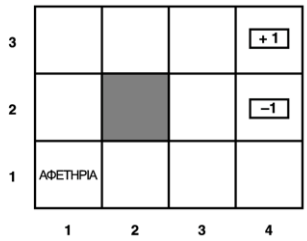
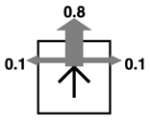
MSc in Artificial Intelligence

Grid World MDPs

Stavrianos Konstantinos





Algorithm

Markov Decision Processes algorithms are the first approach for reinforcement learning and this project computes the problem of grid world example. The agent starts from a state and we are trying to find the path till one terminal state, based on A star algorithm. The heuristics of this task will be computed from the value iteration algorithm with bellman equations. The main characteristic of MDPs is that states that given the present, the future is conditionally independent of the past.

First of all the parameters of this task are that the environment is stochastic, that means that actions aren't obsolete, means noise added on agent's choices and is fully observable. The computations are in infinite horizon and the transition model is known.

A MDP is defined by these components:

- Set of possible States: S={s0,s1,...,sm}
- Initial State: s0
- Set of possible Actions: A={a0,a1,...,an}
- Transition Model: T(s,a,s')
- Reward Function: R(s)

The discount factor describes the preference of the agent for the current rewards over future rewards. We can use the reward given at each state to obtain a measure of the utility of a state sequence. We define the utility of the states: $U_h=R(s_0)+\gamma R(s_1)+\gamma_2 R(s_2)+...+\gamma_n R(s_n)$

To compare the utility of the states we use the following math type: $U(s)=E[\infty\sum_{t=0}\gamma_t R(s_t)]$

The choice for the best action is to pick the max utility from the states and this value of the s is correlated the utility of his neighbor's. $U(s)=R(s)+\gamma \max_a \sum_{s'} T(s,a,s')U(s')$

This mathematical expression means that the utility of the state is the reward of this state plus discount factor gamma picking the maximum from the sum of transition probabilities multiplied with neighbor's utility. We referred that the iterations are infinite but there is a stopping criteria after we reach to equilibrium $U_{k+1}-U_k||<\epsilon(1-\gamma/\gamma)$. Taking into account the utilities between two consecutive iterations we can stop the algorithm when no state's utility changes by much.

Experiment

Create the transition matrix:

```
[[0.9 0.1 0.  0. ]
 [0.  0.  0.  0. ]
 [0.  0.  0.  0. ]]
[[0.9 0.  0.  0. ]
 [0.1 0.  0.  0. ]
 [0.  0.  0.  0. ]]
[[0.1 0.1 0.  0. ]
 [0.8 0.  0.  0. ]
 [0.  0.  0.  0. ]]
[[0.1 0.8 0.  0. ]
 [0.1 0.  0.  0. ]
 [0.  0.  0.  0. ]]
[[0.1 0.8 0.1 0. ]
 [0.  0.  0.  0. ]
 [0.  0.  0.  0. ]]
[[0.8 0.2 0.  0. ]
 [0.  0.  0.  0. ]
 [0.  0.  0.  0. ]]
[[0.1 0.8 0.1 0. ]
```

Figure 1 Transition Matrix

The environment is fully observable that means that we have to construct every possible state with every possible move with their transition probabilities. So we visualize the picture from the created transition model. From every possible state, each possible action and each possible next state is the shape of the transition array.

The next step of the exercise is to test the experiment with different gammas. So we pass variables of gammas with the values of 0.9, 0.6, 0.2. The experiment holds also the iterations of the utility computation that finishes after the utility doesn't have significant differences, depends on the value of the stopping criteria. After the utilities computations we will select a path who drives on a terminal state. A star holds the utilities as heuristics and select this action as the next state in the path to the end.
The following picture outputs the results of the first value iterations. The algorithm will finish if the state of the utilities aren't affected

```
---------------- ITERATIONS ----------------
[0. 0. 0. 0.]
[0. 0. 0. 0.]
[0. 0. 0. 0.]
---------------- ITERATIONS ----------------
[-0.04 -0.04 -0.04  1.  ]
[-0.04  0.   -0.04 -1.  ]
[-0.04 -0.04 -0.04 -0.04]
---------------- ITERATIONS ----------------
[-0.076  -0.076   0.6728  1.    ]
[-0.076  0.    -0.076 -1.   ]
[-0.076 -0.076 -0.076 -0.076]
---------------- ITERATIONS ----------------
[-0.1084    0.430736  0.733712  1.      ]
[-0.1084    0.        0.347576 -1.      ]
[-0.1084 -0.1084 -0.1084 -0.1084]
---------------- ITERATIONS ----------------
[0.25061792 0.56580512 0.77731592 1.        ]
[-0.13756    0.          0.42955448 -1.        ]
[-0.13756   -0.13756     0.19074272 -0.13756   ]
---------------- ITERATIONS ----------------
[0.3775549  0.62151238 0.78861834 1.        ]
[ 0.1156841   0.          0.46832737 -1.        ]
[-0.163804    0.07257396  0.24451843 -0.00504564]
---------------- ITERATIONS ----------------
[0.45188043 0.63967743 0.79312511 1.        ]
```

