CS6700 : Reinforcement Learning
Written Assignment #1

Intro to RL, Bandits, DP | Deadline: 23 Feb 2020, 11:55 pm

Name: Dodiya Keval | **Roll number:** CS19M023

- This is an individual assignment. Collaborations and discussions are strictly prohibited.
- Be precise with your explanations. Unnecessary verbosity will be penalized.
- Check the Moodle discussion forums regularly for updates regarding the assignment.
- Type your solutions in the provided LaTeXtemplate file.
- **Please start early.**

1. (2 marks) You have come across Median Elimination as an algorithm to get $(\epsilon, \delta)-$PAC bounds on the best arm in a bandit problem. At every round, half of the arms are removed by removing arms with return estimates below the median of all estimates. How would this work if we removed only one-fourth of the worst estimated arms instead? Attempt a derivation of the new sample complexity.

Solution: Likewise median elimination we also need to update the value of ϵ_l for the l^{th} round.so first of all we need to find probability of elimination of best arm atl^{th} round. for that we should consider two cases.

Case (1): Underestimation of true best arm(a^*).

P[Q(a^*) $<$ $q_*(a^*)$ - $\frac{\epsilon_l}{2}$] $\leq e^{-2(\frac{\epsilon_l}{2})^2 s_l}$

$= e^{-\frac{\epsilon_l^2}{2} s_l}$ (Using Hoeffding's Inequality).

Case (2): overestimation of bad arm compare to best true arm(a^*).

P[Q(a) $\geq$ Q(a^*) | Q(a^*) $\geq$ $q_*(a^*)$ - $\frac{\epsilon_l}{2}$] $\leq$ P[Q(a) $\geq$ q_*(a) + $\frac{\epsilon_l}{2}$ | Q(a^*) $\geq$ $q_*(a^*)$ - $\frac{\epsilon_l}{2}$]

$\leq e^{-\frac{\epsilon_l^2}{2} s_l}$

so the number of arms those are better than the true best arm a^* will be..

E[# bad arms | Q(a^*) $\geq$ $q_*(a^*)$ - $\frac{\epsilon_l}{2}$] $\leq n_l \; e^{-\frac{\epsilon_l^2}{2} s_l}$

now if optimal arm comes under last $\frac{n_l}{4}$ arms then optimal arm will be eliminated in $(l+1)^{th}$ round. so that probability will be..

P[# bad arms $\geq \frac{3n_l}{4}$ | Q(a^*) $\geq q_*(a^*)$ - $\frac{\epsilon_l}{2}$] $\leq \frac{4}{3}e^{-\frac{\epsilon_l^2}{2}s_l}$

(by markov's inequality $P(X \geq a) \leq \frac{E(X)}{a}$).

now add both the cases to get number of samples.
$e^{-\frac{\epsilon_l^2}{2}s_l} + \frac{4}{3}e^{-\frac{\epsilon_l^2}{2}s_l} \leq \delta_l$
$= s_l \geq \frac{2}{\epsilon_l^2}log\frac{7}{3\delta_l}$
so each arm will play at least s_l times and every level we have n_l arms.now sum of each ϵ_l should less or equal to zero and same for δ_l. so updation of ϵ_l,δ_l and n_l is given below.

$n_1 =$ n $\implies n_l = (\frac{3}{4})^{l-1}n$

$\epsilon_1 = \frac{\epsilon}{k} \implies \epsilon_l = (\frac{k-1}{k})^{l-1}\frac{\epsilon}{k}$

$\delta_1 = \frac{\delta}{2} \implies \delta_l = \frac{\delta}{2^l}$

since we are eliminating $(\frac{1}{4})^{th}$ of arms. total possible levels until one arm left will be $log_{4/3}n$. so final sample complexity will be..

$$\sum_{l=1}^{log_{4/3}n} n_l(\frac{2}{\epsilon_l^2}log\frac{7}{3\delta_l}) = \sum_{l=1}^{log_{4/3}n}(\frac{3}{4})^{l-1}n((\frac{k}{k-1})^{l-1}\frac{k}{\epsilon})^2log(\frac{2^l}{\delta}\frac{7}{3}) \qquad (1)$$

to this equation (1) converge, the value of k will be 8. and by putting this value to equation (1) we will get..

