ECE - 5984 : Homework 3
Instructor: Thinh T. Doan, TA: Amit Dutta
Due Date: 10/14/2021

Problem 1

[30 points] Consider the following stochastic approximation with a fixed step size $\epsilon \in (0, 1)$

$$\theta_{k+1} = (1-\epsilon)\theta_k + \epsilon X_k,$$

where X_k are i.i.d random variables with mean μ and variance σ^2. In addition, given any constant $a > 0$, the Chebyshev's inequality implies

$$\mathbb{P}(|X - E(X)| \geq a) \leq \frac{Var(X)}{a^2},$$

where X is a random variable with $E(X)$ and $Var(X)$ are its expected value and variance, respectively. Show that for any $c > 0$

$$\limsup_{k\to\infty} \mathbb{P}(|\theta_k - \mu| \geq c\sqrt{\epsilon}) \leq \frac{\sigma^2}{c^2}.$$

Step 1: Showing that $\lim_{k\to\infty} E(\theta_k) = \mu$

$$\theta_{k+1} = (1-\varepsilon)\theta_k + \varepsilon X_k$$

$$\therefore \quad \theta_1 = (1-\varepsilon)\theta_0 + \varepsilon X_0$$

$$\theta_2 = (1-\varepsilon)((1-\varepsilon)\theta_0 + \varepsilon X_0) + \varepsilon X_1$$

$$= (1-\varepsilon)^2\theta_0 + (1-\varepsilon)\varepsilon X_0 + \varepsilon X_1$$

$$\theta_3 = (1-\varepsilon)^3\theta_0 + (1-\varepsilon)^2\varepsilon X_0 + (1-\varepsilon)\varepsilon X_1 + \varepsilon X_2$$

$$\vdots$$

$$\theta_k = (1-\varepsilon)^k\theta_0 + \sum_{j=0}^{k}(1-\varepsilon)^j \varepsilon X_{(k-j)}$$

$$E(\theta_k) = E\left((1-\varepsilon)^k\theta_0 + \sum_{j=0}^{k}(1-\varepsilon)^j\varepsilon X_{(k-j)}\right)$$

$$= E\left((1-\varepsilon)^k\theta_0\right) + E\left(\sum_{j=0}^{k}(1-\varepsilon)^j\varepsilon X_{k-j}\right)$$

$$\lim_{k\to\infty} E(\theta_k) = \lim_{k\to\infty} E\left((1-\varepsilon)^k\theta_0\right) + \lim_{k\to\infty} E\left(\sum_{j=0}^{k}(1-\varepsilon)^j\varepsilon X_{k-j}\right)$$

$$= 0 + \lim_{k\to\infty}\left[\sum_{j=0}^{k} E\left((1-\varepsilon)^j\varepsilon X_{k-j}\right)\right]$$

$$= \lim_{k\to\infty}\left[\sum_{j=0}^{k} E\left(\varepsilon(1-\varepsilon)^j\right) E\left(X_{k-j}\right)\right]$$

$$= \lim_{k \to \infty} \left[\sum_{j=0}^{k} \varepsilon (1-\varepsilon)^j \mu \right]$$ as $E(X_k) = \mu$ given in question

$$= \varepsilon \mu \sum_{j=0}^{\infty} (1-\varepsilon)^j$$

$$= \frac{\varepsilon \mu}{1-(1-\varepsilon)} = \frac{\varepsilon \mu}{\varepsilon} = \mu$$

$$\therefore \lim_{k \to \infty} E(\theta_k) = \mu$$

Step 2: Solving $\lim_{k \to \infty} Var(\theta_k)$

$$\theta_k = (1-\varepsilon)^k \theta_0 + \sum_{j=0}^{k} (1-\varepsilon)^j \varepsilon X_{(k-j)}$$ (from step 1 above)

$$= (1-\varepsilon)^k \varepsilon X_0 + (1-\varepsilon)^{k-1} \varepsilon X_1 + \ldots\ldots (1-\varepsilon) \varepsilon X_{k-1} + \varepsilon X_k$$

as X is iid : $Var(aX_1 + bX_2) = a^2 Var(X_1) + b^2 Var(X_2)$

$$\therefore Var(\theta_k) = Var\left[(1-\varepsilon)^k \varepsilon X_0\right] + Var\left[(1-\varepsilon)^{k-1} \varepsilon X_1\right]$$
$$\ldots\ldots + Var\left[(1-\varepsilon) \varepsilon X_{k-1}\right] + Var\left[\varepsilon X_k\right]$$

