CMPUT 365: Introduction to Reinforcement Learning, Winter 2023
Worksheet #6: Temporal Difference Methods for Prediction

Manuscript version: *#f67bcf* - 2023-04-08 23:36:34-06:00

Question 1. (*Exercise 6.1 S&B* with additional details) Fix an episodic MDP with discount factor γ. Consider an episode, of length T timesteps, generated by the agent following policy π in this MDP:

$$(S_0, A_0, R_1, S_1, \dots, S_{T-1}, A_{T-1}, R_T, S_T.$$

Let V_t represent the agent's estimate of the value function v_π and $\delta_t = R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)$ be the corresponding TD error, at timestep t. Recall that $G_t - V_t(S_t)$ denotes the Monte-Carlo error incurred by the agent at timestep t and in state S_t.

(a) Show that

$$G_t - V_t(S_t) = \sum_{k=t}^{T-1} \gamma^{k-t} \delta_k + \sum_{k=t}^{T-2} \gamma^{k-t+1} (V_{k+1}(S_{k+1}) - V_k(S_{k+1}))\,.$$

(b) Note that your answer is different from Eq. 6.6 from the RL book:

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

This happens because when deriving this relation the book assumed that V was fixed. Now, assume that V_t is updated with the TD method:

$$V_{t+1}(S_t) = V_t(S_t) + \alpha_t \delta_t\,,$$

while $V_{t+1}(s) = V_t(s)$ for $s \neq S_t$. Assume that $V_0(s), R_{t+1} \in [0, r_{\max}]$ and $0 \leq \alpha_t \leq \alpha$, $s \in \mathcal{S}$. Show that it then follows that

$$|G_t - V_t(S_t) - \sum_{k=t}^{T-1} \gamma^{k-t} \delta_k| \leq \alpha \frac{\gamma\, r_{\max}}{1-\gamma}$$

hence, $G_t - V_t(S_t) \approx \sum_{k=t}^{T-1} \gamma^{k-t} \delta_k$ indeed holds when the stepsizes α_t are small, in line with what the book suggests after Eq. 6.6.

Question 2. Let $\mathbb{E}_\pi$ be the expectation operator corresponding to the probability distribution $\mathbb{P}_\pi$ induced by following some memoryless policy π from an arbitrary initial distribution. Let

$$H_t = (S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_t)\,.$$

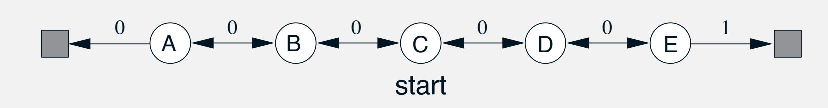
Show that for any $t \geq 1$,

$$\mathbb{E}_\pi[R_{t+1} + \gamma V_t(S_{t+1})|H_t] = (T_\pi V_t)(S_t)\,,$$

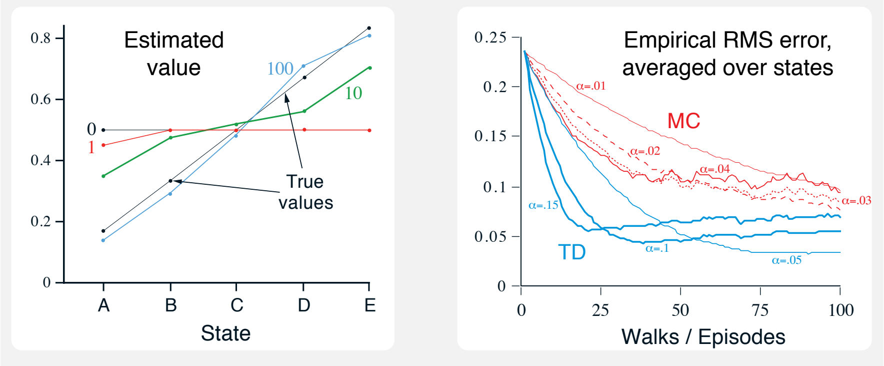
where, as introduced earlier,

$$(T_\pi v)(s) = \sum_{a \in \mathcal{A}} \pi(a|s)\{r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v(s')\}\,.$$

Question 3. (*Exercise 6.3 S&B* with additional details; also see Example 6.2, page 125 of the RL book) Consider the episodic MDP shown in the figure below. The agent starts in state C, and takes the actions `left` or `right` with 50% probability in each of the states. The extreme states are terminal states. The agent receives a reward of $+1$ on the right-most transition and a reward of zero everywhere else.



The left sub-graph shows the value estimates, computed using the TD(0) algorithm with a constant stepsize of $\alpha = 0.1$, for the five states at the beginning of learning (marked by 0), and at the end of 1, 10, and 100 episodes. From the results shown in the left graph, it appears that the first episode results in a change in only $V(A)$. What does this tell you about what happened on the first episode? Why was only the estimate for this one state changed? By exactly how much was it changed?



