ECE - 5984 : Homework 5
Instructor: Thinh T. Doan, TA: Amit Dutta
Due Date: 11/19/2021

Problem 1 [20 points]

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a smooth function, i.e., f is continuously differentiable function and has Lipschitz continuous gradient with constant $L > 0$

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \qquad \forall x, y \in \mathbb{R}^d.$$

Show the following

1. For all $x, y \in \mathbb{R}^d$ [10 points]

$$|f(x) - f(y) - \nabla f(y)^T(x - y)| \leq \frac{L}{2}|x - y|^2$$

Hint: Start from the following equation

$$f(x) = f(y) + \int_0^1 \nabla f(y + \tau(x - y))^T(x - y)d\tau$$

hint given

$$f(x) = f(y) + \int_0^1 \nabla f(y + \tau(x-y))^T (x-y)\, d\tau$$

$$= f(y) + \int_0^1 \langle \nabla f(y + \tau(x-y)), (x-y) \rangle\, d\tau$$

where $\langle v, w \rangle$ = inner product $= v \cdot w^T$

$$\therefore \; |f(x) - f(y) - \nabla f(y)^T(x-y)| = |f(x) - f(y) - \langle \nabla f(y), (x,y) \rangle|$$

$$= \int_0^1 \langle \nabla f(y + \tau(x-y)) - \nabla f(y), (x,y) \rangle\, d\tau$$

$$\leq \int_0^1 \|\nabla f(y + \tau(x-y)) - \nabla f(y)\| \times \|x-y\|\, d\tau$$

via cauchy - schwartz

$$\leq \int_0^1 L\, \|\nabla y + \tau(x-y) - y\| \times \|x-y\|\, d\tau$$

via Lipschitz definition provided.

$$= L\|x-y\|^2 \int_0^1 \tau\, d\tau$$

$$= \underline{\frac{L}{2}\|x-y\|^2}$$

2. In addition to the Lipschitz continuity of ∇f, suppose f is twice continuously differentiable. Show that the Hessian of f is bounded, i.e., [10 points]

$$\|\nabla^2 f(x)\| \leq L \qquad \forall x \in \mathbb{R}^d.$$

- let $c > 0$ and $g(n) = \nabla f(x + ncy)$

$\Rightarrow g'(n) = \nabla^2 f(x + ncy)\, cy$ | equation 1

- By mean value throm there exists a $m \in [0,1]$ such that

$$\nabla f(x+cy) - \nabla f(x) = g(1) - g(0) = g'(m)$$

$$= \nabla^2 f(x + mcy)\, cy \quad \text{| by eq 1 above}$$

$$\Rightarrow \|\nabla f(x+cy) - \nabla f(x)\| = \|\nabla^2 f(x+mcy)\, cy\|$$

$$\leq L\|x + cy - x\| \quad \text{| via Lipschitz condition}$$

$$= L\|cy\|$$

$$\Rightarrow \|\nabla^2 f(x+mcy)\, cy\| \leq L\|cy\|$$

$$\Rightarrow \|\nabla^2 f(x+mcy)\, y\| \leq L\|y\|.$$

letting $c \to 0$.

$$\Rightarrow \|\nabla^2 f(x)\, y\| \leq L\|y\|$$

$$\therefore \|\nabla^2 f(x)\| \leq L$$

Problem 2 [40 points]

In this problem we will derive the policy gradient formulas for different policies. First, we recall a few notation given in the class. Let V_π be the value function associated with the policy π

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

where $\gamma \in (0,1)$ is the discount factor. Let ρ be the initial distribution. With some abuse of notation we denote by

$$V_\pi(\rho) = \mathbb{E}_{s\sim\rho}\left[V_\pi(s)\right].$$

Next, we parameterize the policy π by θ, i.e., $\{\pi_\theta \,|\, \theta \in \Theta\}$. We denote by the discounted state visitation distribution

$$d^\pi_{s_0}(s) := (1-\gamma)\sum_{k=0}^{\infty} \gamma^k \mathbb{P}^\pi(s_k = s \,|\, s_0),$$

and let $d^\pi_\rho(s) = \mathbb{E}_{s_0\sim\rho}\left[d^\pi_{s_0}(s)\right]$. By the policy gradient theorem we have

$$\nabla_\theta V_{\pi_\theta}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_\theta}_{s_0}}\mathbb{E}_{a\sim\pi_\theta(\cdot\,|\,s)}\left[\nabla_\theta \log \pi_\theta(a \,|\, s) Q_{\pi_\theta}(s,a)\right],$$

where

$$Q_\pi(s,a) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^t c(s_k, a_k) \,|\, s_0 = s, a_0 = a\right], \quad a_k \sim \pi(\cdot\,|\,), \ \forall k \geq 1,$$
$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s).$$