$$= \left(\varepsilon (1-\varepsilon)^k\right)^2 Var(X_0)$$
$$+ \left(\varepsilon (1-\varepsilon)^{k-1}\right)^2 Var(X_1)$$
$$+ \ldots\ldots$$
$$+ \left(\varepsilon (1-\varepsilon)^1\right)^2 Var(X_{k-1})$$
$$+ \left(\varepsilon (1-\varepsilon)^0\right)^2 Var(X_k)$$

$$= \sigma^2 \varepsilon^2 \left[(1-\varepsilon)^0 + (1-\varepsilon)^2 \ldots\ldots + (1-\varepsilon)^{2k-2} + (1-\varepsilon)^{2k}\right]$$ as $Var(X_k) = \sigma^2$ given in question

let $S = \lim_{k\to\infty}\left[(1-\varepsilon)^0 + (1-\varepsilon)^2 \dots (1-\varepsilon)^{2k-2} + (1-\varepsilon)^{2k}\right]$

$$S = 1 + (1-\varepsilon)^2 S$$

$$\Rightarrow S - S(1-\varepsilon)^2 = 1$$

$$S(1-(1-\varepsilon)^2) = 1$$

$$\Rightarrow S = \frac{1}{1-(1-\varepsilon)^2} = \frac{1}{2\varepsilon - \varepsilon^2}$$

$$\therefore \lim_{k\to\infty} Var(\theta_k) = S \lim_{k\to\infty} \sigma^2\varepsilon^2 = \frac{\sigma^2\varepsilon^2}{2\varepsilon - \varepsilon^2}$$

Step 3: Chebyshevs equation

$$P(|x - E(x)| \geq a) \leq \frac{Var(x)}{a^2}$$

$$\therefore \lim_{k\to\infty} P(|\theta_k - E(\theta_k) \geq c\sqrt{\varepsilon})$$

$$= \lim_{k\to\infty} P(|\theta_k - \mu| \geq c\sqrt{\varepsilon})$$

as $\lim_{k\to\infty} E(\theta_k) = \mu$ as shown in step 1

$$\leq \lim_{k\to\infty} \frac{Var(\theta_k)}{c^2\varepsilon}$$

$$= \frac{\sigma^2\varepsilon^2}{c^2\varepsilon(2\varepsilon - \varepsilon^2)}$$

as $\lim_{k\to\infty} Var(\theta_k) = \frac{\sigma^2\varepsilon^2}{2\varepsilon - \varepsilon^2}$ as shown in step 2

$$= \frac{\sigma^2}{c^2}\left(\frac{\varepsilon}{2\varepsilon - \varepsilon^2}\right) = \frac{\sigma^2}{c^2(2-\varepsilon)}$$

$$\leq \frac{\sigma^2}{c^2}$$

as $\frac{1}{2-\varepsilon} \in (\frac{1}{2}, 1)$ since $\varepsilon \in (0,1)$

$$\therefore \limsup_{k\to\infty} P(|\theta_k - \mu| \geq c\sqrt{\varepsilon}) \leq \frac{\sigma^2}{c^2}$$

//

Problem 2

We consider a discounted MDP problem with finite state space $\mathcal{S}$ and finite action space $\mathcal{A}$. For any stationary policy μ define the value function V_μ

$$V_\mu(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) \mid s_0 = s\right], \qquad a_k = \mu(s_k),$$

and let $V_\mu(s) \le V^*(s) = V_{\mu^*}(s)$ for the optimal policy μ^* and for all $s \in \mathcal{S}$. Given any value function V_μ we denote by Q_μ the state-action value function

$$Q_\mu(s,a) = r(s,a) + \gamma \sum_{s \in \mathcal{S}} p_{ss'}(a) V_\mu(s').$$

Similarly, the optimal state-action value function Q^* is defined as

$$Q^*(s,a) \triangleq Q_{\mu^*}(s,a) = r(s,a) + \gamma \sum_{s \in \mathcal{S}} p_{ss'}(a) V^*(s').$$

Given a stationary policy μ, defined the Bellman operator T_μ as

$$(T_\mu Q)(s,a) = r(s,a) + \gamma \sum_{s \in \mathcal{S}} p_{ss'}(a) Q(s', \mu(s'))$$
$$(TQ)(s,a) = r(s,a) + \gamma \sum_{s \in \mathcal{S} \in \mathcal{S}} p_{ss'}(a) \max_{a'} Q(s', a').$$

Questions: Let μ be a stationary policy and any state action-value functions Q, Q'.