$\leq 64(\frac{nlog(1/\delta)}{\epsilon^2})\sum_{l=1}^{\infty}(\frac{48}{49})^{l-1}l$

it means..

$= O(\frac{nlog(\frac{1}{\delta})}{\epsilon^2})$

so the sample complexity will be $O(\frac{nlog(1/\delta)}{\epsilon^2})$.

2. (3 marks) Consider a bandit problem in which you know the set of expected payoffs for pulling various arms, but you do not know which arm maps to which expected payoff. For example, consider a 5 arm bandit problem and you know that the arms 1 through 5 have payoffs 3.1, 2.3, 4.6, 1.2, 0.9, but not necessarily in that order. Can you design a regret minimizing algorithm that will achieve better bounds than UCB? What makes you believe that it is possible? What parts of the analysis of UCB will you modify to achieve better bounds?

Solution: since we knows true value of each arm. our improvement in algorithm will be like find out value of maximum number of pulls such that estimate of second best arm become lesser than best arm always. so if we pull arms upto the point at where best arm and second best arm become closer than its middle value($\frac{4.6+3.1}{2} = 3.85$), then there is no need for further exploration(best arm mean = 4.6, second best arm mean = 3.1). so we can say that uncertainty term $u = \sqrt{\frac{logt}{n}} < 4.6 - 3.85 = 0.75$.
$= \sqrt{\frac{logt}{n}} < \frac{3}{4}$
$= n > \frac{16logt}{9}$
so, $n = 1 + \frac{16logt}{9}$, here n maximum number of required pulls.
and regret = $\sum_i E(T_i(n))\triangle i$. $T_i(n)$ is number times i^{th} arm pulled out of 'n' total pulls. so by this approach we are minimizing the expected pulls that decreases regret that is better that UCB.
so algo. will be like...

REPEAT LOOP
explore each arm and compute $Q(a_i) + \sqrt{\frac{logt}{k}}$
UNTILL 'n' TIMES.

t = number of pulls of i^{th} arm. k = total pulls till now.

3. (3 marks) Suppose you face a 2-armed bandit task whose true action values change randomly from time step to time step. Specifically, suppose that, for any time step, the true values of actions 1 and 2 are respectively 0.1 and 0.2 with probability 0.5 (case A), and 0.9 and 0.8 with probability 0.5 (case B).

(a) (1 mark) If you are not able to tell which case you face at any step, what is the best expectation of success you can achieve and how should you behave to achieve it?

Solution: given two cases and both are equiprobable . and assuming arms selection is random from uniform distribution[0,1]. by above policy we will get expectation be like..
$E[CaseA] = (0.1)P(Arm1) + (0.2)P(Arm2)$

$= E[CaseA] = (0.1)\frac{1}{2} + (0.2)\frac{1}{2}$
$E[CaseA] = 0.15$
and same for case B

$E[CaseB] = (0.9)P(Arm1) + (0.8)P(Arm2)$
$E[CaseB] = (0.9)\frac{1}{2} + (0.8)\frac{1}{2}$
$E[CaseB] = 0.85$

The expected Reward would be..
$E[Reward] = (0.15)\frac{1}{2} + (0.85)\frac{1}{2}$
$= E[Reward] = 0.50$

(b) (2 marks) Now suppose that on each step you are told whether you are facing case A or case B (although you still don't know the true action values). This is an associative search task. What is the best expectation of success you can achieve in this task, and how should you behave to achieve it?

Solution: now we already know the case then we just need to find out best(optimal arm) by any bandit algo.(e.x. UCB1, median elimination). so that is arm 2 with payoff 0.2(Case A) and arm 1 with payoff 0.9(case B).
so expected success will be..
$E[Reward] = (0.9)\frac{1}{2} + (0.2)\frac{1}{2} = 0.55$

4. (5 marks) Many tic-tac-toe positions appear different but are really the same because of symmetries.

(a) (2 marks) How might we amend the learning process described above to take advantage of this? In what ways would this change improve the learning process?