Denote the simplex over the set of actions as

$$\Delta(\mathcal{A})^{|\mathcal{S}|} \triangleq \left\{\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \,|\, \theta_{s,a} \geq 0, \sum_{a\in\mathcal{A}} \theta_{s,a} = 1\right\}.$$

Thus, if $\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}$ then we have

$$\nabla_\theta J_{\pi_\theta}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{a\sim d^{\pi_\theta}_{s_0}}\mathbb{E}_{a\sim\pi_\theta(\cdot\,|\,s)}\left[\nabla_\theta \log \pi_\theta(a \,|\, s) A_{\pi_\theta}(a,u)\right],$$

1. Consider direct parameterization, i.e., $\mathcal{S}$ and $\mathcal{A}$ are the set of states and actions, and

$$\pi_\theta(a \,|\, s) = \theta_{s,a}, \qquad \text{where } \theta \in \Delta(\mathcal{A})^{|\mathcal{S}|}.$$

Show that [10 points]

$$\frac{\partial V^{\pi_\theta}(\rho)}{\partial\, \theta_{s,a}} = \frac{1}{1-\gamma} d^{\pi_\theta}_\rho Q^{\pi_\theta}(s,a).$$

assuming this is a typo.

$$V^{\pi_\theta}(\rho) = E_{s_0 \sim \rho}\left[U_{\pi_\theta}(s)\right]$$

$$\frac{\partial V^{\pi_\theta}(s)}{\partial \pi_\theta(a|s)} = \frac{1}{1-\gamma} E_{s \sim d_s^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot|s)}\left[\nabla_\theta \ln\left[\pi_\theta(a|s)\right] Q^{\pi_\theta}(s,a)\right]$$

$$\frac{\partial V^{\pi_\theta}(s)}{\partial \pi_\theta(a|s)} = \frac{\partial}{\partial \pi_\theta(a|s)} E_{s_0 \sim \rho}\left[U^{\pi_\theta}(s)\right]$$

$$= E_{s_0 \sim \rho}\left[\frac{\partial U^{\pi_\theta}(s)}{\partial \pi_\theta(a|s)}\right]$$

$$= E_{s_0 \sim \rho}\left[\frac{1}{1-\gamma} E_{s \sim d_s^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot|s)}\left[\nabla_\theta \ln\left[\pi_\theta(a'|s')\right] Q^{\pi_\theta}(s,a)\right]\right]$$

as $\pi_\theta(a|s) = \theta_{s,a}$

$$= E_{s_0 \sim \rho}\left[\frac{1}{1-\gamma} E_{s \sim d_s^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot|s)}\left[\nabla_{\theta_{s,a}} \ln\left[\theta_{s',a'}\right] Q^{\pi_\theta}(s,a)\right]\right]$$

$$= \frac{1}{1-\gamma} E_{s_0 \sim \rho} E_{s \sim d_s^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot|s)}\left[\nabla_{\theta_{s,a}} \ln\left[\theta_{s',a'}\right] Q^{\pi_\theta}(s,a)\right]$$

$$= \frac{1}{1-\gamma} E_{s_0 \sim \rho} \sum_{s'} d_{s_0}^{\pi_\theta}(s') \sum_{a'} \pi_\theta(a'|s')\left[\nabla_{\theta_{s,a}} \ln\left[\theta_{s',a'}\right] Q^{\pi_\theta}(s,a)\right]$$

$$= \frac{1}{1-\gamma} E_{s_0 \sim \rho}\left[d_{s_0}^{\pi_\theta}(s)\, \pi_\theta(a|s) \left(\frac{1}{\theta_{s,a}} Q^{\pi_\theta}(s,a)\right)\right]$$

$$= \frac{1}{1-\gamma} E_{s_0 \sim \rho}\left[d_{s_0}^{\pi_\theta}(s)\, \theta_{s,a} \left(\frac{1}{\theta_{s,a}} Q^{\pi_\theta}(s,a)\right)\right]$$

$$= \frac{1}{1-\gamma} E_{s_0 \sim \rho}\left[d_{s_0}^{\pi_\theta}(s)\, Q^{\pi_\theta}(s,a)\right]$$

$$= \frac{1}{1-\gamma} Q^{\pi_\theta}(s,a)\, d_\rho^{\pi_\theta}(s)$$

//

2. Consider softmax parameterization

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}.$$

Show that [10 points]

$$\frac{\partial V_{\pi_\theta}(\rho)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\rho^{\pi_\theta} \pi(a \mid s) A^{\pi_\theta}(s,a).$$

[note: here s', a' are used for the gradient calculations $\left(\frac{\partial}{\partial \theta_{s',a'}}\right)$, a'' is used for $\sum_{a'' \in A}$]

