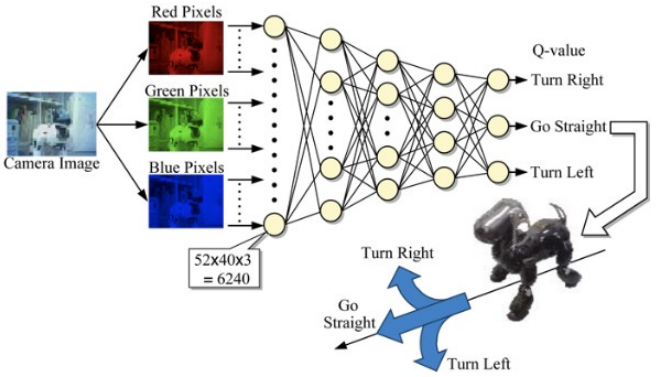
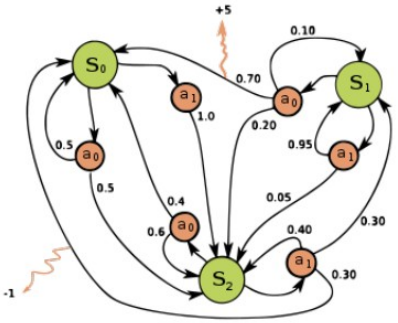
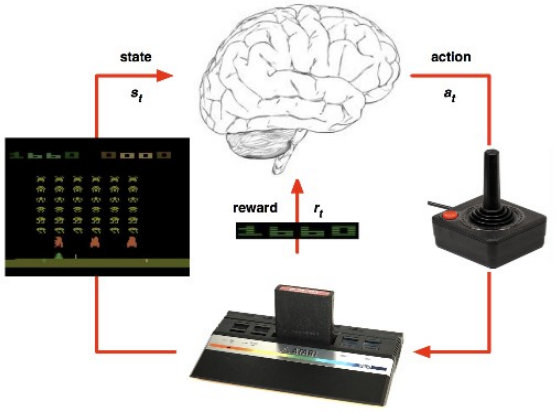
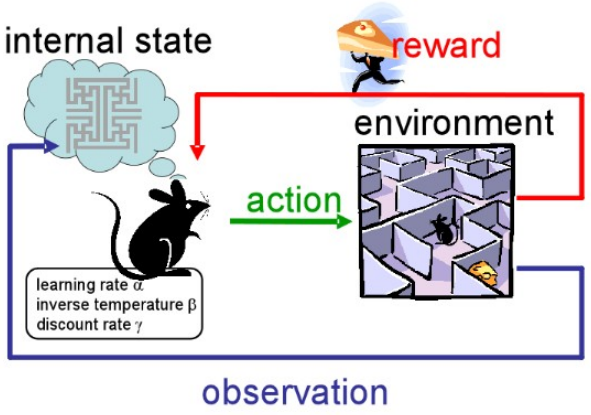
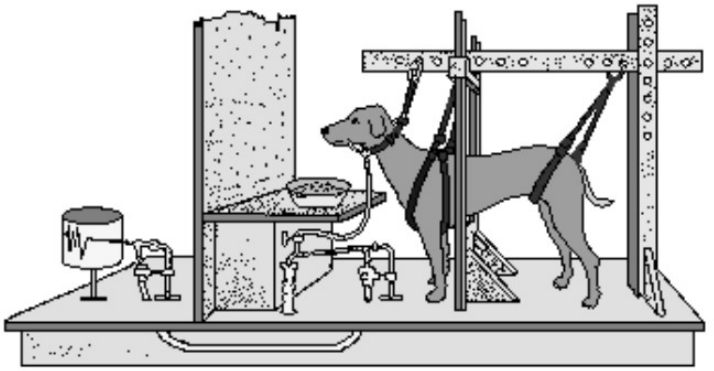
COMP 138: Reinforcement Learning











