CS234 Section 1

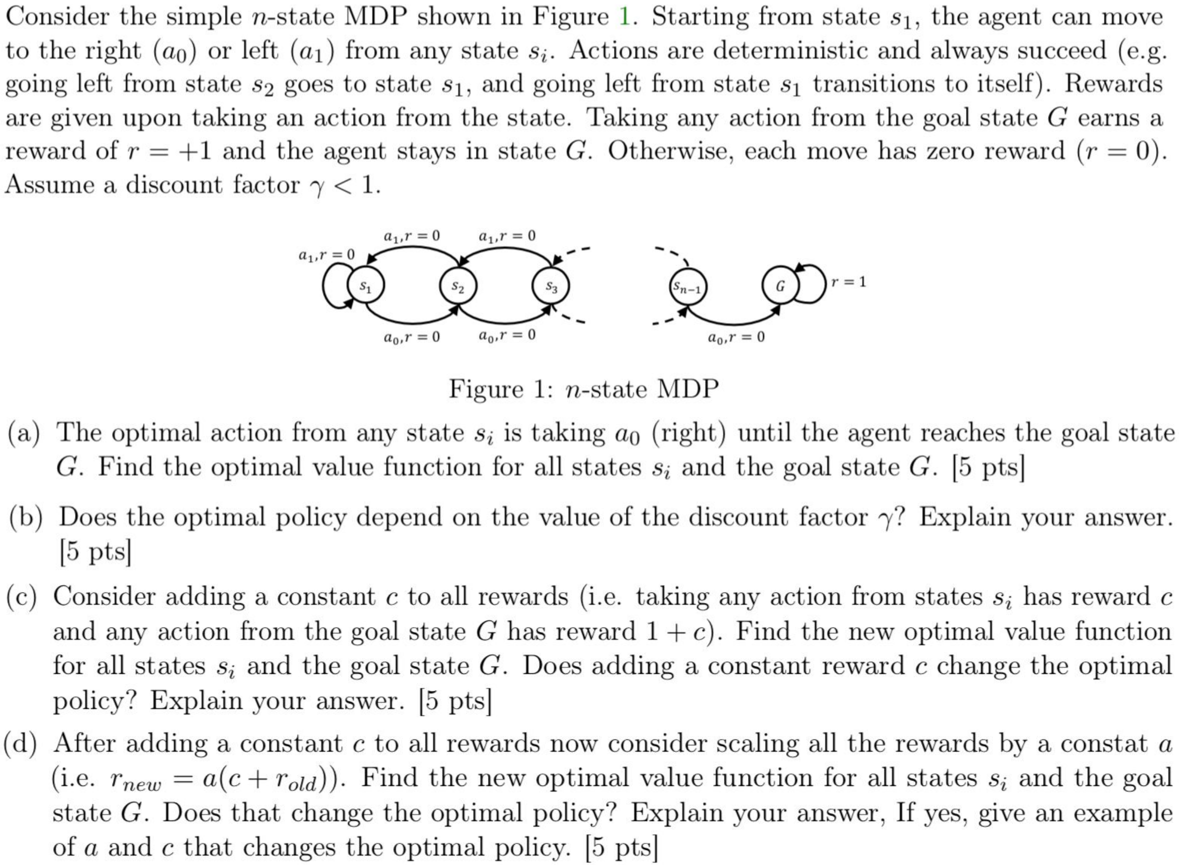
Outline

- Homework 1 Overview

- Exercises

- Key Concept Review

Homework 1 Q1



Consider the simple n-state MDP shown in Figure 1. Starting from state s_1, the agent can move to the right (a_0) or left (a_1) from any state s_i. Actions are deterministic and always succeed (e.g. going left from state s_2 goes to state s_1, and going left from state s_1 transitions to itself). Rewards are given upon taking an action from the state. Taking any action from the goal state G earns a reward of $r = +1$ and the agent stays in state G. Otherwise, each move has zero reward ($r = 0$). Assume a discount factor $\gamma < 1$.

Figure 1: n-state MDP

(a) The optimal action from any state s_i is taking a_0 (right) until the agent reaches the goal state G. Find the optimal value function for all states s_i and the goal state G. [5 pts]

(b) Does the optimal policy depend on the value of the discount factor γ? Explain your answer. [5 pts]

