100/100

ECE5984: Reinforcement Learning

Assignment #3

Nima Mohammadi

nimamo@vt.edu

Problem 1.
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

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

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

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

First, we are going to shows that the stochastic approximator is indeed an estimator of μ:

$$\begin{aligned}\mathbb{E}[\theta_{k+1}] &= \mathbb{E}[(1-\epsilon)\theta_k + \epsilon X_k] \\ &= (1-\epsilon)\mathbb{E}[\theta_k] + \epsilon\mathbb{E}[X_k] \\ &= (1-\epsilon)\mathbb{E}[\theta_k] + \epsilon\mu\end{aligned}$$

Denoting $\mathbb{E}[\theta_k]$ by $\bar{\theta}_k$, we can write the expectation as

$$\bar{\theta}_{k+1} = (1-\epsilon)\bar{\theta}_k + \epsilon\mu$$

Making the assumption that $\theta_0 = 0$, then we have

$$\begin{aligned}\bar{\theta}_1 &= \epsilon\mu \\ \bar{\theta}_2 &= (1-\epsilon)\epsilon\mu + \epsilon\mu \\ \bar{\theta}_3 &= (1-\epsilon)^2\epsilon\mu + (1-\epsilon)\epsilon\mu + \epsilon\mu \\ &\vdots \\ \bar{\theta}_t &= \epsilon\sum_{i=0}^{t-1}(1-\epsilon)^i\mu = \frac{\epsilon\mu\left(1-(1-\epsilon)^t\right)}{1-(1-\epsilon)} \quad = \mu\left[1-(1-\epsilon)^t\right]\end{aligned}$$

✓

Therefore, $\lim_{k\to\infty}\bar{\theta}_k = \mu$.
We also have

$$\text{Var}[\theta_{k+1}] = \text{Var}[(1-\epsilon)\theta_k + \epsilon X_k] = (1-\epsilon)^2\,\text{Var}[\theta_k] + \epsilon^2\,\text{Var}[X_k]$$

Then similarly for the variance,

$$\begin{aligned}
\operatorname{Var}[\theta_0] &= 0 \\
\operatorname{Var}[\theta_1] &= \epsilon^2\sigma^2 \\
\operatorname{Var}[\theta_2] &= \epsilon^2\sigma^2(1+(1-\epsilon)^2) \\
&\vdots \\
\operatorname{Var}[\theta_k] &= \epsilon^2\sigma^2\sum_{i=0}^{k-1}(1-\epsilon)^{2i} = \frac{\epsilon^2\sigma^2\left(1-(1-\epsilon)^{2k}\right)}{1-(1-\epsilon)^2} = \frac{\epsilon\sigma^2\left[1-(1-\epsilon)^{2k}\right]}{2-\epsilon} \qquad \checkmark
\end{aligned}$$

Then with the Chebyshev's inequality we have,

$$\begin{aligned}
\limsup_{k\to\infty}\mathbb{P}\left(\left|\theta_k-\mu\right| \geq c\sqrt{\epsilon}\right) &= \limsup_{k\to\infty}\frac{\epsilon\sigma^2\left[1-(1-\epsilon)^{2k}\right]}{(2-\epsilon)c^2\epsilon} \\
&= \limsup_{k\to\infty}\frac{\delta^2}{c^2(2-\epsilon)} \leq \frac{\delta^2}{c^2} \qquad \checkmark
\end{aligned}$$

where the last inequality comes from $0<\epsilon<1$.

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

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

and let $V_\mu(s) \leq V^*(s) = V_{\mu^*}(s)$ 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

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

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

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

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

$$\begin{aligned}
\left(T_\mu Q\right)(s,a) &= r(s,a)+\gamma\sum_{s\in\mathcal{S}}p_{ss'}(a)Q\left(s',\mu\left(s'\right)\right) \\
(TQ)(s,a) &= r(s,a)+\gamma\sum_{s\in\mathcal{S}\in\mathcal{S}}p_{ss'}(a)\max_{a'}Q\left(s',a'\right)
\end{aligned}$$

