CS234 Problem Session

Week 1: Jan 13

1) [CA session] Problem 1: MDP Design

You are in a Las Vegas casino! You have \$20 for this casino venture and will play until you lose it all or as soon as you double your money (i.e., increase your holding to at least \$40). You can choose to play two slot machines: 1) slot machine A costs \$10 to play and will return \$20 with probability 0.05 and \$0 otherwise; and 2) slot machine B costs \$20 to play and will return \$30 with probability 0.01 and \$0 otherwise. Until you are done, you will choose to play machine A or machine B in each turn. In the space below, provide an MDP that captures the above description.

Describe the state space, action space, rewards and transition probabilities. Assume the discount factor $\gamma = 1$. Rewards should yield a higher reward when terminating with \$40 than when terminating with \$0. Also, the reward for terminating with \$40 should be the same regardless of how we got there (and equivalently for \$0).

2) Problem 2: Contradicting Contractions?

Consider an MDP $M(S, A, P, R, \gamma)$ with 2 states $S = \{S_1, S_2\}$. From each state there are 2 available actions $A = \{stay, go\}$. Choosing *"stay"* from any state leaves you in the same state and gives reward -1. Choosing *"go"* from state S_1 takes you to state S_2 deterministically giving reward -2, while choosing *"go"* from state S_2 ends the episode giving reward 3. Let $\gamma = 1$.

Let $V^*(s)$ be the optimal value function in state s. As you learned in class, value iteration produces iterates $V_1(s), V_2(s), V_3(s), \ldots$ that eventually converge to $V^*(s)$.

(a) [CA Session]

Prove that the ∞-norm distance between the current value function V^k and the optimal value function V^* decreases after each iteration of value iteration.

(b) [Breakout Rooms]

Now let us consider exactly what forms of convergence are ensured.

For the given MDP, let us initialize value iteration as $V_0 = [0, 0]$. Then $V_1 = [-1, 3]$ and $V_2 = [1, 3]$. We also have $V^* = [1, 3]$.

Is there monotonic improvement in the V estimates for all states? If not, does this contradict the result in Q2(a) and why or why not?

3) [Breakout Rooms] Problem 3: Stochastic Optimal Policies

Given an optimal policy that is *stochastic* in an MDP, show that there is always another deterministic policy that has the same (optimal) value.

4) [Breakout Rooms] Problem 4: Parallelizing Value Iteration

During a single iteration of the Value Iteration algorithm, we typically iterate over the states in $\mathcal{S}$ in some order to update $V_t(s)$ to $V_{t+1}(s)$ for all states s. Is it possible to do this iterative process in parallel? Explain why or why not.