For rules on collaboration and late policies, please see the course web page.

**1) 35 points**. An MDP has the following states and rewards $R(s)$:

| State | Reward |
|---|---|
| SCS | 0 |
| Google | 80 |
| Grad School | 20 |
| Startup | 30 |
| Hedge Fund | 150 |
| Flipping Burgers | 10 |
| Professor | 100 |
| Yachts and Bottle Service | 500 |

And the following actions: Work Hard, Goof Off, Web 1.0, Web 2.0, and Insider Trade. You may abbreviate these as long as what you mean is clear.

The MDP has the following positive state transition probabilities $T(s, a, s')$. All other tuples have zero probability. "*" indicates "any action".

| State | Action | Next State | Probability |
|---|---|---|---|
| SCS | Work Hard | SCS | .10 |
| SCS | Work Hard | Google | .30 |
| SCS | Work Hard | Startup | .20 |
| SCS | Work Hard | Hedge Fund | .10 |
| SCS | Work Hard | Grad School | .30 |
| SCS | Goof Off | SCS | .25 |
| SCS | Goof Off | Flipping Burgers | .60 |
| SCS | Goof Off | Grad School | .15 |
| Google | Work Hard | Google | .95 |
| Google | Work Hard | Grad School | .05 |
| Google | Goof Off | Google | .8 |
| Google | Goof Off | Grad School | .1 |
| Google | Goof Off | Flipping Burgers | .1 |
| Grad School | Work Hard | Professor | .2 |
| Grad School | Work Hard | Hedge Fund | .2 |
| Grad School | Work Hard | Startup | .2 |
| Grad School | Work Hard | Grad School | .4 |
| Grad School | Goof Off | Grad School | .8 |
| Grad School | Goof Off | Professor | .1 |
| Grad School | Goof Off | Flipping Burgers | .1 |
| Startup | Web 1.0 | Grad School | .25 |
| Startup | Web 1.0 | Flipping Burgers | .1 |
| Startup | Web 1.0 | Yachts and Bottle Service | .05 |
| Startup | Web 1.0 | Startup | .6 |
| Startup | Web 2.0 | Grad School | .1 |
| Startup | Web 2.0 | Flipping Burgers | .05 |
| Startup | Web 2.0 | Yachts and Bottle Service | .1 |
| Startup | Web 2.0 | Startup | .75 |
| Hedge Fund | Work Hard | Grad School | .1 |
| Hedge Fund | Work Hard | Yachts and Bottle Service | .1 |
| Hedge Fund | Work Hard | Hedge Fund | .8 |
| Hedge Fund | Goof Off | Flip Burgers | .1 |
| Hedge Fund | Goof Off | Yachts and Bottle Service | .1 |
| Hedge Fund | Goof Off | Grad School | .3 |
| Hedge Fund | Goof Off | Hedge Fund | .5 |
| Hedge Fund | Insider Trade | Flipping Burgers | .25 |
| Hedge Fund | Insider Trade | Hedge Fund | .5 |
| Hedge Fund | Insider Trade | Yachts and Bottle Service | .25 |
| Flipping Burgers | * | Flipping Burgers | 1 |
| Professor | * | Professor | 1 |
| Yachts and Bottle Service | * | Yachts and Bottle Service | 1 |

**a.** Use value iteration to solve for the values of all states when $\gamma = .8$. Recall that in value iteration, the value of state $s$ at iteration $t$ is

$$V_t(s) = \max_a R(s) + \gamma \sum_{s'} T(s, a, s') V_{t-1}(s')$$

and set $V_0(s) = 0$ for every state. Loop until convergence.

**(Values approximate) SCS = 476.5 Google = 428.2 Grad School = 569.1 Startup = 693.8 Hedge Fund = 1100 Flipping Burgers = 50 Professor = 500 Yachts and Bottle Service = 2500.**

**b.** What is the optimal policy?

**"Work Hard" for all states except Hedge Fund (Insider Trade) and Startup (Web 2.0).**

**c.** What are the optimal policy and values when $\gamma = 0$? In general, what happens when we solve an MDP with $\gamma = 0$?

**Every policy is optimal, because when $\gamma = 0$ you only care about the present state and in this MDP actions do not impact rewards. In general, when rewards depend on states and action choices, you will do whatever action corresponds to the highest immediate payout.**

**2) 25 points**. Imagine an agent is navigating a 3x3 grid, with a total of 9 states, arranged as follows:

| | | |
|---|---|---|
| $s_1$ | $s_2$ | $s_3$ |
| $s_4$ | $s_5$ | $s_6$ |
| $s_7$ | $s_8$ | $s_9$ |

There are some basic rules to be aware of:

- One of the states is the goal state. If an agent begins the time step in the goal state, she receives a reward of 1.
- The 3x3 grid is surrounded by a barrier, which repels moves and keeps agents inside the grid. **Three reflective barriers are present on the grid's interior**.
- Agents can only move up, down, left, or right. If an agent moves into a barrier, the barrier reflects them and they remain in their current state. If an agent is not obstructed, they move to the next state with probably .9, and remain in their current state with probability .1.

