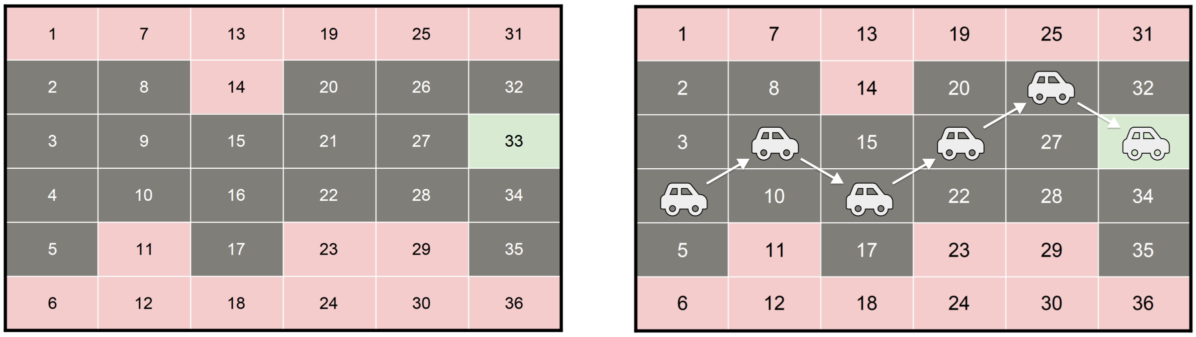
1 Efficient Routing MDP [35 pts]

You are leading a routing and planning team at a self-driving car company and have decided to model your latest urban navigation problem as an MDP. Consider the following environment (Fig. 1). Your car must navigate along the road (gray squares) while avoiding obstacles (red squares) to reach the rider's destination (the green square). Because the road is gridlocked, your car must change lanes whenever it wishes to move forward. From any gray square, your car can either move right & up, or right & down. For example, starting from state 3, your car can move to state 8 or 10. Note that it is not be possible to reach the green square from every state. Actions are deterministic and always succeed unless they will cause you to run into an impassible barrier. The thick outer edge indicates an impassible barrier, and attempting to move in the direction of a barrier from a gray square results in your car moving up one square (e.g. taking any action from state 32 moves the car to state 31).



(a) Grid World

(b) A successful run in Grid World.

Figure 1

A successful run in Grid World 1 is shown in Figure 1b. Taking any action from the green destination square (no. 33) earns a reward of r_g and ends the episode. Taking any action from the red squares that depict obstacles (no. 1, 7, 13...) earns a reward of r_r and ends the episode.

Otherwise, from every other square, taking any action is associated with a reward r_s. Assume the discount factor $\gamma = 0.9$, $r_g = +5$, and $r_r = -5$ unless otherwise specified. Notice the horizon is technically infinite.

(a) Let $r_s \in \{-5, -0.5, 0, 2\}$. Starting in **square 2**, for each of the possible values of r_s, briefly explain what the optimal policy would be in Grid World. In each case is the optimal policy unique and does the optimal policy depend on the value of the discount factor γ? Explain your answer. [5 pts]

(b) Which values of $r_s \in \{-5, -0.5, 0, 2\}$ will yield a policy that returns the shortest path to the green square? (Hint: At least one does.) Explain which ones do, then, pick the minimum of this set of rewards that does, and then find the optimal value function for **states 2, 13, 21 and 32**. [5 pts]

Rather than finding the shortest path between two points, suppose our car is low on gas, so we want to take the path that uses the least fuel. In the real world, navigation optimized for fuel consumption may take more steps to reach a destination [1].

Consider the same MDP, but with two new "efficient actions" – move right or move down. For example, starting from state 3, you can either move to state 4 or 9. Once again, the actions are deterministic and always succeed unless you run into a wall. Attempting to move in the direction of a wall from a gray square using an efficient action results in you moving *down* one square. For clarity, we will use separate symbols r_s for the reward associated with an inefficient action (right & up, or right & down) and r_e for the reward associated with an efficient action.

(c) Let $r_e \in \{-5, -0.5, 0, 2\}$. Starting in **state 2**, for each of the possible values of r_e, briefly explain what the optimal policy would be in Grid World *using only efficient actions*. In each case is the optimal policy unique and does the optimal policy depend on the value of the discount factor γ? Explain your answer. Which values of r_e would cause the optimal policy to return the shortest path to the green destination square? [5 pts]

(d) Consider now that $r_s = 0$. Derive a relation for r_e such that the optimal path from **state 2** to the destination square using only efficient actions is strictly more rewarding than the optimal path using only inefficient actions. [5 pts]

(e) Compare the set of gray states that can reach the goal using only efficient actions, and the set of gray states that can reach the goal using only inefficient actions. Which states are part of one set but not part of the other? Explain your answer. [5 pts]

(f) Consider a general MDP with rewards and transitions. Consider a discount factor of γ. Assume that the horizon is infinite (so there is no termination). Can adding a constant c to all rewards ($r_{new} = c + r_{old}$) change the optimal policy of the MDP? If yes, give an example for Grid World with efficient actions using the r_g, r_s and r_e such that the optimal policy changes for a specific constant. [5 pts]

(g) Imagine your efficient routing MDP is to be used in a popular maps app or website. Choosing what route options will be available and which will be default present inevitable value judgements. The shortest path and the sustainable path are each optimal given their choices of

