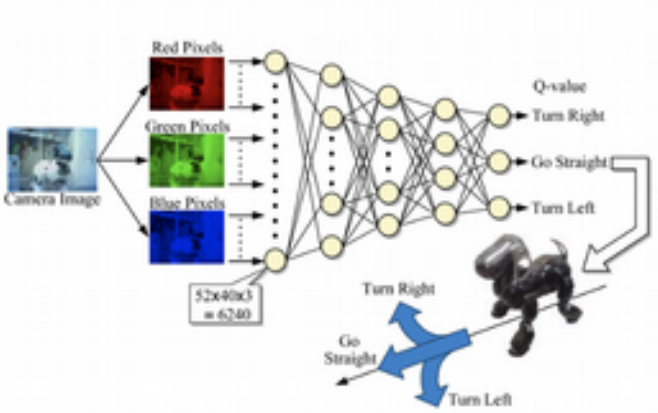
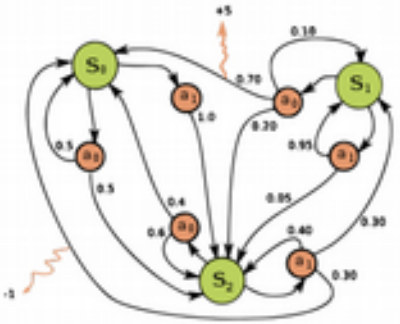
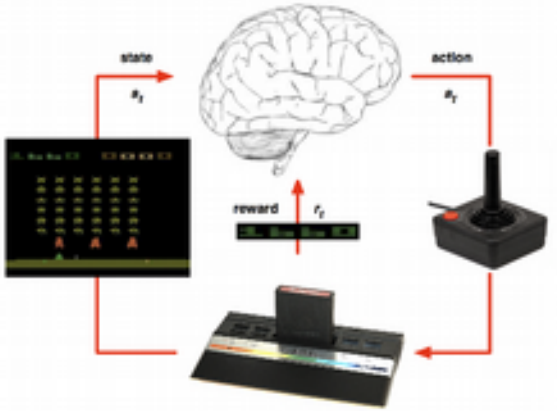
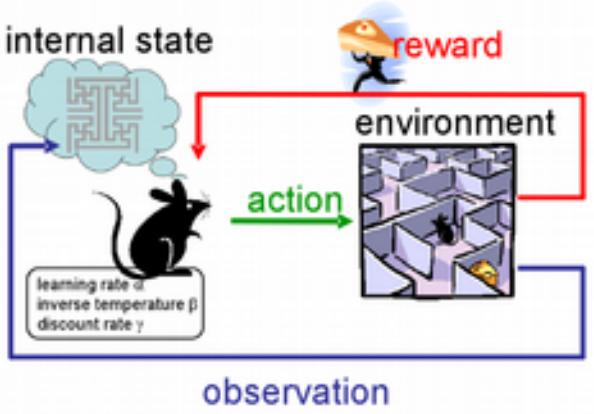
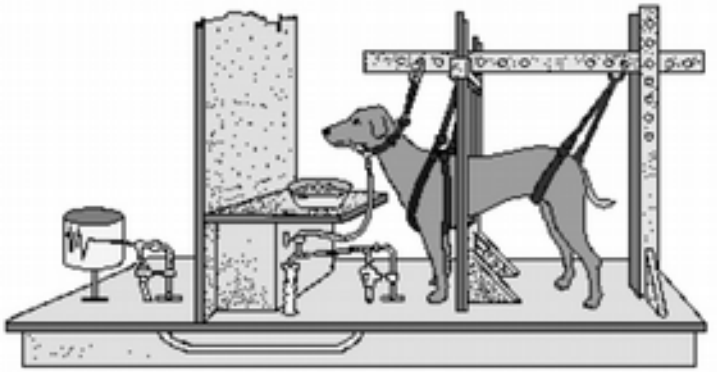
COMP 138: Reinforcement Learning











Instructor: Jivko Sinapov
Webpage: https://www.eecs.tufts.edu/~jsinapov/teaching/comp150_RL_Fall2020/

