Reinforcement Learning 2014/2015

Tutorial 2 (week 4)

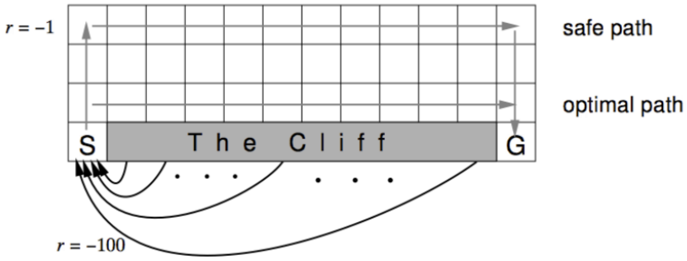
1. Discuss on-policy and off-policy RL for the 2D walker problem. Strengthen your argument by a simulation of a agent moving in an open space of $N \times N$ squares ($3 < N \lessapprox 10$) as well as in a maze (containing a fraction of inaccessible or strongly negatively rewarded squares) of the same dimensions. How do

   - type of algorithm (i.e. whether on-policy or off-policy)
   - initialisation
   - exploration
   - parameters and parameter decay schemes
   - problem size

   influence the solution? Trying out several combinations of the setting of the algorithm should be a group effort. What variant of the algorithm turns out (or is likely) to be most efficient?

2. In the grid-world example, rewards are positive for goals, negative for running into the edge of the world, and zero otherwise. Are the signs of rewards important? Prove using the equation for expected discounted return (Eq. 3.2 in S+B) that adding a constant, C, to all the rewards adds a constant, K, to the values of all states. So, it does not affect the relative values of any states under any policies. What is K in terms of C and γ?
   Now, consider adding a constant C to all rewards in an *episodic* task, such as running a maze. Would this differ from the above case? Why or why not? Give an example to make your point.

3. Discuss how on-policy and off-policy reinforcement learning algorithms solve the cliff-walking problem (see the figure below and lecture RL05).



4. Consider the grid world example in Chapter 4 of the Sutton and Barto book (Example 4.1). Suppose a new state 15 is added just below state 13, and its actions, *left, up, right, down*, take the agent to states 12, 13, 14 and 15 respectively. Assume that the transitions from the original states are unchanged. What then is $V^{\pi}(15)$ for the equiprobable random policy?
Now suppose the dynamics of state 13 are also changed, such that action *down* from state 13 takes the agent to the new state 15. What is $V(15)$ for the equiprobable random policy in this case?

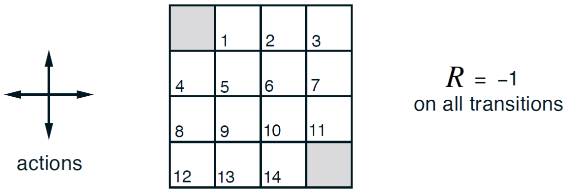


Figure from Sutton and Barto, 2nd edition

5. Recall the set-up of an RL algorithm. Considering algorithmic details, general feasibility and computational complexity as well as the options implied by the previous problem, discuss, in as much details as reasonably possible, applications of RL to problems such as (but not excluding further examples)

   (a) Finding your way in a maze

   (b) Finding your way inside the Forum

   (c) Balancing a pole

   (d) Playing Tetris

   (e) Playing Kendama

   (f) Playing football

   (g) Part-painting problem

   (h) Routing network traffic

   (i) Running a shop

   (j) Running a big company

   (k) Guiding a robot arm to a target

   (l) Coaching a team of soccer robots

   (m) Driving a car

   (n) Driving assistance at intersections

   (o) HCI (Spoken Dialogue Systems)

   (p) Equity trading

   (q) Organising your day

   (r) Biological modelling