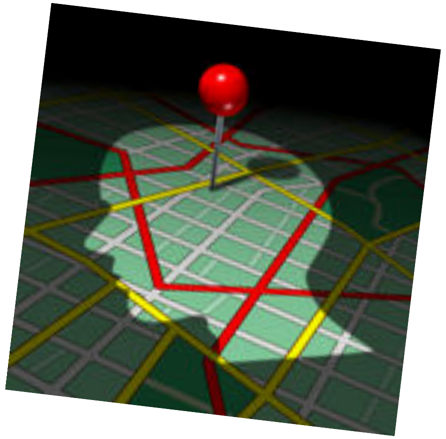
Computational Psychiatry Course – Zürich 2015

Partially Observable Markov Decision Processes

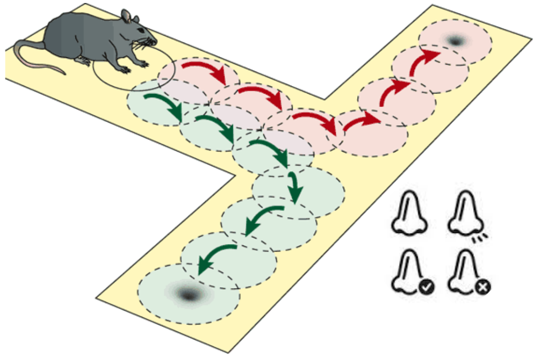
Lionel Rigoux & Frederike Petzschner

Introduction

- MDP >> Full observability: the agent always knows the state of the world
- This might often not be true in real life
  - *Imperfect memory*
    // navigation: “turn left on the seventh street”
    > what if you loose track of the number of streets already passed?

