Documentation: I did not collaborate with anyone on this assignment. I watched the office hour recording to guide my answers in Problem 2.

# 1.f

I found the following cutoff values for discount factor such that an optimal agent starting in the initial far-left state does NOT swim up the river:

WEAK: $\gamma = 0.65$
MEDIUM: $\gamma = 0.76$
STRONG: $\gamma = 0.92$

The increase in discount factor cutoff with increase in current strength makes intuitive sense for the formulation of the problem. Fundamentally, a high $\gamma$ leads the agent to prioritize short-term rewards. Using the algorithms we have developed and evaluated against the grader, higher gamma values lead the agent to swim up the river (i.e. go RIGHT). As the current strength increases, the transition probabilities for successfully swimming RIGHT also improve. This leads the agent to chase the higher reward value in state 6 even at higher discount factors because the long-term reward is easier to achieve.

## 2.a

Yes - it is possible to choose a value of $H$ that results in the optimal policy for finite time horizons. This is because the reward function leaves no ambiguity for the agent and there is always a strategy for maximizing the reward. The problem gives an example for $H = 4$. If, for instance, $H = 5$, then the agent would buy once and sell four times to optimize the policy for starting at $s = 3$.

## 2.b

Since the reward is so highly weighted toward reaching a fully stocked inventory, the agent should always look to reach $s = 10$ if it is possible. Therefore, for $s_0 = 3$, a fully stocked inventory is only possible for $H \geq 7$. The range for which the optimal policy reaches a fully stocked inventory is therefore $H \in [7, \infty)$.

## 2.c

With an infinite time horizon and no discount factor, the agent does not need to terminate the problem to reach the $+100$ reward. It will therefore buy and sell indefinitely because it can maximize the reward by oscillating to generate r$+1$ with every two actions. At $s = 3$, the optimal policy can either buy or sell - it does not matter since the agent can subsequently sell to earn a reward and short term gains are not a priority. At $s = 9$, the agent must sell before it can oscillate between buying and selling ad infinitum because reaching $s = 10$ would terminate the problem and stop it from earning infinite reward.

## 2.d

Yes - it is possible to choose a discount factor such that the optimal policy never fully stocks the inventory. As we just proposed in 2(c), $\gamma = 0$ results in the prioritization of infinite short-term rewards from buying and selling indefinitely. For sufficiently small $\gamma$, this will continue to be the case because the discounted sum of rewards will outweigh the fully stocked incentive according to proposed value function. That is, low $\gamma$ will weigh the immediate reward of selling over the long-term goal of fully stocking the inventory.

## 3.a

As Pan, Bhatia, and Steinhardt specify in their paper, the optimal policy for a large model is to park the AI car and not merge onto the highway. This instance of reward hacking results from the higher mean velocity for the four cars since the human-operated vehicles do not need to slow down, increasing the proxy reward. The gray cars maintain their high speed, while the AI car idles at zero. Since this is only one car, the overall effect on the mean behavior is lower because zero velocity is a singular outlier. Merging would bring each speed down marginally and have a worse overall effect on the proxy. The paper notes that idling may improve the proxy reward, but it also increases the overall mean commute significantly since the merge would be transient.

## 3.b

Minimal mean commute time is essentially a group measure of the efficiency of traveling from points to points. There are other ways to optimize efficient travel than commute time or velocity. I think that you could also reward minimizing application of the brake pedal or minimizing mean wait time length for merging vehicles only. The first would be difficult for the agent to assess for the human-operated vehicles, but brake lights in a congested zone might be usable for sensor data. The latter would practically force the AI car to merge, which may increase the safety risk associated with the proxy. The best solution would likely be a combination of the various efficiency measures to make the agent resilient to reward hacking and unexpected human behaviors.