SI252 Reinforcement Learning 2020/04/08

Homework 5

Professor: Ziyu Shao *Due: 2020/05/15 11:59am*

1. **Optimal Policy for Simple MDP**

   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 $\gamma < 1$.

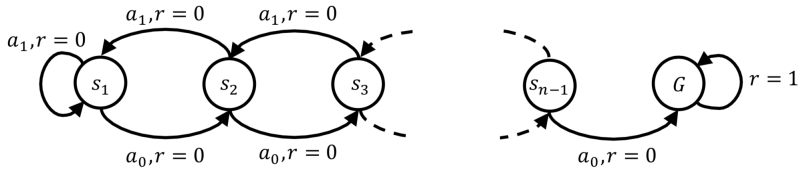


Figure 1: n-state MDP

   (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.

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

   (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.

   (d) After adding a constant c to all rewards now consider scaling all the rewards by a constant 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.

2. **Running Time of Value Iteration**

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).

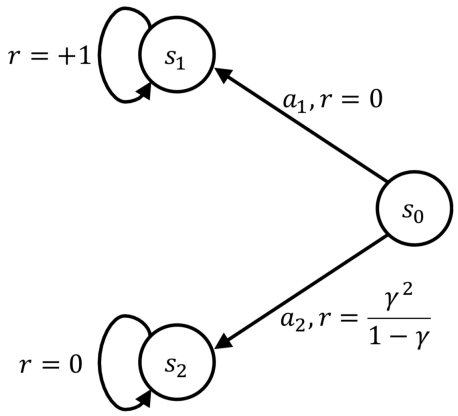


Figure 2: infinite 3-state MDP

(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 $t = 0$?

(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?

(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)

3. **Approximating the Optimal Value Function**

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:

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

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

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

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

Where $V_\pi(s)$ is the value function of the greedy policy π and $V^*(s) = max_{a \in \mathcal{A}} Q^*(s,a)$ 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 $\pi(s) = argmax_{a \in \mathcal{A}} \tilde{Q}(s,a)$. Show the following bound holds for all states $s \in S$.

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

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

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.

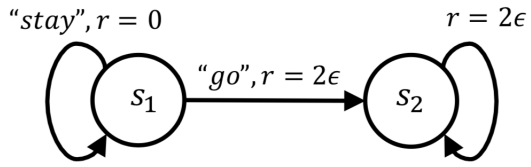


Figure 3: 2-state MDP

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

(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 $\pi(s) = argmax_{a \in \mathcal{A}} \tilde{Q}(s,a)$. (You may need to define a consistent tie break rule)

4. **Bellman Equations**

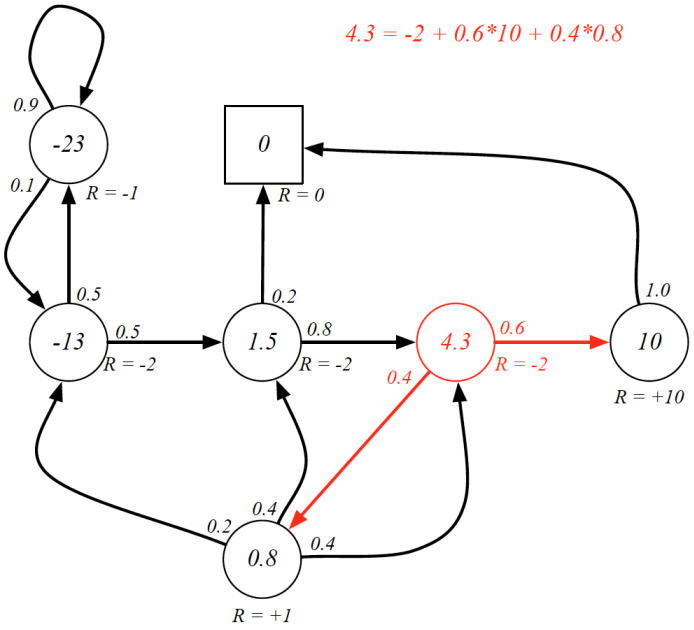
   (a) Reproduce the proof of Bellman equation for Markov Reward Processes (MRPs).