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

1. Let $V_{Q,\mu}\left(s'\right) = Q\left(s',\mu\left(s'\right)\right)$ and $V_Q\left(s'\right) = \max_a Q\left(s',a\right)$. Show that

$$\begin{aligned}
\left\|V_{Q,\mu}-V_{Q',\mu}\right\|_\infty &\leq \left\|Q-Q'\right\|_\infty \\
\left\|V_Q-V_{Q'}\right\|_\infty &\leq \left\|Q-Q'\right\|_\infty
\end{aligned}$$

In this problem, we employ the lemma that

$$\left|\max_a f(a)-\max_a g(a)\right| \leq \max_a\left|f(a)-g(a)\right|$$

By definition,

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

Then we have

$$\begin{aligned}
\left|Q(s,\mu(s))-Q'(s,\mu(s))\right| &\leq \max_{a'}\left|Q(s,a')-Q'(s,a')\right| \\
&\leq \max_{s',a'}\left|Q(s',a')-Q'(s',a')\right|
\end{aligned}$$

which proves

$$\left\| V_{Q,\mu} - V_{Q',\mu} \right\|_\infty \leq \left\| Q - Q' \right\|_\infty$$

✓

For second part we have

$$V_Q - V_{Q'} = \max_a Q(s,a) - \max_{a'} Q'(s,a')$$

Then,

$$\begin{aligned} |\max_a Q(s,a) - \max_{a'} Q'(s,a')| &\leq \max_a |Q(s,a) - Q'(s,a)| \\ \Rightarrow \max_s |\max_a Q(s,a) - \max_{a'} Q'(s,a')| &\leq \max_{s,a} |Q(s,a) - Q'(s,a)| \\ \Rightarrow \left\| V_Q - V_{Q'} \right\|_\infty &\leq \left\| Q - Q' \right\|_\infty \end{aligned}$$

✓

2. Show that

$$\begin{aligned} \left\| T_\mu Q - T_\mu Q' \right\|_\infty &\leq \gamma \left\| Q - Q' \right\|_\infty \\ \left\| TQ - TQ' \right\|_\infty &\leq \gamma \left\| Q - Q' \right\|_\infty \end{aligned}$$

By definition we have

$$\begin{aligned} \left| T_\mu Q - T_\mu Q' \right| &= \left| \left(r(s,a) + \gamma \sum_{s \in \mathcal{S}} p_{ss'}(a) Q\left(s', \mu\left(s'\right)\right) \right) - \left(r(s,a) + \gamma \sum_{s \in \mathcal{S}} p_{ss'}(a) Q'\left(s', \mu\left(s'\right)\right) \right) \right| \\ &= \gamma \sum_{s \in \mathcal{S}} p_{ss'}(a) \left| Q\left(s', \mu(s')\right) - Q'\left(s', \mu(s')\right) \right| \\ &\leq \gamma \sum_{s \in \mathcal{S}} p_{ss'}(a) \max_{a'} \left| Q\left(s', a'\right) - Q'\left(s', a'\right) \right| \\ &\leq \gamma \sum_{s \in \mathcal{S}} p_{ss'}(a) \max_{a', s''} \left| Q\left(s', a'\right) - Q'\left(s', a'\right) \right| \\ &= \gamma \left\| Q - Q' \right\|_\infty \end{aligned}$$

Then for second part,

$$\begin{aligned} (TQ - TQ')(s,a) &= \gamma \sum_{s \in \mathcal{S}} p_{ss'}(a) \left[\max_{a'} Q(s',a') - \max_{a''} Q(s',a'') \right] \\ &\leq \gamma \sum_{s \in \mathcal{S}} p_{ss'}(a) \left[\max_{a'} Q(s',a') - Q(s',a') \right] \\ &\leq \gamma \max_{s'',a'} \left| Q(s'',a') - Q'(s'',a') \right| \\ &= \gamma \left\| Q - Q' \right\|_\infty \end{aligned}$$

✓

3. Show that

$$\begin{aligned} \left\| Q - Q_\mu \right\|_\infty &\leq \frac{\left\| Q - T_\mu Q \right\|_\infty}{1 - \gamma} \\ \left\| Q - Q^* \right\|_\infty &\leq \frac{\left\| Q - TQ \right\|_\infty}{1 - \gamma} \end{aligned}$$

