Computational Intelligence in Games Summer 2017
Prof. Dr. Sanaz Mostaghim, Alexander Dockhorn

2. Exercise Sheet

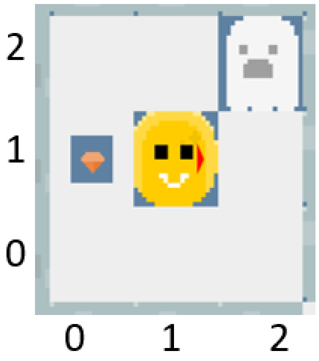
Assignment 8 Agent-Environment Interface

Consider the Pac Man game in a small environment as in the picture. An agent can win the game, if all items are collected. The game consists of the elements:

1) **Agent:** Pac Man
2) **Environment:** ghosts, items, walls
3) **Actions:** left, right, up, down (neutral)

a) What is the reward for the agent for learning to win the game? Give an example.

b) What could be the appropriate state observations? Give an example.



Assignment 9 Player Policies

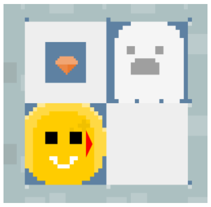
How does the policy-mapping look like, given the states (Pac Man's position, ghost's position, item's position), actions and rewards from Assignment 1? Give some example combinations.

Action a	Own pos.	Ghost's pos.	Item's pos.	$\pi(a\|s)$

What problem could arise using these states and this policy?

Assignment 10 Markov Decision Process

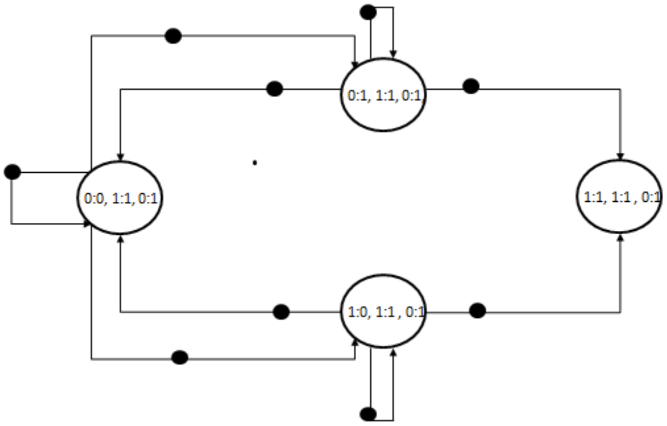
a) Complete the table with probabilities and expected rewards for the finite MDP of the Pac Man example using the state representation (Pac Man's pos., ghost's pos., item's pos.) and a map of size 2x2. Assume that the ghost is moving in each direction with the same probability.



s	s’	a	$p(s' \mid s, a)$	$r(s, a, s')$
(0,0); (1,1); (0,1)	(1,0); (0,1); (0,1)	Right		
(0,0); (1,0); (0,1)	(0,1); (1,1); (0,1)	Up		
(0,0); (1,0); (0,1)	(0,1); (0,0); (0,1)	Up		
(0,0); (1,0); (0,1)	(0,1); (1,0); (0,1)	Up		

b) Complete the transition graph below assuming that the ghost is not moving. Why are some transitions missing?

c) How would the graph change if the ghost would be able to move?

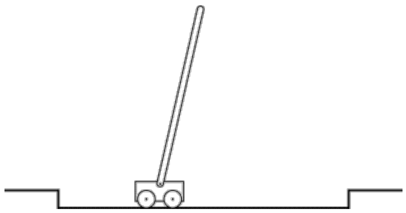


Assignment 11 (Discounted) Return

The objective here is to apply forces to a cart moving along a track so as to keep a pole hinged to the cart from falling over. A failure is said to occur if the pole falls past a given angle from vertical or if the cart runs off the track. The pole is reset to vertical after each failure.

a) This task could be treated as episodic, where the natural episodes are the repeated attempts to balance the pole. If the reward would have been +1 for every time step on which failure did not occur, what would be the meaning of the return at each time?

b) Alternatively, we could treat pole-balancing as a continuing task, using discounting. In this case the reward would be -1 on each failure and zero at all other times. What would be the meaning of the return at each time?