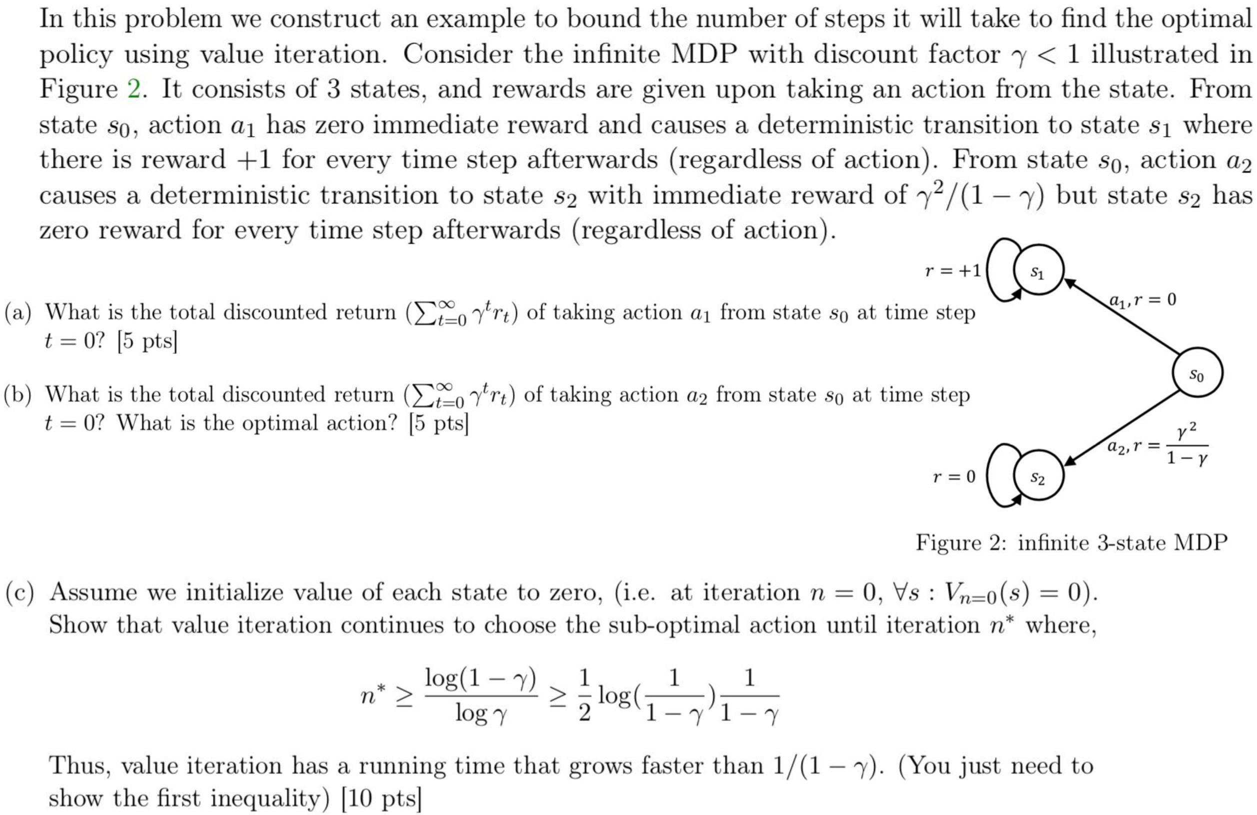
(c) Consider adding a constant c to all rewards (i.e. taking any action from states s_i has reward c and any action from the goal state G has reward $1 + c$). Find the new optimal value function for all states s_i and the goal state G. Does adding a constant reward c change the optimal policy? Explain your answer. [5 pts]

(d) After adding a constant c to all rewards now consider scaling all the rewards by a constat a (i.e. $r_{new} = a(c + r_{old})$). Find the new optimal value function for all states s_i and the goal state G. Does that change the optimal policy? Explain your answer, If yes, give an example of a and c that changes the optimal policy. [5 pts]

Hint:

Deterministic environment

Homework 1 Q2



In this problem we construct an example to bound the number of steps it will take to find the optimal policy using value iteration. Consider the infinite MDP with discount factor $\gamma < 1$ illustrated in Figure 2. It consists of 3 states, and rewards are given upon taking an action from the state. From state s_0, action a_1 has zero immediate reward and causes a deterministic transition to state s_1 where there is reward $+1$ for every time step afterwards (regardless of action). From state s_0, action a_2 causes a deterministic transition to state s_2 with immediate reward of $\gamma^2/(1-\gamma)$ but state s_2 has zero reward for every time step afterwards (regardless of action).

(a) What is the total discounted return ($\sum_{t=0}^{\infty} \gamma^t r_t$) of taking action a_1 from state s_0 at time step $t = 0$? [5 pts]

(b) What is the total discounted return ($\sum_{t=0}^{\infty} \gamma^t r_t$) of taking action a_2 from state s_0 at time step $t = 0$? What is the optimal action? [5 pts]

Figure 2: infinite 3-state MDP

(c) Assume we initialize value of each state to zero, (i.e. at iteration $n = 0$, $\forall s : V_{n=0}(s) = 0$). Show that value iteration continues to choose the sub-optimal action until iteration n^* where,

$$n^* \geq \frac{\log(1-\gamma)}{\log \gamma} \geq \frac{1}{2} \log(\frac{1}{1-\gamma}) \frac{1}{1-\gamma}$$

