

CS7642 Project 2 – Solving LunarLander problem with Double Deep Q-Learning

Qinghui Ge

Abstract—git hash:

I. SOCCER GAME

The soccer game is a two-player zero-sum game. On a 4×2 grid, there are two players A,B and a ball. For each player, the objective is to carry the ball to his or her designated goal area and receive +100 reward. Carrying the ball to opponent's area or allow the opponent to score would receive -100 reward. There are five possible actions for each player: N, S, E, W and Stick.

There is certain ambiguity in the game rules described in Greenward's paper. We implement the rule as following: First determine which player (A or B) moves first, then execute two players' actions based on the random order. For each player, if the next position do not collide with other player's current position, he can move to the new position, otherwise, if he losses ball to the other player(if he has it). After executing two actions, check if the termination criterion is met and assign rewards. The game is implemented like an Opengym environment, with the "step"-like function defined in Alg.1. The *step* function takes a tuple of action $\mathbf{a}=(a_A, a_B)$, and returns the state index based on two players position and ball ownership, reward as a tuple $\mathbf{R}=(R_A, R_B)$, and boolean variable *done* to indicator whether the game has ended. For the soccer game, there are $8 \times 7 \times 2 = 112$ states.

II. LEARNER IMPLEMENTATION

We follow a general off-policy learning template (Alg. 2) for all training process, each learner (Q, Friend-Q, Foe-Q, CE-Q) has to implement it's own way of action selection (for ϵ -greedy to call if greedy action is needed), and the function to update Q values. The general form of updating Q values of agent i is:

$$Q_i(s, \mathbf{a}) = (1 - \alpha)Q_i(s, \mathbf{a}) + \alpha((1 - \gamma)R_i + \gamma V_i(s')) \quad (1)$$

where α is learning rate, γ is discount factor. $\mathbf{a} = (a_1, a_2 \dots a_N)$ is a combination of all agent actions (although we will see in Q-Learning each agent only need to track the Q-table of its own action, while for other learners each agent also keep track of other agent's actions). Each learner defines state value function $V_i(s)$ differently.

Our implementation of Foe-Q is restricted to only two agents, while Q-Learning, Friend-Q, and CE-Q are generalized to handle multi-agents, with each agent potentially having different action choices. The joint action space of all agent is $\mathbf{A} = \{A_1, A_2 \dots A_N\}$.

Algorithm 1 soccer game environment implementation

```

function ENV.STEP( $\mathbf{a}=(a_A, a_B)$ )
    determine player move order
    get first player proposed move
    if no collision then:
        first player move
    else if first player has ball then
        pass ball to second player
    end if
    get second player proposed move
    if no collision then:
        second player move
    else if second player has ball then
        pass ball to first player
    end if
    if ball in player A's goal area then: ▷check game status
         $\mathbf{R}$ , done = (100,-100), True
    else if ball in player B's goal area then:
         $\mathbf{R}$ , done = (-100,100), True
    else
         $\mathbf{R}$ , done = (0,0), False
    end if
    Return  $s(pos_A, pos_B, ball)$ ,  $\mathbf{R}$ , done
end function

```

In Greenward's 2003 paper, it is unclear how learning rate is decayed, and whether Friend/Foe-Q, CE-Q used ϵ -greedy for action exploration. In Greenward's 2005 extented paper, the authors decay learning rate according to the number of times each state-action pair is visited, i.e. $\alpha(s, \mathbf{a}) = 1/n(s, \mathbf{a})$. The 2005 paper also employs $\epsilon = 1$ (random action selection) for Friend/Foe-Q, CE-Q. We implemented this procedure described in Greenward's 2005 paper, as well as other possibilities such as decaying ϵ and α exponentially.

A. Q-Learning

In Q-Learning implementation, each agent i maintains its own Q-table of size $nS \times nA_i$. The greedy action is each agent taking it's own greedy action, i.e:

$$\mathbf{a}^* = (a_1^*, a_2^* \dots a_N^*)$$

$$a_i^* = \operatorname{argmax}_{a_i} Q_i(s, a_i), a_i \in A_i$$

The value function used in Eq.1 is:

$$V(s) = \max_{a_i} Q_i(s, a_i) = Q_i(s, a_i^*) \quad (2)$$

Algorithm 2 off_policy_learning template

```

function OFF_POLICY_LEARNING(learner, env)
  for each episode do
    reset environment, get current state  $s$ 
    for  $t$  in episode do
       $\mathbf{a} = \epsilon\text{-greedy}(\text{learner}, \text{env}, s, \epsilon)$ 
      take action  $\mathbf{a}$ , observe  $s'$ ,  $\mathbf{R}$ , done
      learner.update( $s, s', \mathbf{a}, \mathbf{R}$  done,  $\alpha(s, \mathbf{a}), \gamma$ )
      decay  $\alpha$ , e.g.  $\alpha(s, \mathbf{a}) = 1/n(s, \mathbf{a})$ 
      optional: decay  $\epsilon$ , e.g.  $\epsilon *= \epsilon_{\text{decay}}$ 
       $s = s'$ 
    end for
  end for
end function

```

B. Friend-Q

In Friend-Q, each agent maintain a copy of Q-table of size $nS \times nA_1 \times \dots \times nA_N$. Each agent assumes other agents will cooperate to maximize the Q values, and the greedy action is defined as:

$$\mathbf{a}^* = (a_1^*, a_2^* \dots a_N^*)$$

$$a_i^* = \arg\max_{\mathbf{a}} Q_i(s, \mathbf{a})[i], \mathbf{a} \in \mathbf{A}$$

The value function is defined as:

$$V(s) = \max_{\mathbf{a}} Q_i(s, \mathbf{a}) = Q_i(s, \mathbf{a}^*) \quad (3)$$

C. Foe-Q

In Foe-Q learning, each agent maintain a of Q-table of size $nS \times nA_{\text{opponent}} \times nA_{\text{self}}$. The agents can follow a mixed policy: each agent i takes action a_i with probability $\sigma_i(a_i)$. The greedy policy is obtained from solving the minimax problem of σ :

$$\sigma_i^*(a) = \max_{\sigma_i} \min_{a_o \in A_{\text{opponent}}} \sum_{a_i \in A_{\text{self}}} Q_i(a_o, a_i) \sigma_i(a_i) \quad (4)$$

and the greedy action can be sampled from the minimax probability: $a_i^* \sim \sigma_i^*(a)$.

Let the value function $V_i = \min_{a_o} \sum_{a_i} Q_i(a_o, a_i) \sigma_i(a_i)$, the minimax equation can be formulated in the following linear programming problem:

$$\begin{aligned} & \text{maximize } V_i \\ & 1. nA_o \text{ inequality constraint:} \\ & \quad V_i - \sum_{a_i \in A_i} Q_i(a_o, a_i) \sigma_i(a_i) \leq 0, \forall a_o \in A_{\text{opponent}} \\ & 2. nA_i \text{ inequality constraint: } \sigma_i(a_i) \geq 0, \forall a_i \in A_i \\ & 3. 1 \text{ equality constraint: } \sum_{a_i \in A_i} \sigma_i(a_i) = 1 \end{aligned} \quad (5)$$

The second inequality constraint and equality constraint are simply probability restrictions. Foe-Q follows Alg.3 to solve the minimax problem and update Q values. Note we first update $Q(s)$ with $V(s')$ then solve for $V(s)$ with the updated

$Q(s)$, alternatively we can solve $V(s')$ from $Q(s')$ first, then update $Q(s)$. I believe the order shouldn't matter – more like a chicken or egg problem. But with the first way we can initialize $V = 0$ and no need to treat terminal state specially, the second way we don't need to save V in memory.

Algorithm 3 foe-Q update

```

function FOE-Q.UPDATE( $s, s', \mathbf{a}, \mathbf{R}$ )
   $Q_1(s, a_2, a_1) = (1 - \alpha)Q_1(s, a_2, a_1) + \alpha((1 - \gamma)R_1 + \gamma V_1(s'))$ 
   $Q_2(s, a_1, a_2) = (1 - \alpha)Q_2(s, a_1, a_2) + \alpha((1 - \gamma)R_2 + \gamma V_2(s'))$ 
   $V_1(s), \sigma_1(s) = \text{minimax}(Q_1(s))$ 
   $V_2(s), \sigma_2(s) = \text{minimax}(Q_2(s))$ 
end function