Figure 2 Utilities in first value iterations

Figure 3 Utilities of the states and path to terminal state

The final results are visualized in the upwards figure. With parameter of gamma with 0.9 takes 16 iterations to apply the stopping criteria. Delta variable compared with epsilon so in the sixtieth iteration the utilities aren't affected from the next state's utility. The final step prints the path to the terminal state following the maximum utilities who is the opposite of the cost as used in a star algorithm.

```
---------------- ITERATIONS ----------------
[0.50863317 0.64953825 0.79535124 1.        ]
[ 0.39612266  0.          0.48640935 -1.        ]
[0.29071505 0.25162024 0.34389758 0.12818272]
---------------- ITERATIONS ----------------
[0.50909556 0.64956978 0.79535845 1.        ]
[ 0.39751796  0.          0.48642973 -1.        ]
[0.29401849 0.2528979  0.344397   0.1291427 ]
---------------- ITERATIONS ----------------
[0.50928546 0.64958065 0.79536094 1.        ]
[ 0.39810204  0.          0.48643676 -1.        ]
[0.29543541 0.25348746 0.34461306 0.12958868]
---------------- ITERATIONS ----------------
[0.50936294 0.64958439 0.79536179 1.        ]
[ 0.3983439   0.          0.48643918 -1.        ]
[0.29603653 0.25374915 0.34471132 0.12978439]
---------------- ITERATIONS ----------------
[0.50939438 0.64958568 0.79536209 1.        ]
[ 0.39844322  0.          0.48644002 -1.        ]
[0.29628832 0.253867   0.34475423 0.12987275]
---------------- FINAL RESULT ----------------
Iterations: 16
Delta: 0.00010477963854704786
Gamma: 0.9
Epsilon: 0.001
------------------------------------------------
[0.50939438 0.64958568 0.79536209 1.        ]
[ 0.39844322  0.          0.48644002 -1.        ]
[0.29628832 0.253867   0.34475423 0.12987275]
================================================
[0.29628831545548107, 0.39844321783500447, 0.5093943765842497, 0.649585681261095, 0.7953620878466678, 1.0]
```

The second parameter of gamma is 0.6. That means that the agents decreases his preferences to discover the environment. Also is noticeable that the iterations are fewer in this experiment with lower gamma because the agent doesn't need to explore the environment. The path remains the same of a star to terminal state but the utilities of every state are changing.

```
---------------- ITERATIONS ----------------
[0.05304965 0.21168742 0.47613402 1.        ]
[-0.03701647  0.         0.13523068 -1.        ]
[-0.092224   -0.04995574  0.01108828 -0.08694047]
---------------- ITERATIONS ----------------
[0.06257195 0.21394682 0.47668188 1.        ]
[-0.01897814  0.         0.13665817 -1.        ]
[-0.06629869 -0.04067231  0.01669695 -0.08628256]
---------------- ITERATIONS ----------------
[0.0653101  0.21448092 0.4768004  1.        ]
[-0.01224284  0.         0.13700679 -1.        ]
[-0.05552777 -0.03686614  0.01797863 -0.08559076]
---------------- ITERATIONS ----------------
[0.06613488 0.2146019  0.47682843 1.        ]
[-0.01012029  0.         0.1370846  -1.        ]
[-0.0514202  -0.03579419  0.01841585 -0.08514029]
---------------- ITERATIONS ----------------
[0.06636979 0.21462988 0.47683478 1.        ]
[-0.00946969  0.         0.13710272 -1.        ]
[-0.0500906  -0.0354557   0.01854454 -0.08487081]
---------------- FINAL RESULT ----------------
Iterations: 10
Delta: 0.00041237274160005333
Gamma: 0.6
Epsilon: 0.001
------------------------------------------------
[0.06636979 0.21462988 0.47683478 1.        ]
[-0.00946969  0.         0.13710272 -1.        ]
[-0.0500906  -0.0354557   0.01854454 -0.08487081]
================================================
[-0.050090603694653436, -0.009469693570318325, 0.06636978910055633, 0.2146298757413765, 0.476834781970559, 1.0]
```

Figure 4 value iteration with gamma 0.6 and A star path

The final part of the project is the computation of utilities with gamma parameter values as 0.2. This value expects the agent to be not willing to explore the environment.

```
---------------- ITERATIONS ----------------
[0. 0. 0. 0.]
[0. 0. 0. 0.]
[0. 0. 0. 0.]
---------------- ITERATIONS ----------------
[-0.04 -0.04 -0.04  1.  ]
[-0.04  0.   -0.04 -1.  ]
[-0.04 -0.04 -0.04 -0.04]
---------------- ITERATIONS ----------------
[-0.048  -0.048   0.1184  1.    ]
[-0.048  0.    -0.048 -1.   ]
[-0.048 -0.048 -0.048 -0.048]
---------------- ITERATIONS ----------------
[-0.0496   -0.022976  0.121408  1.      ]
[-0.0496    0.       -0.042016 -1.      ]
[-0.0496 -0.0496 -0.0496 -0.0496]
---------------- FINAL RESULT ----------------
Iterations: 4
Delta: 0.00393984
Gamma: 0.2
Epsilon: 0.001
------------------------------------------------
[-0.0496   -0.022976  0.121408  1.      ]
[-0.0496    0.       -0.042016 -1.      ]
[-0.0496 -0.0496 -0.0496 -0.0496]
================================================
```

Figure 5 value iteration gamma factor 0.2

As we notice the utilities aren't affected after the forth iteration so finishes in this round. The last observation is that in this environment the we don't have path to terminal state because the utilities as heuristics aren't optimized in neighbors states.