Computational Intelligence in Games Summer 2017

Prof. Dr. Sanaz Mostaghim, Alexander Dockhorn

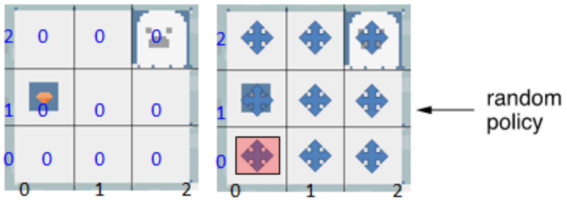
Assignment 12 Iterative Policy Evaluation

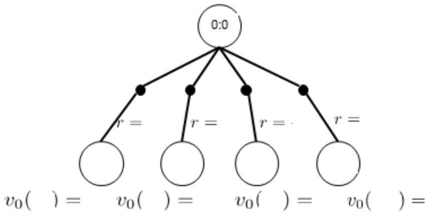
Consider the following for the Pac Man game:

- Nonterminal states: empty cells
- Four possible actions in each state: up, down, left, right
- the actions that take the agent off the grid, leave the state unchanged and give a reward of -1
- This is an undiscounted task: $\gamma = 1$
- The reward is 0 on all transitions, except on collision with the ghost (-999), and item collection ($+100$)
- The ghost is not moving.
- Pac Man follows the equi-probable random policy (all actions equally likely), for all s $\pi(a|s) = {}^1\!/_4$

a) Complete the values for the value function for k = 1 and the given backup diagram.

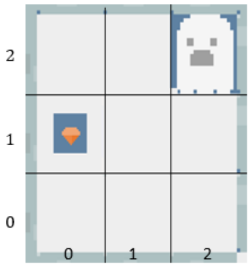
- $k = 0$:





$$v_{k+1}(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) \left[r + \gamma v_k(s')\right]$$

b) Compute the values for k=1:



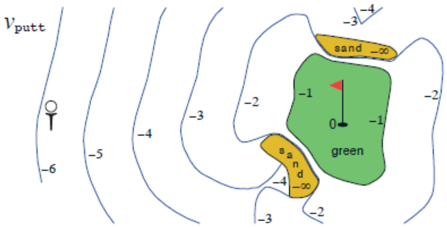
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG INF FAKULTÄT FÜR INFORMATIK

Computational Intelligence in Games Summer 2017
Prof. Dr. Sanaz Mostaghim, Alexander Dockhorn

Assignment 13 Optimal Value Functions

To formulate playing a hole of golf as a reinforcement-learning task, we count a penalty (negative reward) of -1 for each stroke until we hit the ball into the hole. The state is the location of the ball. The value of a state is the negative of the number of strokes to the hole from that location.

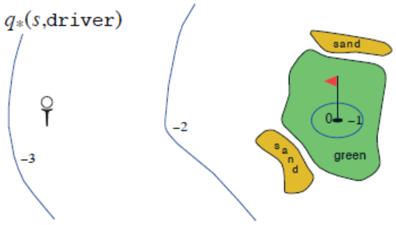
Our actions are how we aim and swing at the ball, of course, and which club we select. Let us take the former as given and consider just the choice of club, which we assume is either a putter or a driver.



The figure on the right shows a possible state-value function, *vputt*(s), for the policy that always uses the putter. The terminal state in-the-hole has a value of 0. From anywhere on the green we assume we can make a putt; these states have value of -1. Off the green we cannot reach the hole by putting, and the value is greater.

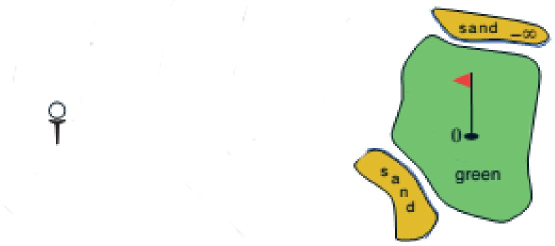
We can hit the ball farther with driver than with putter, but with less accuracy. Putt succeeds anywhere on the green. $Q^*(s, \textit{driver})$ is a possible optimal action-value function for using driver first, then using whichever actions are best.

We can reach the hole in one shot using the driver only if we are already very close; thus the -1 contour covers only a small portion of the green.



If we have two strokes we can reach the hole from much farther away, as shown by the -2 contour. In this case we don't have to drive all the way to within the small -1 contour, but only to anywhere on the green; from there we can use the putter. The -3 contour is still farther out and includes the starting tee. From the tee, the best sequence of actions is two drives and one putt, sinking the ball in three strokes.

Draw or describe the contours of the optimal action-value function for putting, $Q^*(s, \textit{putter})$,for the golf example (using putter first, then using whichever actions are best.)


sand −∞
0
green
sand