Step 1: softmax differentiation

$$\pi_\theta(a|s) = \frac{e^{\theta_{s,a}}}{\sum_{a' \in A} e^{\theta_{s,a'}}}$$

$$\log(\pi_\theta(a|s)) = \log(e^{\theta_{s,a}}) - \log \sum_{a' \in A} \left[e^{\theta_{s,a'}}\right]$$

$$\Rightarrow \nabla_\theta\left[\log(\pi_\theta(a|s))\right] = \nabla_\theta\left[\log(e^{\theta_{s,a}})\right] - \nabla_\theta\left[\log \sum_{a' \in A}\left[e^{\theta_{s,a'}}\right]\right]$$

$$= \nabla_\theta \theta_{s,a} - \nabla_\theta\left[\log \sum_{a'' \in A} e^{\theta_{s,a''}}\right]$$

$$= \nabla_\theta \theta_{s,a} - \frac{\nabla_\theta \sum_{a'' \in A} e^{\theta_{s,a''}}}{\sum_{a'' \in A} e^{\theta_{s,a''}}}$$

as $\nabla_x \log(f(x)) = \frac{\nabla_x f(x)}{f(x)}$ by chain rule

as $\frac{\partial}{\partial \theta_{s',a'}} \sum_{a'' \in A}\left(e^{\theta_{s,a'}} + e^{\theta_{s,a^2}} \dots e^{\theta_{s,a}} \dots\right) = \frac{\partial}{\partial \theta_{s,a}} e^{\theta_{s,a}} = e^{\theta_{s,a}}$ | only a single $a'=a$ and $s'=s$

* as there is only a singular $s = s'$ and $a'' = a'$

$$\Rightarrow \frac{\partial}{\partial \theta_{s',a'}} \sum_{a'' \in A} e^{\theta_{s,a''}} = 1\{s=s'\} e^{\theta_{s,a'}}$$

$$\Rightarrow \frac{\partial}{\partial \theta_{s',a'}}\left[\log(\pi_\theta(a|s))\right] = \nabla_\theta \theta_{s,a} - \frac{1\{s=s'\} e^{\theta_{s,a'}}}{\sum_{a' \in A} e^{\theta_{s,a'}}}$$

$$= \frac{\partial}{\partial \theta_{s',a'}} \theta_{s,a} - 1\{s=s'\} \pi_\theta(a|s)$$

(again as there is only one value of $s=s'$, and $a=a'$ all other partial derivatives $\to 0$)

$$= 1\{s=s'\}\{a=a'\} - 1\{s=s'\}\pi_\theta(a|s)$$

$$\therefore \frac{\partial}{\partial \theta_{s',a'}} \log(\pi_\theta(a|s)) = 1\{s=s'\}\left(1\{a=a'\} - \pi_\theta(a|s)\right)$$

$$\left[\text{where } 1\{a=a'\} \begin{cases} =1 & \text{if } a=a' \\ =0 & \text{if } a \neq a' \end{cases}\right]$$

Step 2: Calculating $\frac{\partial V^{\pi_\theta}(\rho)}{\partial \theta_{s',a'}}$

As the softmax $= \pi_\theta$ is constrained to the simplex:

$$\pi_\theta \in \Delta(A)^{|S|}$$

$$\Rightarrow \frac{\partial V^{\pi_\theta}}{\partial \theta_{s',a'}} = \frac{1}{1-\gamma} E_{s\sim d_{s_0}^{\pi_\theta}} E_{a\sim \pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a|s) A^{\pi_\theta}(s,a)\right]$$

- Let $P^\pi(\tau|s_0=s)$ = the probability of following trajectory τ, when starting in state s and following policy π.
- Let $B = E_{s\sim\rho}\left[P^\pi(\tau|s_0=s)\right]$, where ρ is a distribution over S.

$$\frac{\partial V_{\pi_\theta}(\rho)}{\partial \theta_{s',a'}} = E_{\tau\sim B}\left[\sum_{t=0}^{\infty} \gamma^t \, 1[s_t = s'] \left(1[a_t = a'] A^{\pi_\theta}(s,a) - \pi_\theta(a|s) A^{\pi_\theta}(s,a)\right)\right]$$

$$= E_{\tau\sim B}\left[\sum_{t=0}^{\infty} \gamma^t \, 1[(s_t, a_t) = (s', a')] \, A^{\pi_\theta}(s,a)\right]$$

$$- \pi_\theta(a|s) \sum_{t=0}^{\infty} E_{\tau\sim B}\left[1[s_t = s'] \, A^{\pi_\theta}(s_t, a_t)\right]$$

$$= \frac{1}{1-\gamma} E_{s'\sim d_\rho^{\pi_\theta}} E_{a'\sim \pi_\theta(\cdot|s)}\left[1[(s',a') = (s,a)] \, A^{\pi_\theta}(s,a)\right]$$

$$= \frac{1}{1-\gamma} d_\rho^{\pi_\theta}(s) \, \pi_\theta(a|s) \, A^{\pi_\theta}(s,a)$$

$$\therefore \frac{\partial V_{\pi_\theta}(\rho)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\rho^{\pi_\theta}(s) \, \pi_\theta(a|s) \, A^{\pi_\theta}(s,a)$$