Question 4. (*Exercise 6.4 S&B*) The specific results shown in the right graph of the random walk example are dependent on the value of the step-size parameter, α. *(a)* Do you think the conclusions about which algorithm is better would be affected if a wider range of α values were used? *(b)* Is there a different, fixed value of α at which either algorithm would have performed significantly better than shown? Why or why not?

Question 5. (Challenge Question) (*Exercise 6.5 S&B*) In the right graph of the above figure, the root-mean-squared value error of the TD method seems to go down and then up again, particularly at high values of α. What could have caused this? Do you think this always occurs, or might it be a function of how the approximate value function was initialized?
Hint: Letting $x = x(t)$ denote the angle at which a pendulum deviates from the vertical line, for x small, we have that the acceleration of the pendulum, $\ddot{x}$, satisfies $\ddot{x} = -x$. This creates a periodic motion that never stops. More realistic (physical) pendulums follow an equation such as $\ddot{x} = -x - c\dot{x}$. Here, $\dot{x}$ is the velocity of the pendulum and $-c\dot{x}$ is a force that arises due to air resistance and eventually makes the pendulum to stop with $c > 0$. The discrete time version of this equation can be obtained using Euler discretization. First introduce $v = \dot{x}$. Then, we have $\dot{v} = \ddot{x} = -x - cv$. Discretizing this with timestep $\Delta > 0$, using the notation $v_k = v(\Delta k)$ and approximating $v_{k+1} = v(\Delta(k+1)) = v(\Delta k) + \Delta v'(\Delta k) = v_k + \Delta v'(\Delta k)$ and using $v'(\Delta k) = -x_k - cv_k$, we get

$$v_{k+1} - v_k = -\Delta x_k - c\Delta v_k \,,$$

where we also use $x_k = x(\Delta k)$. We also have $\dot{x} = v$, the discrete time version of which is

$$x_{k+1} - x_k = \Delta v_k \,.$$

Putting together things we have

$$\begin{pmatrix} v_{k+1} \\ x_{k+1} \end{pmatrix} = \begin{pmatrix} 1 - c\Delta & -\Delta \\ \Delta & 1 \end{pmatrix} \begin{pmatrix} v_k \\ x_k \end{pmatrix} \,.$$

For small values of Δ this quite closely approximates the behavior of a pendulum which will swing with decreasing amplitudes. More generally, if we have an equation of the form $e_{k+1} = Ae_k$ with $e_k \in \mathbb{R}^d$ for some matrix A, the values of $e_{k,i}$ can be seen to be oscillating (potentially with decreasing amplitudes) when some of the eigenvalues of the matrix are complex valued. Can you find a connection between this behavior and the TD update equations? For the purpose of this problem consider the simplified update equation $V_{t+1} = (1-\alpha)V_t + \alpha T_\pi V_t$ that we considered in class.

Question 6. (*Exercise 6.7 S&B* with additional details) The TD update can be written as

$$V_{t+1}(S_t) = V_t(S_t) + \alpha_t \Big[R_{t+1} + \gamma V_t(S_t) - V_t(S_{t+1}) \Big],$$

and $V_{t+1}(s) = V_t(s)$ for $s \neq S_t$, where $\alpha_t \in (0, 1]$ is a stepsize.

Design an off-policy version of the TD update that can be used with arbitrary target policy π and a behavior policy b that satisfies that $b(a|s) > 0$ whenever $\pi(a|s) > 0$ for some action $a \in \mathcal{A}$ in some state $s \in \mathcal{S}$. The "goal" of the update is still to estimate v_π, but now the data is generated from the MDP while following policy b. (Hint: use the importance sampling ratio $\rho := \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$ to design this update.) Argue that your update is "correct". Denoting by $\mathbb{E}_b$ the expectation operator underlying the probability measure $\mathbb{P}_b$ induced by interconnecting the behavior policy b with the MDP, show that

$$\mathbb{E}_b[V_{t+1}(S_t) - V_t(S_t) | H_t] = \alpha_t ((T_\pi V_t)(S_t) - V_t(S_t)) \, ,$$

where as before $H_t = (S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_t)$.

Question 7. Modify the Tabular TD(0) algorithm for estimating v_π, to estimate q_π instead.

