CCIS 4100: in-class exercise on TDL

Byron C. Wallace

Complete the following either on your own or in small groups.

Temporal-difference learning

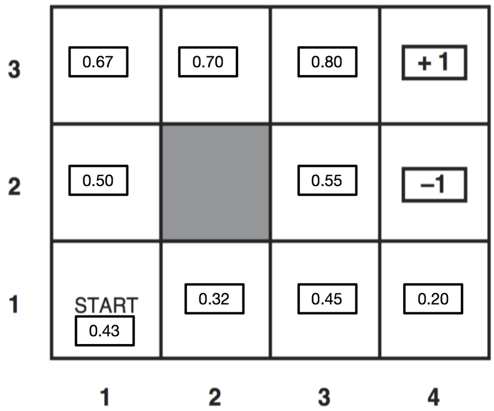

Figure 1: Exciting gridworld (again!) from the text (Figure 17.1). Here we assume that we have populated the grid with V^π estimates after some number of observations/trials.

Assume $R = -0.1$ (i.e., the 'living penalty' – instantaneous reward – is -0.1). This time *we don't know the transition probabilities* and we *don't know the rewards at the outset! – they have to be observed.* Further assume that we do not impose a discount ($\gamma = 1$). Remember, the general form of the update is:

$$V^\pi(s) \leftarrow (1-\alpha)V^\pi(s) + \alpha\{R(s, \pi(s), s') + \gamma V^\pi(s')\} \quad (1)$$

1. Assume the agent has run some number of trials already and, using TDL, come up with the V estimates depicted in Figure 2. Suppose you begin now at state (3,3) and observe:

   (3,3) $\rightarrow$ (3,4)

   Suppose $\alpha = 0.1$. Update the V^π estimate for state (3,3).

2. Now assume we observe a transition under π from (2,3) to (2,4); update V^π for (2,3).

3. Finally, using the updated V^π estimates calculated above, assume we observe a transition from (2,3) $\rightarrow$ (3,3). Update the V^π estimate for (2,3) once more.

4. To think about: what would have happened to the V^π estimate for (2,3) in steps 2 and 3 if γ were 1? More generally, what sort of properties does a 'good' γ value have for TDL learning?