1. Let $V_{Q,\mu}(s') = Q(s', \mu(s'))$ and $V_Q(s') = \max_a Q(s', a)$. Show that [10 points]

$$1)\quad \|V_{Q,\mu} - V_{Q',\mu}\|_\infty \le \|Q - Q'\|_\infty$$
$$2)\quad \|V_Q - V_{Q'}\|_\infty \le \|Q - Q'\|_\infty$$

Part 1:

• given any $s \in S$:

$$|V_{Q,\mu}(s) - V_{Q',\mu}(s)| = |Q(s,\mu(s)) - Q'(s,\mu(s))|$$

as $\{\mu(s) : s \in S\} \subseteq \{a : a \in A\}$ | $\mu(s)$ subset of all possible actions a

$$\Rightarrow |Q(s,\mu(s)) - Q'(s,\mu(s))| \le \max_a |Q(s,a) - Q'(s,a)| \quad \Big|\, a \in A$$
$$= \|Q(s,a) - Q'(s,a)\|_\infty$$

$$\therefore \quad \|V_{Q,\mu} - V_{Q',\mu}\|_\infty \le \|Q - Q'\|_\infty \quad //$$

Part 2

note: given $x = \{x_1, x_2 \dots x_n\}$, $y = \{y_1, y_2 \dots y_n\}$, w.l.o.g: $\max_i x_i \geq \max_j y_j$

$$\Rightarrow \left| \max_i x_i - \max_j y_j \right| = \max_i x_i - \max_j y_j$$

$$= x_{i_2} - \max_j y_j \qquad \text{letting } i_2 = \arg\max x_i$$

$$\leq x_{i_2} - y_{i_2}$$

$$= \left| x_{i_2} - y_{i_2} \right|$$

$$\leq \max_i \left| x_i - y_i \right| \qquad \text{(eq 1)}$$

$$\therefore \left| \max_i x_i - \max_j y_j \right| \leq \max_i \left| x_i - y_i \right|$$

given any $s \in S$:

$$\left| V_Q(s) - V_{Q'}(s) \right| = \left| \max_{a_1} Q(s,a_1) - \max_{a_2} Q'(s,a_2) \right| \qquad (a_1, a_2 \in A)$$

$$\leq \max_a \left| Q(s,a) - Q'(s,a) \right| \qquad \text{using eq (1) from above.}$$

$$= \| Q(s,a) - Q'(s,a) \|_\infty$$

$$\therefore \quad \underline{\underline{\| V_Q - V_{Q'} \|_\infty \leq \| Q - Q' \|_\infty}}$$

2. Show that [10 points]

1) $\|T_\mu Q - T_\mu Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$
2) $\|TQ - TQ'\|_\infty \leq \gamma \|Q - Q'\|_\infty$

<u>Part 1</u>

given any (s,a):

$$(T_\mu Q - T_\mu Q')(s,a) = \gamma \sum_{s' \in S} P_{ss'}(a) \left[Q(s',v) - Q'(s',u) \right] \Big|_{\substack{u \in A \\ v \in A}}$$

$$\leq \gamma \sum_{s' \in S} P_{ss'}(a) \left| Q(s',v) - Q'(s',u) \right|$$

$$\leq \gamma \sum_{s' \in S} P_{ss'}(a) \max_{v,u} \left| Q(s',v) - Q'(s',u) \right|$$

$$\leq \gamma \sum_{s' \in S} P_{ss'}(a) \max_{v} \left| Q(s',v) - Q'(s',v) \right|$$

$$\leq \gamma \sum_{s' \in S} P_{ss'}(a) \max_{v,y} \left| Q(y,v) - Q'(y,v) \right| \Big|_{y \in S}$$

$$= \gamma \max_{v,y} \left| Q(y,v) - Q'(y,v) \right|$$

$$= \gamma \| Q - Q' \|_\infty$$

$$\therefore \quad \| T_\mu Q - T_\mu Q' \|_\infty \leq \gamma \| Q - Q' \|_\infty$$

Part 2

given any (s,a):

$$TQ(s,a) - TQ'(s,a) = \gamma \sum_{s' \in S} P_{ss'}(a) \left[\max_{v} Q(s',v) - \max_{u} Q'(s',u) \right]$$ $\Big|$ $v \in A$, $u \in A$

$$\Rightarrow \left| TQ(s,a) - TQ'(s,a) \right| = \left| \gamma \sum_{s' \in S} P_{ss'}(a) \left[\max_{v} Q(s',v) - \max_{u} Q'(s',u) \right] \right|$$

$$\leq \gamma \sum_{s' \in S} P_{ss'}(a) \left| \left[\max_{v} Q(s',v) - \max_{u} Q(s',u) \right] \right|$$ $\Big|$ via triangle inequality.

