Principles of Online Decision-Making (CS-303)

Problem Set 5 - Solutions

Problem 1

In the Gridworld example, we computed the optimal policy π_* when the reward for $A \to A'$ is $+10$, and for $B \to B'$ the reward is $+5$.

Now suppose we change the reward for $B \to B'$ to some new reward x (instead of $+5$). We want to understand to what extent the optimal policy π_* is affected by such a change.

Identify the ranges of x where the optimal policy π_* changes, and determine the optimal policy for each range.

Let us call x the reward for the $B \to B'$-transition. The optimal policy in this example "tries to" go through the high-reward transitions $A \to A'$ and $B \to B'$ as often as possible. This is most efficiently achieved by going through these transitions, and then walk back up to A resp. B.

Without discounting ($\gamma = 1$), the average reward per step of the A-cycle $(A, A', (3,1), (2,1), (1,1), A, \dots)$ would be $(10-4)/5$, compared to the average reward per step of the B-cycle $(B, B', (1,3), B, \dots)$, which is $(x-2)/3$ (note that this is less than that of A when $x = 5$).

Without discounting, all the other states would simply "point in the direction" of the winning cycle, avoiding the other cycle.

However, discounting (here with $\gamma = 0.9$) complicates this picture considerably, because the two cycles can coexist for some values of x. Different states may point to one or the other cycle, depending on their distance to that cycle.

The optimal policies for the ranges of x are given in Figure 1.

Problem 2

(a) Suppose we are given an MDP, and we add a constant c to every reward (even to transitions that had reward 0 before). Show that for a continuing task, the resulting optimal policy π_* is unaffected by this.

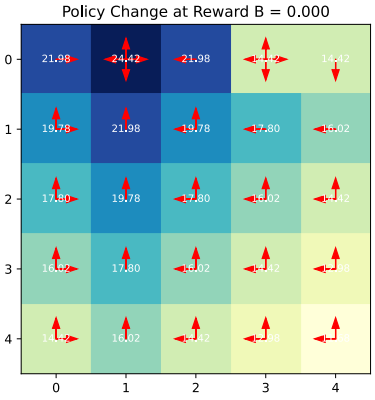
This adds a constant $\sum_{k=0}^{\infty} c\gamma^k = c/(1-\gamma)$ to the value function of every state ($v_*(s)$) and every state-action pair ($q_*(s,a)$). Given that the optimal policy, at every state s, selects the action(s) a that maximizes $q_*(s,a)$, the policy is unaffected.

(b) Suppose we add a constant c to every reward in an episodic task. Show that this may change the optimal policy π_*. You may find a concrete counterexample.

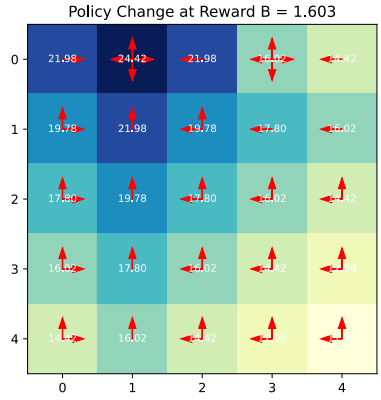
See the maze-exiting example in class: -1 for every time step makes it find shortest path out, $+1$ for every time step would actually encourage the agent not to exit and stay in the maze forever.

The formal reason why the argument for (a) does not apply is this: in this scenario, we are not actually adding the constant to all rewards: the "virtual rewards" $R_{T+1} = R_{T+2} = \dots = 0$ after the end of the episode remain zero.

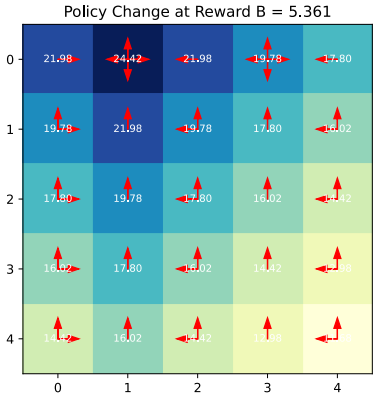
Optimal policies for Gridworld for different values of x.



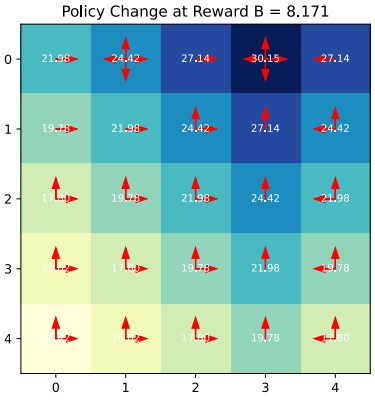
(a) $x < 1.6$: avoid B entirely, go to cycle A.



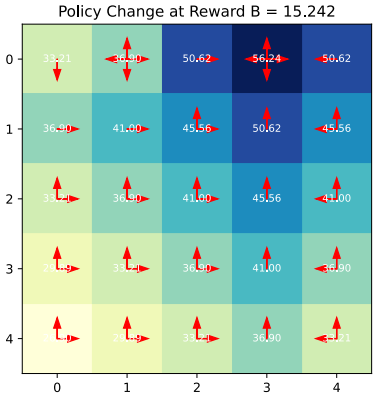
(b) $1.6 < x < 5.36$: avoid cycle B except from $(0, 4)$.



(c) $5.36 < x < 8.17$: columns 3 and 4 now committed to cycle B.



(d) $8.17 < x < 15.24$: column 1 and 2 and state $(1, 0)$ now switched to cycle B.



(e) $x > 15.24$: avoid A entirely (including from state $(0, 0)$), get to cycle B.

Problem 3

Our starting point for the Bellman equations were that the system specification is in terms of a conditional joint distribution over the next state and reward, i.e., $p(s', r|s, a)$. Now suppose that we have a different (but equivalent) specification: we are given the conditional distribution $p(s'|s, a)$ of the next state, and the expected reward given the current state and action $r(s, a)$. Rewrite the four Bellman equations for the four value functions (v_π, v_*, q_π, and q_*) in terms of the three-argument function $p(s'|s, a)$ and the two-argument function $r(s, a) = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a]$.

---

1. $v_\pi(s) = \sum_a \pi(a|s) \left[r(s, a) + \gamma \sum_{s'} p(s'|s, a) v_\pi(s')\right]$
2. $q_\pi(s, a) = r(s, a) + \gamma \sum_{s'} \sum_{a'} p(s'|s, a)\pi(a'|s')q_\pi(s', a')$
3. $v_*(s) = \max_a \left[r(s, a) + \gamma \sum_{s'} p(s'|s, a) v_*(s')\right]$
4. $q_*(s, a) = r(s, a) + \gamma \sum_{s'} p(s'|s, a) \max_{a'} q_*(s', a')$