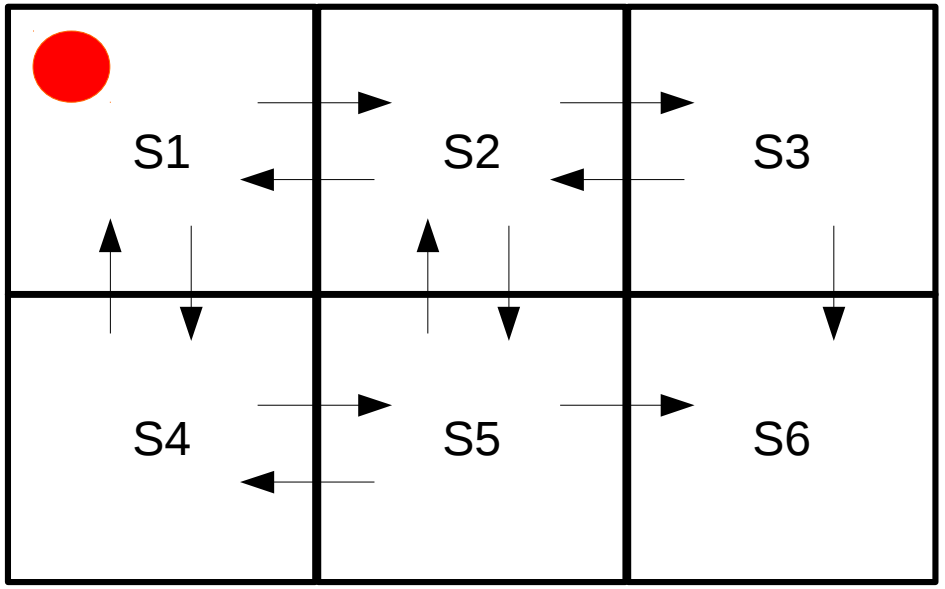
Announcements

- Homework 1 is due this Friday
- Need Help? Let Andre or I know

Reading Assignment

- Chapters 4 and 5
- Reading Responses due Tuesday before class

