CS1305 Summer 2015 Reinforcement learning lab
Due July 21st 10pm

> *Adhere to the Code of Academic Integrity.* You may discuss background issues and general strategies with others and seek help from course staff, but the implementations that you submit must be your own. In particular, you may discuss general ideas with others but you may not work out the detailed solutions with others. It is never OK for you to see or hear another student's code and it is never OK to copy code from published/Internet sources. If you feel that you cannot complete the assignment on you own, seek help from the course staff.

Applicability of reinforcement learning

For each of the following problems, state if it can be modeled as a reinforcement learning problem, and if so, what are the environment, the state, the goal, the actions and the reward.

1. A mobile robot must find a source of light in its environment.
2. A robotic arms is used to play "ball-in-a-cup". See figure below: [1]

Figure 1: Ball-in-a-cup

3. A robotic delivery truck must deliver parcels to all the institutes of NYC.
4. A robotic helicopter needs to take-off from, hover and land on its launch pad.

[1] `https://www.youtube.com/watch?v=xyJAvghtqIM`

Temporal difference learning

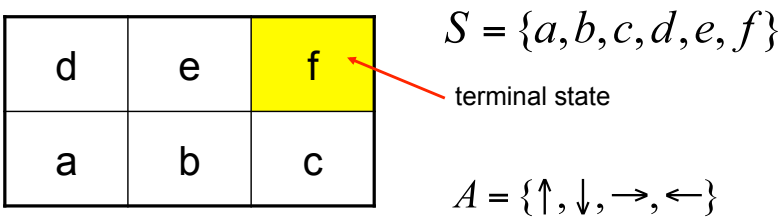

Figure 2: An environment map

An agent stays in the world shown in Figure 2. State f is the terminal state or goal state. The agent may start at any location on the grid. Here are the legal action for each state:

- $A(a) = \{\uparrow, \rightarrow\}$
- $A(b) = \{\uparrow, \rightarrow, \leftarrow\}$
- $A(c) = \{\uparrow, \leftarrow\}$
- $A(d) = \{\downarrow, \rightarrow\}$
- $A(e) = \{\downarrow, \rightarrow, \leftarrow\}$

The reward for the agent's action is

- $R = 100$ if $c \rightarrow f$ or $e \rightarrow f$
- Otherwise $R = 0$

Please answer the following questions:

1. List out three possible paths where an agent starts at state a and ends at state f
2. Assume we assign the initial value for each state as the following
   - $V(f) = 1$
   - $V(a) = V(b) = V(d) = 0$
   - $V(e) = V(c) = 0.5$

   Update $V(a), V(b), V(c), V(d), V(e)$ according to the three paths you provided above. For each step, apply temporal difference learning algorithm where $V(S) \leftarrow V(S) + \alpha(V(S') - V(S))$. Step-size parameter $\alpha = 0.6$. Show each step of your update.
3. List out all optimal policy for starting at state a, b or d.