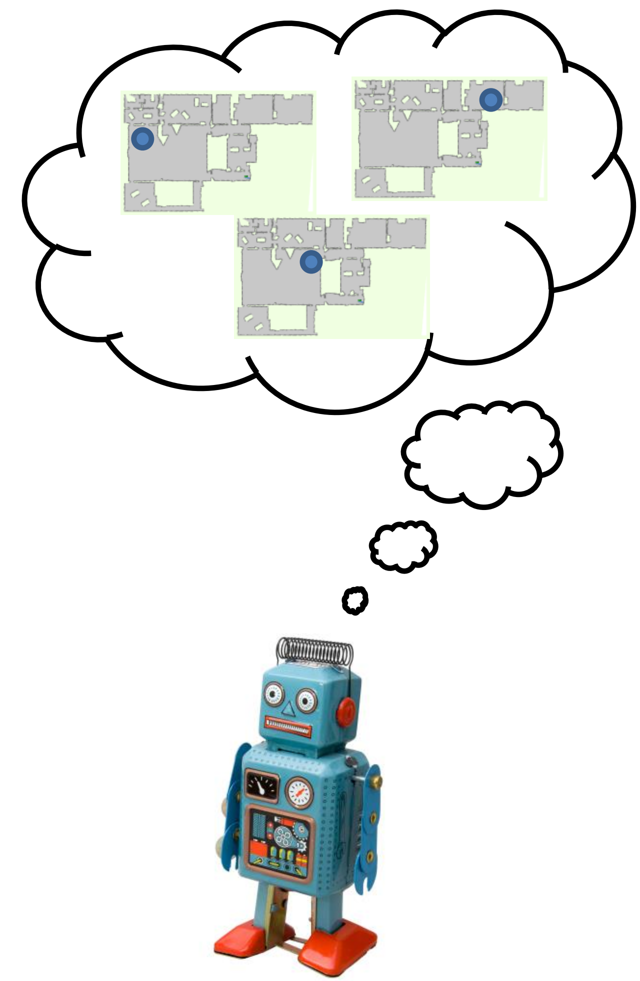
Partially Observable Markov Decision Processes

