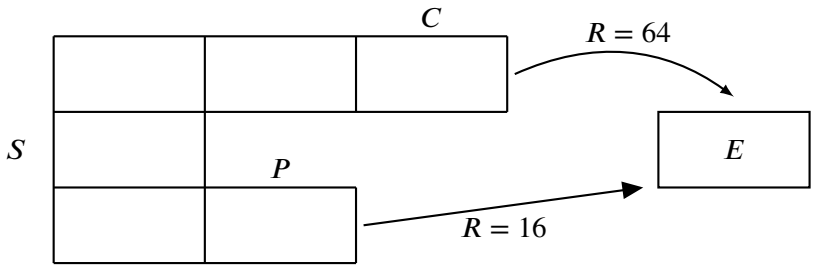
Q1. MDP Exploration

Pacman is now a CS student at UW. He finds himself in a (very) simplified, deterministic grid world MDP representation of UW depicted below, starting at state S. Pacman can take actions up, down, left or right. If an action moves him into a wall, he will stay in the same state. States C and P represent the CS building and the Party building respectively (their labels both appear above the relevant grid square). Pacman will study at the CS building or he will party at the Party building. At states C and P Pacman can take the exit action to receive the indicated reward and enter the terminal state, E. Note $R(s, a, s') = 0$ otherwise. Once in the terminal state the game is over and no actions can be taken. Let the discount factor $\gamma = \frac{1}{2}$ for this problem, unless otherwise specified.



(a) **(i)** Draw the optimal policy for Pacman on the grid world above.

(ii) How many steps k of Value Iteration will it take before $V_k(S) = V^*(S)$?

$k =$ ☐

(iii) Select all values that $U^{\pi_k}(S)$ will take on during the entire process of Policy Iteration given every possible initial policy.

☐ 0 ☐ 2 ☐ 4 ☐ 8 ☐ 16 ☐ 32 ☐ 64 ☐ None of these

(b) Let us mess with Pacman's ability to study. Your task is to change some of the MDP parameters so that Pacman no longer desires to visit the CS building. S is where Pacman starts (the square to the right of the label S). **All subquestions are independent** of each other so consider each change on its own.

(i) What discount factor forces Pacman to be indifferent between studying and partying given that he starts at state S?

$\gamma =$ ☐

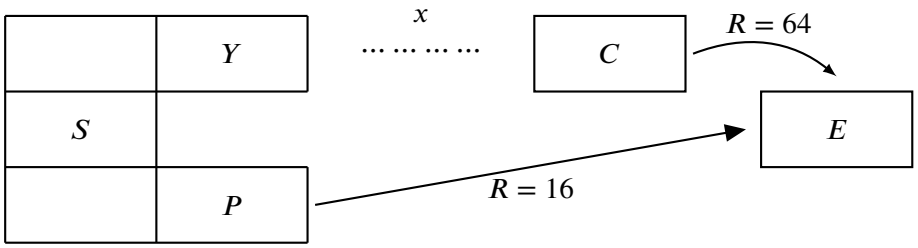
(ii) Tweak the reward function such that Pacman will always choose partying over studying. Write a bound on $R(P, exit, E)$, that guarantees Pacman exits from P instead of C, leaving $R(C, exit, E)$ unchanged. Choose an inequality symbol and fill in the answer box. Select nothing and write **N/A** in the box if this is not possible.

$R(P, exit, E)$ ○ (A) $>$ ○ (B) $<$ ☐

(iii) Consider the following reward functions, $R'(s, a, s')$. Select the new reward functions that cause Pacman not to exit from state C. Note $R(s, a, s')$ is the original reward function from the problem description.

- ☐ $R'(s, a, s') = 1 + R(s, a, s')$
- ☐ $R'(s, a, s') = 100 + R(s, a, s')$
- ☐ $R'(s, a, s') = -1 + R(s, a, s')$
- ☐ $R'(s, a, s') = -100 + R(s, a, s')$
- ☐ $R'(s, a, s') = -R(s, a, s')$
- ☐ $R'(s, a, s') = 2R(s, a, s')$
- ☐ None of these

(iv) Let us make the reward of studying so distant that Pacman no longer exits from C. We'll accomplish this by adding a certain number of grid positions, x of them, in between Y and C as depicted below. Give a lower bound for x that **guarantees** Pacman does not exit from C.



$x \geq$ ☐

Q2.

In this problem, we will consider the task of managing a fishery for an infinite number of days. (Fisheries farm fish, continually harvesting and selling them.) Imagine that our fishery has a very large, enclosed pool where we keep our fish.