   (b) Reproduce the proofs of Bellman Expectation Equations for Markov Decision Processes (MDPs).

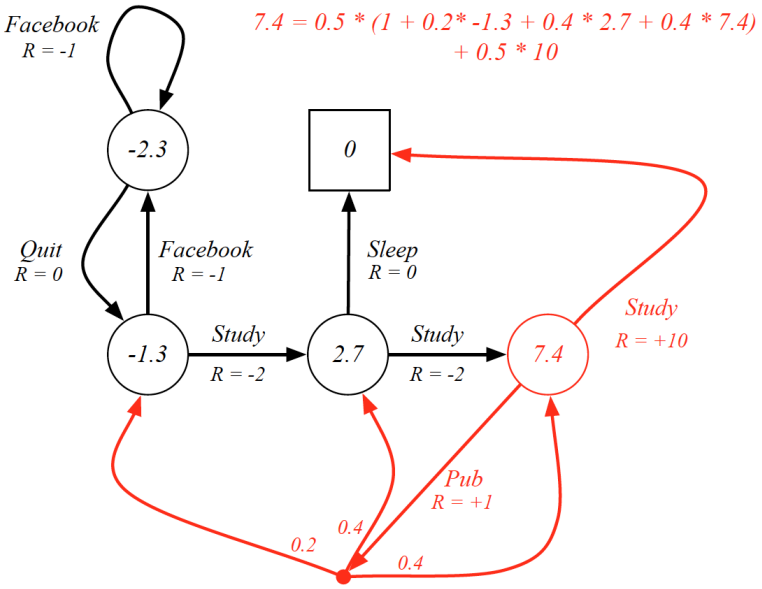
   (c) Reproduce the proofs of Bellman Optimality Equations for Markov Decision Processes (MDPs).

5. **Student MRP & MDPs**

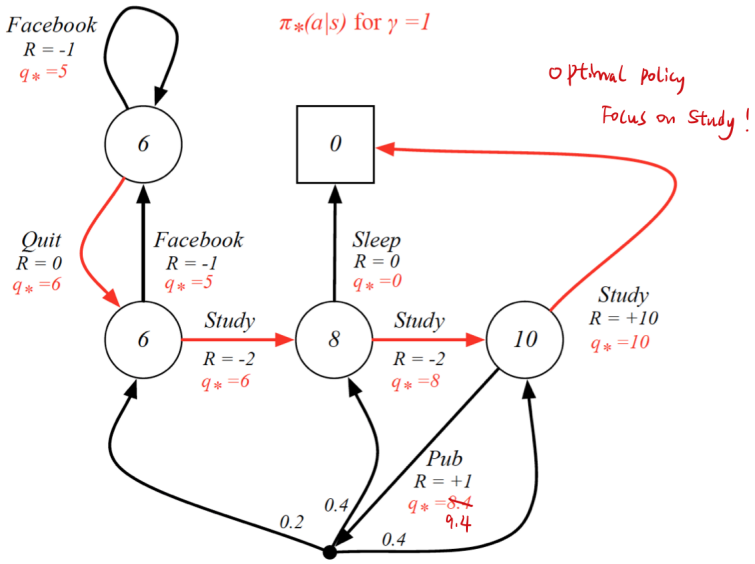
   (a) Reproduce the state values for student MRP



   (b) Reproduce the state values & state-action values for student MDP

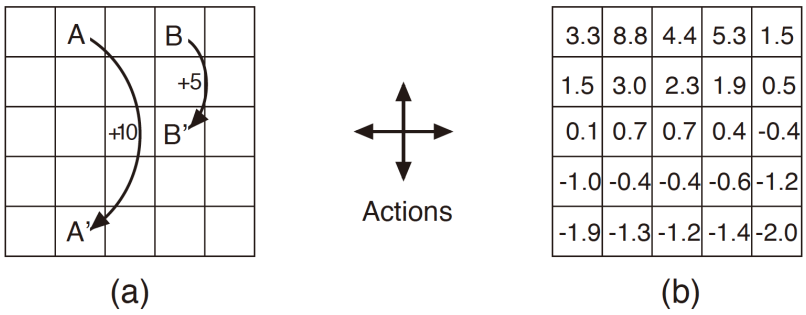


(c) Reproduce the optimal state values, optimal state-action values, and optimal policy for student MDP



