This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the —README.md— for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.f

The largest discount factor to now swim upward at s_1 is 0.65. It makes sense because there is a small immediate reward to swim down where it takes some effort to get the bigger reward to get to the top. Logically, it makes sense to continue moving in order to get the largest reward. If the agent is very lucky, the expected return to go straight from s_1 to the top is $(0.6 \times 0.31 \times 0.31 \times 0.31 \times 0.31 \times 0.6) \times 1 = 0.003$. It is less than the immediate reward of giving up. Therefore, it makes sense for the discount factor to be slightly larger than 0.5 to let the agent start taking risks.

For MEDIUM and STRONG current, the largest discount factors are 0.76 and 0.92. This also makes sense because the stronger the current is, the less chance the agent can get to the top. We can do the same upper bond calculation as WEAK current. The lucky expected return for MEDIUM and STRONG is 0.0008 and 0.0001.

2.a

Yes, if H is large enough, the agent can loop between two states buying and selling stocks to gain more rewards than the full stock reward.

2.b

It is impossible to reach full stock for $H < 7$. For any $H > 200$, the agent can just buy and sell repeatedly to gain more than 100 rewards. For any value in $H \in [7, 200)$, it's impossible to reach 100 rewards by repeating buy and sell so the optimal policy is keep buying to get the full stock.

2.c

For $\gamma = 0$, the agent will only care about the instant reward. So at $s = 3$, only selling can earn reward while at $s = 9$, buying earns more than selling by reaching full stock.

2.d

Because the MDP policy is only relevant to its current state and actions always succeed, the optimal policy will either trap in a buy-sell loop or keep buying until full stock. Even when γ close to 1, because the horizon is infinite, it is very likely that the policy will trap the agent in the buy-sell loop because that can easily yield more rewards than the full stock reward. Thus, I believe that every value in $[0, 1)$ will result in an optimal policy that never fully stock the inventory.

3.a

Because multiple gray cars need to slow down to let the red car merge. If the red car chooses not to merge, all gray cars can continue to go at high speed with a very small decrease in the mean velocity caused by the red car stopping.

3.b

Because all cars can still reach the speed limit after merging. Mean velocity is still a good metrics to monitor, but we can add a penalty when the agent stops the car for too long, forcing it to keep moving in the long run.