Using the results we obtained in part 2,

$$\begin{aligned} \left\| Q - Q_\mu \right\|_\infty &\leq \left\| Q - T_\mu Q \right\|_\infty + \left\| T_\mu Q - T_\mu Q_\mu \right\|_\infty \\ &\leq \left\| Q - T_\mu Q \right\|_\infty + \gamma \left\| Q - Q_\mu \right\|_\infty \\ \Rightarrow \left\| Q - Q_\mu \right\|_\infty &\leq \frac{\left\| Q - T_\mu Q \right\|_\infty}{1 - \gamma} \end{aligned}$$

Then for the second part,

$$\left\| Q - Q^* \right\|_\infty \leq \| Q - TQ \|_\infty + \left\| TQ - Q^* \right\|_\infty$$

We know that $Q^* = TQ^*$, then

$$\|Q - Q^*\|_\infty \le \|Q - TQ\|_\infty + \|TQ - TQ^*\|_\infty$$

And from part 2,

$$\begin{aligned} \|TQ - TQ^*\|_\infty &\le \gamma \|Q - Q^*\|_\infty \\ \Rightarrow \|Q - Q^*\|_\infty &\le \|Q - TQ\|_\infty + \gamma\|Q - Q^*\|_\infty \\ \Rightarrow (1-\gamma)\|Q - Q^*\|_\infty &\le \|Q - TQ\|_\infty \\ \Rightarrow \|Q - Q^*\|_\infty &\le \frac{\|Q - TQ\|_\infty}{(1-\gamma)} \end{aligned}$$

✓

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

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

We had $V_{Q^*} = V^*$, and also $V_{Q_\mu} = V_\mu$ due to μ being a greedy policy. Then,

$$\begin{aligned} \|V^* - V_\mu\|_\infty &\le \|V^* - V_Q\|_\infty + \|V_Q - V_\mu\|_\infty \\ &= \|V_{Q^*} - V_Q\|_\infty + \left\|V_Q - V_{Q_\mu}\right\| \infty \end{aligned}$$

Then part 1 and 3 of the problem gives us

$$\begin{aligned} \|V^* - V_\mu\|_\infty &\le \|Q^* - Q\|_\infty + \|Q - Q_\mu\|_\infty \\ &\le \frac{\|Q - TQ\|_\infty}{1-\gamma} + \frac{\|Q - T_\mu Q\|_\infty}{1-\gamma} \end{aligned}$$

From greedy-ness of the policy μ we have $T_\mu Q = TQ$.

$$\begin{aligned} \|V^* - V_\mu\|_\infty &\le \frac{2\|Q - TQ\|_\infty}{1-\gamma} \\ &= \frac{2\beta}{1-\gamma} \end{aligned}$$

Moreover, we have $V^*(s) \ge V_\mu(s); \forall s \in S$

$$\begin{aligned} V^*(s) - V_\mu(s) &\le \frac{2\beta}{1-\gamma} \\ \Rightarrow V_\mu(s) &\ge V^*(s) - \frac{2\beta}{1-\gamma} \quad \forall s \in S \end{aligned}$$

✓

Problem 3. Coding question

Consider the gridworld shown. 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.

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.