Consider the following $Q$-table, where the values are run to convergence with $\gamma = .5$.

| State x Action | $Q$-value | State x Action | $Q$-value |
|---|---|---|---|
| $s_1$ down | .213 | $s_5$ right | .947 |
| $s_1$ right | .106 | $s_5$ down | .474 |
| $s_2$ down | .449 | $s_6$ up | 1.53 |
| $s_2$ left | .225 | $s_6$ left | 1.53 |
| $s_2$ right | .449 | $s_6$ down | 2 |
| $s_3$ left | .249 | $s_7$ up | .215 |
| $s_3$ down | .947 | $s_7$ right | .056 |
| $s_4$ up | .118 | $s_8$ left | .101 |
| $s_4$ down | .118 | $s_8$ up | .051 |
| $s_4$ right | .449 | $s_8$ right | .027 |
| $s_5$ up | .249 | $s_9$ left | .048 |
| $s_5$ left | .249 | $s_9$ up | .024 |

Feel free to draw pictures to answer the following questions.

**a.** For each state $s$ describe the optimal policy $\pi^*(s)$.
**Here is an optimal policy:**

| | | |
|---|---|---|
| ↓ | ↓ | ↓ |
| → | → | ↓ |
| ↑ | ← | ← |

**b.** Which state is the goal state?
$s_6$.

**c.** Between which states are the three internal barriers?
$(s_1, s_2)$, $(s_5, s_8)$, $(s_6, s_9)$.

**3) 25 points**. Imagine two drivers playing *chicken*, a game where they drive towards one another with their cars. Each driver has three actions — they can choose to go straight, or to turn left or right. For simplicity, we standardize directions according to the perspective of an overhead observer. Thus, if both drivers select the same action, they will crash. At the same time, each driver wants to go straight, to seem tough and fearless. Utilities are given by the following table:

| Utility | Turn Left | Straight | Turn Right |
|---|---|---|---|
| Turn Left | (-20,-20) | (-5,10) | (0,0) |
| Straight | (10,-5) | (-10,-10) | (10,-5) |
| Turn Right | (0,0) | (-5,10) | (-20,-20) |

These are in the format (row player, column player).

**a.** What are the pure strategy Nash equilibria of this game?
**Four pure strategy equilibria corresponding to one player going straight and the other player going left or right.**

**b.** Does the game have any mixed strategy Nash equilibria? What are they?

**Yes. Both players go** $(L, S, R) = (.1, .8, .1)$**.**

**c.** Does either player have a dominant strategy?

**No.**

**d.** Imagine that the row player has the ability to credibly choose to go straight (say, by removing their steering wheel entirely), effectively removing actions from their choice set. Would the row player choose to do this? Why or why not? What does your answer say about the differences between single-agent and multi-agent contexts?

**Yes, the row player would choose to do this, because the column player will then go left or right and secure the highest payoff for the row player. In a single-agent setting, removing actions from your choice set never makes you better off (you might remove the optimal action). However, in a multi-agent setting you can be better off by having fewer choices.**

**4) 15 points**. In a *second-price auction*, the item is given to the highest bidder at the price of the second-highest bid. For instance, imagine Alice is auctioning off a pair of sneakers, Bob bids 10 dollars, and Carl bids 5 dollars. Bob would win the sneakers and pay Alice 5 dollars. With small tweaks to accommodate discretization, this is the rule used by eBay.

We can generalize a second-price auction as follows. Assume that all bidders have *quasi-linear utility*, so that bidder $i$'s utility for receiving the good and paying $\pi$ is

$$u_i \equiv v_i - \pi$$

and that bidders have zero utility for receiving nothing and paying nothing.

Argue from first principles that it is a dominant strategy to bid your true value in a second-price auction, so that you would always want to reveal your true value to the auctioneer, regardless of the actions of the other bidders. HINT: Why would a bidder never gain from bidding higher than her true value? Why would a bidder never gain from bidding lower?

**If an agent bids more than their value, they will either continue to win the auction at the same price, continue to lose the auction, or win the auction at a price higher than their valuation (negative utility). If an agent bids less than their true value, they will either continue to win the auction at the same price, continue to lose the auction, or lose an auction they were winning at a price less than their true valuation (thereby forgoing positive utility for zero utility). Since in all cases the agent's utility does not improve by manipulation, and in some cases utility is lowered, reporting truthfully is a dominant strategy in a second-price auction.**