Reinforcement Learning WS 2024
TAs: Pavel Kolev, Rene Geist, Pierre Schumacher
Due-date: **05.11.2024, 14:00** (upload to ILIAS as one file)
Filename: homework3-NAME1-NAME2-...zip
Homework Sheet 3
October 26, 2024

Regarding submission: Please submit *code* (the changed files) and a *pdf* with the answers in a zip/tgz file and remember the naming convention.

1 Recap

Try to anwser without looking at the lecture notes

(a) What is an Markov reward process (MRP)?

(b) How can I reduce a Markov Decision Process (MDP) to a MRP?

(c) MRPs can be solved in closed form. Why can I not do it with an MDP?

2 TD(λ) updates

Just as the return can be written recursively in terms of the first reward and itself one-step later ($G_t = R_{t+1} + \gamma G_{t+1}$), so can the λ-return G_t^λ.

(a) Start by Derive the analogous recursive relationship from the equation on slide 25, and slide 28 of the lecture notes for n-step return $G_t^{(n)}$. *Hint:*[1]

(b) Now derive the analogous recursive relationship for the λ-return G_t^λ. *Hint:*[2] (This is a difficult exercise)

3 Policy Iteration

Use the codebase from last week's exercise (gridworld.zip) on value iteration. Short descriptions of all the files contained in the archive and an instruction for how to run the code can be found in exercise sheet 1 and 2.

(a) Create a new class *PolicyIterationAgent* in `agent.py` and implement policy iteration. You can use the *ValueIterationAgent* class as a template for the new class. To select the *policy iteration agent* in the command line interface using the agent switch (`-a` or `--agent`), you have to edit `gridworld.py` and add an appropriate option after line 237.

[1] you need $G_{t+1}^{(n-1)}$

[2] you will need the solution to a geometric series, and that $G_{t+1}^{(0)} = V(S_{t+1})$

(b) How many rounds of policy iteration are needed before the start state of `MazeGrid` becomes non-zero? What happens if you switch off the noise (`-n 0,0`)?

(c) How many iterations of policy iteration do we need for the algorithm to converge to an optimal policy? Also try without noise.

(d) Compared with value iteration, what are the advantages and disadvantages of policy iteration? Give a detailed list of the pros and cons.