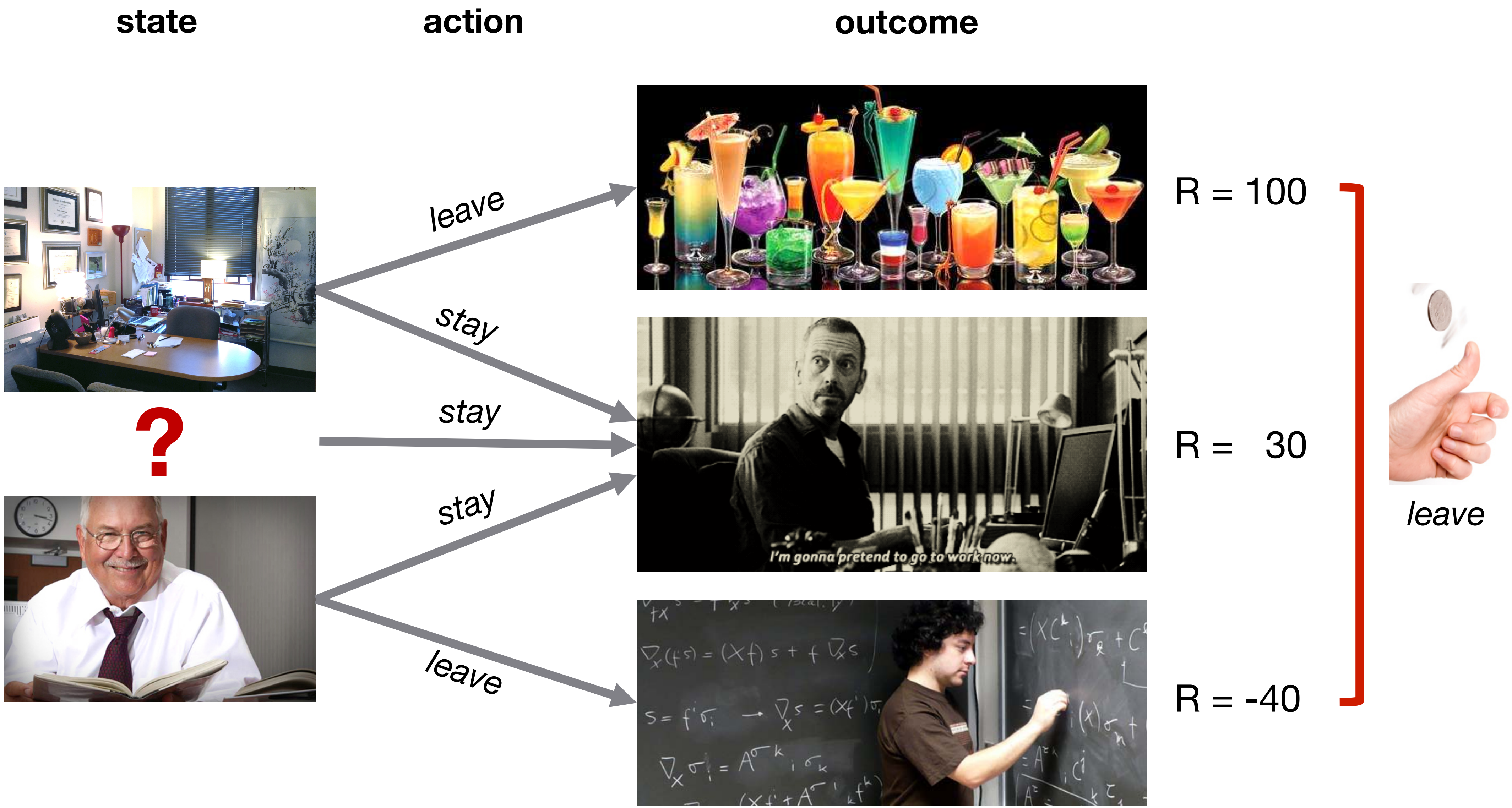


  - *Changing environment*
    // reward selection in a T-maze
    > reward location changes every trials, as cued by a smell



Outline

- Extend the MPD framework to account for state uncertainty
  - Beliefs representation
  - Observation function
  - Belief updating and state chaining
- Formalization
- Solution
- Conclusion
- Perspectives


state
action
outcome
leave
stay
stay
stay
leave
?
I'm gonna pretend to go to work now.
R = 100
R = 30
R = -40
leave