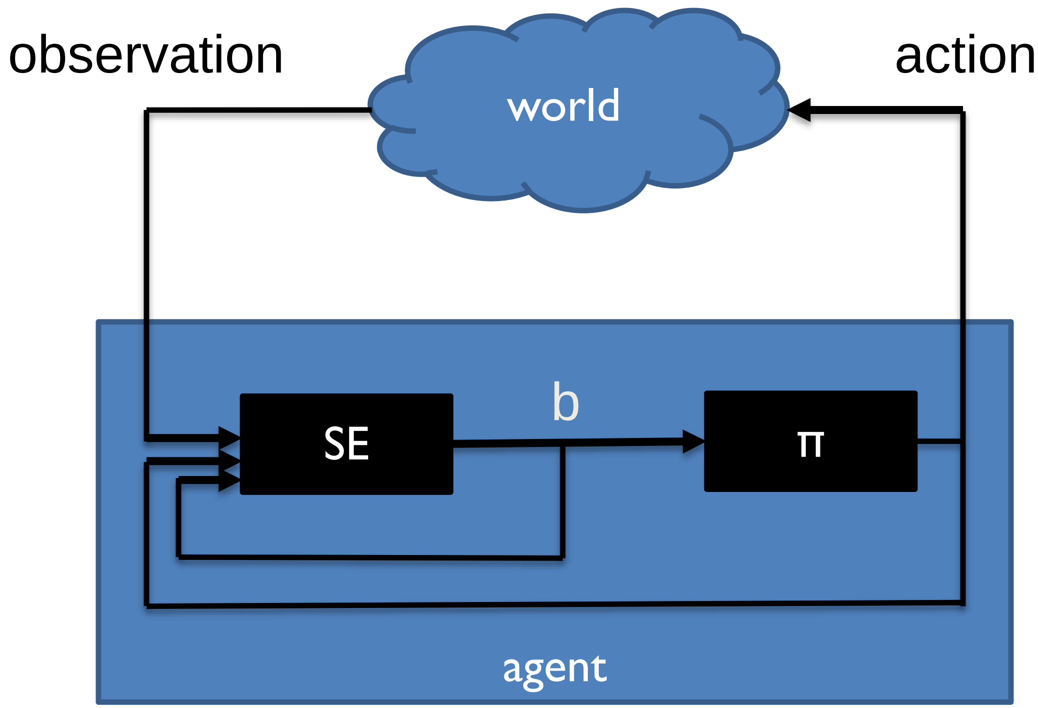
PAH 2015

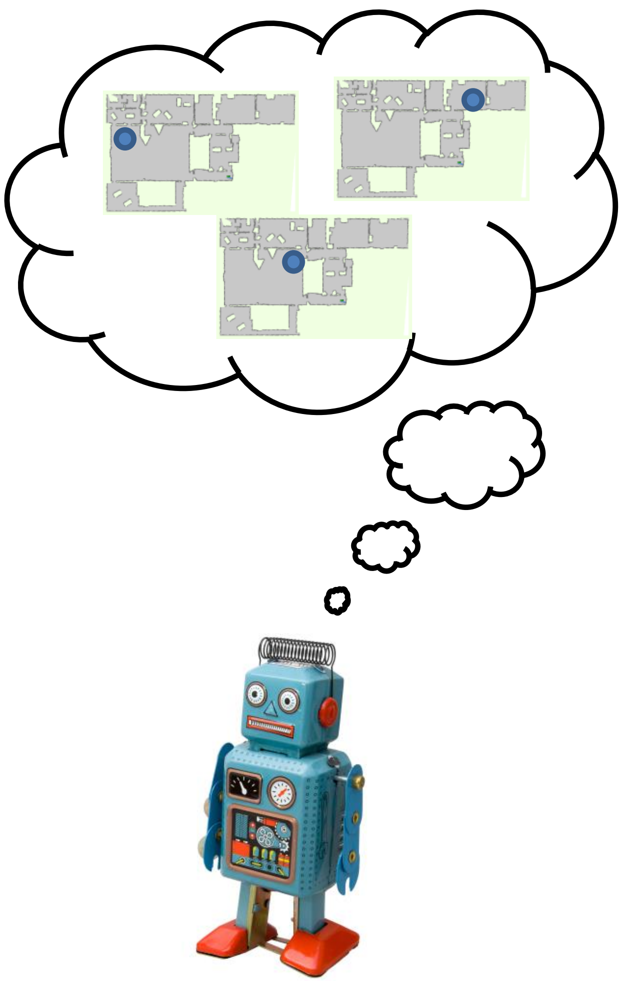
Partial Observability

- the world is not perfect
  - actions take some time to execute
  - actions may fail or yield unexpected results
  - the environment may change due to other agents
  - the agent does not have knowledge about whole situation
  - sensors are not precise



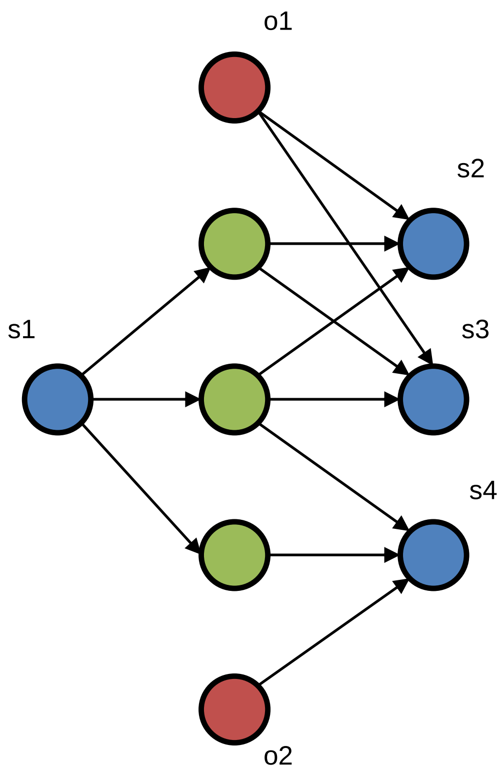
Partial Observability

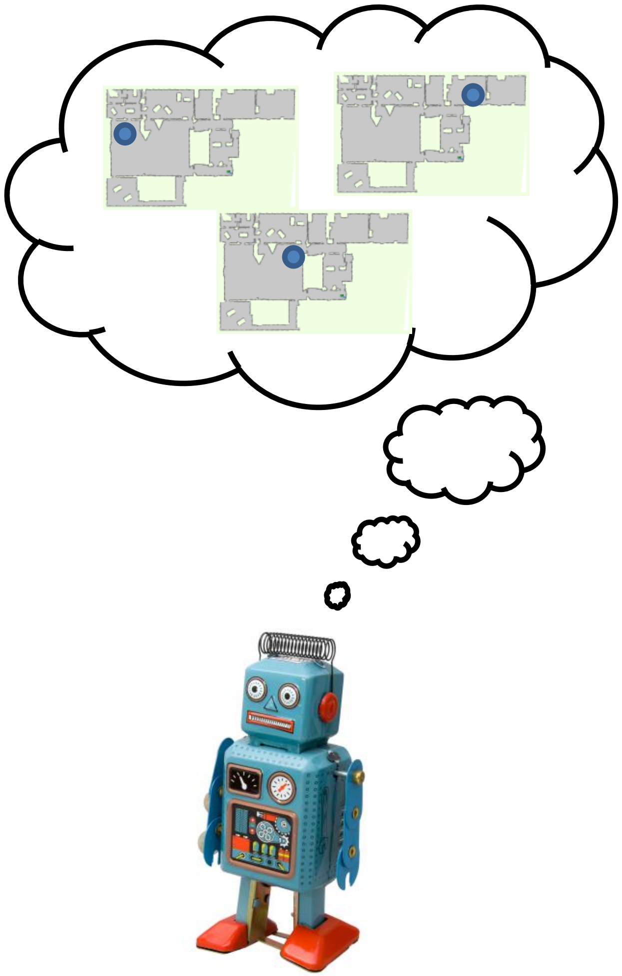




- policy (π)
- belief state (b)
- state estimator (SE), for updating belief state b' based on
  - the current observation o_t
  - the last action a_{t-1}
  - and the previous belief state b_{t-1}