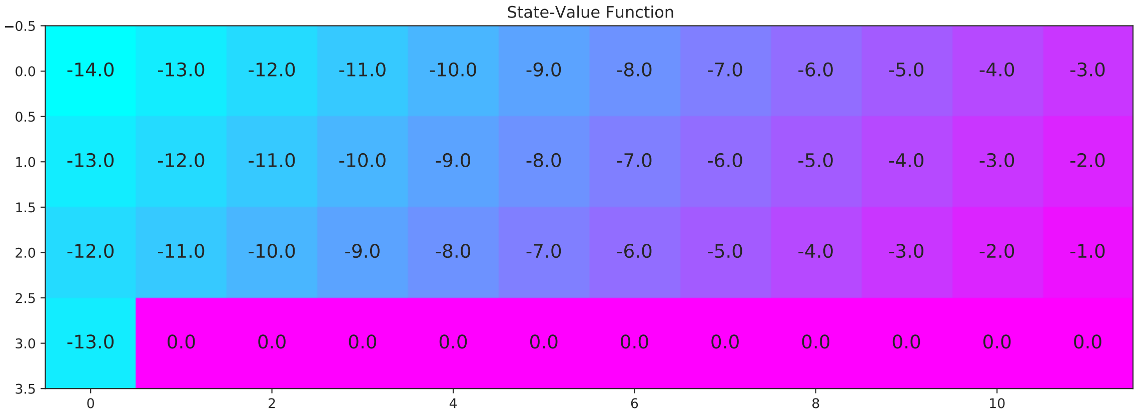
Questions: Write a simulation program to compute V_μ for the cliff-walking problem using TD(λ). In your simulation, consider 1000 episodes, where each episode runs 20 steps of TD(λ) given in class. For each episode, compute the norm of the expected TD update (NEU), the average of temporal difference,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

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

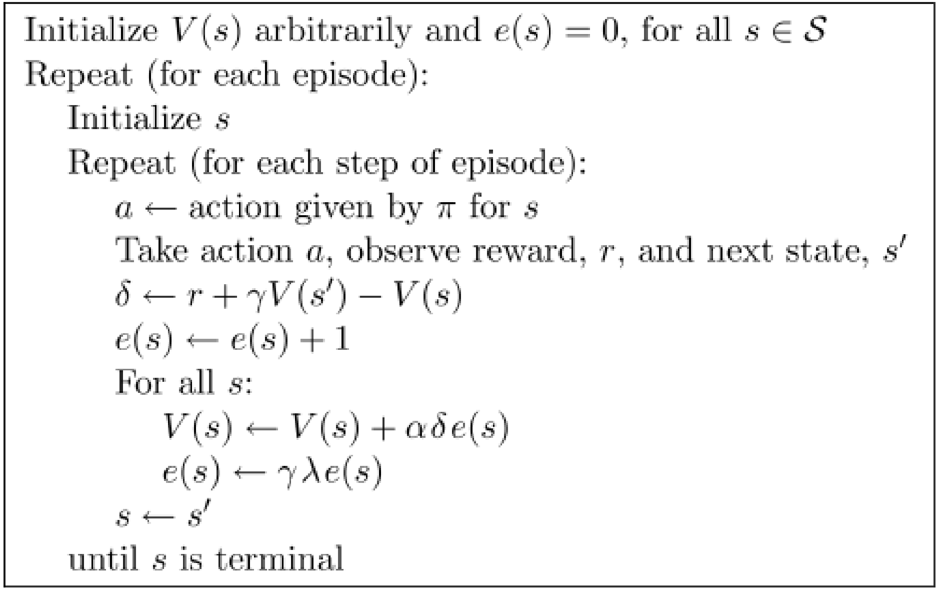
where z_k is the trace vector. Then for every 10 episode, you take the average of the NEU values and plot this average as a function of the number of episodes. Note that for each episode, you should initialize your function values V_μ as the values returned by the previous step.

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.

This part considers the 'prediction problem', that is, estimating the value of an existing policy. To this end we employ the TD(λ) algorithm, for the Cliff-Walking problem. First of all, for the sake of the comparison, I plot the value function for the optimal policy (not demanded in the question):

