Name: Alfred
Student No.: 2020-12788

Exercise 4

Data Science & Reinforcement Learning

Spring 2023

$E\left[\left(\gamma \max_{a'} Q_t(s', a') - (TQ_t)(s,a)\right)^2\right]$

1. Modify the Tabular TD(0) algorithm for estimating V_π below to estimate Q_π.

Tabular TD(0) for estimating v_π

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

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

Algorithm Parameters: step size $\alpha \in (0,1]$, small $\varepsilon > 0$
Initialize $Q(s,a)$, for all $s \in S^+$, $a \in A(s)$ arbitrarily except $Q(terminal, \cdot) = 0$.

Loop for each episode:
  Initialize S.
  Loop for each episode:
    Take action A, observe R, S'
    Choose A' from S' using policy from Q
    $Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma Q(S', A') - Q(S,A)]$
    $S \leftarrow S'$; $A \leftarrow A'$
  until S is terminal

5) $\hookrightarrow \mathrm{Var}(F_t(s,a) \mid H_t) = \mathbb{E}\left[\left(r_t(s,a) + \gamma \max_{a'} Q_t(s', a') - \cancel{Q^*}(s,a) - (TQ_t)(s,a) + \cancel{Q^*}(s,a)\right)^2\right]$

$= \mathrm{Var}\left[\underbrace{r_t(s,a) + \gamma \max_{a'} Q_t(s',a')}_{\text{bonded}} \mid H_t\right) = \mathrm{Var}[r_t(s') + \gamma \max_a Q_t(s',a)]$.

reward $\wedge$ $\max_a Q_t(s', a)$ bonded

then $\mathrm{Var}[F_t(s,a) \mid H_t] \le C(1 + \|\Delta_t\|_\infty^2)$

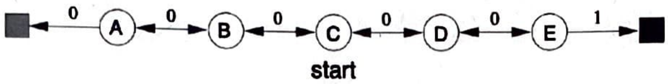
$\max_a Q_t(s', a) - \max_{s,a} Q^*(s,a)$
$\le \max_{s,a} \|Q_t - Q^*\| \le \|\Delta_t\|_\infty$.

then by lemma 2
$\Delta_t(s) = Q_t(s,a) - Q^*(s,a) \to 0$.
Convergence: $Q_t(s,a) = Q^*(s,a)$ w.p. 1

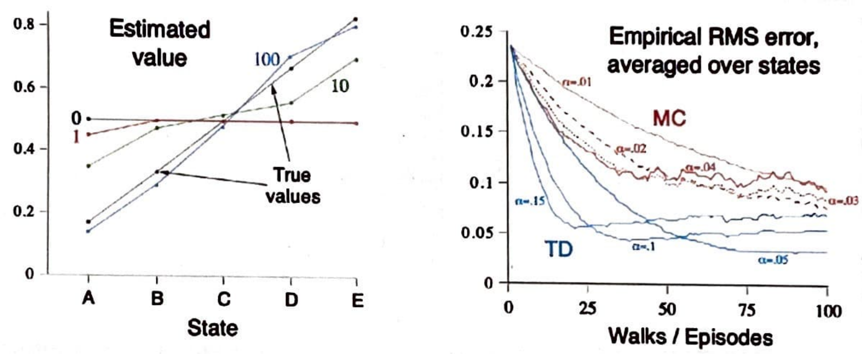
2. Consider the following *Random Walk* example shown in the textbook.

Example 6.2 Random Walk

In this example we empirically compare the prediction abilities of TD(0) and constant-α MC when applied to the following Markov reward process:



A *Markov reward process*, or MRP, is a Markov decision process without actions. We will often use MRPs when focusing on the prediction problem, in which there is no need to distinguish the dynamics due to the environment from those due to the agent. In this MRP, all episodes start in the center state, C, then proceed either left or right by one state on each step, with equal probability. Episodes terminate either on the extreme left or the extreme right. When an episode terminates on the right, a reward of +1 occurs; all other rewards are zero. For example, a typical episode might consist of the following state-and-reward sequence: C, 0, B, 0, C, 0, D, 0, E, 1. Because this task is undiscounted, the true value of each state is the probability of terminating on the right if starting from that state. Thus, the true value of the center state is $v_\pi(\mathsf{C}) = 0.5$. The true values of all the states, A through E, are $\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}$, and $\frac{5}{6}$.



The left graph above shows the values learned after various numbers of episodes on a single run of TD(0). The estimates after 100 episodes are about as close as they ever come to the true values—with a constant step-size parameter ($\alpha = 0.1$ in this example), the values fluctuate indefinitely in response to the outcomes of the most recent episodes. The right graph shows learning curves for the two methods for various values of α. The performance measure shown is the root mean square (RMS) error between the value function learned and the true value function, averaged over the five states, then averaged over 100 runs. In all cases the approximate value function was initialized to the intermediate value $V(s) = 0.5$, for all s. The TD method was consistently better than the MC method on this task.

On the left plot for the estimated values, it appears that the first episode results in a change in only $V(A)$. What does this tell you about what happened in the first episode? Why was only the estimate for $V(A)$ changed? By exactly how much was it changed?

- first episode is until state A without knowing the transition (must have termitated at State A)

$V_A \leftarrow V_A + \alpha(R + \gamma V_T - V_A) = 0{,}5 - 0{,}1(0+1-0{,}5) = 0{,}45$

- only V(A) changed because rewards are 0 ∧ $\gamma = 1$ since initialized at $V(s) = 0{,}5$. TD update doesn't update some V.