Partial Observability

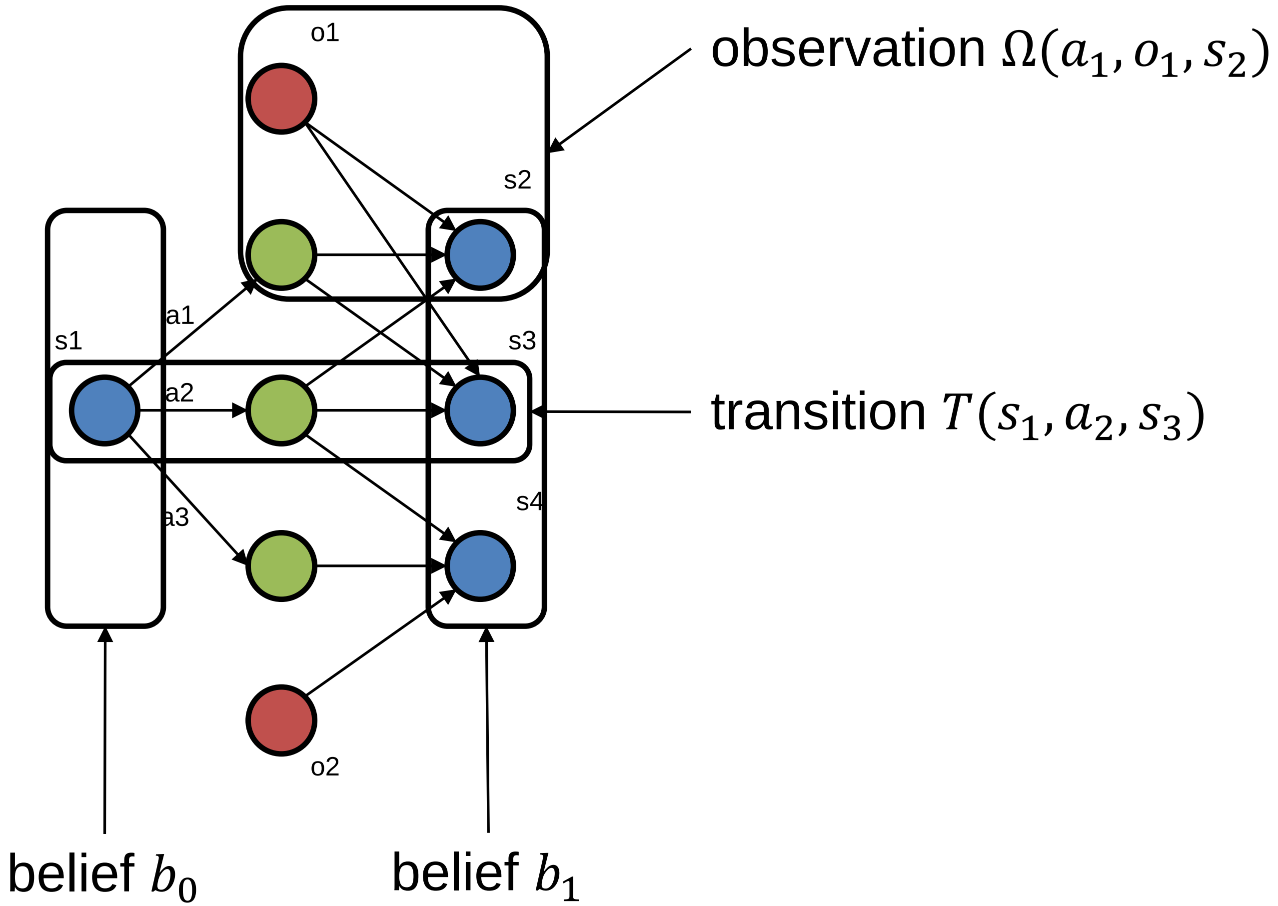




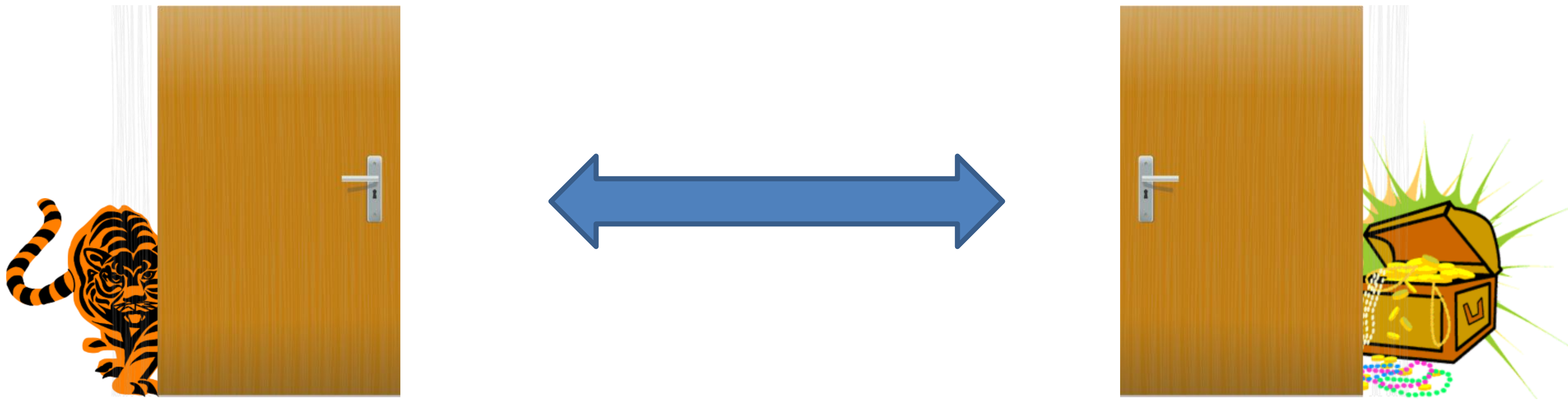
Partially Observable MDPs

- main formal model for scenarios with uncertain observations
- $\langle S, A, D, O, b_0, T, \Omega, R, \gamma \rangle$
  - states – finite set of states of the world
  - actions – finite set of actions the agent can perform
  - time steps
  - observations – finite set of possible observations
  - initial belief function $b_0: S \rightarrow [0,1]$
  - transition function $T: S \times A \times S \rightarrow [0,1]$
  - observation probability $\Omega: A \times O \times S \rightarrow [0,1]$
  - reward function $R: S \times A \rightarrow \mathbb{R}$
  - discount factor $0 \leq \gamma < 1$