3. We denote by

$$Z(s) = \sum_{a \in \mathcal{A}} \pi_\theta(a \mid s) \exp\left(\alpha \frac{A^{\pi_\theta}(s,a)}{1-\gamma}\right),$$

where α is some positive constant. Show that [10 points]

$$\log Z(s) \geq 0, \qquad \forall \pi_\theta \in \Delta(\mathcal{A})^{|\mathcal{S}|}.$$

$$\log Z(s) = \log \sum_{a \in A} \left[\pi_\theta(a|s) \exp\left(\alpha \frac{A^{\pi_\theta}(s,a)}{1-\gamma}\right)\right]$$

$$\geq \sum_{a \in A} \left[\pi_\theta(a|s) \log \exp\left(\alpha \frac{A^{\pi_\theta}(s,a)}{1-\gamma}\right)\right]$$

By jensens inequality. and noting $\log(x)$ is concave

$$= \sum_{a \in A} \left[\pi_\theta(a|s)\, \alpha \frac{A^{\pi_\theta}(s,a)}{1-\gamma}\right]$$

$$= \frac{\alpha}{1-\gamma} \sum_{a \in A} \left[\pi_\theta(a|s)\, A^{\pi_\theta}(s,a)\right]$$

$$= \frac{\alpha}{1-\gamma} \sum_{a \in A} \left[\pi_\theta(a|s) \left[Q_{\pi_\theta}(s,a) - V_{\pi_\theta}(s)\right]\right].$$

$$= \frac{\alpha}{1-\gamma} \times 0 = 0$$

$$\therefore \log Z(s) \geq 0$$

4. Given any two policy π_1 and π_2. Show that [10 points]

$$V_{\pi_1}(\rho) - V_{\pi_2}(\rho) \leq \frac{1}{1-\gamma}\left(\max_s \frac{d_\rho^{\pi_1}(s)}{d_\rho^{\pi_2}(s)}\right) \sum_{s\in\mathcal{S}} d_\rho^{\pi_2}(s) \max_{a'} A_{\pi_2}(s,a')$$

Step 1: Performance difference lemma:

$$V_{\pi_1}(\rho) - V_{\pi_2}(\rho) = E_{\tau\sim P_r^{\pi}(\tau|s_0=s)}\left[\sum_k^\infty \gamma r(s_k,a_k)\right] - V_{\pi_2}(s)$$

* for convienence, let $\tau = \tau\sim P_r^{\pi_1}(\tau|s_0=s)$

$$= E_\tau\left[\sum_{k=0}^\infty \gamma r(s_k,a_k) + V_{\pi_2}(s) - V_{\pi_2}(s_k)\right] - V_{\pi_2}(s)$$

$$= E_\tau\left[\sum_{k=0}^\infty \gamma r(s_k,a_k) - V_{\pi_2}(s_k)\right] + E_\tau\left[\sum_{k=0}^\infty \gamma^k V_{\pi_2}(s_k)\right] - V_{\pi_2}(s)$$

$$= E_\tau\left[\sum_{k=0}^\infty \gamma r(s_k,a_k) - V_{\pi_2}(s_k)\right] + E_\tau\left[\sum_{k=1}^\infty \gamma^k V_{\pi_2}(s_k)\right] + \gamma^0 V_{\pi_2}(s_0) - V_{\pi_2}(s)$$

$$= E_\tau\left[\sum_{k=0}^\infty \gamma r(s_k,a_k) - V_{\pi_2}(s_k)\right] + E_\tau\left[\sum_{k=1}^\infty \gamma^k V_{\pi_2}(s_k)\right]$$

$$= E_\tau\left[\sum_{k=0}^\infty \gamma^k\left[(r(s_k,a_k) - V_{\pi_2}(s_k)) + \gamma V_{\pi_2}(s_{k+1})\right]\right]$$

$$= E_\tau\left[\sum_{k=0}^\infty \gamma^k\left[r(s_k,a_k) + \gamma E\left[V_{\pi_2}(s_{k+1}) \mid s_k,a_k\right] - V_{\pi_2}(s_k)\right]\right]$$

$$= E_\tau \sum_{k=0}^\infty \gamma^k\left[Q_{\pi_2}(s_k,a_k) - V_{\pi_2}(s_k)\right]$$

$$= E_\tau \sum_{k=0}^\infty \gamma^k A_{\pi_2}(s_k,a_k)$$

$$= \frac{1}{1-\gamma} E_\tau\left[A_{\pi_2}(s_k,a_k)\right].$$

$$= \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi_1}(s)\,\pi_1(a|s)\,A_{\pi_2}(a,s)$$ | by τ definition above.