```

D. Correlated-Q

In CE-Q, each agent maintain a copy of Q-table of size $nS \times nA_1 \times \dots \times nA_N$. All agents follows a mixed policy based on joint probability $\sigma(\mathbf{a}) = \sigma(a_1, a_2 \dots a_N)$. The joint probability has to satisfy rational constraint as well as regular probability constraint. For utilitarian CE-Q, the objective function is to maximize the sum of the agents' rewards, and the uCE-Q can be formulated in the following linear programming problem:

$$\begin{aligned} & \text{maximize } \sum_i^{n_{\text{agent}}} \sum_{\mathbf{a} \in \mathbf{A}} Q_i(\mathbf{a}) \sigma(\mathbf{a}) \\ & 1. \text{ Each agent contribute } nA_i(nA_i - 1) \text{ rational constraint:} \\ & \quad \sum_{\mathbf{a}_o} Q_i(\mathbf{a}_o, a'_i) \sigma(\mathbf{a}_o, a_i) \leq \sum_{\mathbf{a}_o} Q_i(\mathbf{a}_o, a_i) \sigma(\mathbf{a}_o, a_i) \\ & \quad \forall a_i \neq a'_i \in A_i \\ & \quad \mathbf{a}_o \in \{A_1, A_2 \dots A_{i-1}, A_{i+1} \dots A_N\} \\ & 2. \Pi_i nA_i \text{ inequality constraint: } \sigma(\mathbf{a}) \leq 0, \forall \mathbf{a} \in \mathbf{A} \\ & 3. 1 \text{ equality constraint: } \sum_{\mathbf{a} \in \mathbf{A}} \sigma(\mathbf{a}) = 1 \end{aligned} \quad (6)$$

The rational probability can be interpret as following: assuming there is a referee recommend all agent to act according to joint probability distribution $\sigma(\mathbf{a})$. Assuming all other agent follows the referee's advise, for agent i , the expected reward for playing the recommended action a_i is $\sum_{\mathbf{a}_o} Q_i(\mathbf{a}_o, a_i) \sigma(\mathbf{a}_o, a_i)$, while for playing another action a'_i is $\sum_{\mathbf{a}_o} Q_i(\mathbf{a}_o, a'_i) \sigma(\mathbf{a}_o, a_i)$. The rational constraint simply says, at correlated equilibrium, an agent receives maximum reward if he follows referee's recommendation, given other agents also follow the referee.

After solving the above LP and obtain the joint probability, the state value function can be calculated as: $V_i(s) = \sum_{\mathbf{a}} Q_i(s, \mathbf{a}) \sigma(s, \mathbf{a})$. The uCE-Q updating procedure is summarized in Alg.4.

III. EXPERIMENT RESULTS

While training all learners, we keep track of the Q value of state s (as show in Fig.), with player A taking action S and

Algorithm 4 Correlated-Q update

```
function CE-Q.UPDATE( $s, s', \mathbf{a}, \mathbf{R}$ )  
  for  $i = 1, 2, \dots, n_{agent}$  do  
     $Q_i(s, \mathbf{a}) = (1 - \alpha)Q_i(s, \mathbf{a}) + \alpha((1 - \gamma)R_i + \gamma V_i(s'))$   
  end for  
   $\sigma(s, \mathbf{a}) = uCE(\{Q_1(s), Q_2(s) \dots Q_N(s)\})$   
  for  $i = 1, 2, \dots, n_{agent}$  do  
     $V_i(s) = \sum_{\mathbf{a}} Q_i(s, \mathbf{a})\sigma(s, \mathbf{a})$   
  end for  
end function
```

player B sticking to reproduce figure.3 in Greenward 2003 paper. Since in state s , taking action N would bump into the wall and result in stick effectively, we will add up the Q value of player B take N and stick together. It is unclear to me whether the iteration in x-axis means each time step or each episode in Greenward's paper, but given the large number I would assume it is every time step – although this doesn't make material difference that affect our conclusion.

From Fig.2(b), we can see the Q values converges quickly for friend-Q. However, the resulting actions are not rational. From the Q-table of state s , we can see for player B, action E dominate other actions, so player B would move right and expect player A to move aside or score for him. The best Q-value for player B is both A and B take action E , so there is 50% chance B can score +100 in the current round. While the Q-table for player A has $Q_A[:, W] = 10$, this means A can take any action and expect B to move W and score for him. Note the maximum possible Q-value ± 10 results from the $(1 - \gamma)R_i$ term when we have $\gamma = 0.9$.