Partially Observable MDPs - probabilities

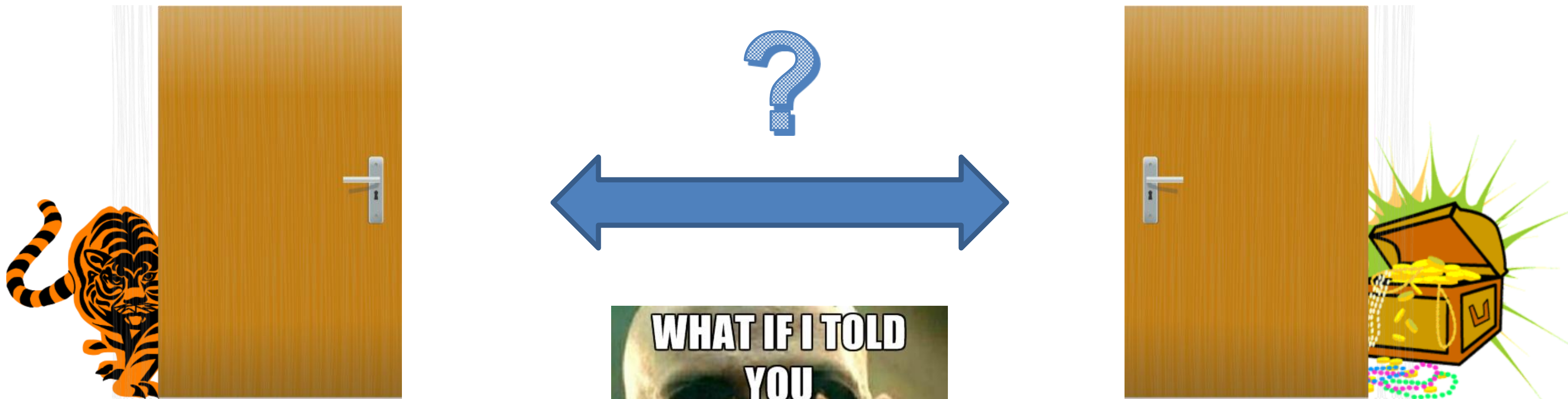


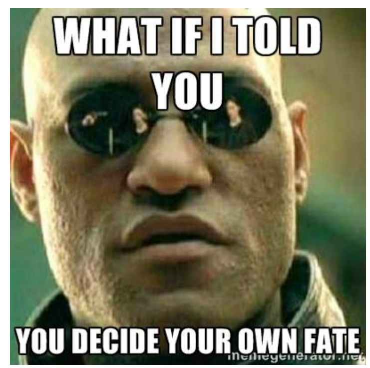
The Tiger Problem



- states = {tiger-left, tiger-right}
- actions = {listen, open-left, open-right}
    - transitions: no change (listen), restart (open-right, open-left)
- observations = {hear-tiger-left (TL), hear-tiger-right (TR)}
- rewards: surprised by tiger, found treasure, listening

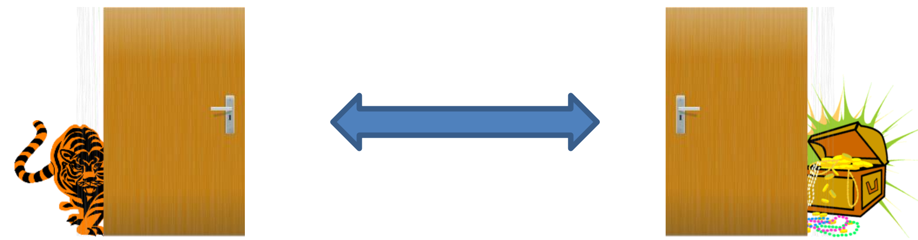
The Tiger Problem





- states = {tiger-left, tiger-right}
- actions = {listen, open-left, open-right}
    - transitions: no change (listen), restart (open-right, open-left)
- observations = {hear-tiger-left (TL), hear-tiger-right (TR)}
- rewards: surprised by tiger, found treasure, listening

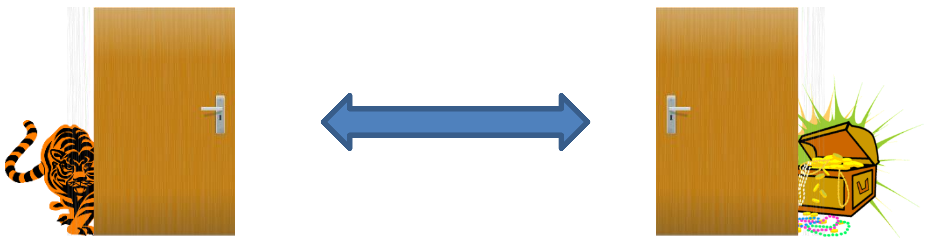
The Tiger Problem (transition prob.)



listen →	tiger-left	tiger-right
tiger-left	1.0	0.0
tiger-right	0.0	1.0

open-left/right →	tiger-left	tiger-right
tiger-left	0.5	0.5
tiger-right	0.5	0.5

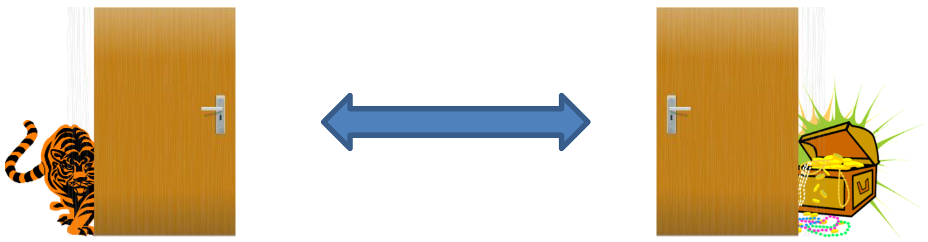
The Tiger Problem (observation prob.)



listen →	hear TL	hear TR
tiger-left	0.85	0.15
tiger-right	0.15	0.85

open-left/right →	hear TL	hear TR
tiger-left	0.5	0.5
tiger-right	0.5	0.5

The Tiger Problem (immediate reward)