[1] Google Maps Blog

rewards, but the rewards formalize different values. How would you present the shortest path and sustainable path options in a maps app and why? Please use 2-4 sentences to explain your answer. There is no single "correct" answer– reasonable explanations of your choice will receive full credit. [5 pts]

Context for the debate: "Green nudges" such as setting the most efficient route as the (changeable) default on mapping apps have been proposed as a way to encourage environmentally beneficial actions.[2] They can help close the gap between the desire to act more sustainably that many people express and their day to day behavior.[3] Some have argued that green nudges do not go far enough and that given the urgency of climate change the sustainable option should be the only one presented. Others argue that nudges infringe autonomy and so users should be explicitly asked which default they would prefer; others that nudges are only acceptable when the intention behind the nudge is transparently presented. For more context, see [3].

[2] A nudge is an "aspect of choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing economic incentives" (Thaler and Sunstein, 2008).

[3] Siipi and Koi, 2021. https://philpapers.org/archive/SIITEO.pdf

2 Value Iteration Theorem [35 pts]

In this problem, we will deal with contractions and fixed points and prove an important result from the value iteration theorem. From lecture, we know that the Bellman backup operator B given below is a contraction with the fixed point as V^*, the optimal value function of the MDP. The symbols have their usual meanings. γ is the discount factor and $0 \leq \gamma < 1$. In all parts, $||v||$ is the infinity norm of the vector.

$$(BV)(s) = \max_a [R(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V(s')]$$

We also saw the contraction operator B_π which is the Bellman backup operator for a particular policy given below:

$$(B_\pi V)(s) = \mathbb{E}_{a \sim \pi}[R(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V(s')]$$

(a) Recall that $||BV - BV'|| \leq \gamma ||V - V'||$ for two random value functions V and V'. Prove that B_π is also a contraction mapping: $||B_\pi V - B_\pi V'|| \leq \gamma ||V - V'||$. [5 pts]

(b) Prove that the fixed point for B_π is unique. What is the fixed point of B_π? [5 pts]

In value iteration, we repeatedly apply the Bellman backup operator B to improve our value function. At the end of value iteration, we can recover a greedy policy π from the value function using the equation below:

$$\pi(s) = \arg\max_a [r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V(s')]$$

Suppose we run value iteration for a finite number of steps to obtain a value function V (V has not necessarily converged to V^*). Say now that we evaluate our policy π obtained using the formula above to get V^π. **Note that here and for the rest of Q2, π refers to the greedy policy.**

(c) Is V^π always the same as V? Justify your answer. [5 pts]

In lecture, we learned that running value iteration until a certain tolerance can bring us close to recovering the optimal value function. Let V_n and V_{n+1} be the outputs of value iteration at the n^{th} and $n + 1^{th}$ iterations respectively. Let $\varepsilon > 0$ and consider the point in value iteration such that $||V_{n+1} - V_n|| < \frac{\varepsilon(1-\gamma)}{2\gamma}$. Let π be the greedy policy given the value function V_{n+1}.
You will now prove that this policy π is ε-optimal. This result justifies why halting value iteration when the difference between success iterations is sufficiently small, ensures the decision policy obtained by being greedy with respect to the value function, is near-optimal.
Precisely if

$$||V_{n+1} - V_n|| < \frac{\varepsilon(1-\gamma)}{2\gamma}$$

then,

$$||V^\pi - V^*|| \leq \varepsilon$$

(d) When π is the greedy policy, what is the relationship between B and B_π? [2 pts]

(e) Prove that $||V^\pi - V_{n+1}|| \leq \varepsilon/2$.
Hint: Introduce an in-between term and leverage the triangle inequality. [6 pts]

(f) Prove $||B^k V - B^k V'|| \leq \gamma^k ||V - V'||$ [3 pts]

(g) Prove that $||V^* - V_{n+1}|| \leq \varepsilon/2$. [7pts]
Hints: Note that $||V^* - V_{n+1}|| = ||V^* + V_{n+2} - V_{n+2} - V_{n+1}||$ and you can repeatedly apply this trick. It may also be useful to leverage part (f) and recall that V^* is the fixed point of the contraction B.

(h) Use the results from parts (e) and (g), to show that $||V^\pi - V^*|| \leq \varepsilon$ [2 pts]

3 Frozen Lake MDP [25 pts]

Now you will implement value iteration and policy iteration for the Frozen Lake environment from OpenAI Gym. We have provided custom versions of this environment in the starter code.

(a) **(coding)** Read through `vi_and_pi.py` and implement `policy_evaluation`, `policy_improvement` and `policy_iteration`. The stopping tolerance (defined as $\max_s |V_{old}(s) - V_{new}(s)|$) is tol $= 10^{-3}$. Use $\gamma = 0.9$. Return the optimal value function and the optimal policy. [10pts]

(b) **(coding)** Implement `value_iteration` in `vi_and_pi.py`. The stopping tolerance is tol $= 10^{-3}$. Use $\gamma = 0.9$. Return the optimal value function and the optimal policy. [10 pts]

(c) **(written)** Run both methods on the Deterministic-4x4-FrozenLake-v0 and

Stochastic-4x4-FrozenLake-v0 environments. In the second environment, the dynamics of the world are stochastic. How does stochasticity affect the number of iterations required, and the resulting policy? [5 pts]