For A, E states only non-zero changes are possible.

$\Delta V_A = -0{,}05$

3. In the *Random Walk* example of Question 2, the curves on the right plot are dependent on the value of the step-size parameter α. Based on these results, can you come up with a strategy for the choice of the step-size parameter α? Justify your answer.

Small parameter steps can definitely reduce error for both MC & TD.
But when at limit it will reach the minimum lowest error,
then arbitrary/strategical choice won't improve the plot.

4. Q-learning is an off-policy algorithm, but there does not seem to be any importance sampling ratio used in the update rule. Why is it the case?

target policy is greedy w.r.t $Q(S,A)$ while b is another different policy.

(Importance Sampling) It's not needed because. $Q(S,A) \leftarrow Q(S,A) + \alpha(r + \gamma \max_a Q(S',a) - Q(S,A))$

learns from taking various actions (even random) which don't need

policy. We use the max Q instead.

a are all actions probed in state s'

from behavior policy, A is sampled the first time only. later weren't so no need of importance sampling.

w.p. with probability

5. **Prove the convergence of Q-learning.**

Lemma 1 (Contraction)

the operator $\mathcal{T}$ defined for generic function $q: S \times A \to \mathbb{R}$

as $(\mathcal{T}q)(s,a) := \sum_{s' \in S} P(s'|s,a)\left[r(s,a) + \gamma \max_{a'} q(s',a')\right]$

this operator is a contraction in sup-norm, i.e. ∞.

$$\|\mathcal{T}q_1 - \mathcal{T}q_2\|_\infty \le \gamma \|q_1 - q_2\|_\infty$$

Proof

$$\|\mathcal{T}q_1 - \mathcal{T}q_2\|_\infty = \max_{s,a} |(\mathcal{T}q_1)(s,a) - (\mathcal{T}q_2)(s,a)|$$

$$= \max_{s,a} \left| \sum_{s' \in S} P(s'|s,a)\left[\cancel{r(s,a)} + \gamma \max_{a'} q_1(s',a') - \cancel{r(s,a)} - \gamma \max_{a'} q_2(s',a')\right]\right|$$

$$= \max_{s,a} \gamma \left| \sum_{s' \in S} P(s'|s,a)\left[\max_{a'} q_1(s',a') - \max_{a'} q_2(s',a')\right]\right|$$

$$\le \max_{s,a} \gamma \sum_{s'} P(s'|s,a) \underbrace{\max_{s',a'} |q_1(s',a') - q_2(s',a')|}_{\|q_1 - q_2\|_\infty}$$

$$= \gamma \|q_1 - q_2\|_\infty \sum_{s'} P(s'|s,a)$$

$\mathbb{E}_x[f(x)] \le \max_x f(x)$

$\mathcal{T}Q^* = Q^*$

Lemma 2 (Stochastic approx.): A random process $\{\Delta_t\}_t$ taking values in $\mathbb{R}^n$ defined as

Jaakkola, Jordan, Singh, 1993

$$\Delta_{t+1}(x) := (1 - \alpha_t(x))\Delta_t(x) + \alpha_t(x) F_t(x) \;;\; \Delta_t(x) \to 0 \text{ w.p. 1 under:}$$

i) $0 \le \alpha_t \le 1$, $\sum_t \alpha_t(x) = \infty$, $\sum_t \alpha_t^2(x) < \infty$

ii) $\max_x |\mathbb{E}[F_t(x) | H_t]| \le \gamma \max_x |\Delta_t(x)| = \gamma \|\Delta_t\|_\infty$

iii) $\mathrm{Var}(F_t(x) | H_t) \le C(1 + \|\Delta_t\|_\infty^2)$ for x, $C > 0$

Q-learning

$$Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t) + \alpha_t(s_t,a_t)\left[r_t + \gamma \max_{a'} Q_t(s_{t+1},a') - Q_t(s_t,a_t)\right]$$

WTS $Q_t \to Q^*$ w.p. 1

under α ($\sum \alpha = \infty \wedge \sum \alpha^2 < \infty$)

[Proof] Define Δ_t, $\Delta_t(s,a) := Q_t(s,a) - Q^*(s,a)$

$$\to \Delta_{t+1}(s,a) = (1 - \alpha_t(s,a))\Delta_t(s,a) + \alpha_t(s,a)\left[r_t + \gamma \max_{a'} Q_t(s',a') - Q^*(s,a)\right]$$

Define $F_t(s,a) = r_t(s,a) + \gamma \max_{a'} Q(x,a) - Q^*(s,a)$, x: random sample from MDP $\sim P(\cdot|s,a)$

$$\mathbb{E}[F_t(s,a) | H_t] = \sum_{s'} P(s'|s,a)\left[r(s,a) + \gamma \max_{a'} Q_t(s',a') - Q^*(s,a)\right]$$

$H_t = \{\Delta_t, \Delta_{t-1}, \dots, F_{t-1}, \alpha_{t-1}\}$

$$= (\mathcal{T}Q_t)(s,a) - Q^*(s,a) = (\mathcal{T}Q_t)(s,a) - (\mathcal{T}Q^*)(s,a)$$

$$\|\mathbb{E}[F_t | H_t]\|_\infty = \|\mathcal{T}Q_t - \mathcal{T}Q^*\|_\infty \le \gamma \|Q_t - Q^*\|_\infty \text{ (by Lemma 1)} = \gamma \|\Delta_t\|_\infty.$$

$\to$ PAGE 1