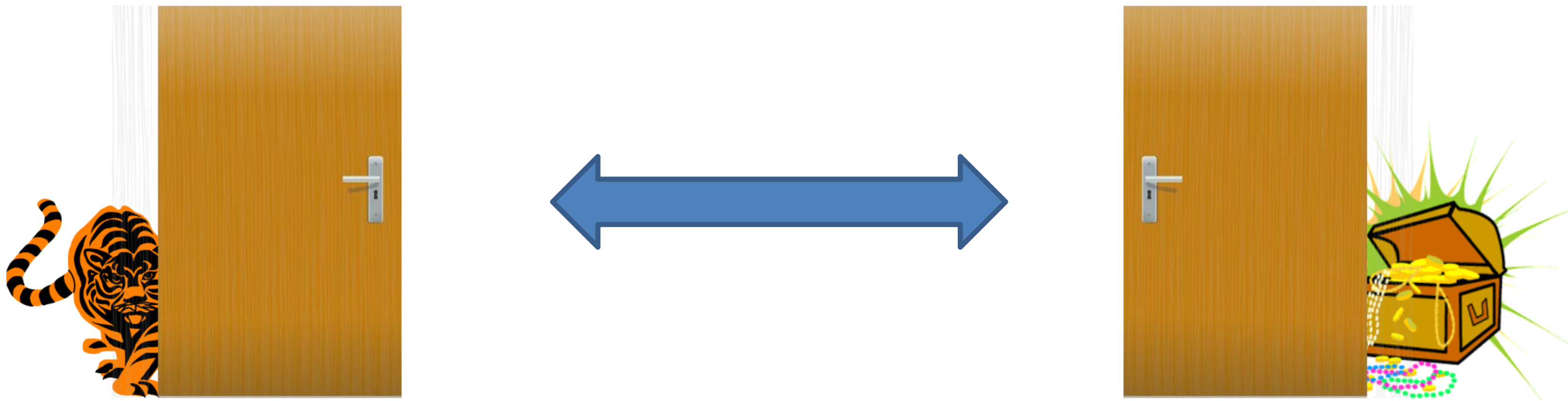


reward	tiger-left	tiger-right
listen	-1	-1
open-left	-100	+10
open-right	+10	-100

Partially Observable MDPs - beliefs

- beliefs represent a probability distribution over states
- beliefs are uniquely identified by the history
  - b_1 - probability distribution over states after playing one action
  - $b_t \leftarrow \Pr(s_t | b_0, a_0, o_1, \ldots, o_{t-1}, a_{t-1}, o_t)$
- we can exploit dynamic programming (define transformation of beliefs)
  - $b_t(s') = \mu\Omega(a, o, s'). \sum_{s \in S} T(s, a, s') b_{t-1}(s)$
  - where
    - o is the last observation
    - a is the last action
    - μ is the normalizing constant

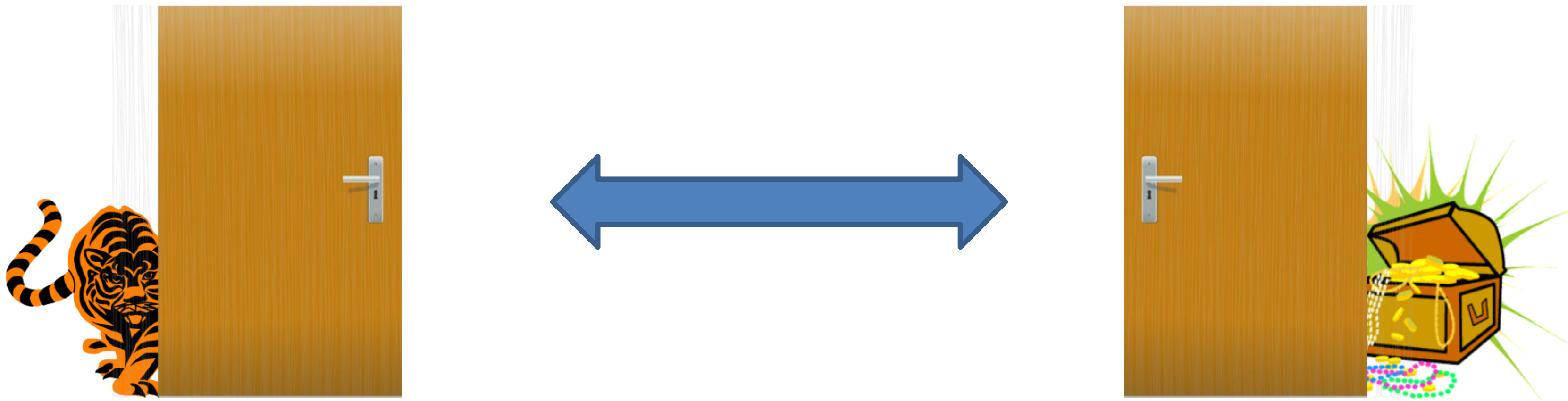
The Tiger Problem (belief update)



s_0: tiger-left　　belief vector　　s_1: tiger-right

0:　　belief b_0

The Tiger Problem (belief update)

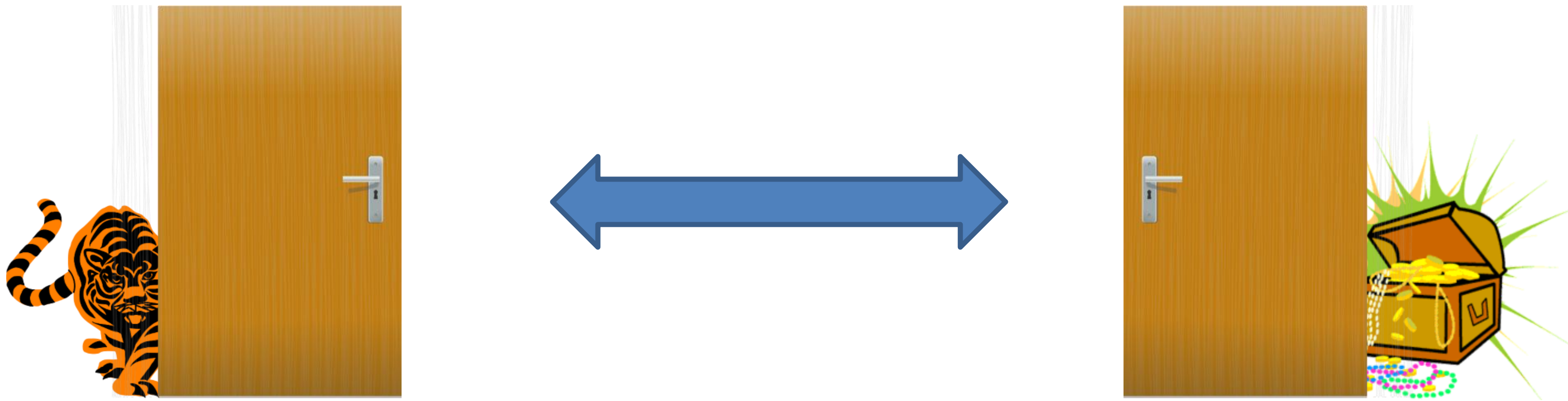


s_0: tiger-left   belief vector s_1: tiger-right

0: belief b_0

1: listen a_0

The Tiger Problem (belief update)