Step 2: Gradient domination

$$V_{\pi_1}(\rho) - V_{\pi_2}(\rho) = \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi_1}(s)\, \pi_1(a|s)\, A_{\pi_2}(s,a)$$

$$\leq \frac{1}{1-\gamma} \sum_{s,a} \left[d_\rho^{\pi_1}(s) \max_{a'} A_{\pi_2}(s,a') \right]$$

$$= \frac{1}{1-\gamma} \sum_{s} \left[\frac{d_\rho^{\pi_1}(s)}{d_\rho^{\pi_2}(s)}\, d_\rho^{\pi_2}(s) \max_{a'} A_{\pi_2}(s,a') \right]$$

$$\leq \frac{1}{1-\gamma} \max_{s} \left[\frac{d_\rho^{\pi_1}(s)}{d_\rho^{\pi_2}(s)} \right] \sum_{s} \left[d_\rho^{\pi_2}(s) \max_{a'} A_{\pi_2}(s,a') \right]$$

$$\therefore\ V_{\pi_1}(\rho) - V_{\pi_2}(\rho) \leq \frac{1}{1-\gamma} \max_{s} \left[\frac{d_\rho^{\pi_1}(s)}{d_\rho^{\pi_2}(s)} \right] \sum_{s} \left[d_\rho^{\pi_2}(s) \max_{a'} A_{\pi_2}(s,a') \right]$$

Problem 4 - Coding question [20 points]

Questions: Write a simulation program to implement policy gradient in the tabular setting for the cliff-walking problem, similar to Hw3 and Hw4. For this task, the initial distribution always assigns a probability 1 to position S (i.e., you always start from S). Moreover, $f(\pi) = \mathbb{E}_{s\sim\rho}[V_\pi(s)]$ and we use the softmax parameterization

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'\in\mathcal{A}} \exp(\theta_{s,a'})}, \qquad \text{where } \theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$

where the gradient is given in problem 2.

In your simulation, consider a number of episodes, where each episode runs until the program terminates or at most 200 steps. Submit your code and explain what's the optimal policy your algorithm can return, e.g., is it a deterministic policy? Is it the same with the policy returned by Q-learning in Hw4?

Implementing the basic policy gradient alogithim with softmax parameterization was unable to solve for the cliffwalking-v0 environment. However the agent was able to converge on a local minimum - learning to avoid the falling off the cliff by moving into walls on the right hand side of the environment.

A heatmap generated from 1000 trajectories of the non determinsitic policy displayed within figure 1 visulizes this behavior. Hyperparameters attained via a gridsearch, with none resulting in a policy capable of returning an average reward of -200 (the maximum number of steps). Runs from one of these sub-optimal policies are displayed within figure 2.

The resultant policy is clearly different to that returned by Q-learning (HW4), which was able to solve for an optimal deterministically policy, and attain a consitant -13 reward by following it.

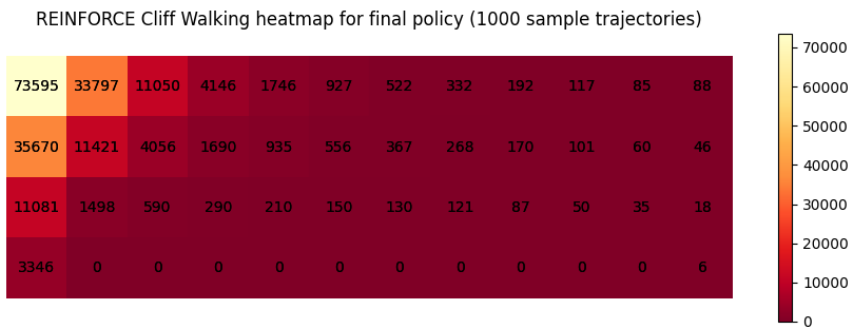

Figure 1: Heatmap of 1000 sample trajectories from the final policy generated via the REINFORCE policy gradient algorithim