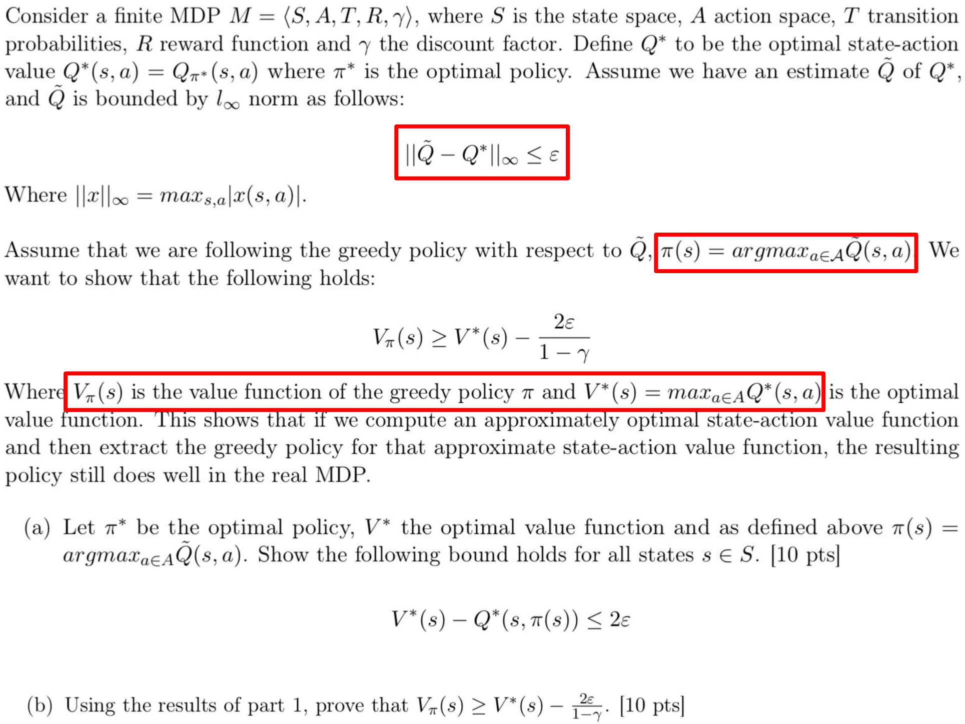
Thus, value iteration has a running time that grows faster than $1/(1-\gamma)$. (You just need to show the first inequality) [10 pts]

Hint:

Deterministic environment

c) before n*, always choose sub-optimal action; after n*, optimal action.

Homework 1 Q3



Consider a finite MDP $M = \langle S, A, T, R, \gamma \rangle$, where S is the state space, A action space, T transition probabilities, R reward function and γ the discount factor. Define Q^* to be the optimal state-action value $Q^*(s,a) = Q_{\pi^*}(s,a)$ where π^* is the optimal policy. Assume we have an estimate $\tilde{Q}$ of Q^*, and $\tilde{Q}$ is bounded by l_∞ norm as follows:

$$\boxed{||\tilde{Q} - Q^*||_\infty \leq \varepsilon}$$

Where $||x||_\infty = max_{s,a}|x(s,a)|$.

Assume that we are following the greedy policy with respect to $\tilde{Q}$, $\boxed{\pi(s) = argmax_{a \in A}\tilde{Q}(s,a)}$. We want to show that the following holds:

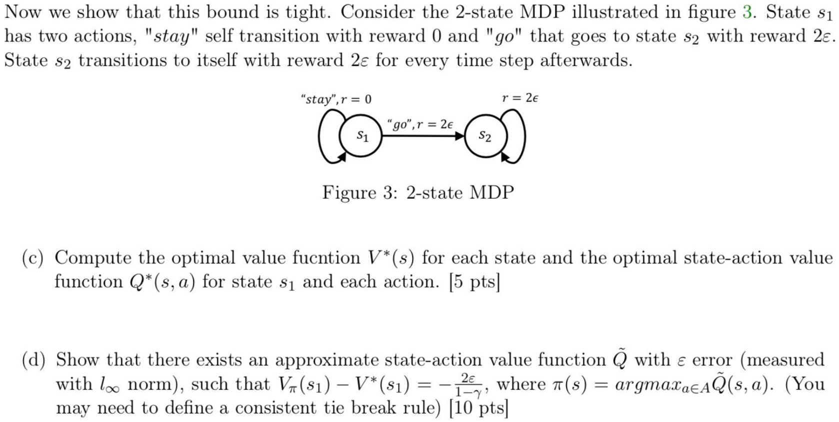
$$V_\pi(s) \geq V^*(s) - \frac{2\varepsilon}{1-\gamma}$$

Where $\boxed{V_\pi(s) \text{ is the value function of the greedy policy } \pi \text{ and } V^*(s) = max_{a\in A}Q^*(s,a)}$ is the optimal value function. This shows that if we compute an approximately optimal state-action value function and then extract the greedy policy for that approximate state-action value function, the resulting policy still does well in the real MDP.

(a) Let π^* be the optimal policy, V^* the optimal value function and as defined above $\pi(s) = argmax_{a\in A}\tilde{Q}(s,a)$. Show the following bound holds for all states $s \in S$. [10 pts]

$$V^*(s) - Q^*(s, \pi(s)) \leq 2\varepsilon$$

(b) Using the results of part 1, prove that $V_\pi(s) \geq V^*(s) - \frac{2\varepsilon}{1-\gamma}$. [10 pts]



Now we show that this bound is tight. Consider the 2-state MDP illustrated in figure 3. State s_1 has two actions, "*stay*" self transition with reward 0 and "*go*" that goes to state s_2 with reward 2ε. State s_2 transitions to itself with reward 2ε for every time step afterwards.