s_0: tiger-left — belief vector — s_1: tiger-right

0: belief b_0

1: listen a_0 → TL

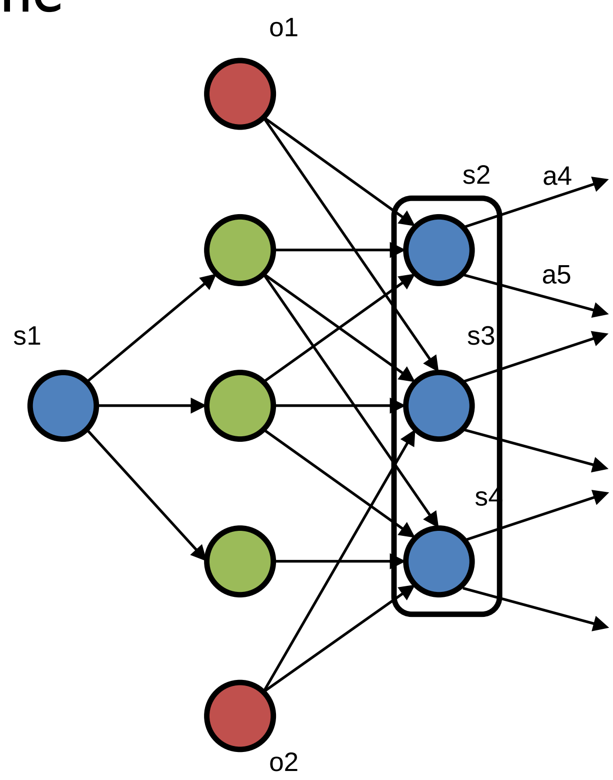
2: belief $b_1(\cdot) = \mu\Omega(a, o, \cdot). \sum_{s \in S} T(s, a, \cdot) b_{t-1}(s)$

Partially Observable MDPs - values

- beliefs determine new values
  - $V(b) = \max_{a \in A}[R(b,a) + \gamma \sum_{b' \in B} T(b,a,b')V(b')]$
- what we have done …
  - we have transformed a POMDP to a continuous state MDP
  - belief state is a simplex
    - $|S| - 1$ dimensions
- in theory we can use all the algorithms for MDPs (value iteration)
  - but B is infinite

Solving Continuous State MDPs

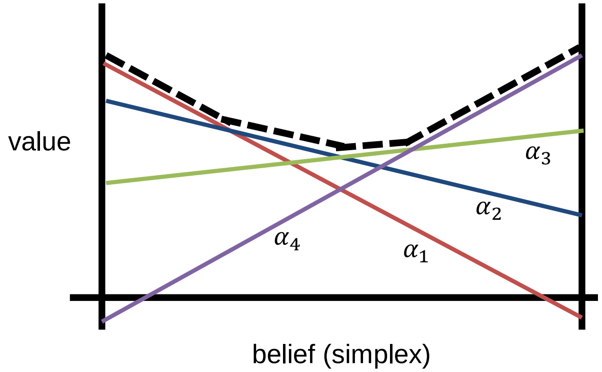
- in value iteration we take max of actions
- the belief space can be partitioned depending on the fact, which action is the best one

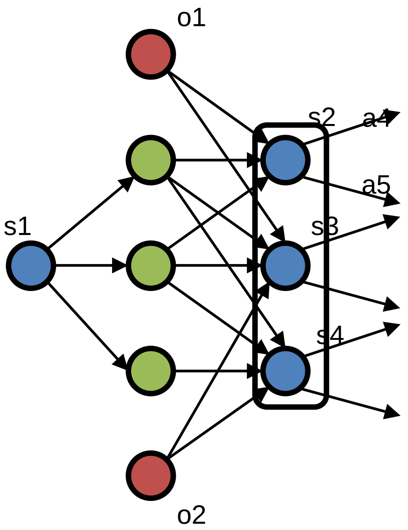


s2	s3	s4	V(a4)	V(a5)
0.2	0.1	0.7	3	2
0.7	0.1	0.2	1	7

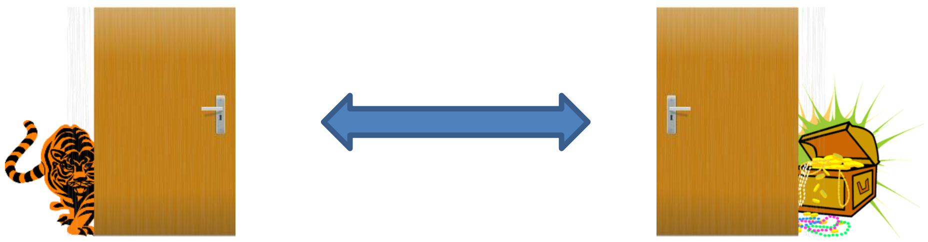
Solving Continuous State MDPs

- values can be compactly represented as a finite set of α vectors; $V = \{\alpha_0, \dots, \alpha_m\}$
  - α vector is an $|S|$ dimensional hyper-plane
    - a linear function representing utility values after selecting some fixed action
  - defines the value function over a bounded region of the belief
  - $V(b) = \max_{\alpha \in V} \sum_{s \in S} \alpha(s) b(s)$
  - V is a piece-wise linear convex function





