The Johns Hopkins University
Department of Electrical and Computer Engineering

Foundations of Reinforcement Learning

Problem Set #4

Homework Rules:

- Whenever a proof is required, please provide a detailed justification of every step.
- Whenever Python/Matlab simulations are required, please attach a copy of the code. The code should be legible by an experienced reader.
- **Reference book**: Reinforcement Learning: An Introduction, 2nd edition, by Richard S. Sutton and Andrew G. Barto.
- Unless explicitly stated otherwise, exercises are expected to be solved by yourself using only Blackboard resources and your own lecture notes. Verbal and piazza discussions are acceptable, provided that do not involve any explicit writing of the solution.

Problems.

1. **Problem 1. (20 points)**

   You are in casino! You start with \$10 and will play until you lose it all or as soon as you get \$30. You can choose to play two slot machines: 1) slot machine A costs \$10 to play and will return \$20 with probability 0.1 and \$0 otherwise; and 2) slot machine B costs \$20 to play and will return \$30 with probability 0.4 and \$0 otherwise. Until you are done, you will choose to play machine A or machine B in each turn.

   (a) **(5 points):** Compute the expected values of the money you gain from playing machine A and B, respectively.

   (b) **(5 points):** We can model this by MDP. Let the state to be the current money you have. Let the action be playing either machine A or B once, and the reward be the money you gain from that play.

   Write down the state space, and the action space. Then draw a diagram for this MDP. (The diagram should look like the one in Example 3.3, p52, Sutton&Barto, please use different node marks for non-terminal and terminal states)

   (c) **(5 points):** Explain that why all possible policies $\pi(a|s)$ for this MDP can be uniquely defined by $\beta \in [0, 1]$, where β is the probability of choosing slot machine A when you have \$20.

   Now consider such a policy π_β. Compute v_{π_β} for all the non-terminal states. What is the optimal policy?

   (d) **(5 points):** If we now assume that "slot machine B costs \$20 to play and will return \$30 with probability $0 < \eta < 1$ and \$0 otherwise". What value of η ensures that any policy is an optimal policy?