Q-Learning



+ 100 reward for getting to S6
0 for all other transitions

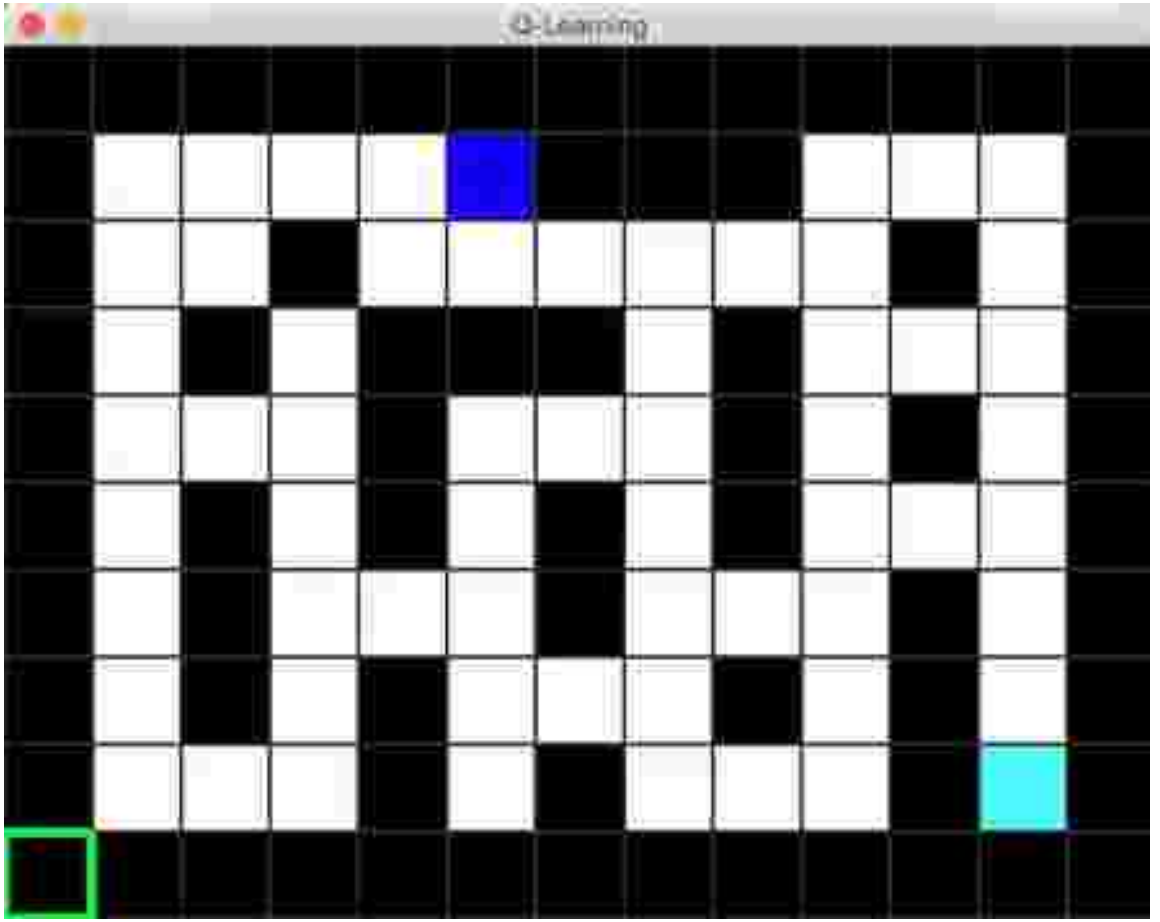
Update rule upon executing action a in state s, ending up in state s' and observing reward r :