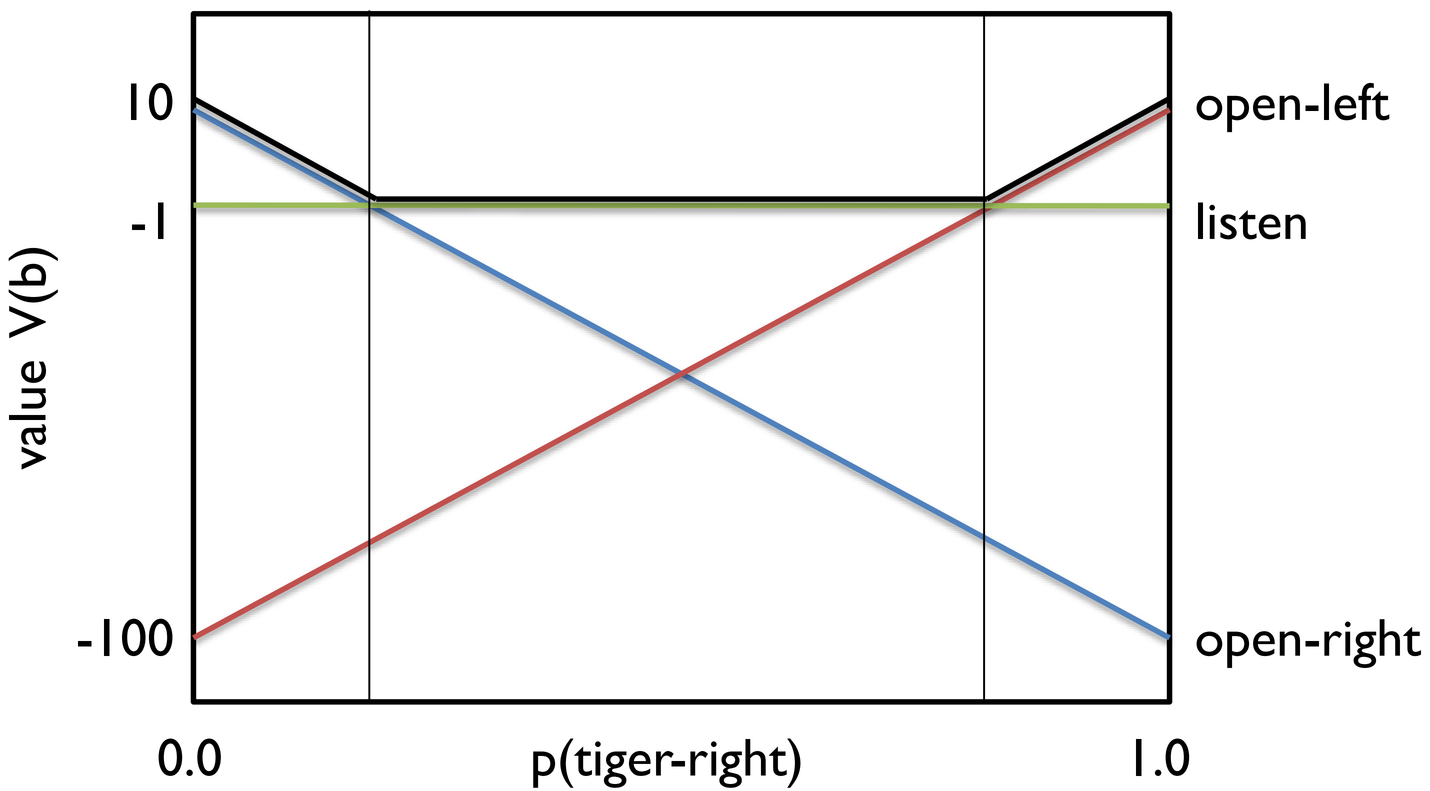
The Tiger Problem (1-step opt. policy)



open-right
[0.0, 0.1]

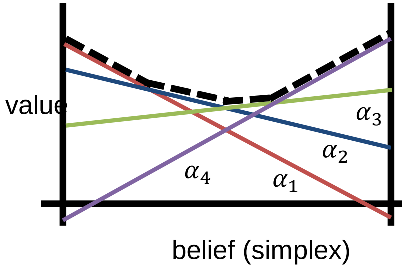
listen
[0.1, 0.9]

open-left
[0.9, 1.0]