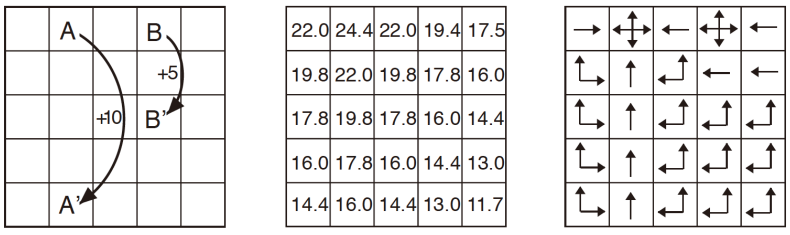
6. **5x5 Grid World**

(a) Reproduce the state values under the uniform random policy.



What is the value function for the uniform random policy?

(b) Reproduce the optimal state values, optimal state-action values, and optimal policy



What is the optimal value function over all possible policies?
What is the optimal policy?

7. **Bonus Problem: A Gambling Model**.
At each play of the game, a gambler can bet any nonnegative amount up to his present fortune and will either win or lose that amount with probabilities p and $q = 1 - p$, respectively. The gambler is allowed to make n bets, and his objective is to maximize the expectation of the logarithm of his final fortune. Fine the optimal strategy for the gambler.

8. **Bonus Problem: Accepting the Best Offer**.
A senior undergraduate student of SIST in ShanghaiTech University have applied many top graduate programs around the world. He is now presented with n offers in a sequential order. After looking at an offer, he must decide whether to accept it (and terminate the process) or to reject it. Once rejected, an offer is lost. Suppose that the only information he has at any time is the relative rank of the present offer compared with previously ones. His objective is to maximize the probability of selecting the best

offer when all $n!$ orderings of the offers are assumed to be equally likely. Please help help him to find the optimal policy and compute the corresponding probability.

9. **Bonus Problem: Bayesian Bandit Process**.
There are two arms which may be pulled repeatedly in any order. Each pull may result in either a success or a failure. The sequence of successes and failures which results from pulling arm i (i=1 or 2) forms a Bernoulli process with unknown success probability θ_i. A success at the t^{th} pull yields a reward γ^{t-1} $(0 < \gamma < 1)$, while an unsuccessful pull yields a zero reward. At time zero, each θ_i has a beta prior distribution with two parameters α_i, β_i and these distributions are independent for different arms. These prior distributions are updated to successive posterior distributions as arms are pulled. Since the class of beta distributions is closed under Bernoulli sampling, these posterior distributions are all beta distributions. How should the arm to pull next at each stage be chosen so as to maximize the total expected reward from an infinite sequence of pulls?

   (a) One intuitive policy suggests that at each stage we should pull the arm for which the current expected value of $\theta_i$is the largest. This policy behaves very good in most cases. However, it is unfortunately not optimal. Please provide an example to show why such policy is not optimal.

   (b) For the expected total reward under an optimal policy, show that the following recurrence equation holds:

$$R_1(\alpha_1, \beta_1) = \frac{\alpha_1}{\alpha_1 + \beta_1}[1 + \gamma R(\alpha_1 + 1, \beta_1, \alpha_2, \beta_2)] + \frac{\beta_1}{\alpha_1 + \beta_1}[\gamma R(\alpha_1, \beta_1 + 1, \alpha_2, \beta_2)]$$

$$R_2(\alpha_2, \beta_2) = \frac{\alpha_2}{\alpha_2 + \beta_2}[1 + \gamma R(\alpha_1, \beta_1, \alpha_2 + 1, \beta_2)] + \frac{\beta_2}{\alpha_2 + \beta_2}[\gamma R(\alpha_1, \beta_1, \alpha_2, \beta_2 + 1)]$$

$$R(\alpha_1, \beta_1, \alpha_2, \beta_2) = \max\{R_1(\alpha_1, \beta_1), R_2(\alpha_2, \beta_2)\}$$

   (c) Discuss the computational complexity of solving the above equation. How to solve it approximately?

   (d) Find the optimal policy. (Hint: Gittins index policy)