Solution: if we use advantage of symmetry, all four corner are same behaving so one state is enough to represent all four corner. hence it will reduces number of different states and total actions. implies computation will goes down and algorithm learning time gets decrease(e.x. algorithm doesn't need to explore all four corners only one corner information is enough to play a game). so the conclusion is amendment will be fine if opponent uses same approach.

(b) (1 mark) Suppose the opponent did not take advantage of symmetries. In that case, should we? Is it true, then, that symmetrically equivalent positions should necessarily have the same value?

Solution: suppose opponent did not take advantage of symmetries. in case we also should not use of symmetry. and there is case when symmetrically equivalent positions does not have same value.(e.x. suppose opponent playing all corners except any two so those two position value might be different from other two.)

(c) (2 marks) Suppose, instead of playing against a random opponent, the reinforcement learning algorithm described above played against itself, with both sides learning. What do you think would happen in this case? Would it learn a different policy for selecting moves?

Solution: if agent plays against another agent as itself. then it is possible that some amount of time both agent converge and reflect same actions to each other or we can say cyclic effect. and it is also possible to one agent learn better than other so one agent achieve higher skill set compare to other. and it might possible that one agent takes wrong actions all the times with better agent.

5. (1 mark) Ego-centric representations are based on an agent's current position in the world. In a sense the agent says, I don't care where I am, but I am only worried about the position of the objects in the world relative to me. You could think of the agent as being at the origin always. Comment on the suitability (advantages and disadvantages) of using an ego-centric representation in RL.

Solution: An ego-centric agent always thinking about area which is near to him. it doesnot think about global world. ego-centric agent may learns faster than non-ego centric agent. and ego-centric agent cannot make action that are beneficial in long run(global environment). so its kind of greedy thinker.

6. (2 marks) Consider a general MDP with a discount factor of γ. For this case assume that the horizon is infinite. Let π be a policy and V^π be the corresponding value function. Now suppose we have a new MDP where the only difference is that all rewards have a constant k added to them. Derive the new value function V^π_{new} in terms of V^π, c and γ.

Solution:

$V_\pi(s) = E[\sum\limits_{i=0}^{\infty} \gamma^i R_{t+i+1} | s_t = s]$.

here return at time t+i $= \sum\limits_{i=0}^{\infty} \gamma^i R_{t+i+1}$, so

now we are adding k constant then $V'_\pi(s) = E[\sum\limits_{i=0}^{\infty} \gamma^i (R_{t+i+1} + k) | s_t = s]$.

$= E[\sum\limits_{i=0}^{\infty} \gamma^i R_{t+i+1} + k \sum\limits_{i=0}^{\infty} \gamma^i | s_t = s]$.

$= V_\pi(s) + \dfrac{k}{1-\gamma}$.

so,

$V^\pi_{new} = V^\pi + \dfrac{k}{1-\gamma}$.

7. (4 marks) An ϵ-soft policy for a MDP with state set $\mathcal{S}$ and action set $\mathcal{A}$ is any policy that satisfies

$$\forall a \in \mathcal{A}, \forall s \in \mathcal{S} : \pi(a|s) \geq \frac{\epsilon}{|\mathcal{A}|}$$

Design a stochastic gridworld where a deterministic policy will produce the same trajectories as a ϵ-soft policy in a deterministic gridworld. In other words, for every trajectory under the same policy, the probability of seeing it in each of the worlds is the same. By the same policy I mean that in the stochastic gridworld, you have a deterministic policy and in the deterministic gridworld, you use the same policy, except for ϵ fraction of the actions, which you choose uniformly randomly.

(a) (2 marks) Give the complete specification of the world.

Solution:

(b) (2 marks) Will SARSA on the two worlds converge to the same policy? Justify.