Instructor: Jivko Sinapov

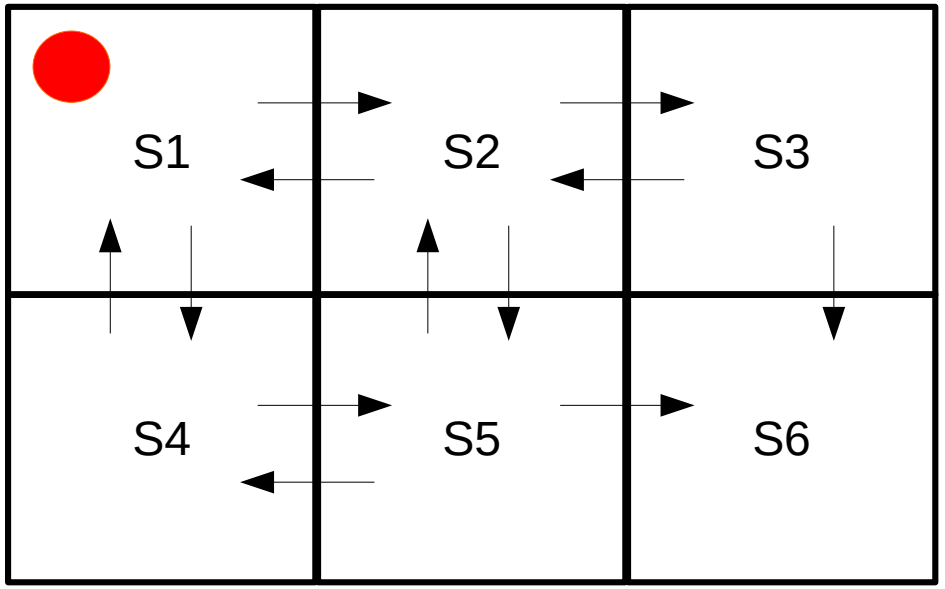
Announcements

- Homework 1 is due soon

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

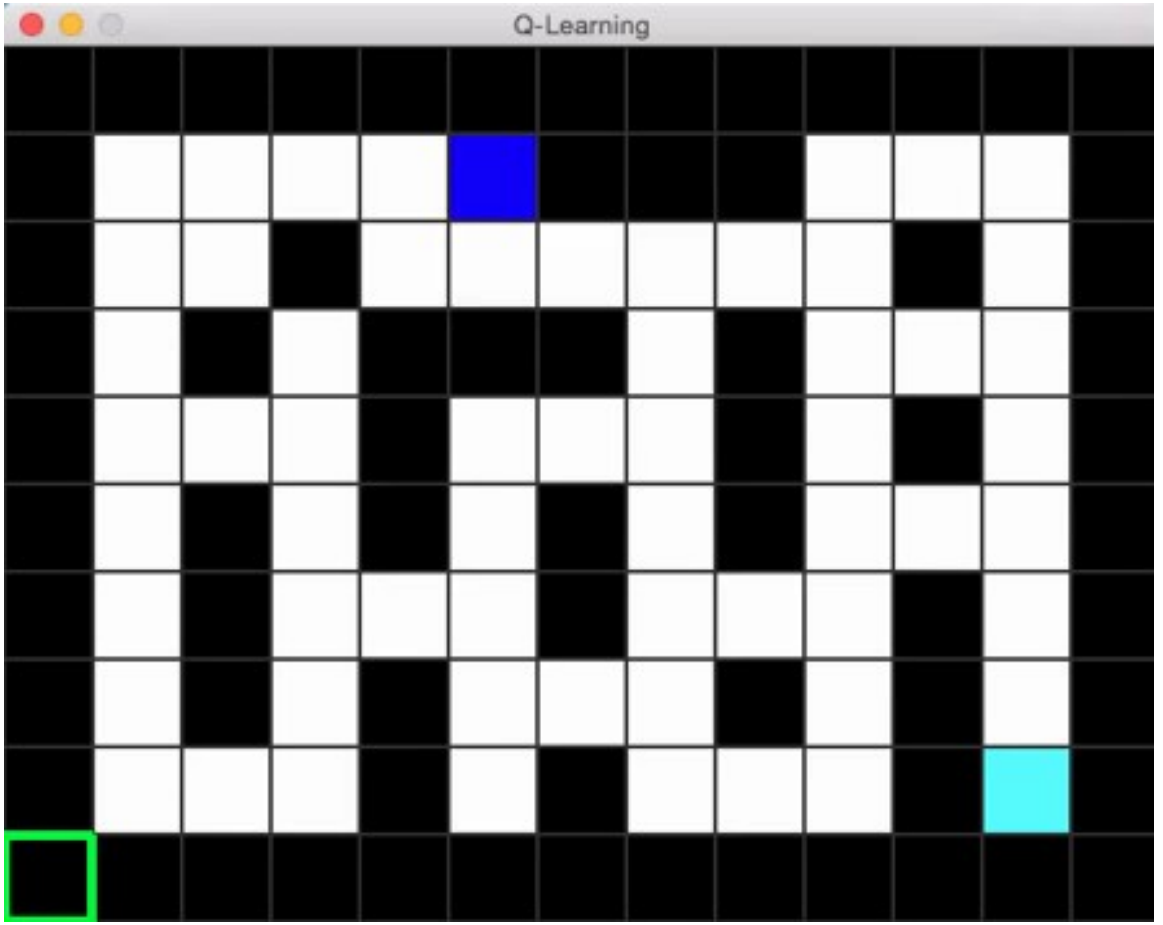
Update rule upon executing action a in state s, ending up in state s' and observing reward r :