$$Q(s , a)=r +\gamma \max a' Q(s' ,a ')$$

γ = 0.5 (discount factor)

Q-Table

S1	right	0
S1	down	0
S2	right	0
S2	left	0
S2	down	0
S3	left	0
S3	down	0
S4	up	0
S4	right	0
S5	left	0
S5	up	0
S5	right	100


Q-Learning

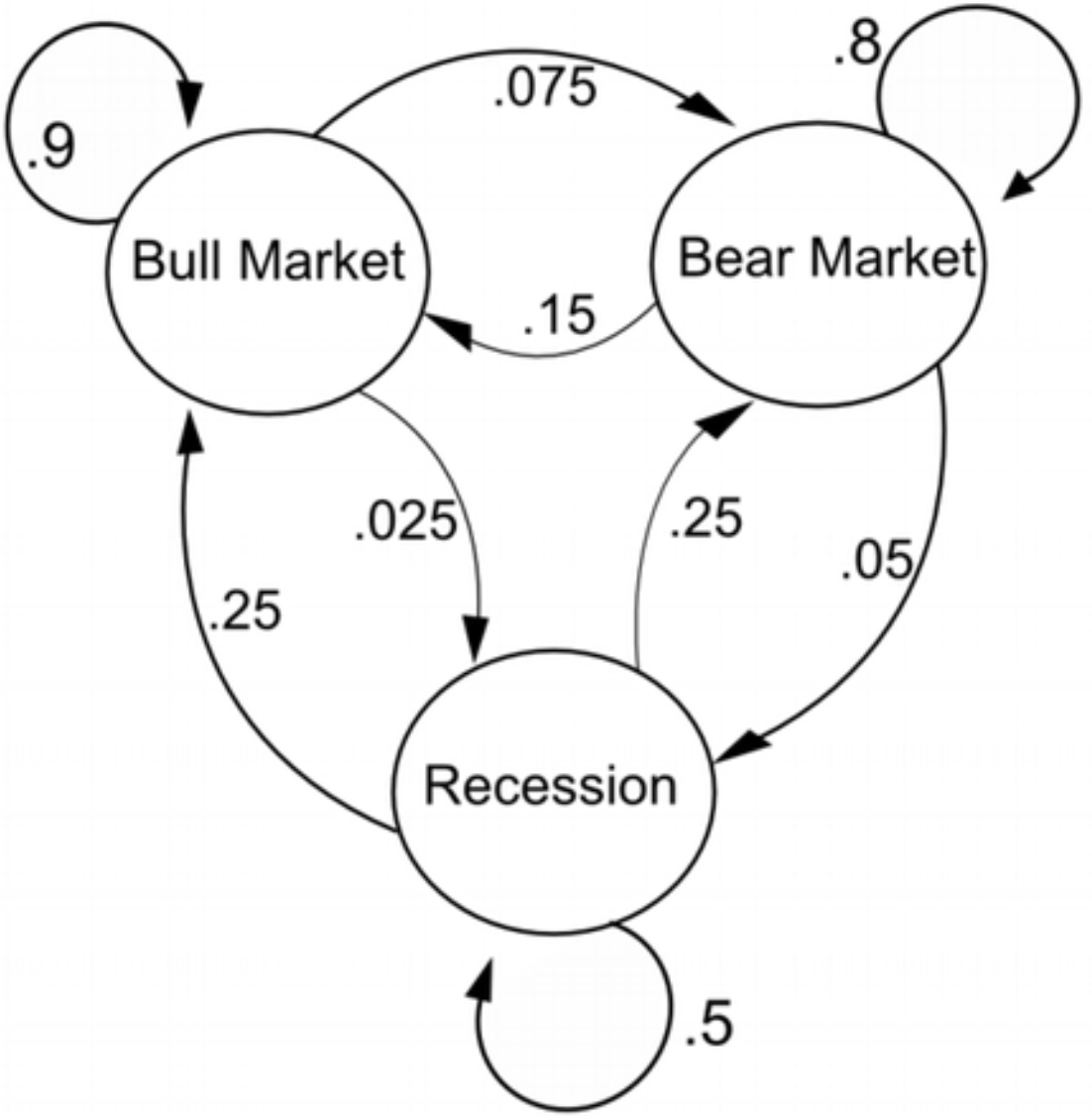
Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
  Initialize S
  Repeat (for each step of episode):
    Choose A from S using policy derived from Q (e.g., ϵ-greedy)
    Take action A, observe R, S'
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$
    $S \leftarrow S'$
  until S is terminal

