Aalto University
School of Electrical
Engineering

ELEC-E8125 Reinforcement learning Partially observable Markov Decision Processes

Ville Kyrki

19.11.2019

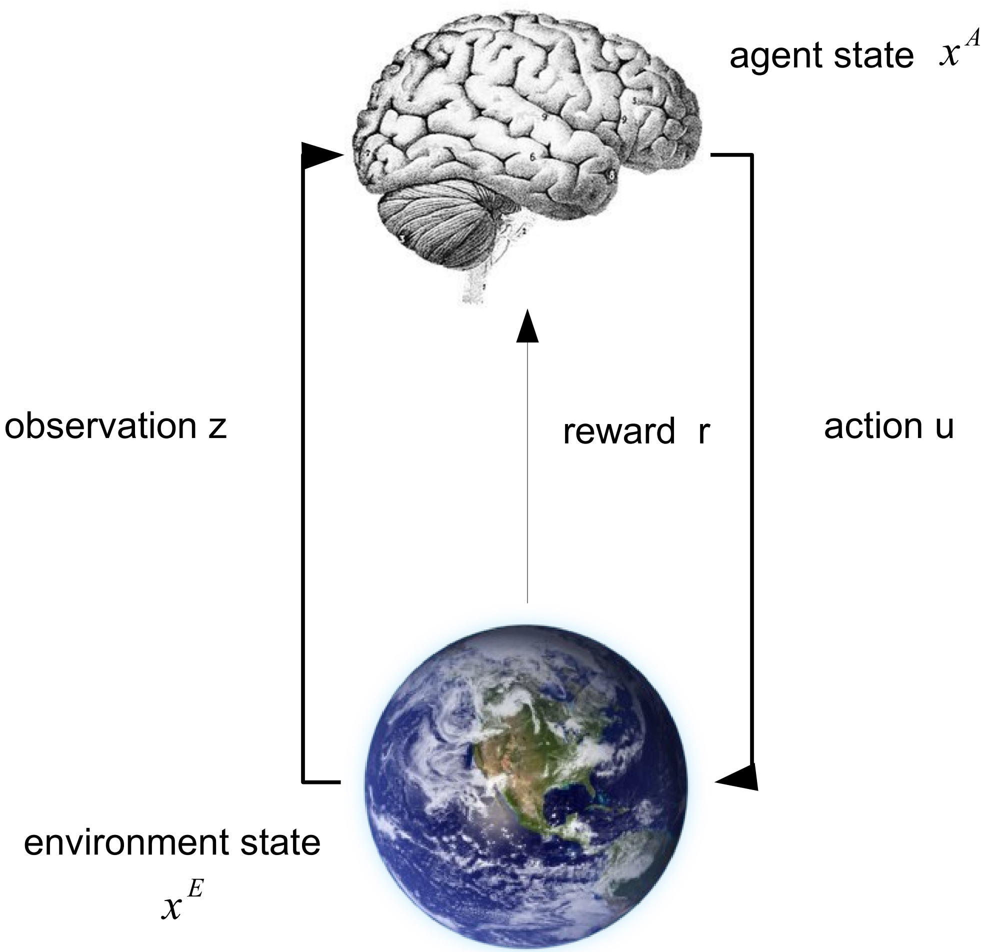
Today

- Partially observable Markov decision processes

Learning goals

- Understand POMDPs and related concepts.
- Be able to explain why solving POMDPs is difficult.

Partially observable MDP (POMDP)



POMDP
Environment not directly observable

Defined by dynamics
$P(x_{t+1}^E | x_t^E, u_t)$

Reward function
$r_t = r(x_{t+1}, x_t)$

$\boxed{\text{Observation model } P(z_t | x_t^E, u_t)}$

Solution similar, eg.
$u_{1,\ldots,T}^* = max_{u_1,\ldots,u_T} E\left[\sum_{t=1}^T r_t\right]$

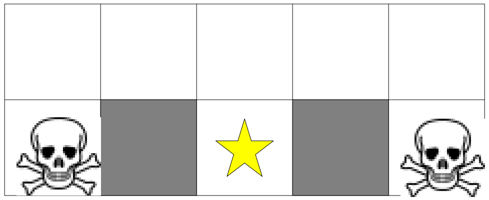
Agent state is not environment state!

Partial observability example

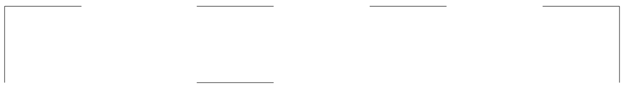
- Observe only adjacent walls.
- Starting state unknown, in upper row of grid.
- Assume perfect actions.

- Give a policy as function of observations!

- Any problems?



Observations:



Can you present a (time-dependent) optimal policy as a tree?