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

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

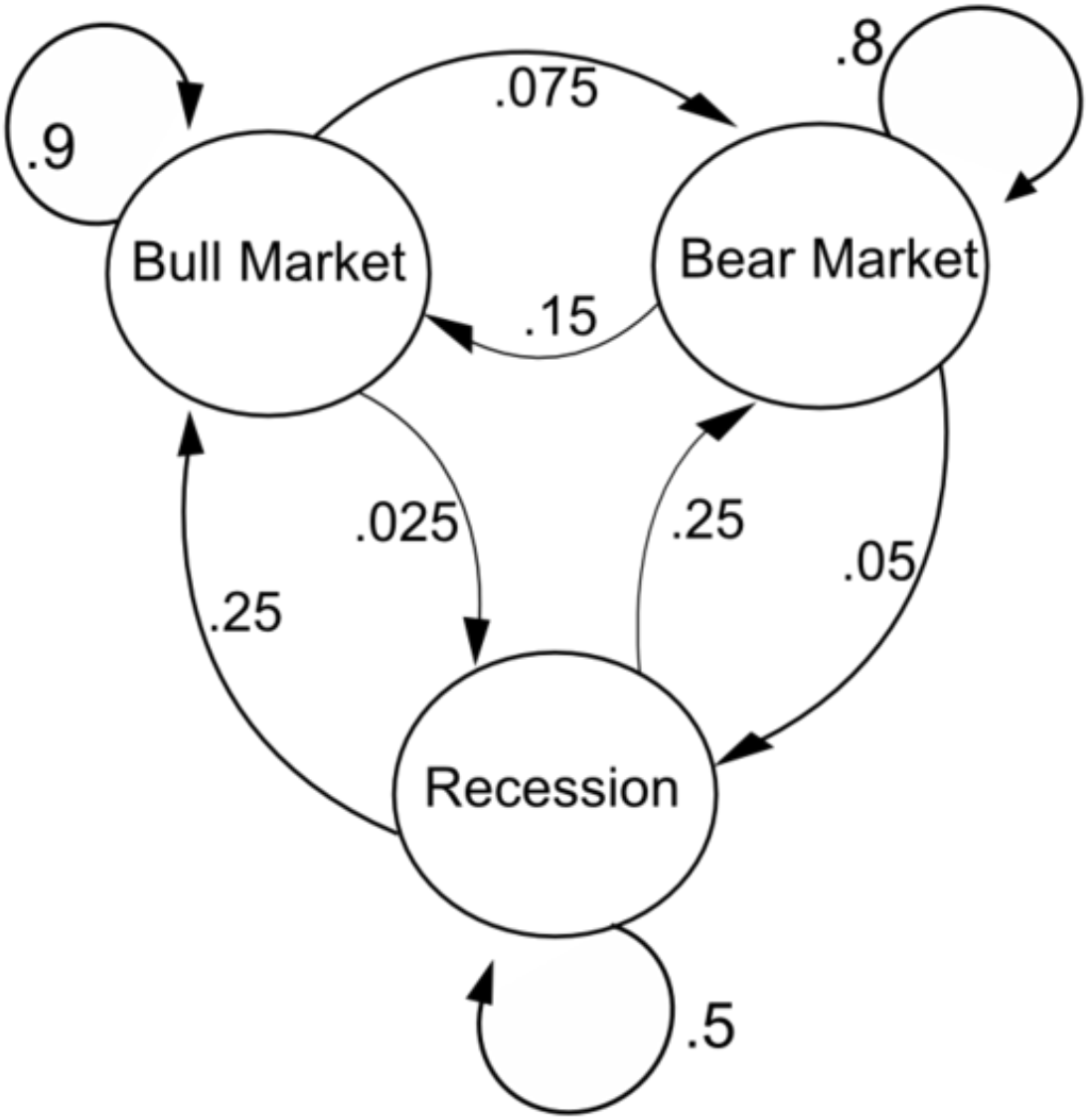
Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