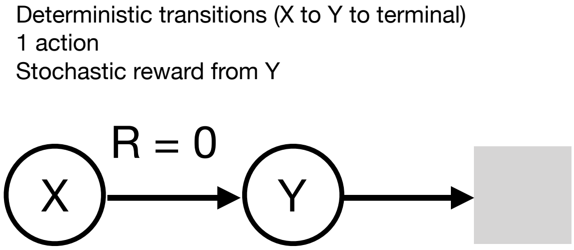
Tabular TD(0) for estimating v_π

```
Input: the policy $\pi$ to be evaluated
Algorithm parameter: step size $\alpha \in (0, 1]$
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(\textit{terminal}) = 0$

Loop for each episode:
    Initialize $S$
    Loop for each step of episode:
        $A \leftarrow$ action given by $\pi$ for $S$
        Take action $A$, observe $R$, $S'$
        $V(S) \leftarrow V(S) + \alpha\big[R + \gamma V(S') - V(S)\big]$
        $S \leftarrow S'$
    until $S$ is terminal
```

Question 8. Suppose that in an environment, state transitions are deterministic and that the reward is bounded, so that $r_{\text{MIN}} = 0$ and $r_{\text{MAX}} = 1$, with the expected reward being $\mathbb{E}[R_{t+1}|S_t = s] = 0.5$ for all timesteps t and states s. Find the maximum and minimum possible TD error $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$, where $\gamma = 0.9$ and $V = v_\pi$ for a deterministic policy π.

Question 9. Assume the agent interacts with a simple two-state MDP shown below. Every episode begins in state X, and ends when the agent transitions from state Y to the terminal state T (denoted by gray box). Therefore, the set of states is $\mathcal{S} = \{X, Y, T\}$. There is only one possible action in each state, which means that there is only one possible policy in this MDP. Let us denote the set of actions $\mathcal{A} = \{A\}$. In state Y, the agent terminates when it takes action A; upon doing so, it sometimes gets a reward of +1000 and sometimes gets a reward of -1000, that is the reward on this last transition is stochastic. Let $\gamma = 1.0$.


Deterministic transitions (X to Y to terminal)
1 action
Stochastic reward from Y

$$P(R = r|Y) = \begin{cases} 0.5 & \text{if } r = -1000 \\ 0.5 & \text{if } r = +1000 \end{cases}$$

(a) Write down $\pi(a|s)$, for all $s \in \mathcal{S}, a \in \mathcal{A}$.

(b) Write down all the possible trajectories (sequence of states, actions, and rewards) in this MDP that start from state X?

(c) What is $v_\pi(X)$ and $v_\pi(Y)$)?

(d) Assume that our estimate is equal to the value of π. That is $V(s) = v_\pi(s)$ for all $s \in \mathcal{S}$. Then compute the TD-error $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ for the transition from state Y to the terminal state, assuming $R_{t+1} = +1000$. Why is the TD-error not zero if we initialize the value estimate with the true value function, that is, $V^{\text{INIT}}(Y) = v_\pi(Y)$?

(e) Based on the above answer, what does this mean for the TD-update, for constant $\alpha = 0.1$? Will the value estimate of state Y remain zero, i.e. $V(Y) = v_\pi(Y) = 0$, after we update the value as well? Recall that the TD-update is $V^{\text{NEW}}(S_t) = V^{\text{OLD}}(S_t) + \alpha\delta_t$. What does this tell us about the updates that TD(0) would make on this MDP?

(f) What is the expected TD-update, from state Y for the given V?

(g) Assume still that $V = v_\pi = \mathbf{0}$. What is the expectation and the variance of the TD update from state X? What is the expectation and the variance of the Monte-carlo update from state X?

Question 10. (**Challenge Question**) In this question we consider the variance of the TD target, $R_{t+1} + \gamma V(S_{t+1})$, compared to the variance of the Monte-Carlo target, G_t. To make the analysis simpler, assume that $V = v_\pi$. Show that, under this simplification, the variance of the Monte-Carlo target is higher than (or equal to) the variance of the TD target, that is,

$$\mathbb{V}[G_t|S_t] \geq \mathbb{V}[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t].$$

(Note that variance of the targets is a factor in learning speed—targets with lower variance typically allow for faster learning. Also note that both these targets are equal in expectation (why?), that is $\mathbb{E}[G_t|S_t] = \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t]$.)
Hint: Recall the tower property of expectation, also known as the law of total expectation:

$$\mathbb{E}[G_t|S_t] = \mathbb{E}\big[\mathbb{E}[G_t|S_{t+1}]\big|S_t\big],$$

where the outer expectation is over S_{t+1} and the inner expectation is over G_t. You might have to use the following decomposition as well, which can be derived using the law of total variance along with the Markov property (how?):

$$\mathbb{V}[G_t|S_t] = \mathbb{E}\big[\mathbb{V}[G_t|S_{t+1}]\big|S_t\big] + \mathbb{V}\big[\mathbb{E}[G_t|S_{t+1}]\big|S_t\big]. \tag{1}$$

One way to use this decomposition could be to simplify $\mathbb{V}\big[\mathbb{E}[G_t|S_{t+1}]\big]$ by showing that $\mathbb{E}[G_t|S_{t+1}] = \mathbb{E}[R_{t+1}|S_{t+1}] + \gamma v_\pi(S_{t+1})$.