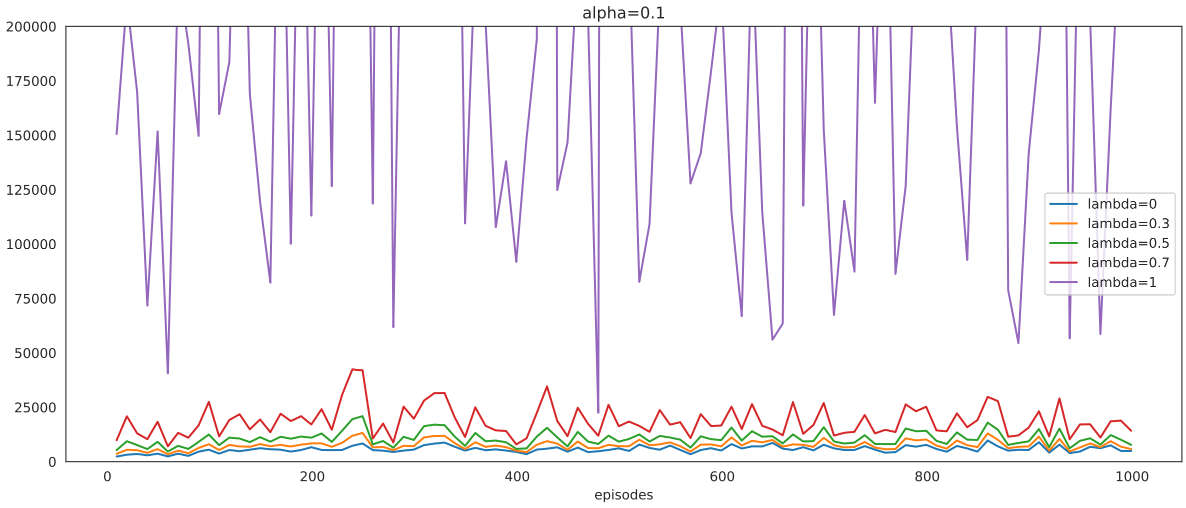


I have also included the Python code for my simulation at the end of this document. I am using the tabular form of the algorithm. The pseudocode of this algorithm is as follow.

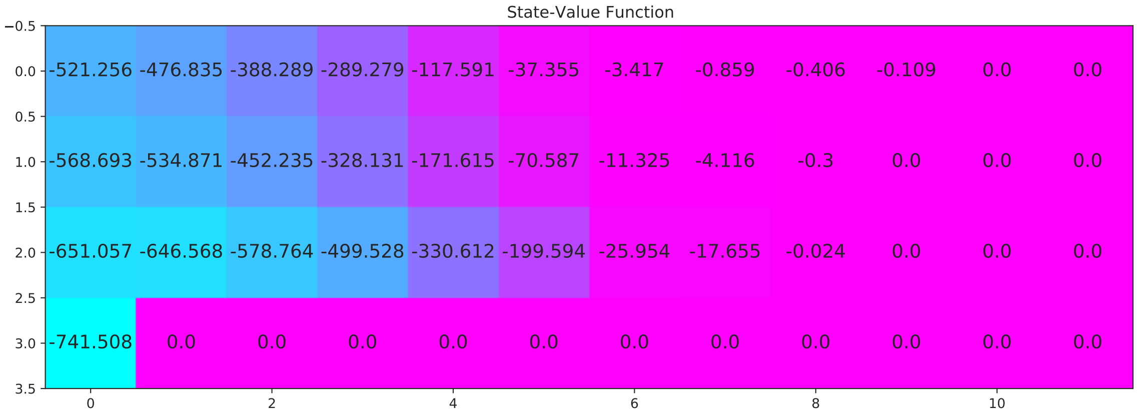


```
Initialize $V(s)$ arbitrarily and $e(s) = 0$, for all $s \in \mathcal{S}$
Repeat (for each episode):
  Initialize $s$
  Repeat (for each step of episode):
    $a \leftarrow$ action given by $\pi$ for $s$
    Take action $a$, observe reward, $r$, and next state, $s'$
    $\delta \leftarrow r + \gamma V(s') - V(s)$
    $e(s) \leftarrow e(s) + 1$
    For all $s$:
      $V(s) \leftarrow V(s) + \alpha\delta e(s)$
      $e(s) \leftarrow \gamma\lambda e(s)$
    $s \leftarrow s'$
  until $s$ is terminal
```

I have run the experiment twice, first with the learning rate α set to 0.1. I have also run the experiment with a decaying schedule for α, where as we expected the TD error can diminish to some extent for later episodes. The NEU plot below depicts the results for constant learning rate, following the instructions of the problem:

