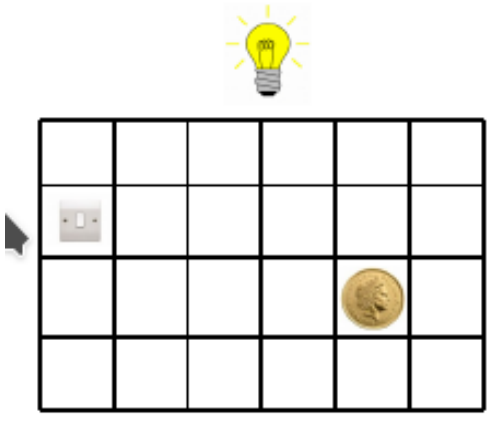
MDPs and RL

Useful Formulas

- Q-learning update: $Q_{k+1}(s,a) = Q_k(s,a) + \alpha\left(R_{t+1} + max_{a'}\gamma\, Q_k(s',a') - Q_k(s,a)\right)$.
- Sarsa update: $Q_{k+1}(s,a) = Q_k(s,a) + \alpha\left(R_{t+1} + \gamma\, Q(s',a') - Q_k(s,a)\right)$.

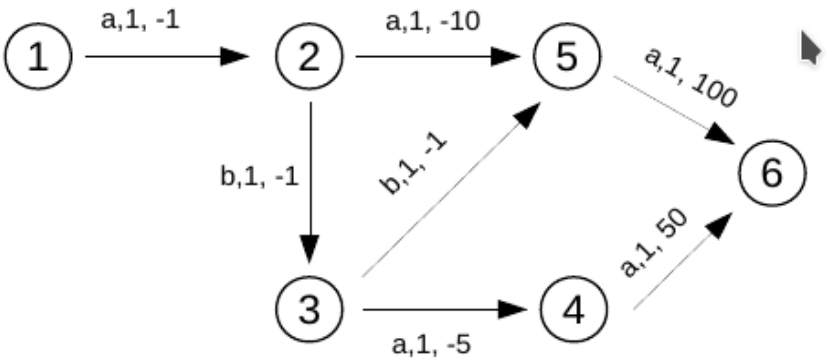
Questions

1. What is an MDP? What are the elements that define an MDP?
2. What makes a transition system Markovian?
3. What does it mean that an RL method bootstraps? Provide an example of an RL algorithm that bootstraps and one that does not.
4. An agent has to find the coin in the MDP below, and pick it up. The actions available to the agent are move up, down, left, right, toggle switch and pick up. The action toggle switch turns on and off the light in the room, and succeeds only if executed in the square with the switch, while it does not do anything anywhere else. The action pick up picks up the coin if executed in the square with the coin and if the light is on, while does nothing anywhere else, or with the light off. How would you model this domain so that the representation is Markovian?



Note on notation: in the following MDPs, each state is labeled with an id. Each transition is labeled with the name of the corresponding action, the probability of landing in the next state, and the reward for that transition. If a state has no outgoing edges, it is an absorbing state.

5. Calculate the action-value function that Sarsa and Q-learning would compute on the following MDP, while acting with an ε-greedy policy with and ε = 0.1 and γ = 0.5.



6. Calculate the action-value function that Q-learning and Sarsa would compute on the following MDP, with $\gamma = 0.5$ and $\varepsilon=0.1$.