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

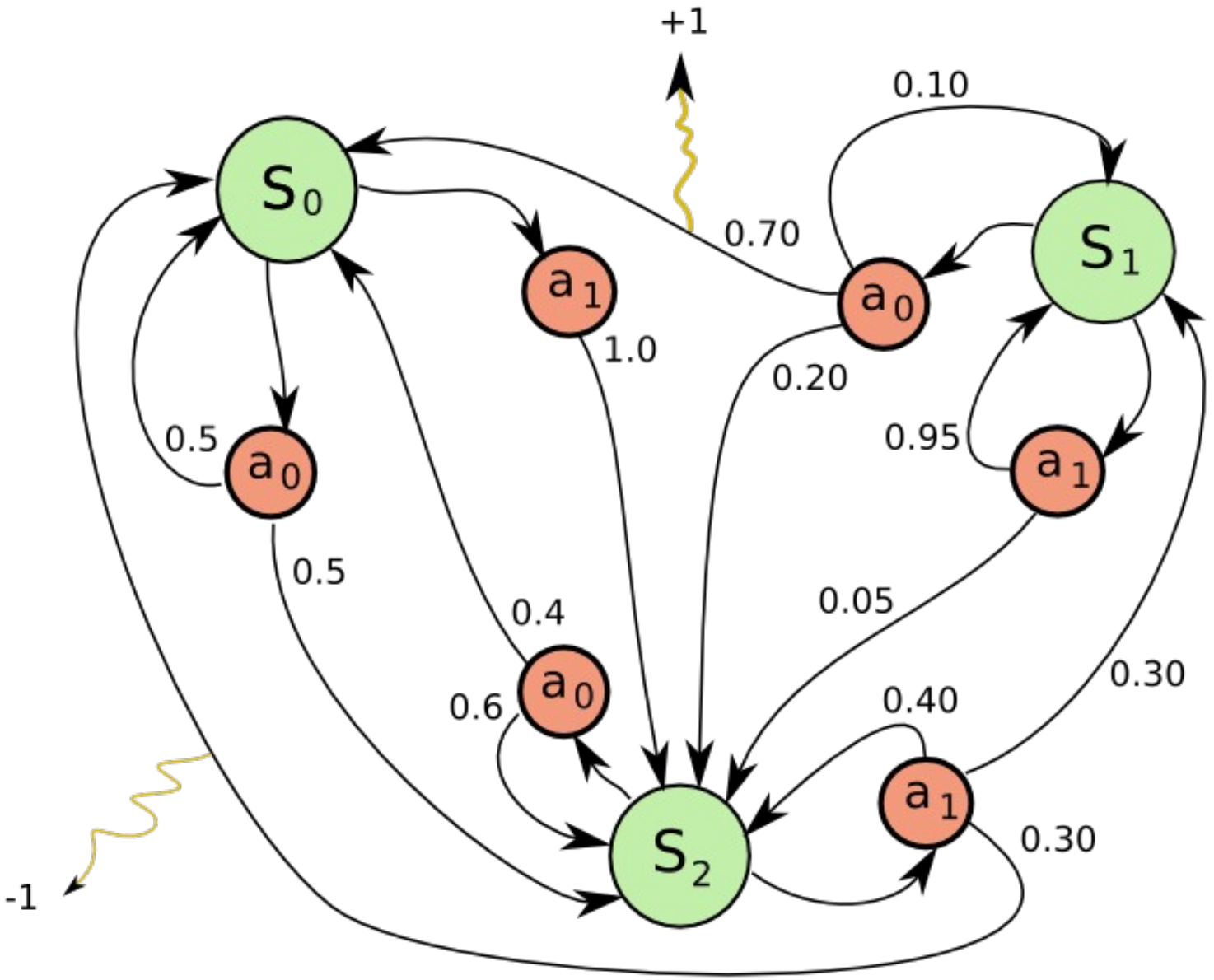
Andrey Andreyevich Markov (1856 – 1922)