state

not known

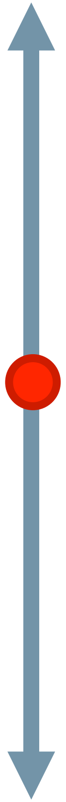
S_0

S_1

belief

$b = p(s=S_1)$

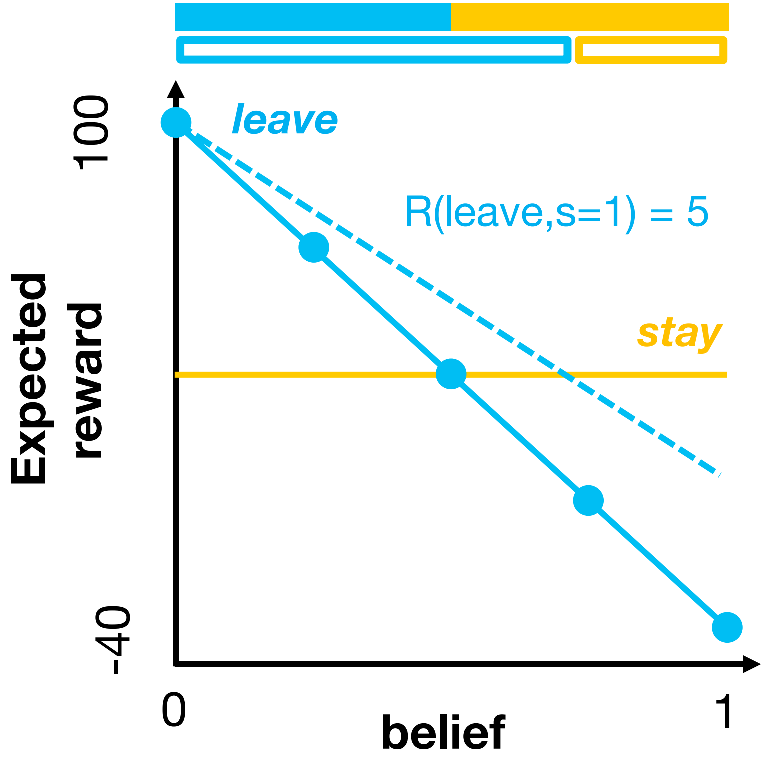
$p(s=S_1) = 0$



$p(s=S_1) = 1$

actions and payoff function

optimal policy



$$E[R](a) = p(x=0)\, R_0(a) + p(x=1)\, R_1(a)$$

state

observation function

provide information about state

belief

S_0

	leave	stay	listen
noises	0	0.5	0.15
no one	1	0.5	0.85

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

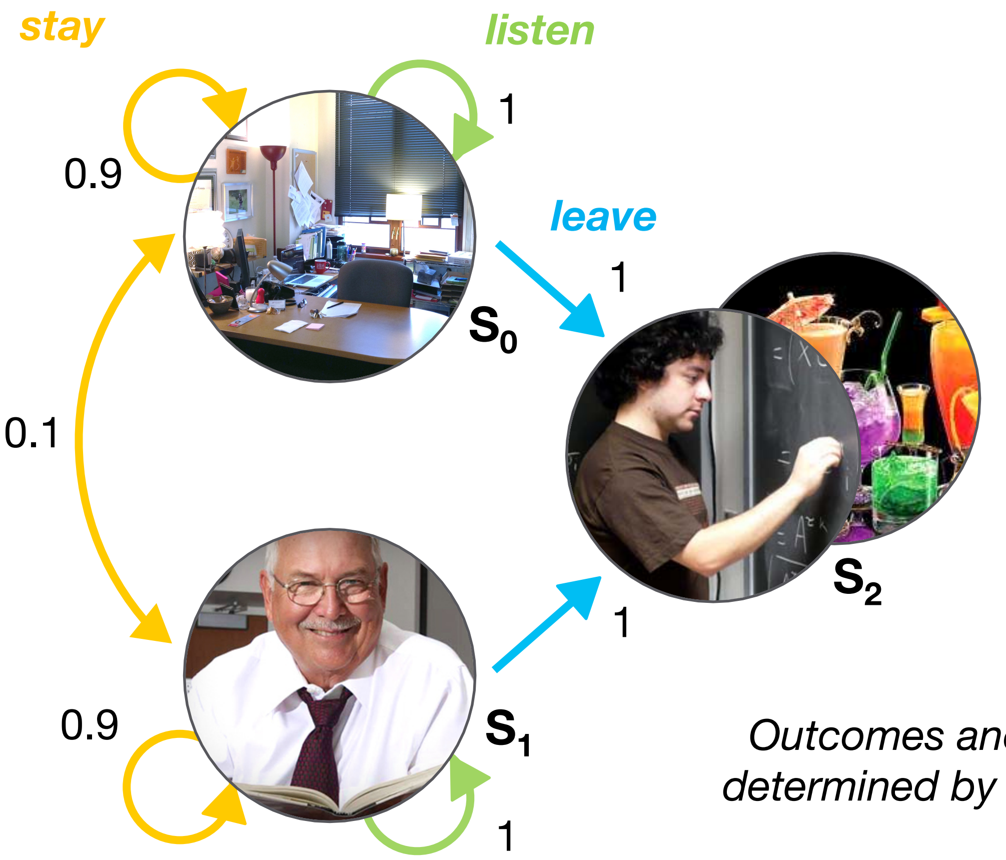
S_1

	leave	stay	listen
noises	1	0.5	0.85
no one	0	0.5	0.15

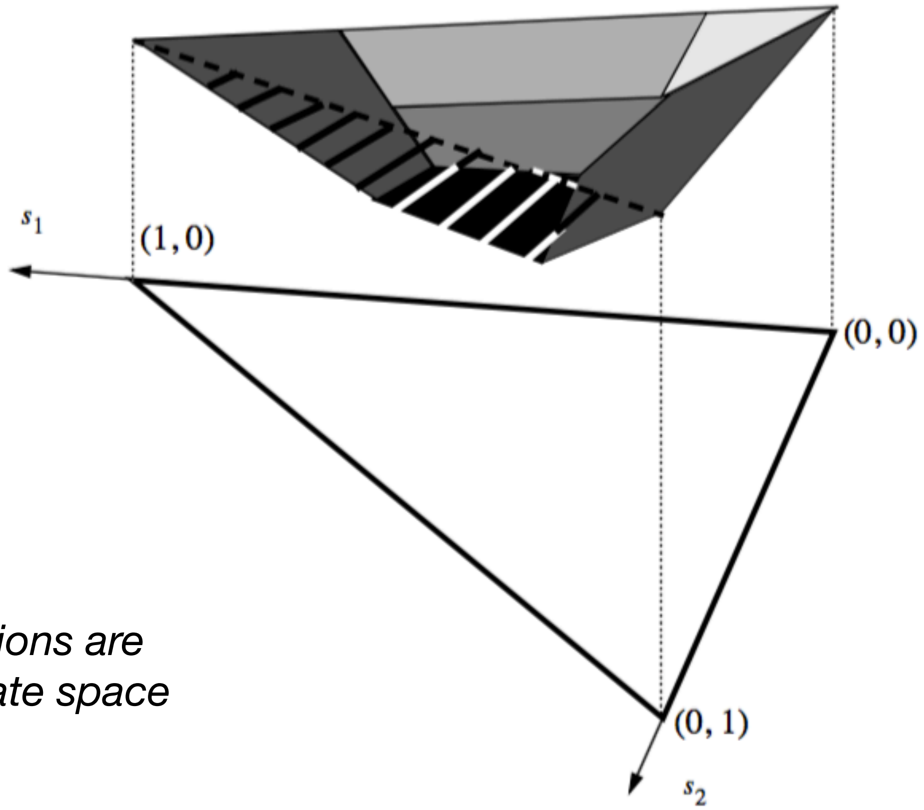
$p(s=S_1) = 0$

$p(s=S_1) = 1$

state space



belief space



Outcomes and observations are determined by the real state space

Policy relies on the belief state

POMDP Formalism