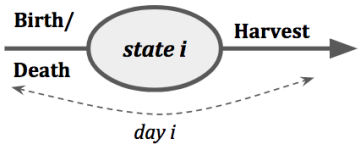
Harvest (11pm): Before we go home each day at 11pm, we have the option to harvest some (possibly all) of the fish, thus removing those fish from the pool and earning us some profit, x dollars for x fish.

Birth/death (midnight): At midnight each day, some fish are born and some die, so the number of fish in the pool changes.

An ecologist has analyzed the ecological dynamics of the fish population. They say that if at midnight there are x fish in the pool, then after midnight there will be exactly $f(x)$ fish in the pool, where f is a function they have provided to us. (We will pretend it is possible to have fractional fish.)

To ensure you properly maximize your profit while managing the fishery, you choose to model it using a Markov decision problem.

For this problem we will define States and Actions as follows:
State: the number of fish in the pool that day (before harvesting)
Action: the number of fish you harvest that day

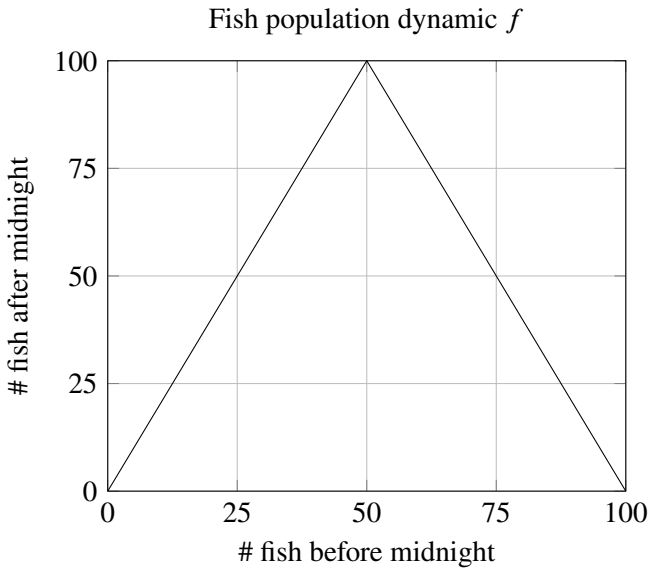


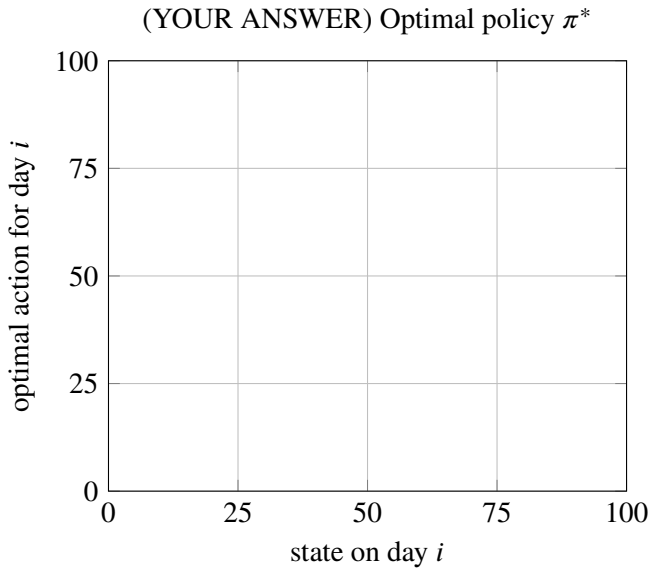
1. How will you define the transition and reward functions?