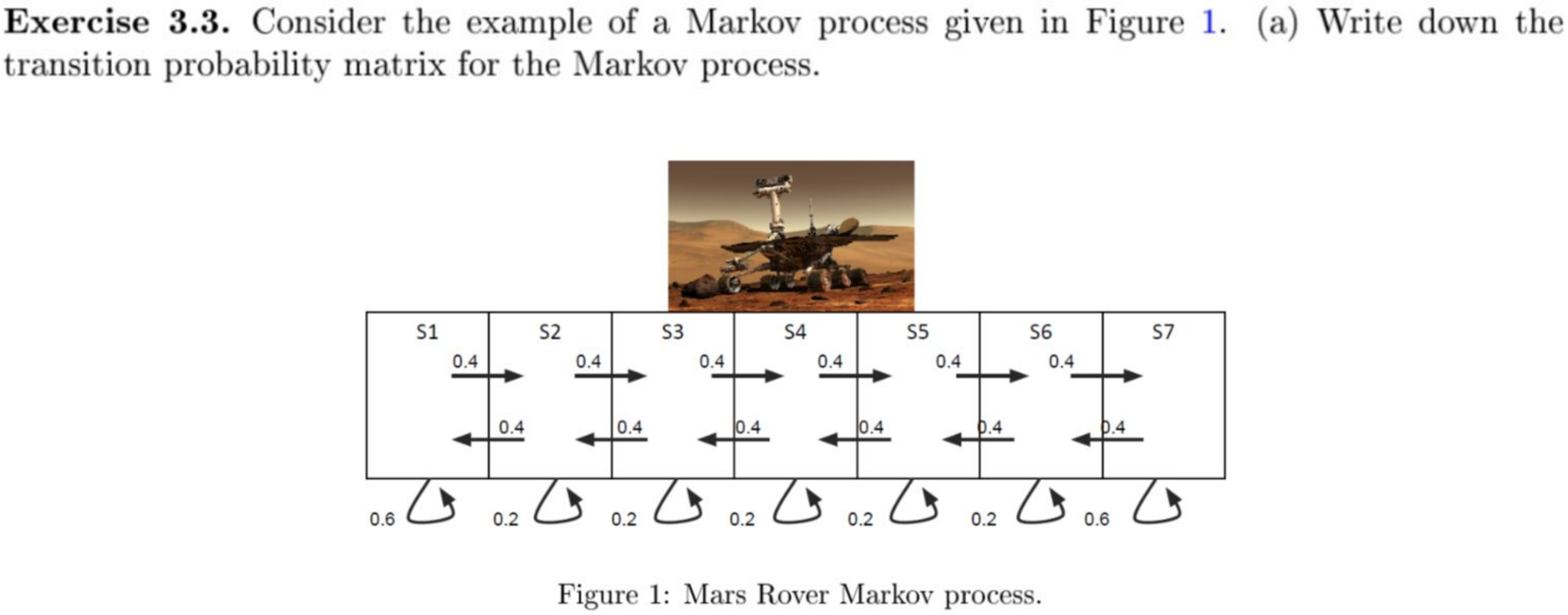
"stay", r = 0
"go", r = 2ϵ
r = 2ϵ
s_1
s_2

Figure 3: 2-state MDP

(c) Compute the optimal value fucntion $V^*(s)$ for each state and the optimal state-action value function $Q^*(s,a)$ for state s_1 and each action. [5 pts]

(d) Show that there exists an approximate state-action value function $\tilde{Q}$ with ε error (measured with l_∞ norm), such that $V_\pi(s_1) - V^*(s_1) = -\frac{2\varepsilon}{1-\gamma}$, where $\pi(s) = argmax_{a\in A}\tilde{Q}(s,a)$. (You may need to define a consistent tie break rule) [10 pts]

Exercise 1



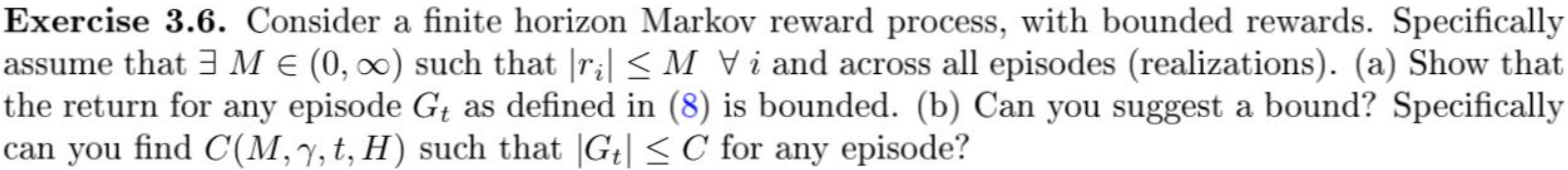
Exercise 3.3. Consider the example of a Markov process given in Figure 1. (a) Write down the transition probability matrix for the Markov process.

Figure 1: Mars Rover Markov process.