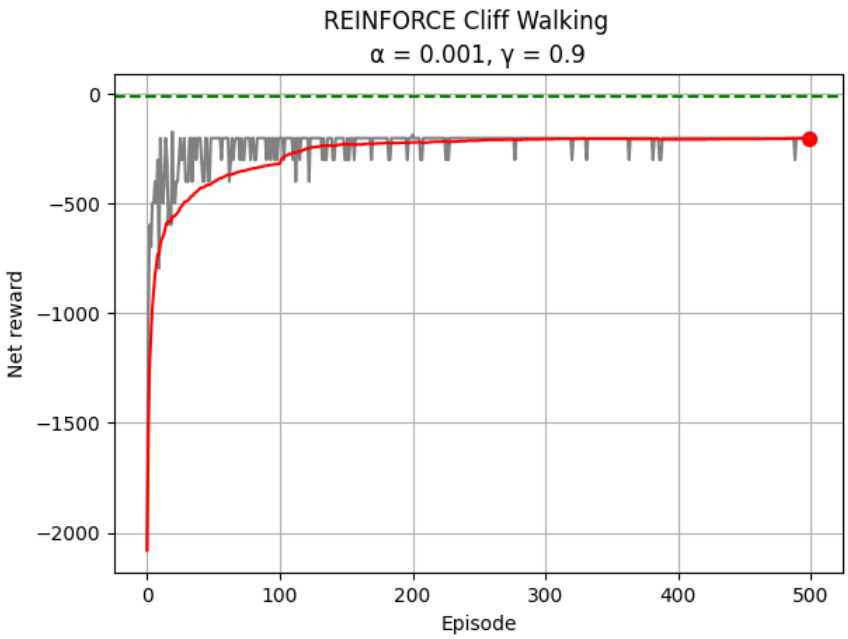

Figure 2: Episodic net reward and 100 episode moving average displayed in red. The Policy convereges on on -200 reward, learning to not walk off the cliff yet never reaching the terminal node. The maximum reward (-13) is displayed in green.

Expanding upon the REINFORCE policy gradient through the addition of a baseline function generated a policy that could constantly reach the terminal node. The Q Actor-Critic architecture was implemented with the resultant heatmap of 1000 sample runs following the final policy displayed within figure 3,and episode rewards displayed within figure 4 and 5. The resultant policy is suboptimal as it does not hug the cliff, with the agent instead opting to take a cautious path. This suboptimal cautious path results in an average net reward of -15.3.

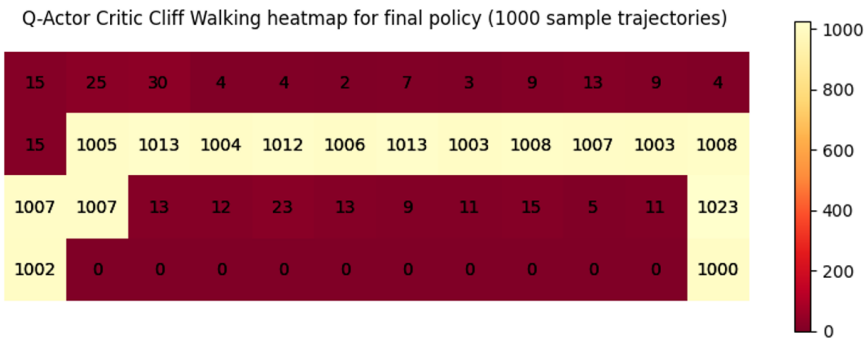

Figure 3: Heatmap of 1000 sample trajectories from the final policy generated via the Q-Actor Critic policy gradient algorithm

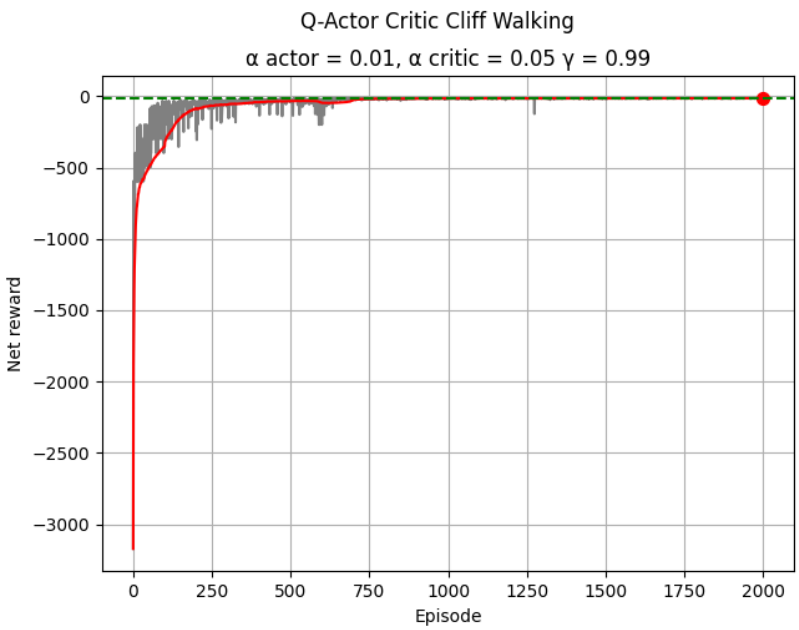

Figure 4: Episodic net reward for Q-Actor Critic (Zoomed out)

