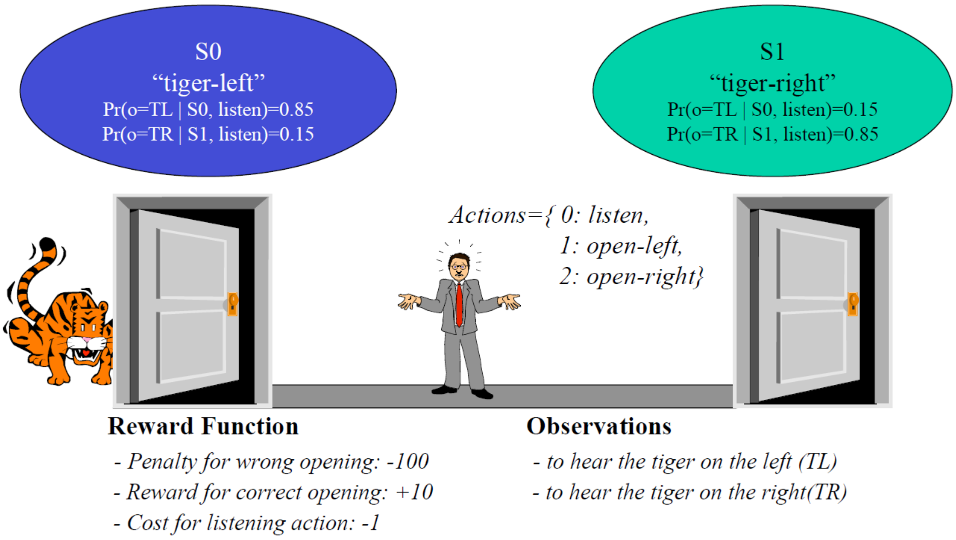
AI & Robotics: Research
Exercise 5

Marc Toussaint
Learning & Intelligent Systems Lab, TU Berlin
Marchstr. 23, 10587 Berlin, Germany

Summer 2020

Please prepare written solutions to the following exercises that you can share by screen in our session. The notes can be brief, but esp. equations and derivations should be written precisely. Try to use LaTeX.

1 The Tiger Problem POMDP


S0
"tiger-left"
Pr(o=TL | S0, listen)=0.85
Pr(o=TR | S1, listen)=0.15
S1
"tiger-right"
Pr(o=TL | S0, listen)=0.15
Pr(o=TR | S1, listen)=0.85
Actions={ 0: listen,
1: open-left,
2: open-right}
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)

The Tiger Problem is a small (educational) POMDP. There are two doors. When the agent opens one of them s/he either gets reward +10 if opening the right door, or reward -100 (eaten by a tiger) when opening the wrong door. The agent is initially uncertain which door is the right one. The agent has a third action, *listening*, which produces a noisy observation $o \in \{\text{TL}, \text{TR}\}$ (Tiger-is-Left, Tiger-is-Right), but gives reward -1.

a) In the lecture we formalized a POMDP as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathcal{T}, \mathcal{O}, \mathcal{R}, b_0, \gamma \rangle$, where I added the initial belief state b_0, and the discount factor γ, as part of the POMDP specification. Briefly define each of these quantities $\langle \mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathcal{T}, \mathcal{O}, \mathcal{R}, b_0, \gamma \rangle$ for the tiger POMDP (taking numbers from the figure above). Assume $b_0 = (\frac{1}{2}, \frac{1}{2})$.

b) Assume that the tiger is truly behind the left door (but the agent does not know). Consider an agent that always chooses to listen. Compute the belief state after t observations, assuming that the agent observed m-times TL, and n-times TR, which $m + n = t$.

c) For that same agent, provide the Q-value $Q(b_t, a)$ for the alternative 2 actions $a \in \{\text{open-left}, \text{open-right}\}$, given the belief state b_t. Assume that the episodes terminate after opening a door (no rewards beyond this point).

d) For which belief states should the agent stop listening and open a door for a discount factor of $\gamma = 1$? (How would this change if there were zero costs for listening?)

e) Explain the DESPOT tree for this particular POMDP. Specifically, sketch the tree up to $t = 2$ (where the agent experiences a_1, o_1, a_2, o_2) (drawing by hand and adding a photo is absolutely fine). Which nodes are visited by which "scenarios" (in the DESPOT sense)?

2 UCT and Optimism

When “walking down a tree” using A*, we choose the action branching with

$$a^* = \operatorname*{argmax}_{a \in \mathcal{A}(v)} \; g(v) + h(v)$$

where $\mathcal{A}(v)$ is the set of actions (children) at node v, $g(v)$ is the cost-so-far, and $h(v)$ is a lower bound of the true cost-to-go from node v (aka. admissible heuristic).

When walking down a tree using UCT, we choose the action branching with

$$\operatorname*{argmax}_{a \in \mathcal{A}(v)} \; \frac{Q(v,a)}{n(v,a)} + \beta \sqrt{\frac{2 \ln n(v)}{n(v,a)}}$$

where $n(v)$ counts how often we visited node v, $n(v,a)$ counts how often we've selected a at v, and $Q(v,a)$ is the sum of all returns ($\sim$ neg. path costs) we got on paths that selected a at v.

Provide an argument for why the UCT is an “admissible heuristic with high probability”. That is, try to show that, with high probability, the term after argmax over-estimates the mean return.

Simplification: To make your argument, you may assume that the returns of each rollout is truely Gaussian distributed with some known variance σ. Note that (in the context of UCT) is defined to be the sum of returns of all experience rollouts. We simplify the UCT policy to

$$\operatorname*{argmax}_{a \in \mathcal{A}(v)} \; \frac{Q(v,a)}{n(v,a)} + \beta \sqrt{\frac{1}{n(v,a)}} \; ,$$

which drops the $2 \ln n(v)$. How can one choose β to ensure that this term is an “admissible heuristic with high probability $(1-\epsilon)$”? (No explicit solution necessary, just an explanation how in principle.)