Exercise 1

$$\mathbf{P} \;=\; \begin{array}{cc} \begin{array}{ccccccc} S1 & S2 & S3 & S4 & S5 & S6 & S7 \end{array} & \\ \left(\begin{array}{ccccccc} 0.6 & 0.4 & 0 & 0 & 0 & 0 & 0 \\ 0.4 & 0.2 & 0.4 & 0 & 0 & 0 & 0 \\ 0 & 0.4 & 0.2 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.4 & 0.2 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0.2 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0.4 & 0.2 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0.4 & 0.6 \end{array}\right) & \begin{array}{c} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \\ S7 \end{array} \end{array}$$

Exercise 2

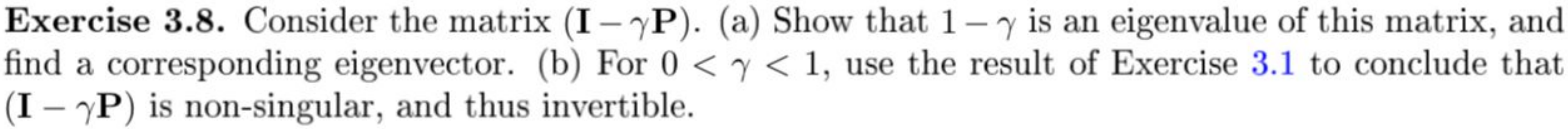


Exercise 3.6. Consider a finite horizon Markov reward process, with bounded rewards. Specifically assume that $\exists\ M \in (0, \infty)$ such that $|r_i| \leq M \ \ \forall\ i$ and across all episodes (realizations). (a) Show that the return for any episode G_t as defined in (8) is bounded. (b) Can you suggest a bound? Specifically can you find $C(M, \gamma, t, H)$ such that $|G_t| \leq C$ for any episode?

Exercise 2

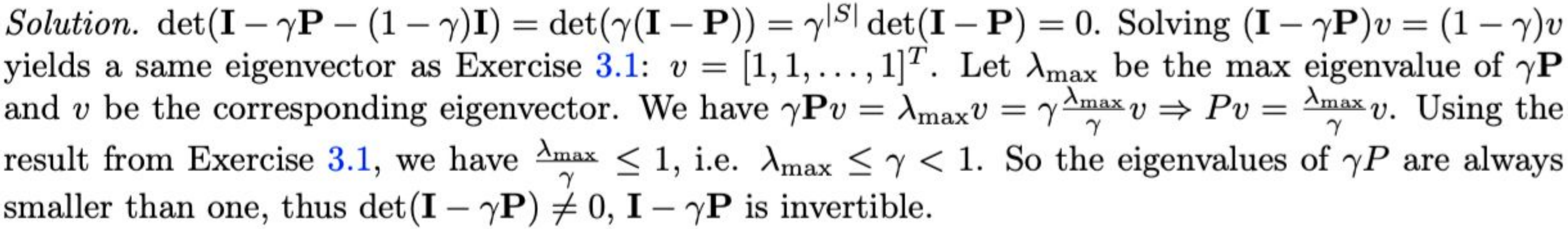
$$
\begin{aligned}
|G_t| &= |\sum_{i=t}^{H-1} \gamma^{i-t} r_i| \\
&\leq \sum_{i=t}^{H-1} \gamma^{i-t} |r_i| \\
&\leq \sum_{i=t}^{H-1} \gamma^{i-t} M \\
&= \frac{1-\gamma^{H-t}}{1-\gamma} M \\
&\leq \frac{1}{1-\gamma} M
\end{aligned}
$$

Exercise 3



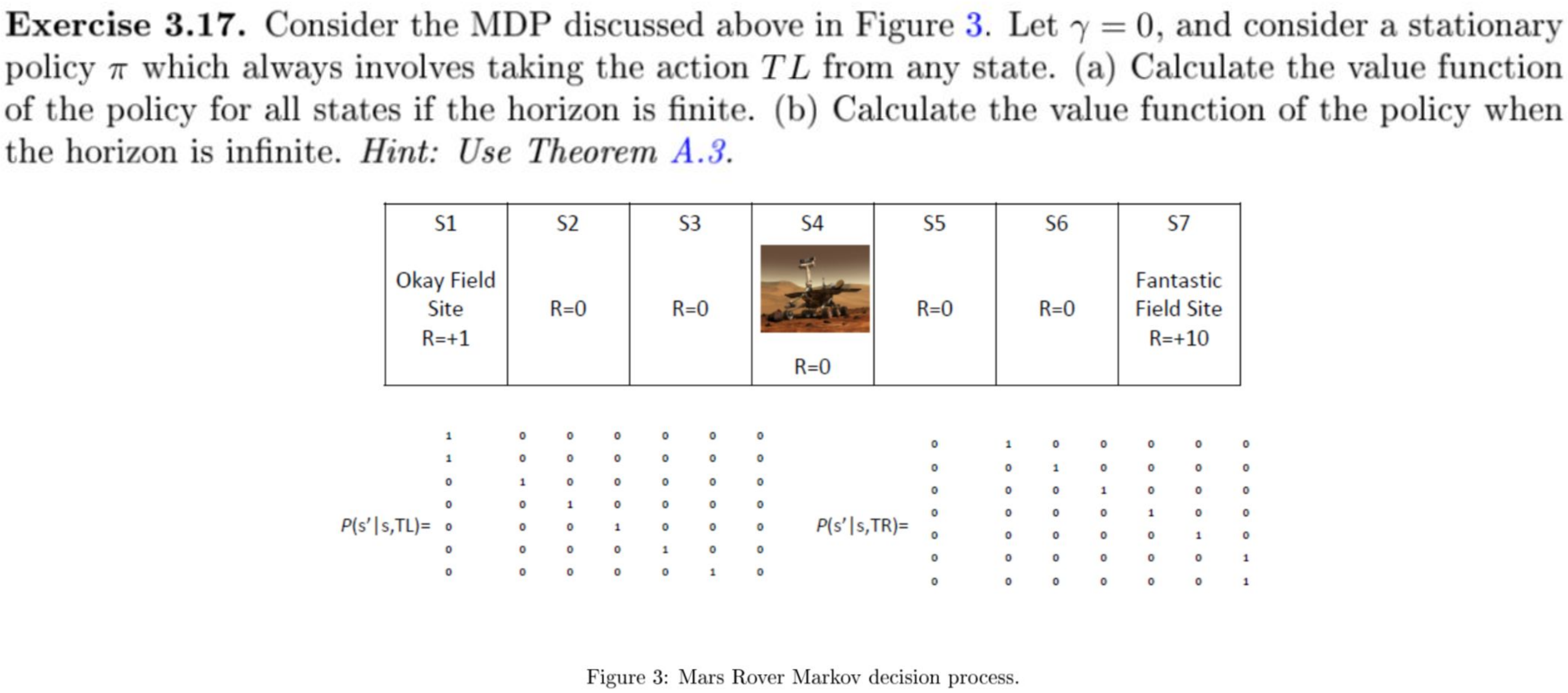
Exercise 3.8. Consider the matrix $(\mathbf{I} - \gamma \mathbf{P})$. (a) Show that $1 - \gamma$ is an eigenvalue of this matrix, and find a corresponding eigenvector. (b) For $0 < \gamma < 1$, use the result of Exercise 3.1 to conclude that $(\mathbf{I} - \gamma \mathbf{P})$ is non-singular, and thus invertible.