MDP

- S set of states
- A set of actions
- T transition matrix $S \times A \rightarrow S$
- R reward function $S \times A \rightarrow \mathbb{R}$
- γ discount factor

POMDP extension

- Ω set of observations
- O observation probabilities
  $\mathrm{S} \times A \times \Omega \rightarrow [0, 1]$
- B belief space
- r reward function $B \times A \rightarrow \mathbb{R}$
- τ belief update function $B \times A \times \Omega \rightarrow B$

Simulation workflow

Initial state (s,b)
- Select action $a = \pi(b)$
- Update state $s\uparrow' = T(s,a)$
- Receive outcome $R(s,a)$
- Get observation $o = O(s\uparrow', a)$
- Update belief $b\uparrow' = \tau(b,a,o)$
- Start over

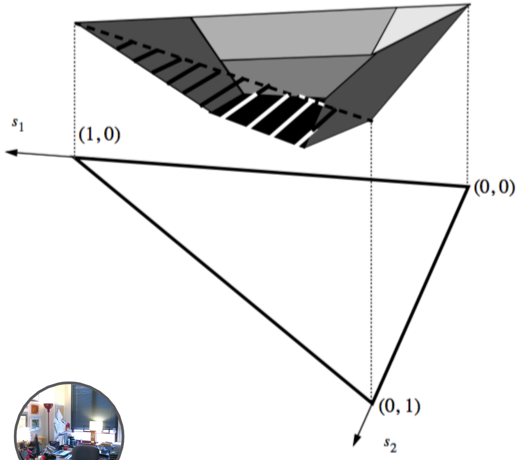
$$V^{\pi}(b) = \sum_{t=0}^{\infty} \gamma^{t}\, r(b_t, a_t)$$