   $T(s, a, s') =$ _____

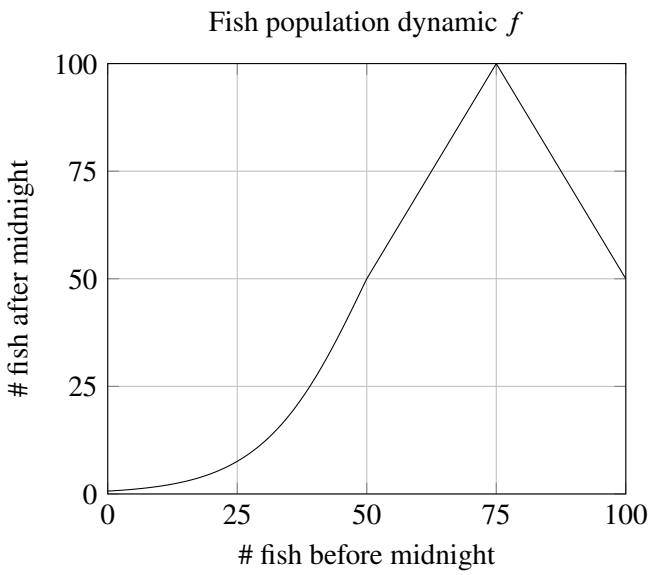
   $R(s, a) =$ _____

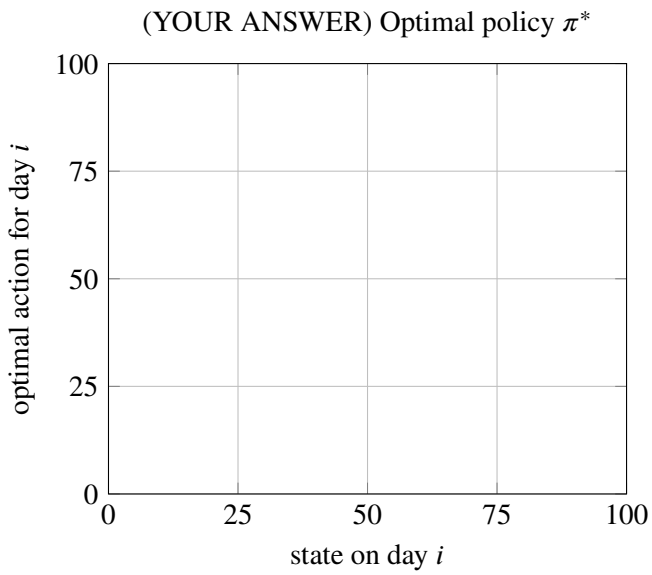
2. Suppose the discount rate is $\gamma = 0.99$ and f is as below. Graph the optimal policy π^*.





3. Suppose the discount rate is $\gamma = 0.99$ and f is as below. Graph the optimal policy π^*.

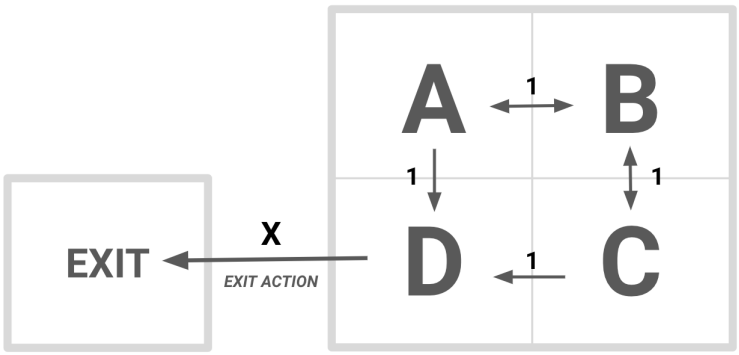




Q3. MDP Exploration

For this problem, you are lost in Padelford hall and are late to your next class. Various areas in the building are represented by the states **A, B, C, D**. At **D**, you can see the way out so you will only exit and make your way to your next class, getting some reward x. Otherwise any state transition gives you renewed hope that you may one day yet escape, giving a small reward.

In MDP terms, the available actions at **state A, B, C** are *LEFT, RIGHT, UP,* and *DOWN* unless there is a wall in that direction. The only action at **state D** is the *EXIT ACTION* and gives the agent a **reward of** x. The **reward for non-exit actions is always 1**.



(a) Let all actions be deterministic. Assume $\gamma = \frac{1}{2}$. Express the following in terms of x.

$V^*(D) =$

$V^*(A) =$

$V^*(C) =$

$V^*(B) =$

(b) Let any non-exit action be successful with probability $= \frac{1}{2}$. Otherwise, the agent stays in the same state with reward $= 0$. The *EXIT ACTION* from the **state D** is still deterministic and will always succeed. Assume that $\gamma = \frac{1}{2}$.

For which value of x does $Q^*(A, DOWN) = Q^*(A, RIGHT)$? Box your answer and justify/show your work.

(c) We now add one more layer of complexity. Turns out that the reward function is not guaranteed to give a particular reward when the agent takes an action. Every time an agent transitions from one state to another, once the agent reaches the new state s', a fair 6-sided dice is rolled. If the dices lands with value x, the agent receives the reward $R(s, a, s') + x$. The sides of dice have value $1, 2, 3, 4, 5$ and 6.

Write down the new bellman update equation for $V_{k+1}(s)$ in terms of $T(s, a, s')$, $R(s, a, s')$, $V_k(s')$, and γ.

Q4. Hours Worked

(a) How many hours did you spend on this homework? Any reasonable answer (number greater than zero) will receive credit. This will not affect your score on any other problem.