Exercise 3



Solution. $\det(\mathbf{I} - \gamma\mathbf{P} - (1-\gamma)\mathbf{I}) = \det(\gamma(\mathbf{I} - \mathbf{P})) = \gamma^{|S|}\det(\mathbf{I} - \mathbf{P}) = 0$. Solving $(\mathbf{I} - \gamma\mathbf{P})v = (1-\gamma)v$ yields a same eigenvector as Exercise 3.1: $v = [1, 1, \ldots, 1]^T$. Let $\lambda_{\max}$ be the max eigenvalue of $\gamma\mathbf{P}$ and v be the corresponding eigenvector. We have $\gamma\mathbf{P}v = \lambda_{\max}v = \gamma\frac{\lambda_{\max}}{\gamma}v \Rightarrow Pv = \frac{\lambda_{\max}}{\gamma}v$. Using the result from Exercise 3.1, we have $\frac{\lambda_{\max}}{\gamma} \leq 1$, i.e. $\lambda_{\max} \leq \gamma < 1$. So the eigenvalues of γP are always smaller than one, thus $\det(\mathbf{I} - \gamma\mathbf{P}) \neq 0$, $\mathbf{I} - \gamma\mathbf{P}$ is invertible.

Exercise 4



Exercise 3.17. Consider the MDP discussed above in Figure 3. Let $\gamma = 0$, and consider a stationary policy π which always involves taking the action TL from any state. (a) Calculate the value function of the policy for all states if the horizon is finite. (b) Calculate the value function of the policy when the horizon is infinite. *Hint: Use Theorem A.3.*

S1	S2	S3	S4	S5	S6	S7
Okay Field Site R=+1	R=0	R=0	R=0	R=0	R=0	Fantastic Field Site R=+10

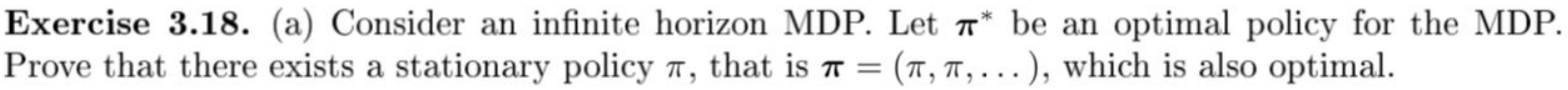
$$P(s'|s,TL)=\begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{matrix} \qquad P(s'|s,TR)=\begin{matrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{matrix}$$

Figure 3: Mars Rover Markov decision process.

Exercise 4

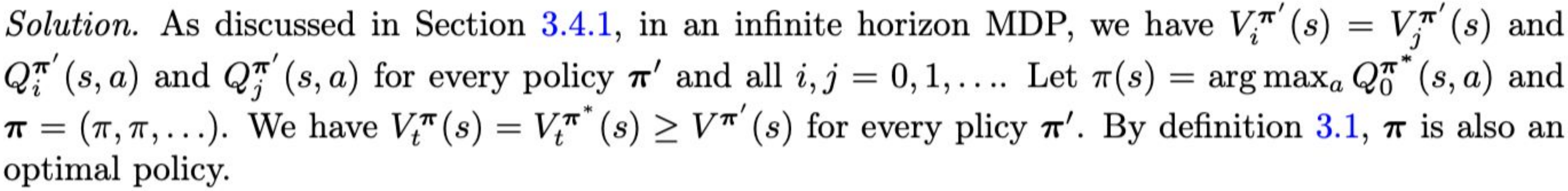
$$V^{\pi} = [1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 10]^{T}$$