Foe-Q and uCE-Q also converges. If the initialization are the same and apply Greenward 2005 paper settings (decay α according to $n(s, a)$, random action exploration), the learning curves are nearly identical for both learners. The final Q-value difference are with in 10^{-7} and both learner gives a mixed policy with both player randomly choose between stick and heading south: $\sigma_A(stick/N) = 0.423, \sigma_A(S) = 0.577, \sigma_B(stick/N) = 0.606, \sigma_B(S) = 0.394$. Note for uCE-Q each players policy is marginal probability computed from joint probability $\sigma_i(a_i) = \sum_{\mathbf{a}_o} \sigma(\mathbf{a}_o, a_i)$.

REFERENCES

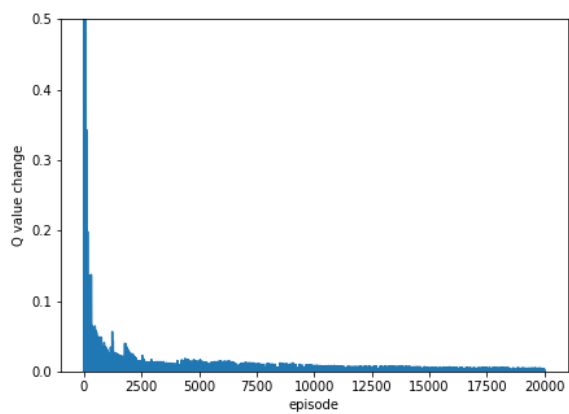
player B	stick	W	E	N	S
player A					
stick	9.0	10.0	8.551	9.000	8.999
W	9.0	10.0	8.776	9.000	8.992
E	9.0	10.0	-0.935	9.000	9.000
N	9.0	10.0	8.553	9.000	9.000
S	9.0	10.0	8.085	8.993	9.000

(a) player A

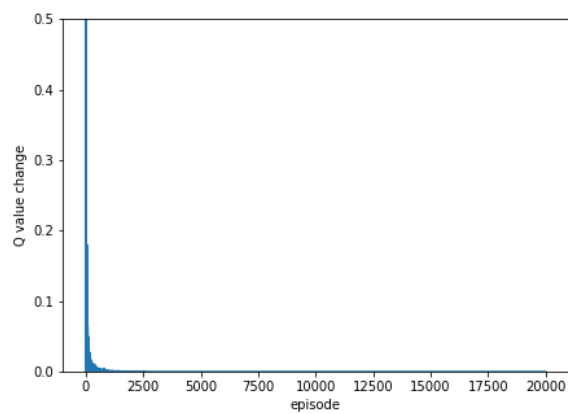
player B	stick	W	E	N	S
player A					
stick	8.527	-10.0	9.000	8.525	8.087
W	8.517	-10.0	8.762	8.524	8.088
E	8.094	-10.0	9.495	8.096	8.100
N	8.526	-10.0	8.996	8.523	8.100
S	8.099	-10.0	9.000	8.092	8.487

(b) player B

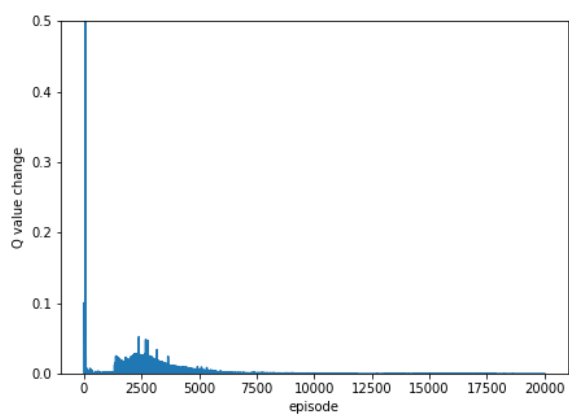
Fig. 1. Q table of state s of trained friend-Q learner



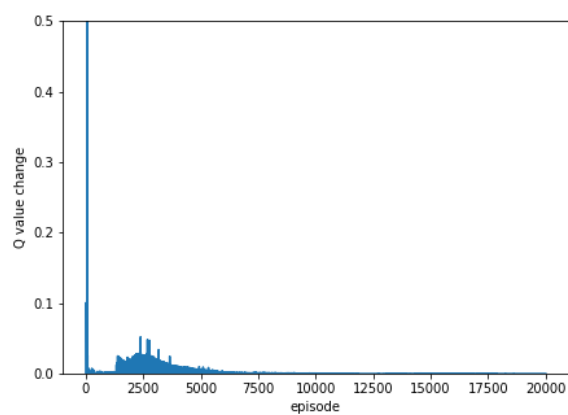
(a) Q-Learning



(b) Friend-Q



(a) Foe-Q



(b) uCE-Q

Fig. 2.