note: given $x = \{x_1, x_2 \ldots\ldots x_n\}$, $y = \{y_1, y_2 \ldots\ldots y_2\}$. , w.l.o.g: $\max_i x_i \geq \max_j y_j$

$$\Rightarrow \left| \max_i x_i - \max_j y_j \right| = \max_i x_i - \max_j y_j$$

$$= x_{i_2} - \max_j y_j$$ $\Big|$ letting $i_2 = \arg\max x_i$

$$\leq x_{i_2} - y_{i_2}$$

$$= \left| x_{i_2} - y_{i_2} \right|$$

$$\leq \max_i \left| x_i - y_i \right| \qquad (\text{eq } 1)$$

• applying eq 1 from above:

$$\gamma \sum_{s' \in S} P_{ss'}(a) \left| \left[\max_{v} Q(s',v) - \max_{u} Q(s',u) \right] \right|$$

$$\leq \gamma \sum_{s' \in S} P_{ss'}(a) \max_{v} \left| Q(s',v) - Q'(s',v) \right|$$ $\Big|$ via (eq 1)

$$\leq \gamma \sum_{s' \in S} P_{ss'}(a) \max_{v,y} \left| Q(y,v) - Q'(y,v) \right|$$ $\Big|$ $y \in S$

$$= \gamma \max_{v,y} \left| Q(y,v) - Q'(y,v) \right|$$

$$= \gamma \left\| Q - Q' \right\|_\infty$$

$$\therefore \quad \left\| TQ - TQ' \right\|_\infty \leq \gamma \left\| Q - Q' \right\|_\infty$$ //

3. Show that [10 points]

$$1)\quad \|Q - Q_\mu\|_\infty \le \frac{\|Q - T_\mu Q\|_\infty}{1-\gamma}$$

$$2)\quad \|Q - Q^*\|_\infty \le \frac{\|Q - TQ\|_\infty}{1-\gamma}$$

Part 1

$$Q(s,a) - T_\mu Q(s,a) \leq \| Q - T_\mu Q \|_\infty$$

let $\| Q - T_\mu Q \|_\infty = r$.

$$\Rightarrow \quad T_\mu Q \leq Q + r\mathbb{1}$$

$$\Rightarrow \quad T_\mu^2 Q = T_\mu(T_\mu Q) \leq T_\mu(Q + r\mathbb{1}) = T_\mu Q + \gamma r \mathbb{1}$$

$$\Rightarrow \quad T_\mu^3 Q \leq T_\mu^2 Q + \gamma^2 r \mathbb{1}$$

$$\vdots$$

$$\Rightarrow \quad T_\mu^L Q \leq T_\mu^{L-1} Q + \gamma^{L-1} r \mathbb{1}$$

$$\Rightarrow \quad T_\mu^L Q - T_\mu^{L-1} Q \leq \gamma^{L-1} r \mathbb{1}$$

note that: $$T_\mu^m Q - Q = \sum_{L=1}^{m} (T_\mu^L Q - T_\mu^{L-1} Q)$$

$$\leq \sum_{L=1}^{m} \gamma^{L-1} r \mathbb{1}$$

<u>letting $m \to \infty$:</u>

- $$\lim_{m\to\infty} (T_\mu^m Q - Q) = Q_\mu - Q$$
- $$\lim_{m\to\infty} \left(\sum_{L=1}^{m} \gamma^{L-1} r \mathbb{1} \right) = \frac{r\mathbb{1}}{1-\gamma}$$

$$\therefore \quad \lim_{m\to\infty} (T_\mu^m Q - Q) \leq \lim_{m\to\infty} \left(\sum_{L=1}^{m} \gamma^{L-1} r \mathbb{1} \right)$$

$$\Rightarrow \quad Q_\mu - Q \leq \frac{r\mathbb{1}}{1-\gamma}$$

$$\Rightarrow \quad Q_\mu - Q \leq \frac{\| Q - T_\mu Q \|_\infty \mathbb{1}}{1-\gamma}$$

$$\Rightarrow \quad \| Q_\mu - Q \|_\infty \leq \frac{\| Q - T_\mu Q \|_\infty}{1-\gamma}$$