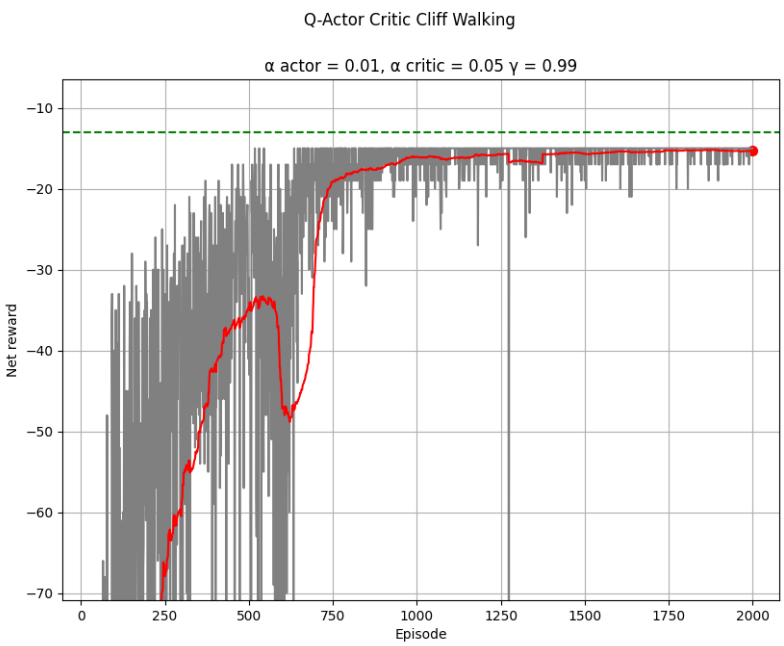

Figure 5: Episodic net reward for Q-Actor Critic (Zoomed in)

Problem 5 - Coding question [20 points]

Consider again the Cartpole environment from OpenAI Gym, which can be found here:
https://github.com/openai/gym/blob/master/gym/envs/classic_control/cartpole.py

In this question, we will consider policy gradient with linear function approximation using Fourier basis [1]. For this problem, consider discount factor is $\gamma = 0.9$ and the initial distribution can be any probability distribution.

[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 implement policy gradient with Gaussian policy for the cart-pole problem. That is, the action is generated by

$$a \sim \mathcal{N}(\phi(s)^T\theta, \sigma^2),$$

where σ is your choice (e.g., 0.5) and $\phi(s)$ is your feature vector. The parameter $\theta \in \mathbb{R}^d$, where you can select the dimension d. The gradient of $f(\pi_\theta) = \mathbb{E}_{s\sim\rho}[V_{\pi_\theta}(s)]$ is then given as

$$\nabla_\theta f(\pi_\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^\pi}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\Big(\nabla_\theta \log \pi_\theta(a \mid s)\Big)Q_{\pi_\theta}(s,a)\right],$$

and

$$\nabla_\theta \log \pi_\theta(a \mid s) = \frac{(a - \phi(s)^T\theta)\phi(s)}{\sigma^2}.$$

Note: softmax was used instead per lecturers announcement

In your simulation, consider a number of episodes, where each episode runs until the program terminates. Submit your code and show whether the policy returned by your algorithm work? Try your code with different dimension d and σ, and explain the impacts of these choices on the performance of your algorithms.

The REINFORCE algorithm with softmax parameterization and Fourier basis function approximation was able to solve the CartPole-V0 environment as displayed within figure 6 and 7.

As expected, REINFORCE has very high variance and is exceptionally sensitive to hyperparameters. Hyperparameters were found via a gridsearch, and are listed below:

- alpha = 0.0005
- gamma = 0.99
- Fourier order = 6 (with full coupling)

In addition, the state space velocity variables were bounded between -1.11 and 1.11 for training.

To ensure robustness, the algorithim and hyperparameters were validated with 10 unseen random seeds, with 8 reaching a steady state reward of 200 in less then 10 episodes, and the remaining 2 eventually converging after ~400 and ~3000 episodes as seen in figure 6.