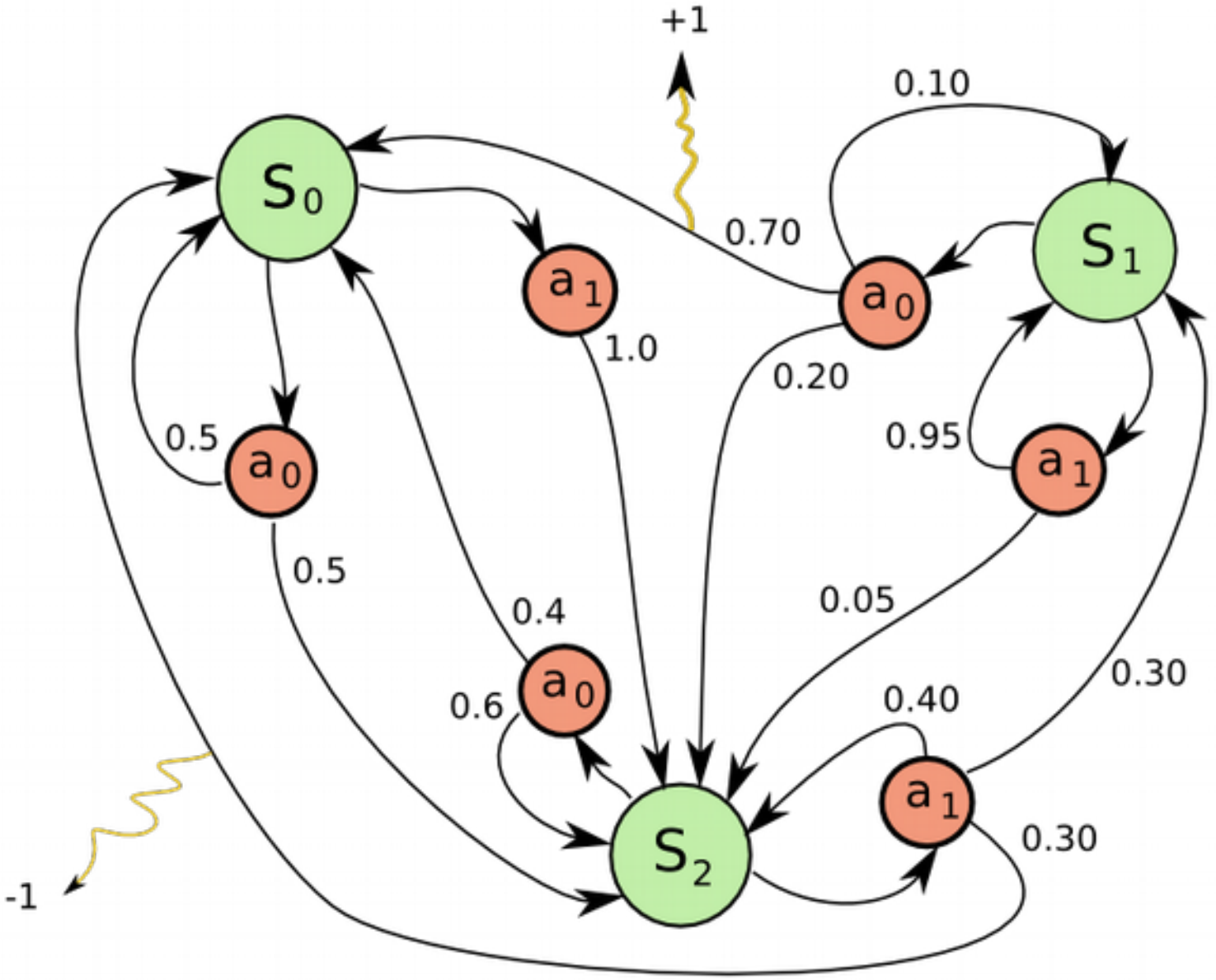
Andrey Andreyevich Markov (1856 – 1922)

[http://en.wikipedia.org/wiki/Andrey_Markov]

