Homework 1

ELEN E6885: Introduction to Reinforcement Learning

September 15, 2022

Problem 1 (**2-Armed Bandit, 20 Points**)

Consider the following 2-armed bandit problem: the first arm has a fixed reward 0.3 and the second arm has a 0-1 reward following a Bernoulli distribution with probability 0.6, i.e., arm 2 yields reward 1 with probability 0.6. Assume we selected arm 1 at $t = 1$, and arm 2 four times at $t = 2, 3, 4, 5$ with reward $0, 1, 0, 0$, respectively. We use the sample-average technique to estimate the action-value, and then use it to guide our choices starting from $t = 6$.

1. [5 pts] Which arm will be played at $t = 6, 7$, respectively, if the greedy method is used to select actions?

2. [10 pts] What is the probability to play arm 2 at $t = 6, 7$, respectively, if the ϵ-greedy method is used to select actions ($\epsilon = 0.1$)?

3. [5 pts] Why could the greedy method perform significantly worse than the ϵ-greedy method in the long run?

Problem 2 (**Softmax, 15 Points**)

For the softmax action selection, show the following.

1. [5 pts] In the limit as the *temperature* $\tau \to 0$, softmax action selection becomes the same as greedy action selection.

2. [5 pts] In the limit as $\tau \to \infty$, softmax action selection yields equiprobable selection among all actions.

3. [5 pts] In the case of two actions, the softmax operation using the Gibbs distribution becomes the logistic (or sigmoid) function commonly used in artificial neural networks.

Problem 3 (**Incremental Implementation, 15 Points**)

Suppose we have a sequence of returns $G_1, G_2, \cdots, G_{n-1}$, all starting in the same state and each with a corresponding random weight W_i, $i = 1, 2, \cdots, n-1$. We wish to form the estimate

$$V_n = \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}, \; n \geq 2,$$

and keep it up-to-date as we obtain an additional return G_n. In addition to keeping track of V_n, we must maintain for each state the cumulative sum C_n of the weights given to the first n returns. Show that the update rule for $V_{n+1}, n \geq 1$ is

$$V_{n+1} = V_n + \frac{W_n}{C_n}\Big[G_n - V_n\Big],$$

and

$$C_{n+1} = C_n + W_{n+1},$$

where $C_0 = 0$ (and V_1 is arbitrary and thus need not be specified).

Problem 4 (**MDP: State-Value Function, 15 Points**)

Compute the state value for S_t in all the MDPs in Fig. 1 – 3. The decimal number above lines refers to the probability of choosing the corresponding action. The value r refers to reward, which can be deterministic or stochastic. Assume $\gamma = 1$ for all questions and all terminal states (i.e., no successors in the graph) always have zero values.

Problem 5 (**MDP: Optimal Policy, 10 Points**)

Given an arbitrary MDP with reward function $\mathcal{R}_s^a$ and constants α and $\beta > 0$, prove that the following modified MDPs have the same optimal policy as the original MDP.

1. [5 pts] Everything remains the same as the original MDP, except it has a new reward function $\alpha + \mathcal{R}_s^a$. Assume that there is no terminal state and discount factor $\gamma < 1$.

2. [5 pts] Everything remains the same as the original MDP, except it has a new reward function $\beta \cdot \mathcal{R}_s^a$.

Problem 6 (**MDP: Simple Card Game, 25 Points**)

In a card game, you repeatedly draw a card (with replacement) that is equally likely to be a 2 or 3. You can either *Draw* or *Stop* if the total score of the cards you have drawn is less than 6. Otherwise, you must Stop. When you Stop, your reward is equal to your total score (up to 5), or zero if you get a total of 6 or higher. When you Draw, you receive no reward. Assume there is no discount ($\gamma = 1$). We formulate this problem as an MDP with the following states: $0, 2, 3, 4, 5$, and a *Done* state for when the game ends.

1. [10 pts] What is the state transition function and the reward function for this MDP?

2. [12 pts] What is the optimal state-value function and optimal action-value function for this MDP? (Hint: Solve Bellman optimality equation starting from states 4 and 5.)

3. [3 pts] What is the optimal policy for this MDP?

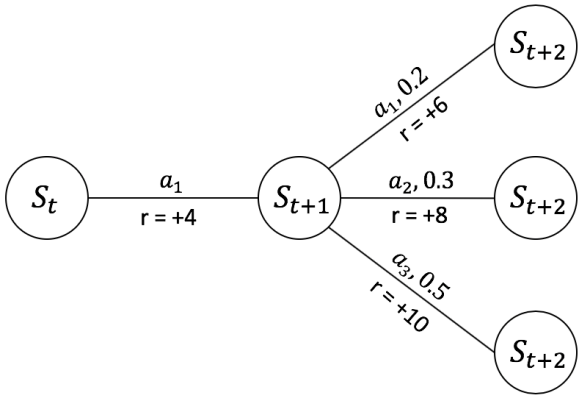

Figure 1: MDP with deterministic transitions

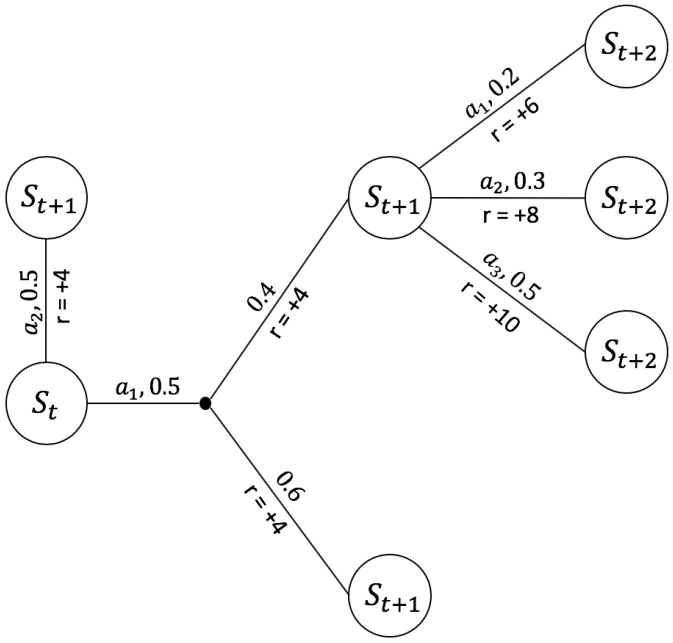

Figure 2: MDP with stochastic transitions

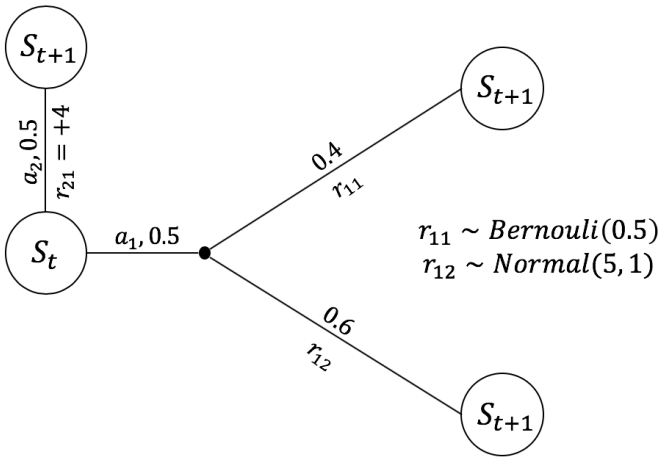

Figure 3: MDP with stochastic rewards