$$\pi^{\star} = \operatorname*{argmax}_{\pi} V^{\pi}$$

Resolution

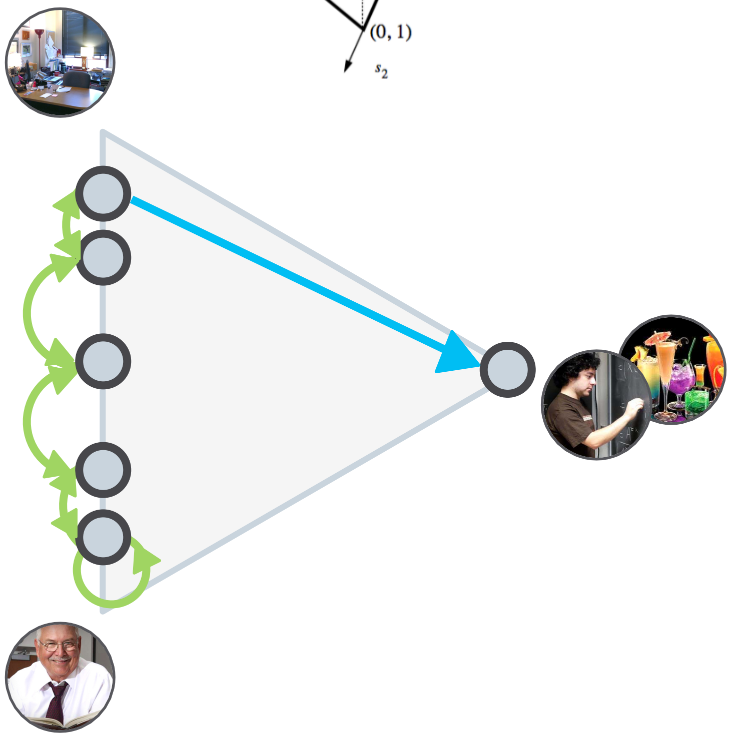
The value function is always convex

- Certainty is preferable to uncertainty
- Gathering information is valuable



The solution can be discretized

- Optimal solution often visit a finite number of belief states
- The POMDP can then be reformulated as a (fully observable) MDP



Take home message

POMDPs allow to model:

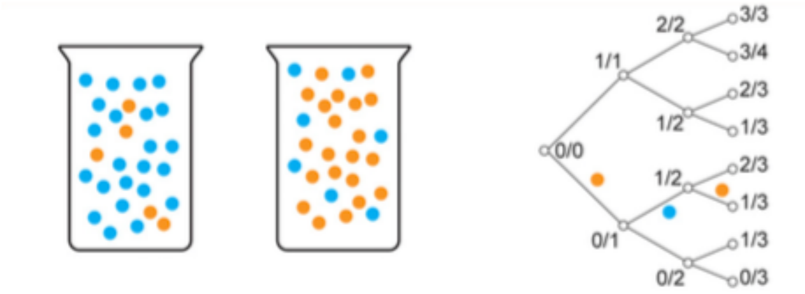
- sequential decision making in a complex environment (MDP)
- subjectivity about the state of the world (PO)

POMDPs can capture:

- information gathering as an economic decision
- irrational behaviour as an optimal policy based on wrong representations

Perspectives

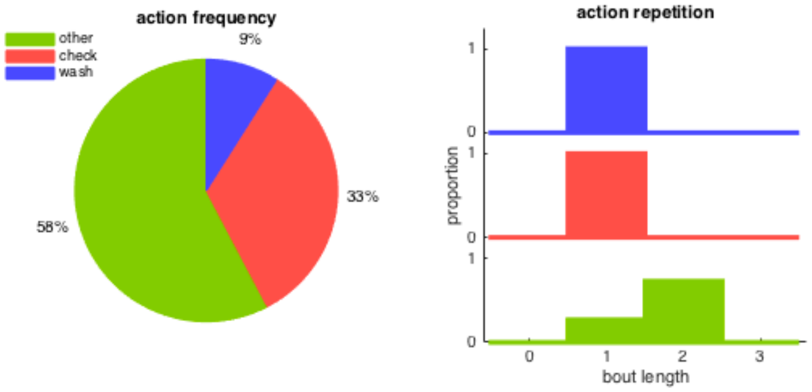
Information sequential sampling with varying payoffs



[Averbeck 2015, PCB]

Errors as exploratory behaviour in reversal learning tasks

Checking behaviours in OCD



Questions?

Thank you for your attention