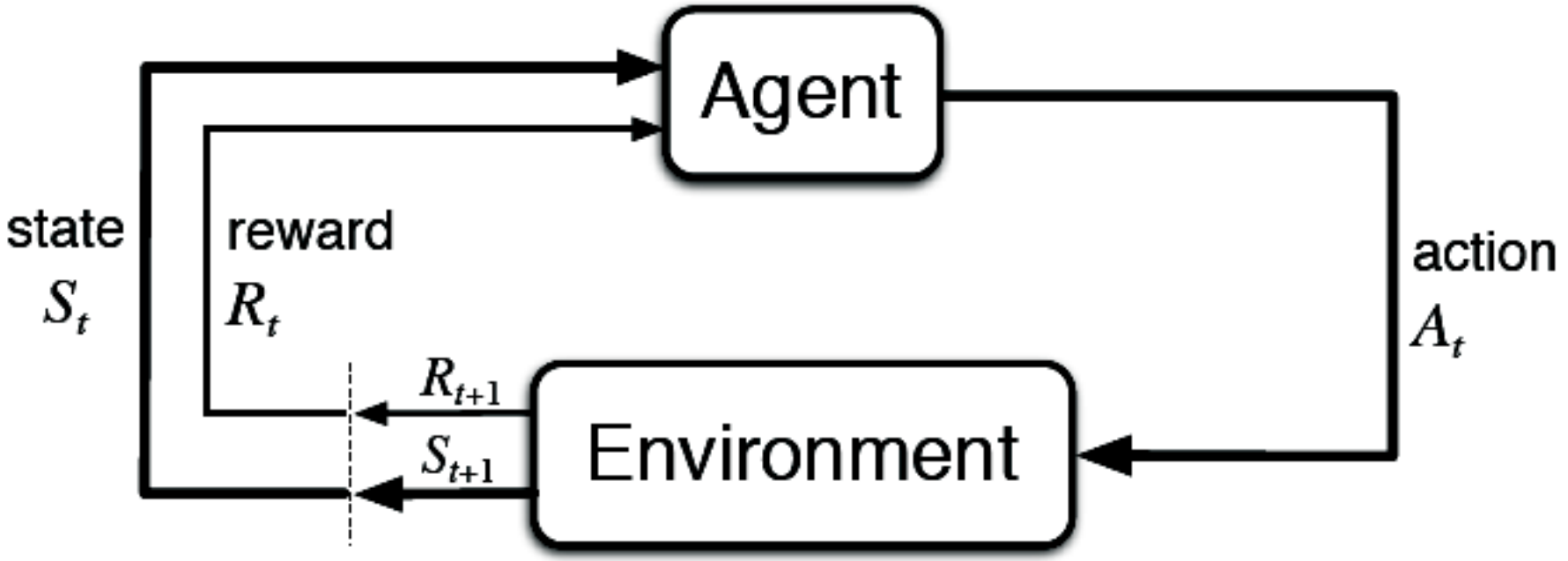
Markov Chain



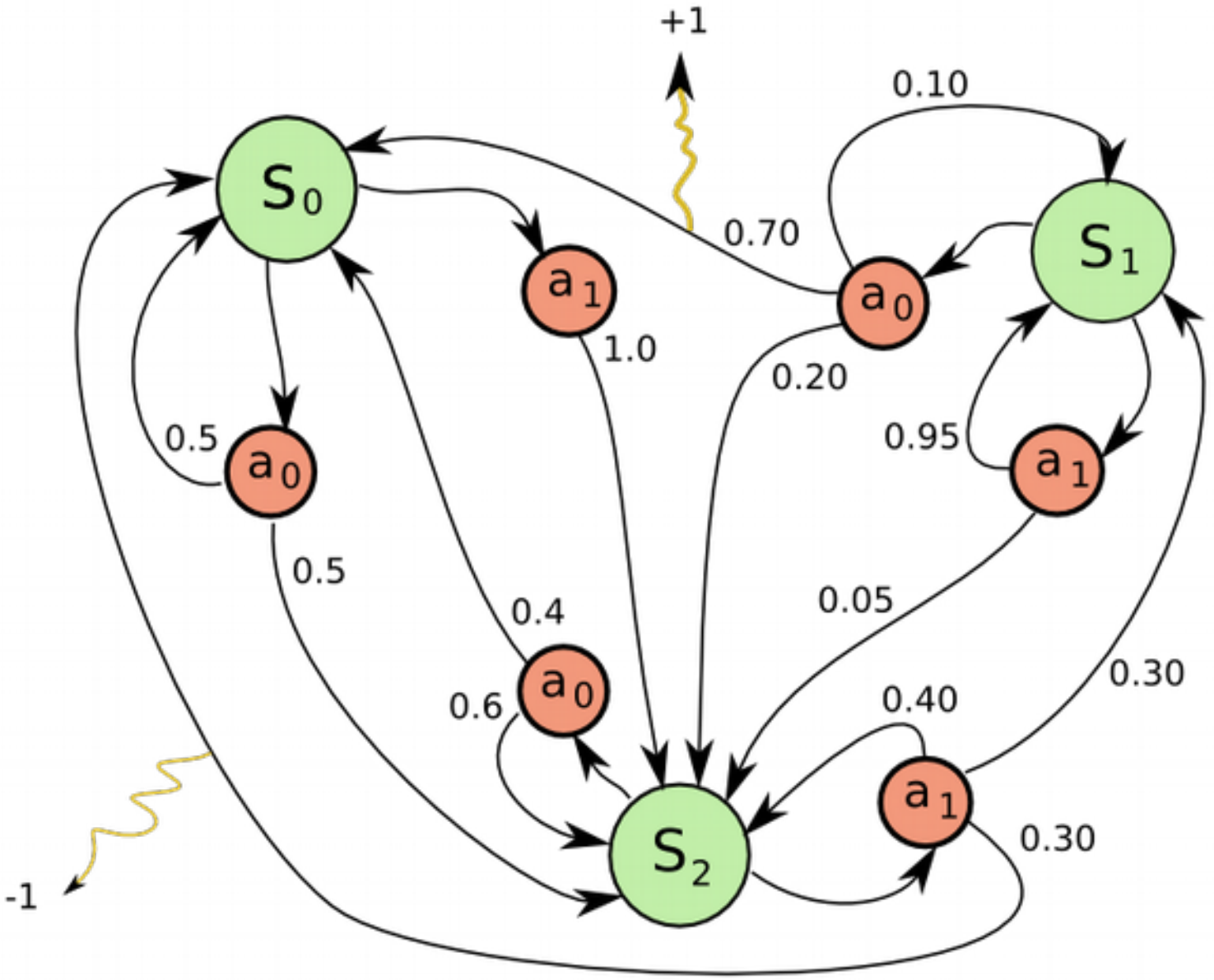
Markov Decision Process



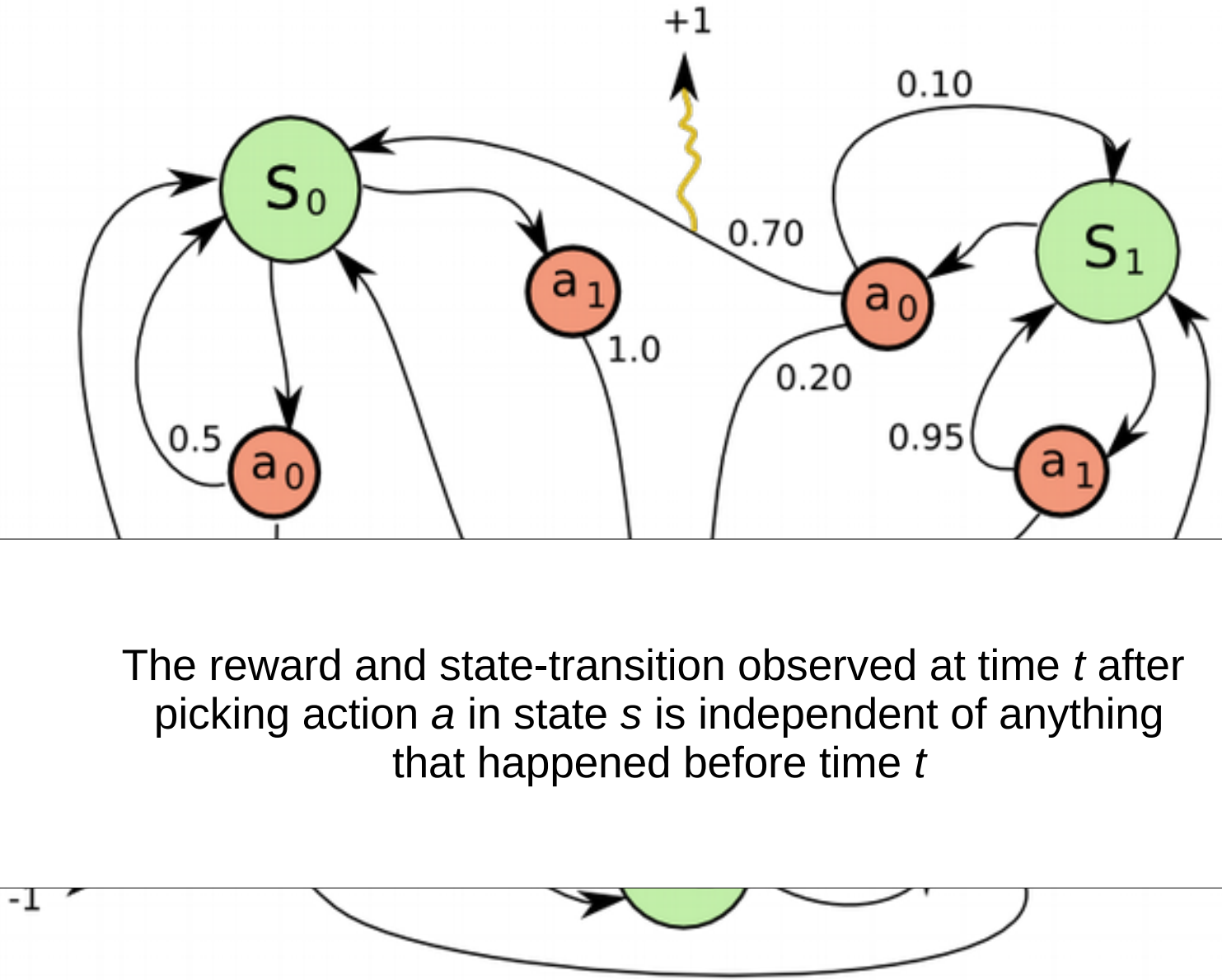
The Reinforcement Learning Problem



RL in the context of MDPs



The Markov Assumption



The reward and state-transition observed at time *t* after picking action *a* in state *s* is independent of anything that happened before time *t*

Formalism and Notation

(section 3.1)

The “boundary” between State and Agent

“… the boundary between agent and environment is typically not the same as the physical boundary of robot’s or animal’s body. Usually, the boundary is drawn closer to the agent than that. For example, the motors and mechanical linkages of a robot and its sensing hardware should usually be considered parts of the environment rather than parts of the agent.”

The “boundary” between State and Agent

“Similarly, if we apply the MDP framework to a person or animal, the muscles, skeleton, and sensory organs should be considered part of the environment. Rewards, too, presumably are computed inside the physical bodies of natural and artificial learning systems, but are considered external to the agent.”

Discussion Questions