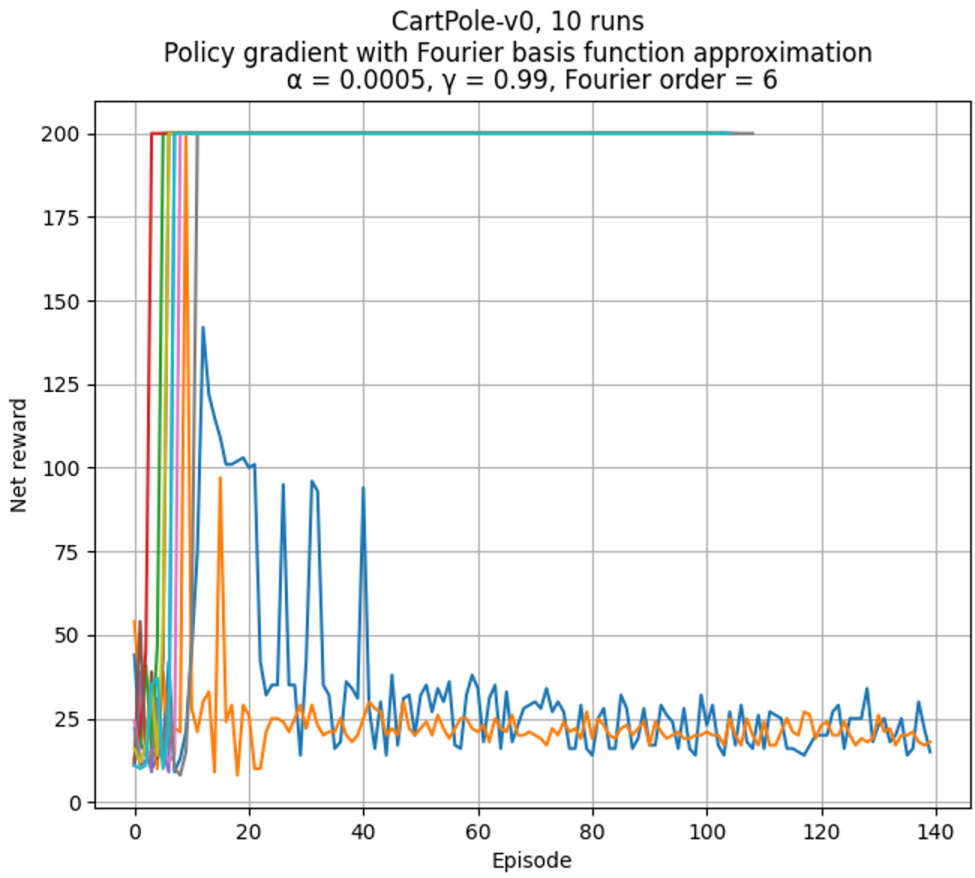

Figure 6: Episodic net reward for REINFORCE (Zoomed in). Note the 8 runs that are able to quickly converge to the maximum reward of 200.

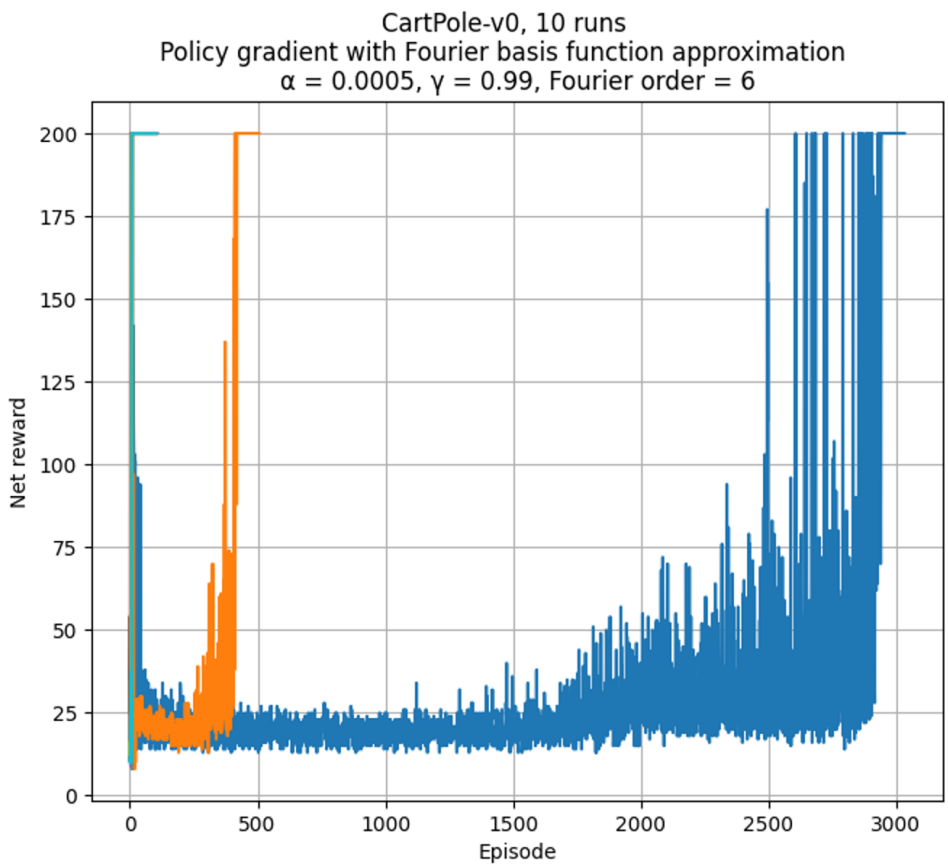

Figure 7: Episodic net reward for REINFORCE (Zoomed out). Note the 2 runs that do not converge quickly in orange and blue, yet are able to within ~400 and ~3000 episodes respectively.