//

Part 2

$$Q(s,a) - TQ(s,a) \leqslant \| Q - TQ \|_\infty$$

let $\| Q - TQ \|_\infty = r$.

$$\Rightarrow \quad TQ \leqslant Q + r\mathbb{1}$$

$$\Rightarrow \quad T^2 Q = T(TQ) \leqslant T(Q + r\mathbb{1}) = TQ + \gamma r \mathbb{1}$$

$$\Rightarrow \quad T^3 Q \leqslant T^2 Q + \gamma^2 r \mathbb{1}$$

$$\vdots$$

$$\Rightarrow \quad T^L Q \leqslant T^{L-1} Q + \gamma^{L-1} r \mathbb{1}$$

$$\Rightarrow \quad T^L Q - T^{L-1} Q \leqslant \gamma^{L-1} r \mathbb{1}$$

note that: $T^m Q - Q = \sum_{L=1}^{m} (T^L Q - T^{L-1} Q)$

$$\leqslant \sum_{L=1}^{m} \gamma^{L-1} r \mathbb{1}$$

letting $m \to \infty$:

- $\lim_{m\to\infty} (T^m Q - Q) = Q^* - Q$
- $\lim_{m\to\infty} \left(\sum_{L=1}^{m} \gamma^{L-1} r \mathbb{1} \right) = \frac{r\mathbb{1}}{1-\gamma}$

$$\therefore \quad \lim_{m\to\infty} (T^m Q - Q) \leqslant \lim_{m\to\infty} \left(\sum_{L=1}^{m} \gamma^{L-1} r \mathbb{1} \right)$$

$$\Rightarrow \quad Q^* - Q \leqslant \frac{r\mathbb{1}}{1-\gamma}$$

$$\Rightarrow \quad Q^* - Q \leqslant \frac{\| Q - TQ \|_\infty \mathbb{1}}{1-\gamma}$$

$$\Rightarrow \quad \| Q - Q^* \|_\infty \leqslant \frac{\| Q - TQ \|_\infty}{1-\gamma}$$

//

4. Let μ be the greedy policy for any state-action value function Q, i.e.,

$$\mu(s) = \arg\max_a Q(s, a)$$

Define the Bellman error for Q as $\beta = \|TQ - Q\|_\infty$. Let V_μ be the value function associated with the greedy policy μ. Show that [10 points]

$$V_\mu(s) \geq V^*(s) - \frac{2\beta}{1-\gamma}, \qquad \forall s \in \mathcal{S}.$$

• from Q2.3: $\|Q - Q_\mu\|_\infty \leqslant \dfrac{\|Q - TQ\|_\infty}{1-\gamma}$ (eq 1)

$\|Q - Q^*\|_\infty \leqslant \dfrac{\|Q - TQ\|_\infty}{1-\gamma}$ (eq 2)

$$\|Q^* - Q_\mu\|_\infty = \|(Q^* - Q) + (Q - Q_\mu)\|_\infty$$

$$\leqslant \|Q^* - Q\|_\infty + \|Q - Q_\mu\|_\infty$$ via triangle inequality.

$$= \|Q - Q^*\|_\infty + \|Q - Q_\mu\|_\infty$$

$$\leqslant \frac{\|Q - TQ\|_\infty}{1-\gamma} + \frac{\|Q - TQ\|_\infty}{1-\gamma}$$ using eq (1) and eq (2) above

$$= \frac{2\|Q - TQ\|_\infty}{1-\gamma}$$

$$\therefore \quad \|Q^* - Q_\mu\|_\infty \leqslant \frac{2\|Q - TQ\|_\infty}{1-\gamma}$$

$$\Rightarrow \quad Q^*(s, \pi^*(s)) - Q(s, \mu(s)) \leqslant \frac{2\|Q - TQ\|_\infty}{1-\gamma}$$ for a given $s \in S$

$$\Rightarrow \quad V^*(s) - V_\mu(s) \leqslant \frac{2\|Q - TQ\|_\infty}{1-\gamma}$$

$$\Rightarrow \quad V_\mu(s) \geqslant V^*(s) - \frac{2\|Q - TQ\|_\infty}{1-\gamma}, \quad \forall s \in S$$

//

Problem 3 - Coding question [30 points]