Exercise 5



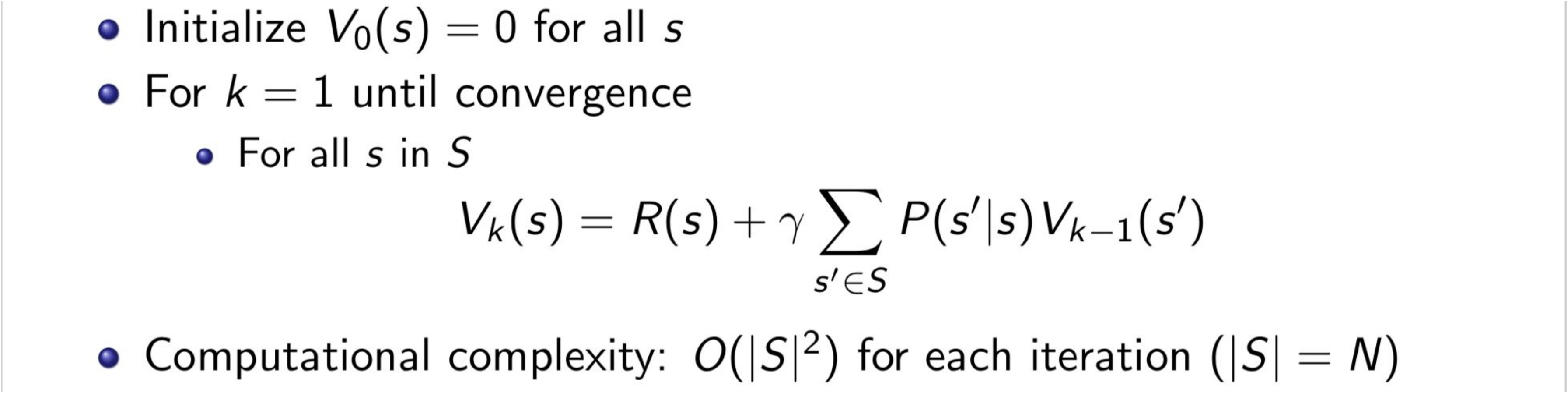
Exercise 3.18. (a) Consider an infinite horizon MDP. Let $\boldsymbol{\pi}^*$ be an optimal policy for the MDP. Prove that there exists a stationary policy π, that is $\boldsymbol{\pi} = (\pi, \pi, \dots)$, which is also optimal.

Exercise 5



Solution. As discussed in Section 3.4.1, in an infinite horizon MDP, we have $V_i^{\boldsymbol{\pi}'}(s) = V_j^{\boldsymbol{\pi}'}(s)$ and $Q_i^{\boldsymbol{\pi}'}(s,a)$ and $Q_j^{\boldsymbol{\pi}'}(s,a)$ for every policy $\boldsymbol{\pi}'$ and all $i, j = 0, 1, \ldots$. Let $\pi(s) = \arg\max_a Q_0^{\boldsymbol{\pi}^*}(s,a)$ and $\boldsymbol{\pi} = (\pi, \pi, \ldots)$. We have $V_t^{\boldsymbol{\pi}}(s) = V_t^{\boldsymbol{\pi}^*}(s) \geq V^{\boldsymbol{\pi}'}(s)$ for every plicy $\boldsymbol{\pi}'$. By definition 3.1, $\boldsymbol{\pi}$ is also an optimal policy.

Policy Evaluation



- Initialize $V_0(s) = 0$ for all s
- For $k = 1$ until convergence
  - For all s in S

$$V_k(s) = R(s) + \gamma \sum_{s' \in S} P(s'|s) V_{k-1}(s')$$

- Computational complexity: $O(|S|^2)$ for each iteration ($|S| = N$)

Policy Iteration

- Set $i = 0$
- Initialize $\pi_0(s)$ randomly for all states s
- While $i == 0$ or $\|\pi_i - \pi_{i-1}\|_1 > 0$ (L1-norm, measures if the policy changed for any state):
  - $V^{\pi_i} \leftarrow$ MDP V function policy **evaluation** of π_i
  - $\pi_{i+1} \leftarrow$ Policy **improvement**
  - $i = i + 1$

Policy evaluation:

- Compute state-action value of a policy π_i
  - For s in S and a in A:

$$Q^{\pi_i}(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^{\pi_i}(s')$$

Policy Improvement:

- Compute new policy π_{i+1}, for all $s \in S$

$$\pi_{i+1}(s) = \arg\max_a Q^{\pi_i}(s,a) \ \forall s \in S$$

Proof: Monotonic Improvement

$$
\begin{aligned}
V^{\pi_i}(s) &\leq \max_a Q^{\pi_i}(s,a) \\
&= \max_a R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^{\pi_i}(s') \\
&= R(s,\pi_{i+1}(s)) + \gamma \sum_{s' \in S} P(s'|s,\pi_{i+1}(s)) V^{\pi_i}(s') \quad //\text{by the definition of } \pi_{i+1} \\
&\leq R(s,\pi_{i+1}(s)) + \gamma \sum_{s' \in S} P(s'|s,\pi_{i+1}(s)) \left(\max_{a'} Q^{\pi_i}(s',a')\right) \\
&= R(s,\pi_{i+1}(s)) + \gamma \sum_{s' \in S} P(s'|s,\pi_{i+1}(s)) \\
&\quad \left(R(s',\pi_{i+1}(s')) + \gamma \sum_{s'' \in S} P(s''|s',\pi_{i+1}(s')) V^{\pi_i}(s'')\right) \\
&\ \ \vdots \\
&= V^{\pi_{i+1}}(s)
\end{aligned}
$$

Value Iteration

- Set $k = 1$
- Initialize $V_0(s) = 0$ for all states s
- Loop until [finite horizon, convergence]:
  - For each state s

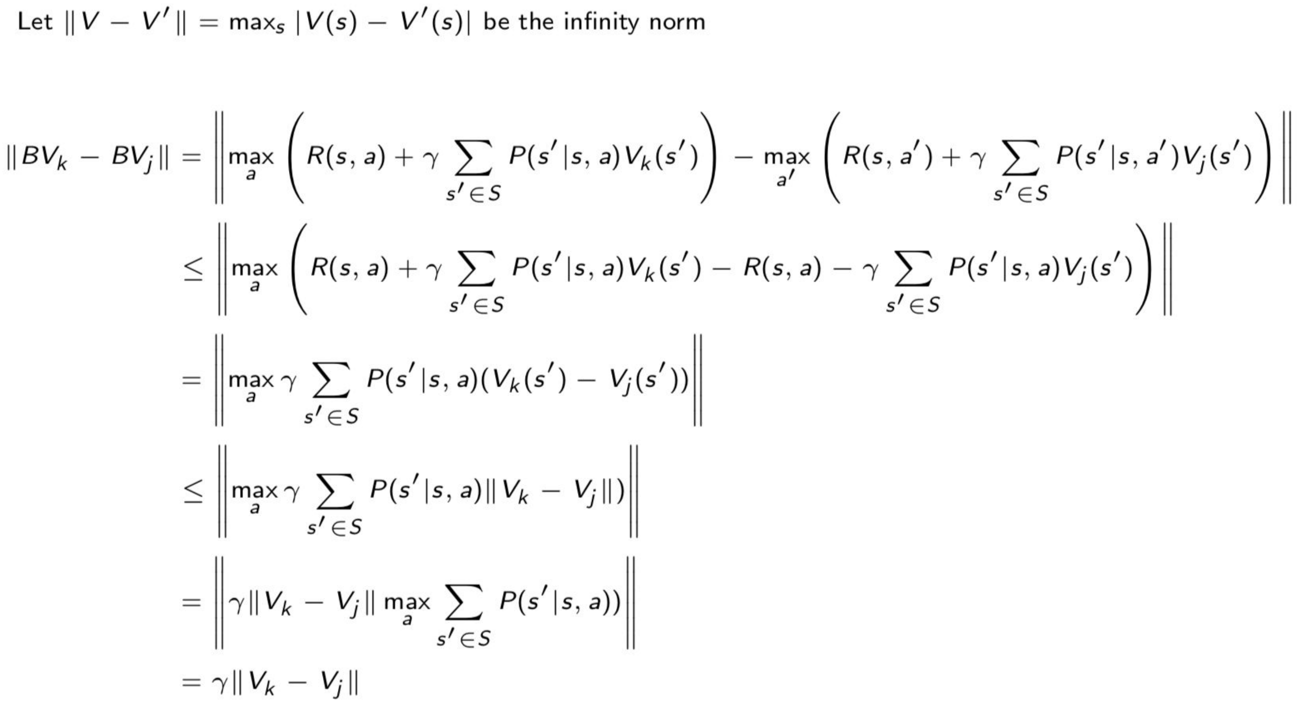
$$V_{k+1}(s) = \max_a R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V_k(s')$$

  - View as Bellman backup on value function

$$V_{k+1} = BV_k$$

$$\pi_{k+1}(s) = \arg\max_a R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V_k(s')$$

Proof: Bellman Backup is a Contraction Operator



Let $\|V - V'\| = \max_s |V(s) - V'(s)|$ be the infinity norm

$$\begin{aligned}
\|BV_k - BV_j\| &= \left\| \max_a \left(R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V_k(s') \right) - \max_{a'} \left(R(s,a') + \gamma \sum_{s' \in S} P(s'|s,a') V_j(s') \right) \right\| \\
&\leq \left\| \max_a \left(R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V_k(s') - R(s,a) - \gamma \sum_{s' \in S} P(s'|s,a) V_j(s') \right) \right\| \\
&= \left\| \max_a \gamma \sum_{s' \in S} P(s'|s,a)(V_k(s') - V_j(s')) \right\| \\
&\leq \left\| \max_a \gamma \sum_{s' \in S} P(s'|s,a) \|V_k - V_j\|) \right\| \\
&= \left\| \gamma \|V_k - V_j\| \max_a \sum_{s' \in S} P(s'|s,a)) \right\| \\
&= \gamma \|V_k - V_j\|
\end{aligned}$$