CMPUT 397
October 23, 2020

Worksheet C2M3

1. Consider the following MDP, with three states B, C and D ($\mathcal{S} = \{B, C, D\}$), and 2 actions ($\mathcal{A} = \{1, 2\}$), with $\gamma = 1.0$. Assume the action values are initialized $Q(s, a) = 0 \; \forall \; s \in \mathcal{S}$ and $a \in \mathcal{A}$. The agent takes actions according to an ϵ-greedy policy with $\epsilon = 0.1$.

(a) What is the optimal policy for this MDP? What are the action-values corresponding to the optimal policy: $q^*(s, a)$?

(b) Imagine the agent experienced a single episode, and the following experience: $S_0 = B, A_0 = 2, R_1 = 0, S_1 = D, A_1 = 2, R_2 = 4$. What are the Sarsa updates during this episode, assuming $\alpha = 0.1$? Start with state B, and compute and apply the Sarsa update. Then compute and apply the Sarsa update for the value of state D.

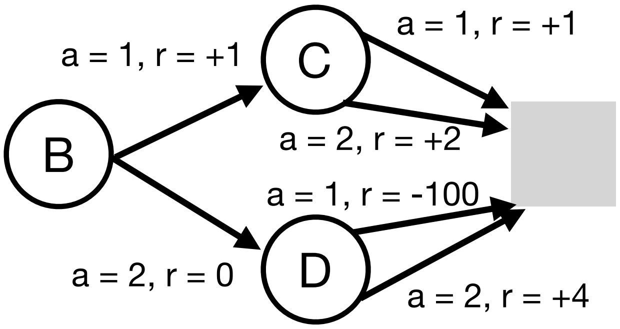
(c) Using the sample episode above, compute the updates Q-learning would make, with $\alpha = 0.1$. Again start with state B, and then state D.

(d) Let's consider one more episode: $S_0 = B, A_0 = 2, R_1 = 0, S_1 = D, A_1 = 1, R_2 = -100$. What would the Sarsa updates be? And what would the Q-learning updates be?

(e) Assume you see one more episode, and it's the same one as in 1d. Once more update the action values, for Sarsa and Q-learning. What do you notice?

(f) What policy does Q-learning converge to? What policy does Sarsa converge to?

Deterministic transitions



2. In Monte Carlo control, we required that every state-action pair be visited infinitely often. One way this can be guaranteed is by using exploring starts. Can we use exploring starts for Sarsa? Further, we have talked about using Sarsa with an ϵ-greedy policy. Can we use Monte Carlo with an ϵ-greedy policy? Does this ensure sufficient exploration?

3. (*Exercise 6.11 S&B*) Why is Q-learning considered an off-policy control method? Why is Sarsa considered on-policy, but Expected Sarsa can be used off-policy?

4. (*Exercise 6.12 S&B*) Suppose action selection is greedy. Is Q-learning then exactly the same algorithm as Sarsa? Will they make exactly the same action selections and weight updates? (**Additional Challenge:** What about Expected Sarsa? Does it have the same or different updates as Q-learning or Sarsa?)

Sarsa (on-policy TD control) for estimating $Q \approx q_*$

```
Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\textit{terminal}, \cdot) = 0$

Loop for each episode:
    Initialize $S$
    Choose $A$ from $S$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
    Loop for each step of episode:
        Take action $A$, observe $R$, $S'$
        Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
        $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma Q(S', A') - Q(S, A)\big]$
        $S \leftarrow S'$; $A \leftarrow A'$;
    until $S$ is terminal
```

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

```
Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\textit{terminal}, \cdot) = 0$

Loop for each episode:
    Initialize $S$
    Loop for each step of episode:
        Choose $A$ from $S$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
        Take action $A$, observe $R$, $S'$
        $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
        $S \leftarrow S'$
    until $S$ is terminal
```

5. In this question we compare the variance of the target for Sarsa and Expected Sarsa. Recall the update for Sarsa is

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\right]$$

and for Expected Sarsa is

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \sum_{a' \in \mathcal{A}} \pi(a'|S_{t+1}) Q(S_{t+1}, a') - Q(S_t, A_t)\right].$$

(a) Start by comparing the part of the update that is different: $Q(S_{t+1}, A_{t+1})$ compared to $\sum_{a' \in \mathcal{A}} \pi(a'|S_{t+1}) Q(S_{t+1}, a')$. Write down the variance for these two terms, given $S_{t+1} = s'$.

$$\mathrm{Var}(Q(s', A_{t+1})) \quad \text{and} \quad \mathrm{Var}\left(\sum_{a' \in \mathcal{A}} \pi(a'|s') Q(s', a')\right)$$

Conclude that the variance is zero for Expected Sarsa, but likely non-zero for Sarsa. Notice that the only random variable is A_{t+1}, which is the action selected according to the target policy π with distribution $\pi(\cdot|S_{t+1})$.

(b) **Challenge Question:** Show that the variance of the Sarsa target is always greater than or equal to the variance of the Expected Sarsa target, given $S_t = s$ and $A_t = a$. Hint: use the Law of Total Variance, which states that for any two random variables X and Y, $\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X])$. This law also applies to conditional distributions: for any random variables X, Y and Z, $\mathrm{Var}(Y|Z) = \mathbb{E}[\mathrm{Var}(Y|X, Z)|Z] + \mathrm{Var}(\mathbb{E}[Y|X, Z]|Z)$.