“I found it particularly interesting when the book was discussing what the cutoff between the agent and the environment is. For example, the quote that "The general rule we follow is that anything that cannot be changed arbitrarily by the agent is considered to be outside of it and thus part of its environment" seems rather interesting. **Perhaps one question I have is what constitutes something that can be "changed arbitrarily"? What might an example of this be for a robot?**” - Jeremy

Example 3.3 and 3.4

In-Class Exercise

- See shared lecture notes document

Policies and Value Functions

Discussion Comments

“Is there an example of RL task which has a different goal rather than maximizing cumulative numerical rewards?”

- Ziyi

Discussion Comments

“Another question I have relates to the authors example of the reward function for a chess playing robot. The author explained a common example is +1 for winning a game, -1 for losing a game, and 0 for non-terminal positions (with the remarks that rewarding intermediate states may cause the agent to become proficient in taking pieces but not actually winning the game). Would it be possible to achieve the same effect by providing a small reward based on taking pieces, with a significant (maybe orders of magnitude larger) reward for winning? Would this help enable faster training of the agent (as generally taking pieces gives a strong position)?”

– Jeremy

Discussion Comments

“How can MDP deal with problems where some of the states are hidden since most of the real-world problems have hidden information?”

- Tung

Discussion Comments

“How can we know what is the most “optimal” policy for a game or an environment if there are multiple policies given the same amount of reward?”

- Tung

Discussion Comments

“I’d like to know more about how the discount factor plays an important role in the efficient training of the agent to learn a task. How different values of the discount factor influence the time needed for the agent to learn the task. I have observed that a value close to 1 is preferred, but how would the learning change if this value is decreased?”

– Yash

Discussion Comments

“As far as I understand, a robot made to solve a real-world problem such as delivering mail over the air would “trigger” a new state and consult its policy with some fixed frequency (or whenever a drastic change happens). I am wondering if it is common for the robot to adjust that frequency. In case of severe weather and strong wind, for instance, the agent might want to consult its policy more often than it would when the wind stays still. Would adjusting the frequency with which a new state is triggered be a kind of action that the agent can make?”

– Sasha