History and information state

- *History* (= Information state) is the sequence of actions and observations until time *t*.

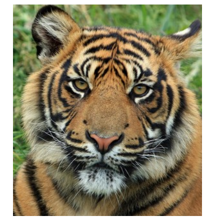
- Information state is Markovian, i.e.,

$$P_I(I_{t+1}|u_t, I_t) = P_I(I_{t+1}|u_t, I_t, I_{t-1}, \ldots, I_0)$$

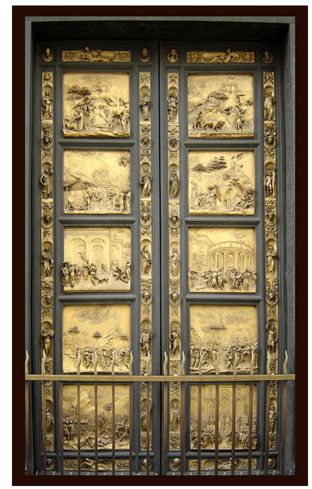
- POMDP thus corresponds to Information state MDP.

Example: Tiger problem

r=10



r=-100



U = {open right, open left, listen}

P(HL|TL)=0.85
P(HR|TL)=0.15
P(HL|TR)=0.15
P(HR|TR)=0.85

?

What kind of policy would be reasonable?

Policy depends on history of observations and actions = information state.

Belief state, belief space MDP

- Belief state = distribution over states.
  - Compresses information state.

- Belief $b_t(x) \equiv p(x_t = x | I_t)$ ← Can be represented as a vector $\boldsymbol{b} = (b(x_1), b(x_2), \ldots)$

- POMDP corresponds to belief space MDP.
- POMDP solution can be structured as
  - State estimation (of belief state) +
  - Policy on belief state.

Belief update

Similar to state estimator, e.g. Kalman filter, particle filter:

= state estimation

"measurement update"

"prediction"

$$b_z^u(x) = b_{t+1} = \frac{\boxed{P(z|x,u)}\boxed{\sum_{x'} P(x|x',u)\, b_t(x')}}{\boxed{\sum_{x',x''} P(x''|x',u)\, P(z|x'',u)\, b_t(x')}}$$

Normalization factor

Single step policies

- Value of belief state for a particular single step policy

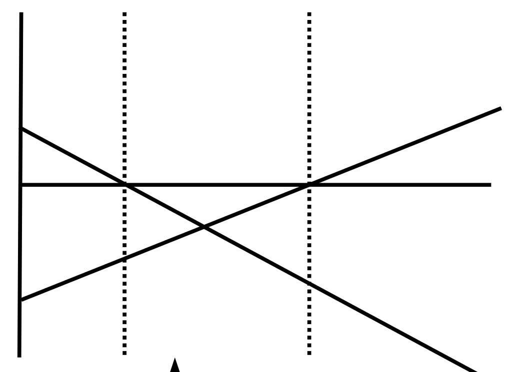
$$V_{\pi}(\boldsymbol{b}) = \sum_x b(x) V_{\pi}(x)$$

- Can be represented as *alpha vector* (consisting of values for each state)

$$V_{\pi}(\boldsymbol{b}) = \boldsymbol{\alpha}^T \boldsymbol{b}$$

- Value of optimal policy is then

$$V^*(\boldsymbol{b}) = max_i \, \boldsymbol{\alpha}_i^T \boldsymbol{b}$$



Maximum over all actions

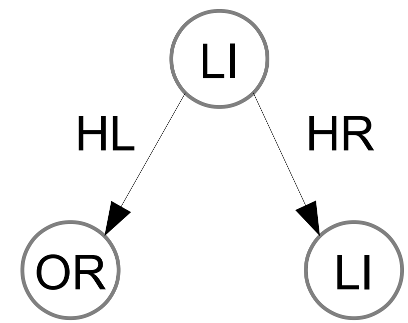
Piecewise linear and convex (PWLC)

Conditional plans and policy trees

- Similar to single step policies, value functions of multi-step policies can be represented as alpha vectors.

- Best policy for a particular belief is then again

$$V^*(\boldsymbol{b}) = max_i \, \boldsymbol{\alpha}_i^T \boldsymbol{b}$$



What's the α-vector of this policy?

