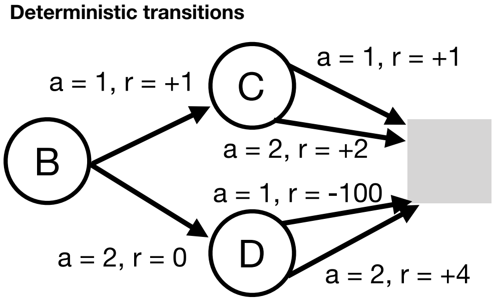
CMPUT 365: Introduction to Reinforcement Learning, Winter 2023
Worksheet #7: Temporal Difference Methods for Control

Manuscript version: *#61d026* - 2023-05-02 16:22:57-06:00

Question 1. Consider an episodic MDP with the states B, C, D, and the terminal state T (i.e. $\mathcal{S} = \{B, C, D, T\}$), and 2 actions ($\mathcal{A} = \{1, 2\}$) with transitions and rewards as shown on the figure below.



Assume that the action values are initialized $Q(s, a) = 0$, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. The agent takes actions according to an ϵ-greedy policy with $\epsilon = 0.1$. Set the discount factor $\gamma = 1.0$.

1. Determine an optimal policy and the optimal action-value function.
2. In the remainder of the problem, we will consider the Sarsa and the Q-learning algorithms where the initial action-values are set to zero. Further, the stepsize is set to $\alpha = 0.1$. Consider the episodic data $S_0 = B, A_0 = 2, R_1 = 0, S_1 = D, A_1 = 2, R_2 = 4$. What are the action-values obtained by running Sarsa on this data for the various states?
3. What are the action-values obtained by running Q-learning on the above data for the various states?
4. 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 are the action-values obtained by running Sarsa on this data for the various states starting from the previous estimates obtained with Sarsa on the first episode's data? What are the action-values obtained by running Q-learning on this data for the various states starting from the previous estimates obtained with Q-learning on the first episode's data?
5. Assume the next episode's data is the same as in Part 4. Again, write down the action-value estimates after running Sarsa on this data, continuing from the previously obtained estimates. Do the same for Q-learning.
6. Do you notice any relationship between the values you obtained by emulating Sarsa and those that you got by emulating Q-learning? What is the relationship and why does it hold?

Question 2. Answer the following:

1. Give at least two conditions that are “necessary” for Sarsa to produce a sequence $(Q_t)_{t\geq 0}$ of value function estimates that converges to q^* with probability one. For each condition, justify why is it “necessary”, that is even if all the other conditions hold and the condition in consideration is violated, then convergence fails to hold.
2. Can any of these conditions be satisfied when Sarsa is used an in episodic MDP with exploring starts (and which one)? Why or why not?

Question 3. (*Exercise 6.11 S&B*) Why is Q-learning considered an *off-policy control* method?

Question 4. (*Exercise 6.12 S&B*, slightly modified) Suppose the action selection is greedy (as opposed to, say, ϵ-greedy).

1. Is Q-learning then exactly the same algorithm as Sarsa? That is, will they make exactly the same action selections and weight updates?
2. Are there any downsides to this choice?

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

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

and the update for expected-Sarsa is

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

where π is a fixed policy that is used to generate the data $S_0, A_0, R_1, S_1, A_1, R_2, \ldots$.

1. Let $H_t = (S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_t)$ and $H'_t = (S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_t, A_t)$. Show that

$$\mathbb{V}[Q(S_{t+1}, A_{t+1}) \,\big|\, H_{t+1}] \geq \mathbb{V}\left[\sum_{a' \in \mathcal{A}} \pi(a'|S_{t+1}) Q(S_{t+1}, a') \,\middle|\, H_{t+1}\right].$$

2. **Challenge Question:** Show that

$$\mathbb{V}[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) \,\big|\, H'_t] \geq \mathbb{V}\left[R_{t+1} + \gamma \sum_{a' \in \mathcal{A}} \pi(a'|S_{t+1}) Q(S_{t+1}, a') \,\middle|\, H'_t\right],$$

   that is, the appropriate conditional variance of the Sarsa target is always at least as large as that of for the Expected Sarsa target. (Hint: Use the law of total variance.)

Question 6. (Challenge Question) (*Exercise 6.13 S&B*) What are the update equations for Double Expected Sarsa with an ϵ-greedy target policy?

Question 7. (*Exercise 6.8 S&B*) Show that an action-value version of the expression

$$G_t - V(S_t) = \sum_{k=t}^{T-1} \gamma^{k-t} \delta_k^{(\text{value})},$$

where $\delta^{(\text{value})} = R_{t+1} + \gamma V(S_{t+1}, A_{t+1}) - V(S_t, A_t)$, holds for the action-value form of the TD error

$$\delta_t = R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)$$

as well. (Assume that the action-values don't change from step to step, that is $Q_t = Q$ for all $t \geq 0$ and some function Q.)

Question 8. Solve the following questions.

1. Let $X_1, X_2, \ldots$ be independent random variables, that take on the values $\{0, 1\}$. Assume that for all $t \geq 1$, $\mathbb{P}(X_t = 1) = p$, where $p \in (0, 1)$. Define

   $$T := \min\{t \geq 1 \,:\, X_t = 1\}\,.$$

   Think of $X_1, X_2, \ldots$ as outcomes of a coin-flip, where 0 corresponds to tails and 1 corresponds to heads. Then T is the number of flips until seeing the first head (including the first timestep). Show that

   $$\mathbb{E}[T] = 1/p.$$

2. Using the answer in the first part, show that there exists a finite, episodic MDP with n states and two actions, such that all of the following hold:

   - The episode lengths are *at least* n.
   - New episodes start from a fixed state s_0 of the MDP.
   - ϵ-greedy with Q-learning (initialized at zero) needs at least $\Omega(2^n)$ episodes before the action-value of the optimal action at s_0 gets higher than the action value of the sub-optimal action at s_0, regardless of the choice of the stepsizes in Q-learning.
   - The value of s_0 under the optimal policy is one.
