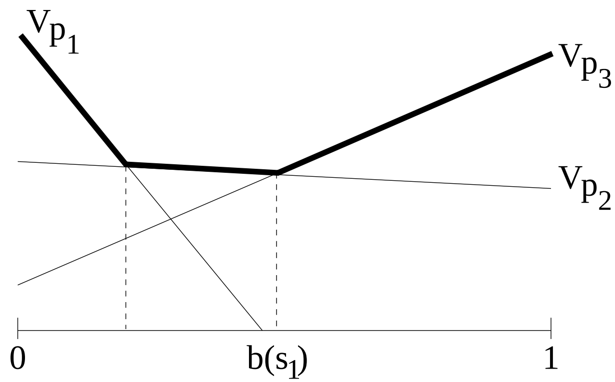
or $V_p(b) = b \cdot \alpha_p$.
The optimal t-step value of starting in belief state b is the value of executing the best policy tree in that belief state:

$$V_p(b) = \max_{p \in \mathcal{P}} b \cdot \alpha_p \ .$$

But there are *many* possible policies p!!

Insights from geometry

Consider a world with only 2 states, $s_1,\ s_2$, and different policies $p_1,\ p_2,\ p_3$

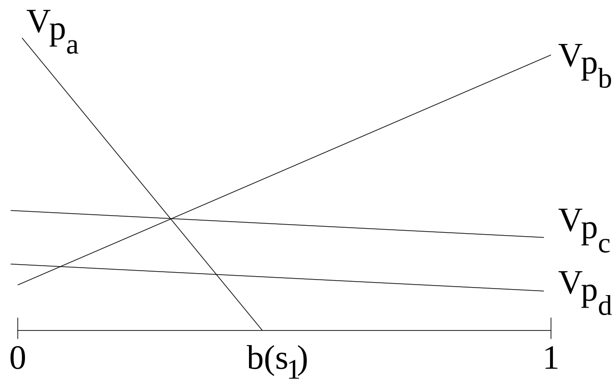


Optimal t-step value forms a piecewise-linear convex surface.

Can define regions (in b) where there is one single policy tree p such that $b \cdot \alpha_p$ is maximal over the entire region.

Insights from geometry (2)

Some policy trees are totally dominated:



Here V_{p_c}, V_{p_d} are dominated and thus not useful. They can be ignored.

POMDP: value iteration

Problem: given a (parsimonious) set of useful policy trees V_{t-1}, how to construct a parsimonious representation of V_t?

Exhaustive enumeration:

1. Construct all possible trees V_t from the given V_{t-1}.
2. Prune out the trees which are not useful.

Exponential in $|\Omega|$!!

Witness

We must avoid exhaustively generating all V_t.

Consider an auxiliary function

$$Q_t^a(b) = \sum_s b(s)R(s,a) + \gamma \sum_o P(o|a,b)V_{t-1}(b'_o) \ .$$

We now have $V_t(b) = \max_a Q_t^a(b)$.
Q-functions are piecewise-linear and convex.
We can define a unique minimal useful set of policy trees for each Q function (finding these is our new problem).

Witness algorithm is used to find the set. It has polynomial complexity.

POMDP: value iteration 2

$\mathcal{V}_1 = [0 \ldots 0]$

$t = 1$

loop

$t = t + 1$

foreach $a \in \mathcal{A}$

$Q_t^a =$witness$(\mathcal{V}_{t-1}, a)$

Prune $\cup_a Q_t^a$ to get $\mathcal{V}_t$

until $\sup_b |V_t(b) - V_{t-1}(b)| < \epsilon$

Witness inner loop

Step 1: Include one tree which is optimal for some belief state b to U_a, the set of minimal useful trees.

Define a new tree p_{new}: If p is a t-step policy tree, o_i an observation, and p' a $(t-1)$ step policy tree, p_{new} is a tree that agrees with p except for observation o_i, where $o_i(p_{new}) = p'$.

Witness theorem: if for some b, we can generate a tree p_{new} such that $V_{p_{new}}(b) > V_{\tilde{p}}(b)$ for all $\tilde{p} \in U_a$, the U_a is not yet a perfect representation of $Q_t^a(b)$. Add p_{new} to U_a.

Replace subtrees of $\tilde{p}$ by $p' \in \mathcal{V}_{t-1}$ until no witness points are found.