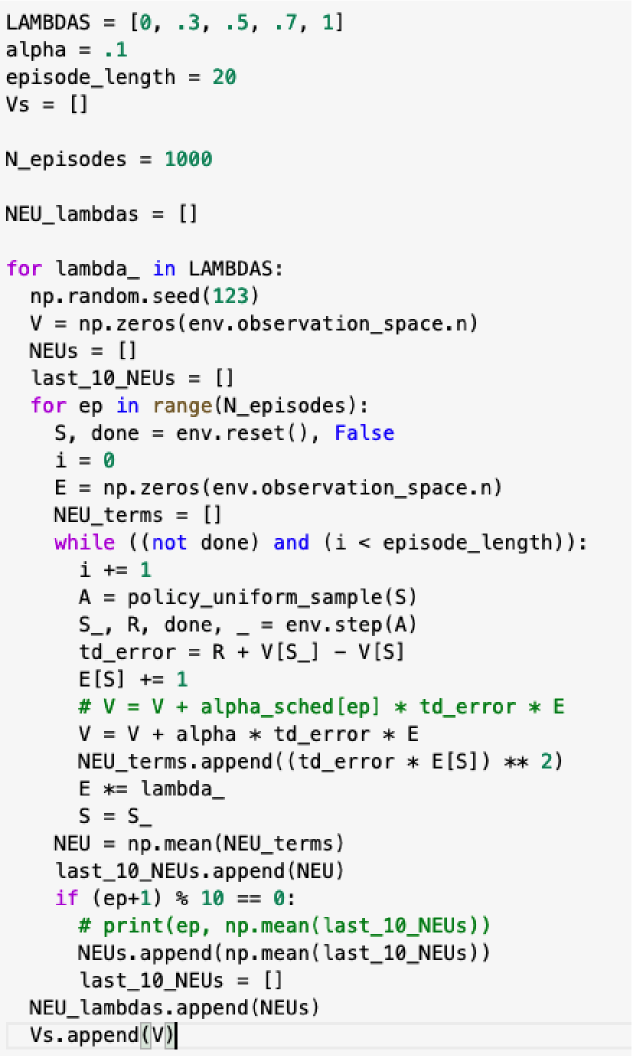


The value function for $\lambda = 0.3$ is depicted below:



-521.256	-476.835	-388.289	-289.279	-117.591	-37.355	-3.417	-0.859	-0.406	-0.109	0.0	0.0
-568.693	-534.871	-452.235	-328.131	-171.615	-70.587	-11.325	-4.116	-0.3	0.0	0.0	0.0
-651.057	-646.568	-578.764	-499.528	-330.612	-199.594	-25.954	-17.655	-0.024	0.0	0.0	0.0
-741.508	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

As shown above, the algorithm is not converged. Checking the number of the times each states is visited also confirms that many of the states, specifically the terminal state, are not visited and the episodes were prematurely terminated. My observations indicate that the number of simulated episodes and the maximum length of the episodes were simply insufficient leading to the unstable fluctuating regime our simulation ends up with.



```
LAMBDAS = [0, .3, .5, .7, 1]
alpha = .1
episode_length = 20
Vs = []

N_episodes = 1000

NEU_lambdas = []

for lambda_ in LAMBDAS:
  np.random.seed(123)
  V = np.zeros(env.observation_space.n)
  NEUs = []
  last_10_NEUs = []
  for ep in range(N_episodes):
    S, done = env.reset(), False
    i = 0
    E = np.zeros(env.observation_space.n)
    NEU_terms = []
    while ((not done) and (i < episode_length)):
      i += 1
      A = policy_uniform_sample(S)
      S_, R, done, _ = env.step(A)
      td_error = R + V[S_] - V[S]
      E[S] += 1
      # V = V + alpha_sched[ep] * td_error * E
      V = V + alpha * td_error * E
      NEU_terms.append((td_error * E[S]) ** 2)
      E *= lambda_
      S = S_
    NEU = np.mean(NEU_terms)
    last_10_NEUs.append(NEU)
    if (ep+1) % 10 == 0:
      # print(ep, np.mean(last_10_NEUs))
      NEUs.append(np.mean(last_10_NEUs))
      last_10_NEUs = []
  NEU_lambdas.append(NEUs)
  Vs.append(V)
```