Solution:

Yes, SARSA on the two worlds converge to the same policy, because probability of actions taking is same for both the worlds so state transition will also be same. so whatever difference in worlds SARSA will converges to the same policy for both worlds

8. (7 marks) You receive the following letter:
Dear Friend, Some time ago, I bought this old house, but found it to be haunted by ghostly sardonic laughter. As a result it is hardly habitable. There is hope, however, for by actual testing I have found that this haunting is subject to certain laws, obscure but infallible, and that the laughter can be affected by my playing the organ or burning incense. In each minute, the laughter occurs or not, it shows no degree. What it will do during the ensuing minute depends, in the following exact way, on what has been happening during the preceding minute: Whenever there is laughter, it will continue in the succeeding minute unless I play the organ, in which case it will stop. But continuing to play the organ does not keep the house quiet. I notice, however, that whenever I burn incense when the house is quiet and do not play the organ it remains quiet for the next minute. At this minute of writing, the laughter is going on. Please tell me what manipulations of incense and organ I should make to get that house quiet, and to keep it so.
Sincerely,
At Wits End

(a) (3 marks) Formulate this problem as an MDP (for the sake of uniformity, formulate it as a continuing discounted problem, with $\gamma = 0.9$. Let the reward be +1 on any transition into the silent state, and -1 on any transition into the laughing state.) Explicitly give the state set, action sets, state transition, and reward function.

Solution:
State Set:
there will be two states: (1) Laughing , (2) Silent.
State Set(K) = {L,S} where L = Laughing , S = Silent.

Action Set:
there will be 4 actions . suppose burning of incense represents as 'I', and organ playing represent as 'O' then actions will be like..
(1) $I'O'$ = Not playing organ with not burning incense.
(2) $I'O$ = Playing organ with not burning incense
(3) IO' = Not playing organ with burning incense
(4) IO = Playing organ with burning incense .
So action set(A) = $\{I'O', I'O, IO', IO\}$

State Transition:

Current State(S)	Action(A)	Reward(a)	Next State(S)
S	$I'O'$	-1	L
S	$I'O$	-1	L
S	IO'	+1	S
S	IO	-1	L
L	$I'O'$	-1	L
L	$I'O$	+1	S
L	IO'	-1	L
L	IO	+1	S

Reward function:
Reward function is represented as $E[Reward|s, s', a]$ where
s = current state , s' = next state , a = action
$E[Reward|s = L, s' = S, a = I'O'] = +1$,
$E[Reward|s = S, s' = L, a = I'O'] = -1$

$E[Reward|s = L, s' = S, a = I'O] = +1$,
$E[Reward|s = S, s' = L, a = I'O] = -1$

$E[Reward|s = L, s' = S, a = IO'] = +1$,
$E[Reward|s = S, s' = L, a = IO'] = -1$

$E[Reward|s = L, s' = S, a = IO] = +1$,
$E[Reward|s = S, s' = L, a = IO] = -1$

(b) (2 marks) Starting with simple policy of **always** burning incense, and not playing organ, perform a couple of policy iterations.

Solution: Given that $\pi_0(L) = IO' = \pi_0(S)$ and $\gamma = 0.9$

(a) Round 1
Evaluation :
V(L) = 0, V(S) = 0

$Q(L, IO') = E[Reward|L, L, IO'] + \gamma(V(L))$
$Q(L, IO') = -1 + \gamma(0)$
$Q(L, IO') = -1$

$Q(S, IO') = E[Reward|s = S, s' = S, a = IO'] + \gamma(V(S))$
$Q(S, IO') = 1 + \gamma(0)$
$Q(S, IO') = 1$
so at nth iteration this function convergences to -10 because of G.p., V(L) $= -\dfrac{\gamma^{n+1} - 1}{\gamma - 1}$ = -10.
and by same procedure $V(S) = 10$

Policy Improvement : V(L) = -10, V(S) = 10.

$Q(L, IO) = E[Reward|s = L, s' = S, a = IO] + \gamma(V(S))$
$= 1 + 0.9(10) = 10$

$Q(L, I'O') = E[Reward|s = L, s' = S, a = I'O'] + \gamma(V(L))$
$= -1 + 0.9(-10) = -10$

$Q(L, IO') = E[Reward|s = L, s' = S, a = IO'] + \gamma(V(L))$
$= -1 + 0.9(-10) = -10$

$Q(L, I'O) = E[Reward|s = L, s' = S, a = I'O] + \gamma(V(S))$
$= 1 + 0.9(10) = 10$

$Q(S, IO) = -1 + 0.9(10) = 8$
$Q(S, I'O') = -1 + 0.9(10) = 8$
$Q(S, IO') = 1 + 0.9(10) = 10$
$Q(S, I'O) = -1 + 0.9(10) = 8$
so best actions for state L are IO and $I'O$, and for state S is IO' only(e.g. V(L) = 10, V(S) = 10.).