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

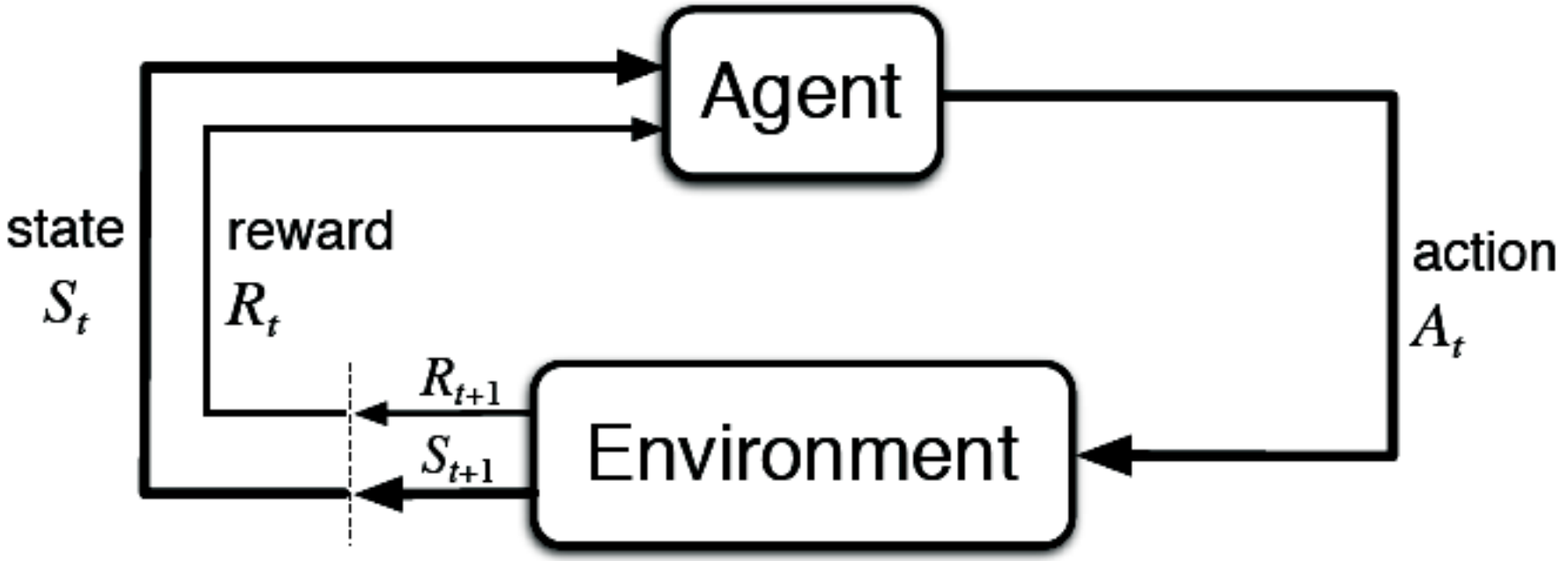
Markov Chain



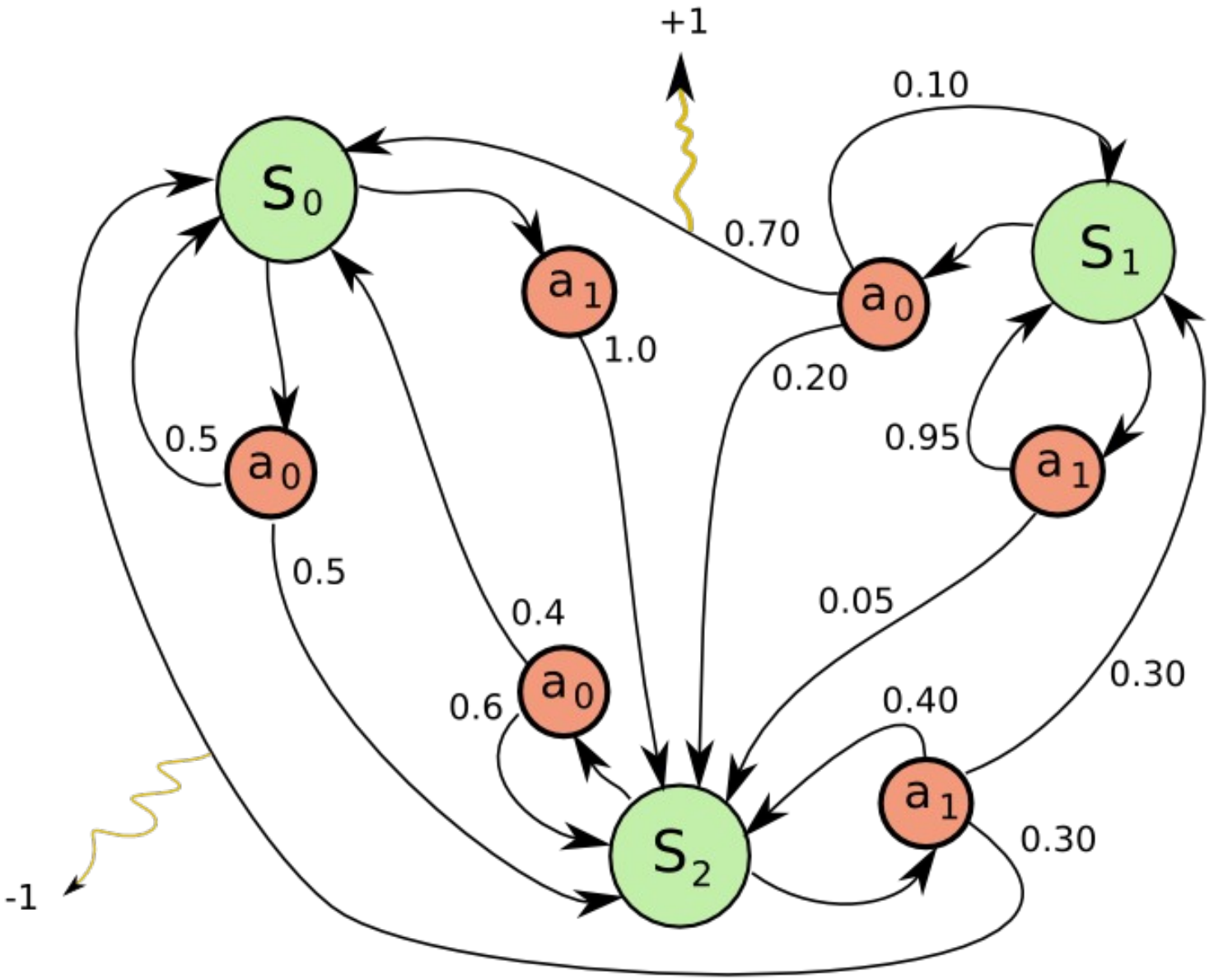
Markov Decision Process



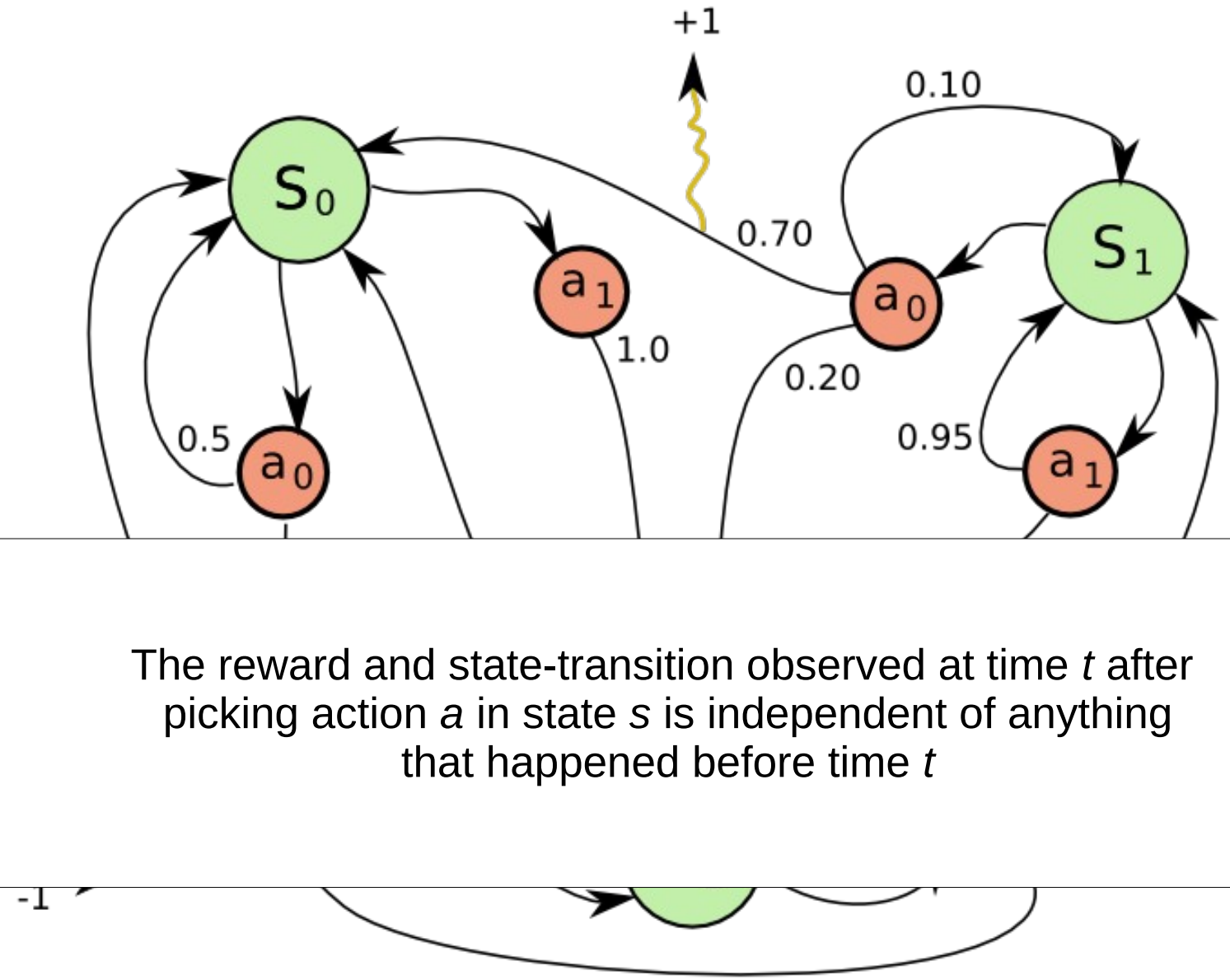
The Reinforcement Learning Problem



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.”

Example 3.3 and 3.4

In-Class Exercise

- How do we modify the MDP for the robot if the robot can get “full” after picking up some number of cans and then needs to deposit the cans before picking up more?

Policies and Value Functions

Discussion Comments

“ While it makes sense that rewards should be directly connected to goals, it seems like that might also inhibit learning. For example, the book gave the example of chess, where an agent received +1 for winning a game, -1 for losing a game, and 0 for anything else. Given the sheer number of possible states and actions in a chess game, it seems to me a like an agent would take forever to learn to play at even a passable level- many of its actions would be completely random. What tools and techniques can be applied on top of a reward formulation like this to help agents train efficiently?”

- Grayson

Discussion Comments

“The book additionally introduces a "p function", p(s', r|s,a), which is used to calculate transition probabilities. There is no name for this function given, is there one?”

- Hayley

Discussion Comments

“According to Eq.(3.2), the probability of each possible value for S_t and R_t depends on the immediately preceding state and action, not at all on earlier states and actions. But in the definition of Markov property, it is required that the state must include information about ALL aspects of the past agent-environment interaction that make a difference for the future. Is there a contradiction?”

- Qing

Discussion Comments

“What are examples of states that do not satisfy the Markov property, essentially states that have all the information that affect future? I am considering situations like those in finance where today's stock price does not contain enough data to forecast tomorrow's stock price. If we were to enhance this data with a rich history of the stock's behavior and details about the company, could this potentially transform it into a state satisfying the Markov property? How can we know that the state has all the necessary information that affects the future?”