Value iteration on belief states

- Bellman equation

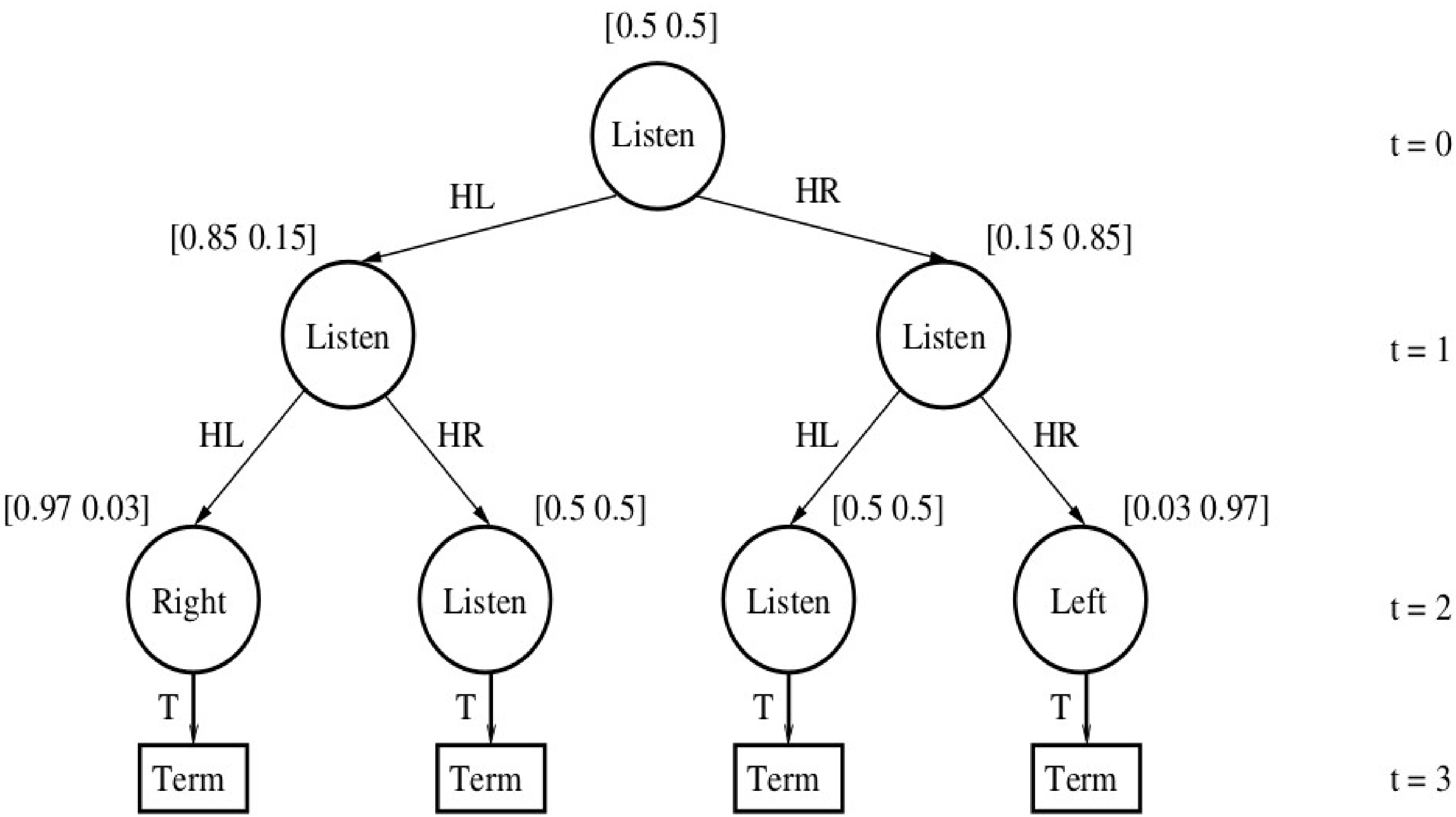
$$V_{n+1}^{*}(b) = max_u \left[\sum_x b(x) r(x,u) + \gamma \sum_z \sum_{x'} P(z|x',u) \sum_x P(x'|x,u) b(x) V_n^{*}(b_z^u) \right]$$

- No trivial closed form solution (similar to MDP tabulation) because *V(b)* is a function of a continuous variable.
- At each iteration, each plan of previous iteration is combined with each possible action/observation pair to generate plans of length *n*+1.
  - At each iteration number of conditional plans increases by

$$|V_{n+1}| = |U| |V_n|^{|Z|}$$

- Some conditional plans often not optimal for any belief.
  - Corresponding alpha-vectors never dominant.
  - Alpha-vectors (/conditional plans) can be pruned at each iteration.

Starting from known belief state



Computational complexity

- Number of possible policy trees of horizon H is

$$|U|^{\frac{|Z|^H - 1}{|Z| - 1}} \approx |U|^{|Z|^{H-1}}$$

- Infinite horizon POMDPs thus not possible to construct in general.

Summary

- Partially observable MDPs are MDPs with observations that depend stochastically on state.
- POMDP = belief-state estimation + belief-state MDP.
- POMDPs computationally untractable in general situations.
  - Approximations are needed for larger than toy problems.

Next week: Larger POMDPs