(b) Round 2
Evaluation :
V(L) = 10, V(S) = 10
here we found best action at state L are $I'O$ and IO so we will find value for both the actions and take maximum of it. so
$Q(L, IO) = E[Reward|s = L, s' = S, a = IO] + \gamma(V(S))$
$Q(L, IO) = 1 + \gamma(10)$
$Q(L, IO) = 10$

$Q(L, I'O) = E[Reward|s = L, s' = S, a = I'O] + \gamma(V(S))$
$Q(L, I'O) = 1 + \gamma(10)$
$Q(L, I'O) = 10$
so $V(L) = max\{10, 10\} = 10$

$Q(S, IO') = E[Reward|s = S, s' = S, a = IO'] + \gamma(V(S))$
$Q(S, IO') = 1 + \gamma(10)$
$Q(S, IO') = 10$
so at nth iteration this function convergences to 10 because of G.p., V(L) $= -\dfrac{\gamma^{n+1} - 1}{\gamma - 1} = 10.$
and by same procedure $V(S) = 10$

Policy Improvement : V(L) = 10, V(S) = 10.

$Q(L, IO) = E[Reward|s = L, s' = S, a = IO] + \gamma(V(S))$
$= 1 + 0.9(10) = 10$

$Q(L, I'O') = E[Reward|s = L, s' = S, a = I'O'] + \gamma(V(L))$
$= -1 + 0.9(10) = 8$

$Q(L, IO') = E[Reward|s = L, s' = S, a = IO'] + \gamma(V(L))$
$= -1 + 0.9(10) = 8$

$Q(L, I'O) = E[Reward|s = L, s' = S, a = I'O] + \gamma(V(S))$
$= 1 + 0.9(10) = 10$

$Q(S, IO) = -1 + 0.9(10) = 8$
$Q(S, I'O') = -1 + 0.9(10) = 8$
$Q(S, IO') = 1 + 0.9(10) = 10$
$Q(S, I'O) = -1 + 0.9(10) = 8$
so best actions for state L are IO and $I'O$, and for state S is IO' only(e.g. V(L) = 10, V(S) = 8) same action as previous round.

(c) (2 marks) Finally, what is your advice to "At Wits End"?

Solution: Seeing as policy iteration converges after the first two iterations, the policy obtained is optimal. Therefore, my final advice would be:

- Play the Organ(may or may not burn Incense) when you hear Laughter(State L).

- Burn the Incense and **do not** play the Organ when there is Silence(State S).

(d) (2 marks) Finally, what is your advice to "At Wits End"?

Solution: since policy iterations converges at second round its best optimal action. so play the organ when "At Wits End" hear the ghost laugh. and if house is silent then just burn incense to preserve silence.

9. (4 marks) Consider the task of controlling a system when the control actions are delayed. The control agent takes an action on observing the state at time t. The action is applied to the system at time $t + \tau$. The agent receives a reward at each time step.

(a) (2 marks)What is an appropriate notion of return for this task?

Solution: Reward for an action at time t can be write as $R_{t+\tau+1}$. hence, return will be...
$G_t = R_{t+\tau+1} + \gamma R_{t+\tau+1} + ... = \sum\limits_{i=0}^{\infty} \gamma^i R_{t+\tau+i+1}$,

(b) (2 marks) Give the TD(0) backup equation for estimating the value function of a given policy.

Solution: TD(0) backup equation would be..
$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+\tau+1} + \gamma V(S_{t+\tau+1} - V(S_t)]$

hereα = learning rate,
Error would be $R_{t+\tau+1} + \gamma V(S_{t+\tau+1} - V(S_t)$.