- Song

Discussion Comments

“As I reflect on this chapter, a question that comes to mind is: How do researchers handle situations where the Markov assumption is violated, i.e., when the current state alone does not contain all the information necessary for decision-making?”

- Arya

Discussion Comments

“I thought the part about the discount rate parameter was quite interesting. With the gamma term, you can change the thinking process of the agent to either invite it to be farsighted by keeping gamma close to 1 (so there is little discount) or have it be more present-focused by changing gamma to around 0. The book states that this gamma term is in the interval [0,1], meaning that it includes 0 and 1. What would be the purpose of having gamma = 1, meaning there is no discount? Is this ever done in practice?”

- Aidan

Discussion Comments

“I do have some questions regarding the relationship between policy, value-function and state-value function for a policy. I somehow get the conceptual idea of them, but I'm still quite confused in terms of math. What's the difference between value function and state-value function? What's the difference between their functionalities? Also, how is Bellman Equation involved in getting the optimal policy? Should we just maximize the Bellman Equation?”

- Zichen

Discussion Comments

“The k-armed bandit problem was Markov decision processes but without the notion of state. Are there perhaps other useful ways to model things as Markov decision processes, but leaving out some other part (the actions and/or rewards)? This would probably be outside the scope of Reinforcement Learning.”

- Randy

Discussion Comments

“To what extent would solving the Bellman optimality equation be more or less desirable compared to more widely used Reinforcement Learning algorithms in situations where there are not as many state-action pairs (and thus, solving the Bellman optimality equation would be computationally feasible)?”

- Randy

Discussion Comments

“What are parameterized functions? What alternative do you have for maintaining value functions per state if there are too many states? A function that aggregates information about similar states together?”

- Chami

Discussion Comments

“1. How might MDPs be applied in real-world scenarios, such as business decision-making or public policy?

2. How can MDPs adapt when the environment is constantly changing or when there's a large degree of uncertainty?”

- Jianan

Discussion Comments

“Firstly, with all the Bellman Equations, we seem to already know about every possible state and their transition probabilities. But in some practice, we can't fully understand the environment from the start, so how can we derive Bellman Optimality Equations if we don't have knowledge of the states and probabilities? Secondly, to satisfy Bellman Optimality Equations, we seem to need a deterministic policy. What would happen if we introduce methods such as epsilon-greedy, in which the policy itself has some uncertainties? Can we calculate the ideal policy in that case?”

- Shijie

Discussion Comments

“In 3.6 it’s said that “A policy π is defined to be better than or equal to a policy π’ if its expected return is greater than or equal to that of π’ for all states.”. Can the condition that the returns of π are greater than π’ in most states or the sum of returns of π are greater than π’ means that π is defined to be better than or equal to a policy π’?”

- Tian

Discussion Comments

“In exercise 3.13 and 3.26, what is the four-argument p?”

- Sean

Discussion Comments

“My main question throughout the chapter was if there were perhaps ways for an agent to generate the Markov model on-the fly as it experienced the environment. From my understanding this chapter mainly focuses on a system where the agent already “knows” such a model and is then attempting to gauge the optimal path through states. What would an agent’s setup procedure for such a system look like?”

- Tanay

Discussion Comments

“I have some questions about how MDP could apply to RL problems. In this chapter, the actions are made by one agent. If we were to model chess decisions, two agents' action are both required for knowing the next state we are in. However, we need to make decisions without knowing what opponent will do, which makes us uncertain about the expected return since it depends on opponent's action. Is MDP sufficient to model this scenario or we need another RL method?”

- Zixiao

Discussion Comments

“How is the discounting factor typically determined in real-world applications?”

- Angus

Discussion Comments

“Could an embodied agent that is deployed for a long period of time still separate its tasks episodically? For example, could an agent that is delivering packages treat each delivery as a separate episode, and if so would this be desirable over treating it like a continuing task?”

- Brennan

Discussion Comments

How are reinforcement learning and control systems related ? is one concept inspired from the other? I feel like both are attempting to make a "agent " behave in an intending way . Control systems seems more focus on mathematical modeling when RL is on experimenting. Are they cases where you would combine both methods ?

- Kendrick

Discussion Comments

“I was wondering, what other approaches are there besides reinforcement learning to approximately solve MDPs?”

- Rusny