Principles of Online Decision-Making (CS-303)

Problem Set 5

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.

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.

(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.

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 $r(s, a) = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a]$.