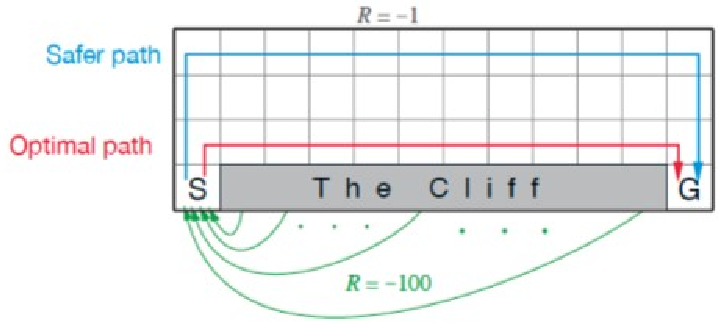


Figure 1: Cliff-walking or gridworld problem (Example 6.6 in Sutton and Barto's book)

Consider the gridworld shown in Fig. 1. This is a standard undiscounted ($\gamma = 1$), episodic task, with start (S) and goal (G) states, and the usual actions causing movement up, down, right, and left. Reward is -1 on all transitions except those into the region marked "The Cliff." Stepping into this region incurs a reward of -100 and sends the agent instantly back to the start. There are 48 states (positions in the grid) and 4 actions. This environment can be found here https://github.com/caburu/gym-cliffwalking.

Let μ be a fixed stochastic policy, which assigns uniform distribution on $\mathcal{A}$, i.e., given any state i, we have the probability of taking action a is $\mu(a \mid i) = 1/4$ for all $a \in \mathcal{A}$. Your job is to implement TD(λ) for finding V_μ. You can either implement TD(λ) with tabular settings or with linear function approximations [1]. Note that this is an undiscounted and episodic problem.

[1.] Value Function Approximation in Reinforcement Learning using the Fourier Basis. George Konidaris and Sarah Osentoski and Philip Thomas.

Questions: Write a simulation program to compute V_μ for the cliff-walking problem using TD(λ). In your simulation, consider 1000 episodes, where each episode runs 20 steps of TD(λ) given in class. For each episode, compute the norm of the expected TD update (NEU), the average of temporal difference, i.e.,

$$\text{NEU} = \frac{1}{20} \sum_{k=1}^{20} (d_k z_k)^2,$$

where z_k is the trace vector. Then for every 10 episode, you take the average of the NEU values and plot this average as a function of the number of episodes. Note that for each episode, you should initialize your function values V_μ as the values returned by the previous step.

You are asked to submit a pseudo code to explain your simulation and a plot which shows 5 curves of the average of NEU values as a function of the number of episodes for $\lambda = 0, .3, .5, .7, 1$. Finally, briefly explain the impacts of λ on the performance of TD learning.

Pseudo Code:

```
On-line Tabular TD(λ)

Input: a policy π, trace decay rate λ ∈ [0, 1]
Output: NEU_episodes (List of NEU values corresponding to each episode)
Algorithm parameters: step size α > 0, learning rate γ ∈ [0, 1], number of episodes, length of episodes
Initialize: V(s) = 0, Z(s) = 0, ∀ s ∈ S, NEU_episodes = empty list, NEU_steps = empty list

Loop for each episode:
   s ← 0
   Z(s) ← 0 ∀ s
   NEU_steps ← empty list
   Loop for each step of episode:
      a ← π(s)
      Take action a, observe reward r, and next state s'
      δ ← r + γV(s') – V(s)
      Z(s) ← Z(s) + 1
      For all s:
         V(s) ← V(s) + αδZ(s)
         Z(s) ← γλZ(s)
      NEU_steps.append((δZ(s)²)
      s ← s'
      If s is terminal or end of episode
         NEU_episodes.append(sum(NEU_steps)/number of steps in episode)
         Go to next episode
```

```
HW3 Generate Plots

Algorithm parameters: lambda_list = [0, 0.3, 0.5, 0.7, 1]

For lambda in lambda_list:
   NEU_episodes ← TD(π = random policy,  λ=lambda)
   NEU_averaged ← Place entries of NEU_episodes into bins of length 10, average each bin
   Plot NEU_averaged vs episode counts for current lambda
Display plot
```