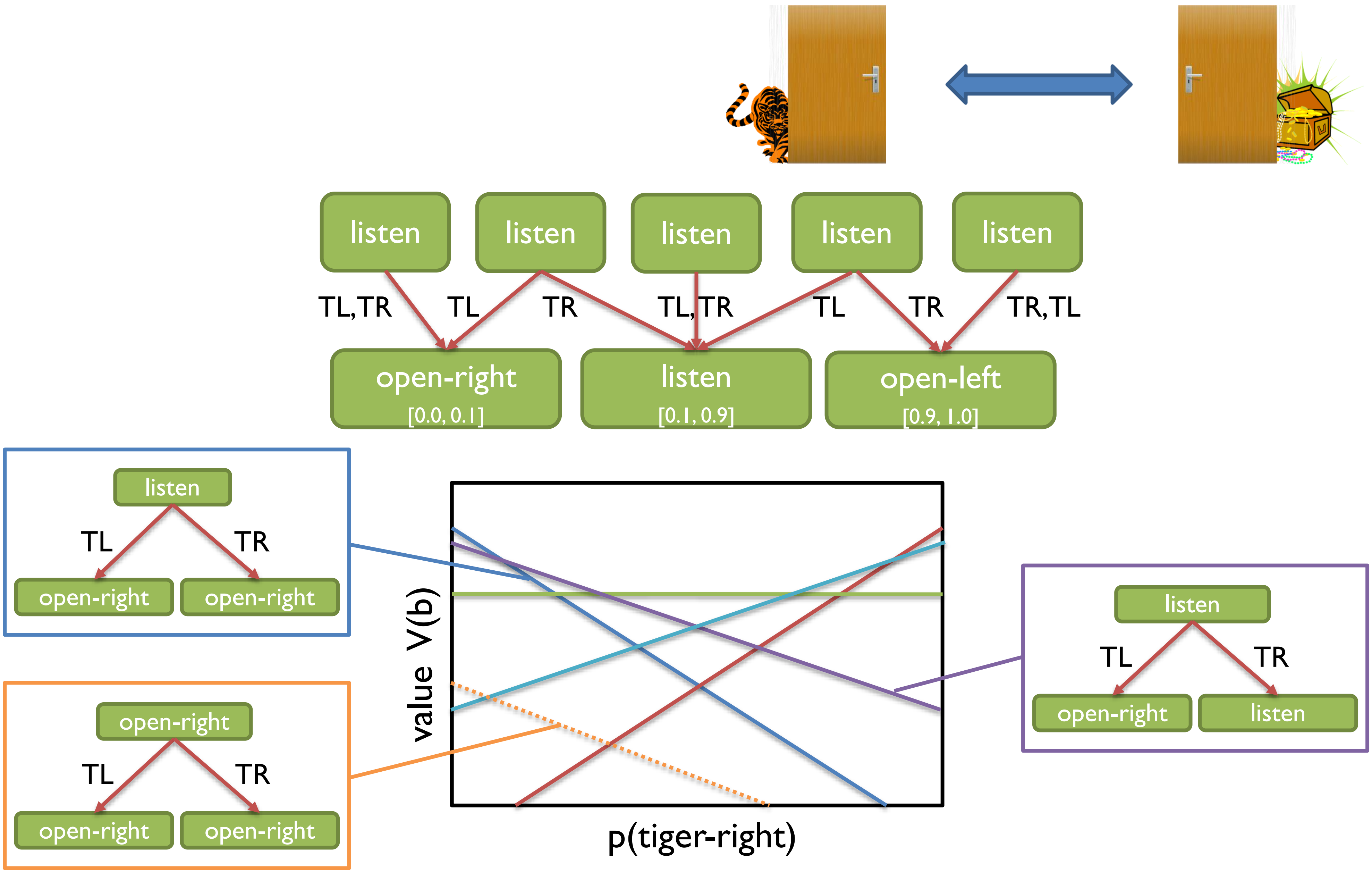
Solving Continuous State MDPs

- **Q: Can we modify value iteration algorithm to work with α functions?**

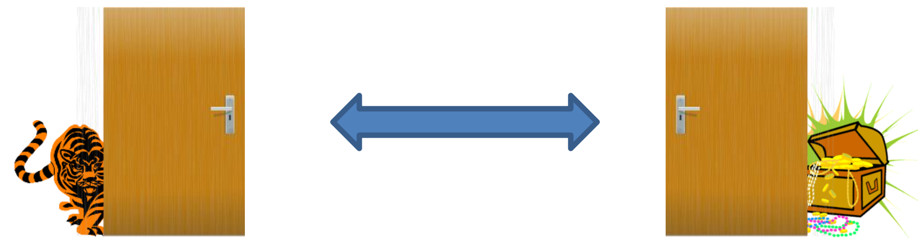


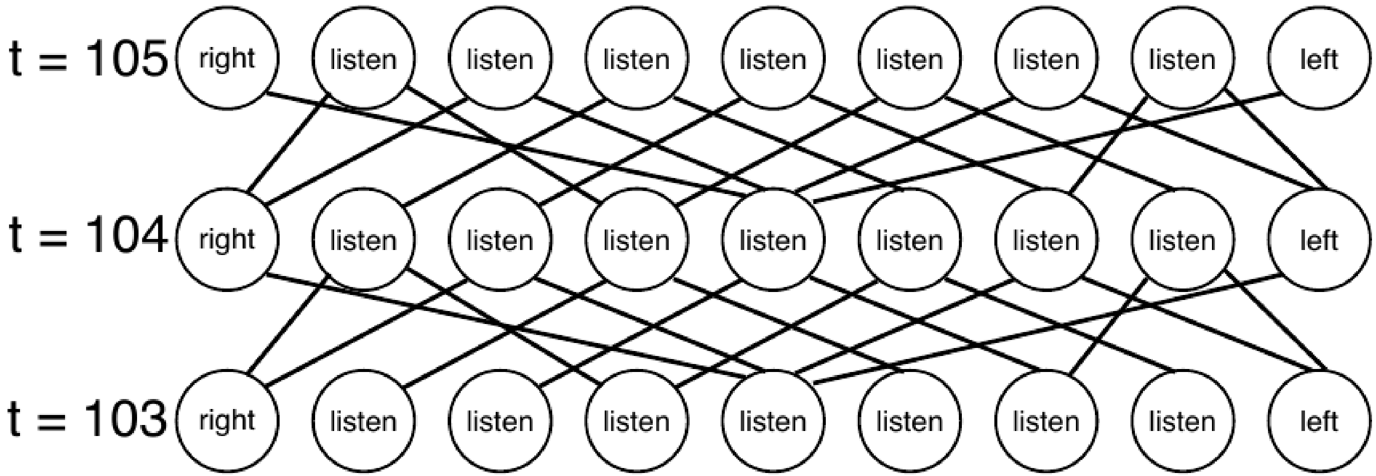
- exact value iteration for POMDPs
  - $V^t(b) = \max_{a \in A}[\sum_{s \in S} R(s,a)b(s) +$
  - $+\gamma \sum_{o \in O} \max_{\alpha' \in V^{t-1}} \sum_{s \in S} \sum_{s' \in S} T(s,a,s')\Omega(o,s',a)\alpha'(s')b(s)]$
- the above formula compute values (we need α-vectors)
  - $\alpha^{a,*}(s) = R(s,a)$
  - $\alpha_i^{a,o}(s) = \gamma \sum_{s' \in S} T(s,a,s')\,\Omega(o,s',a)\alpha_i'(s') \qquad \forall \alpha'_i \in V'$
  - $V^a = \alpha^{a,*} \bigoplus \alpha^{a,o_1} \bigoplus \alpha^{a,o_2} \bigoplus \cdots$
  - $V = \bigcup_{a \in A} V^a$

The Tiger Problem (2-step opt. policy)



The Tiger Problem (opt. policy)





After enough iterations The Tiger Problem solution converges to a **stationary policy** (not in general!).