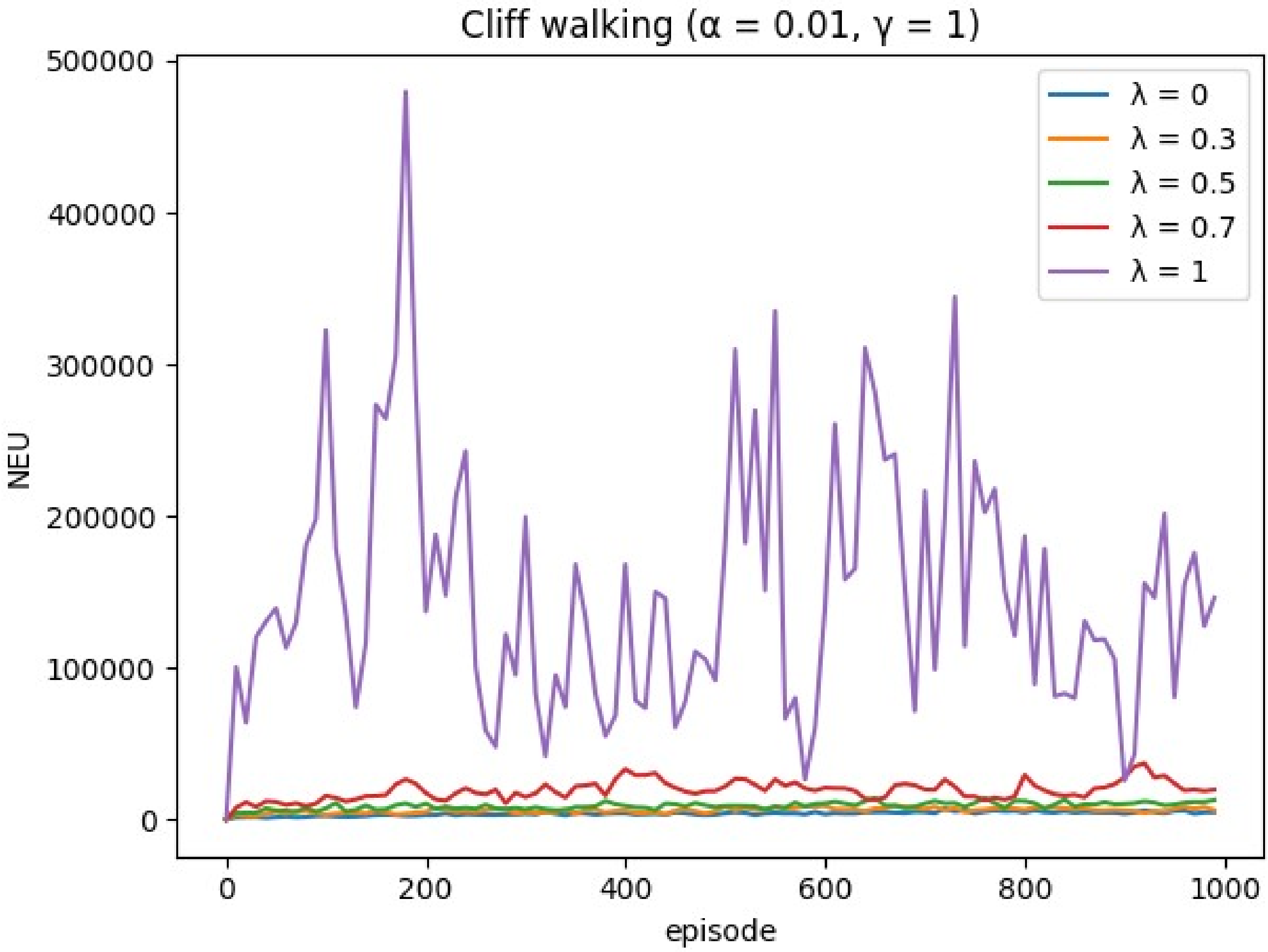
Briefly explain the impacts of λ on the performance of TD learning:

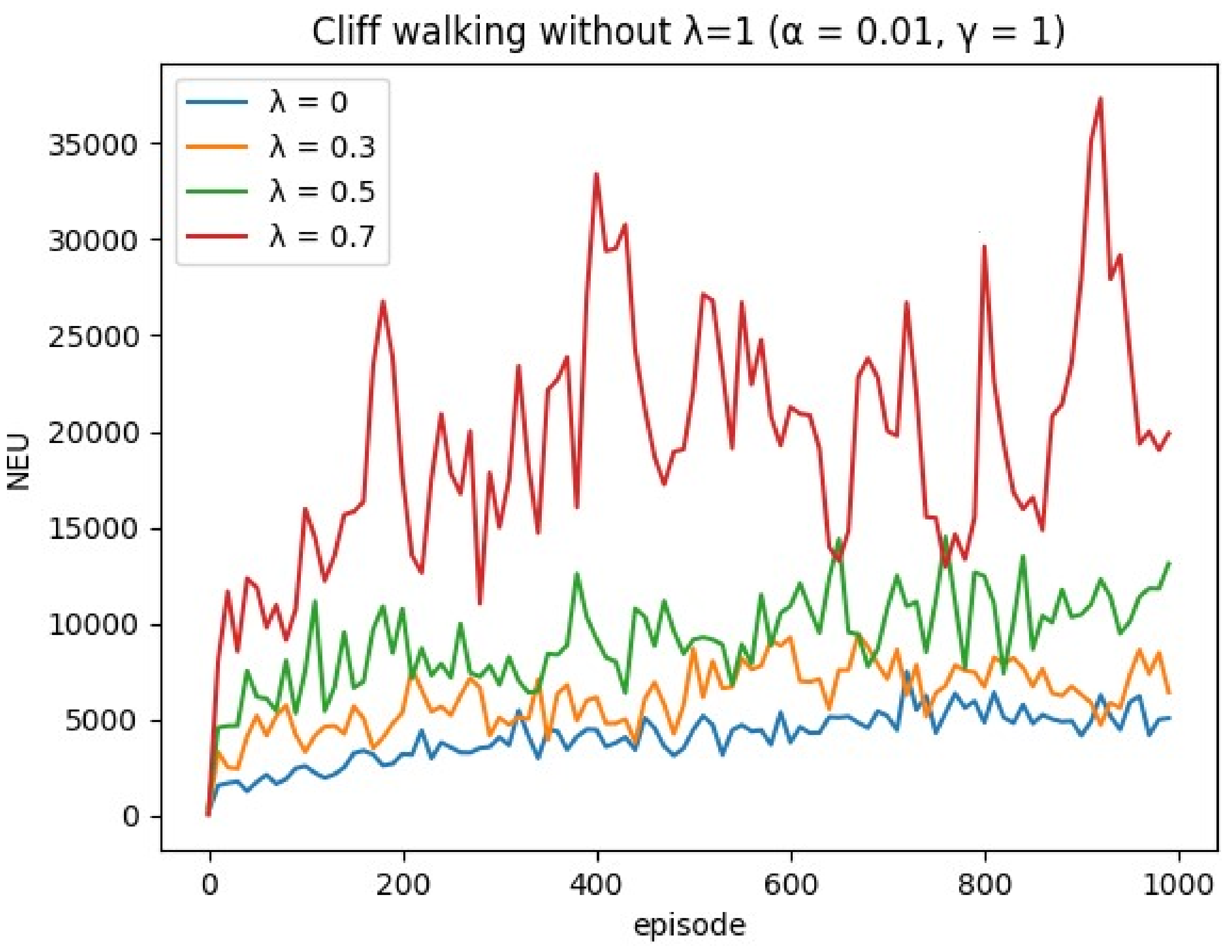
The value of lambda ($\lambda \in [0, 1]$) governs the decay rate of the trace vector. These trace vectors determine the degree to which previous predictions at a given state are eligible for updates. Values of λ closer to zero cause the trace vector to decay faster, whereas λ values closer to one create a longer lasting trace. As such, high values of λ result in value predictions of states visited further into the past getting assigned credit/blame (based on currently observed errors) to a greater extent than if λ were lower.

When $\lambda = 0$ you have Temporary Difference Learning (TD(0)), and when $\lambda = 1$ you effectively have Monte Carlo (MC). TD(0) and MC have different characteristics, with one being preferable over the over for a given problem. Lambda provides the ability to pick a point between these two extremes, maximizing the cumulative benefit.

For example, MC has high variance, zero bias, is not very sensitive to the initial values chosen, and does not exploit the Markov property (and therefore usually more effective in non-Markov environments). Whereas TD(0) has low variance, some bias, is sensitive to the initial value, and exploits the Markov property via bootstrapping (and therefore usually more effective in Markov environments). Tuning the value of λ for a given problem and environment can therefore speed up convergence through utilizing these respective properties.

Note that the increasing variance can be observed within the following graphs.


Cliff walking (α = 0.01, γ = 1)
λ = 0
λ = 0.3
λ = 0.5
λ = 0.7
λ = 1
NEU
episode
500000
400000
300000
200000
100000
0
0
200
400
600
800
1000


Cliff walking without λ=1 (α = 0.01, γ = 1)
λ = 0
λ = 0.3
λ = 0.5
λ = 0.7
NEU
episode
35000
30000
25000
20000
15000
10000
5000
0
0
200
400
600
800
1000