21CS91R14_HW2

November 5, 2021

0.1 # RL HW2

0.2 *Name*: Sunandan Adhikary

0.3 Roll no: 21CS91R14

1. ϵ-greedy policy is implemented in the code. ϵ-greedy policy is used to select the next action.

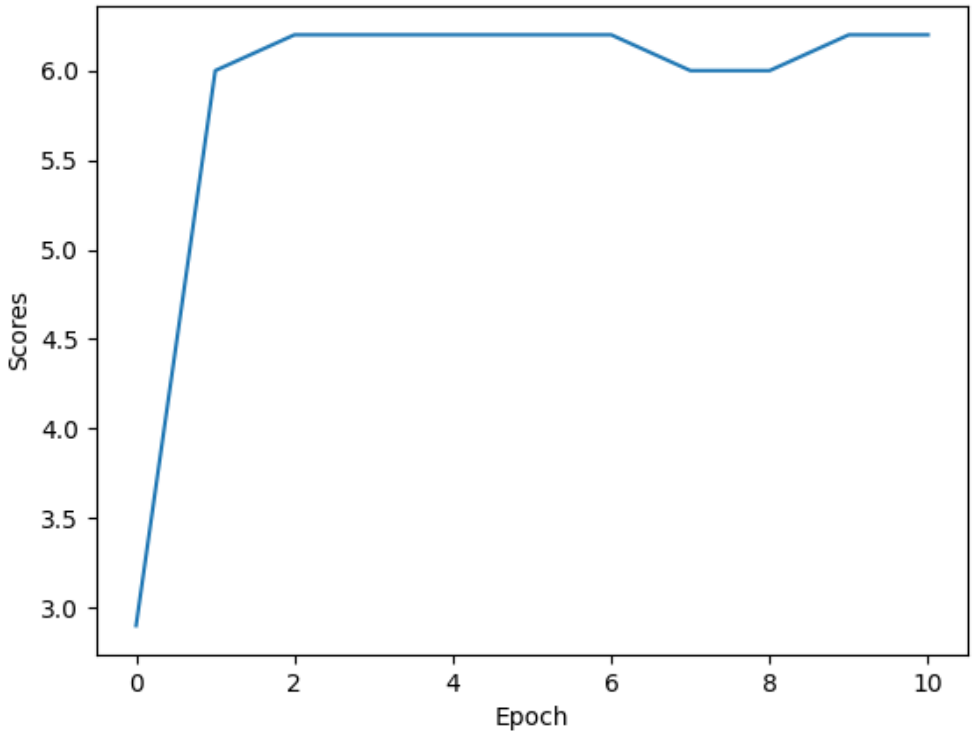
```
> Test1: ok
> Test2: ok
> Test3: ok
```

2. **a)** Completed `initialize_models()`, `get_q_values()`, `update_target()`, `calc_loss()`, `add_optimizer()` functions in

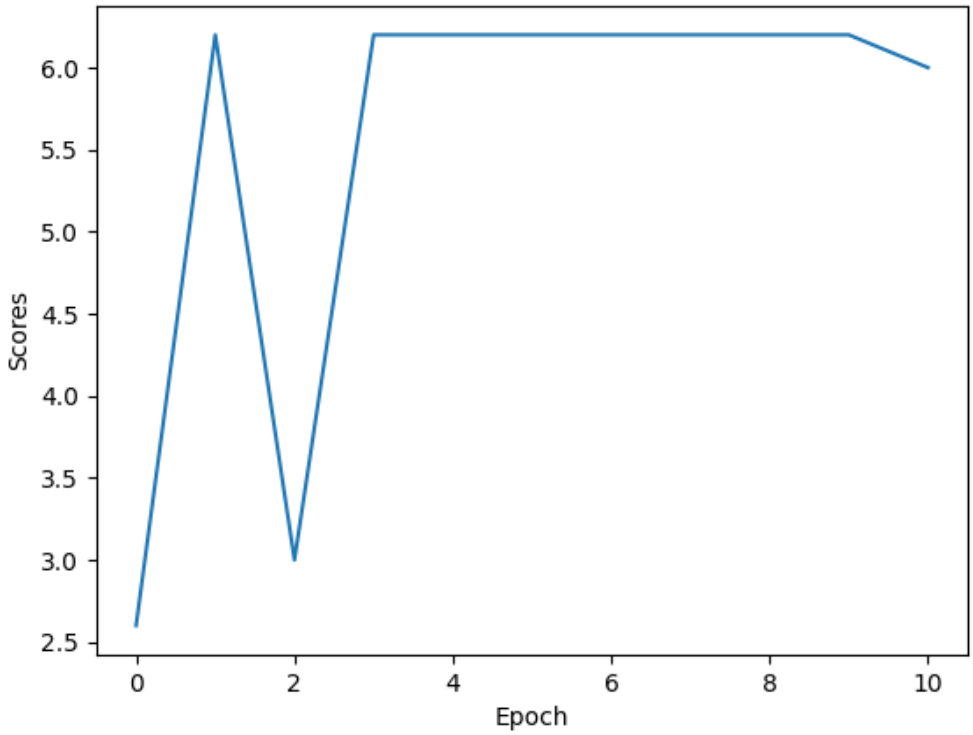
`q2_1_linear_torch.py`. After training the agent was able to achieve the maximum *~6.2 average reward* by training with the provided

parameters within *5.533073902130127 ~ 6.31389594078064 s*.

- sample training score vs epoch plot 1



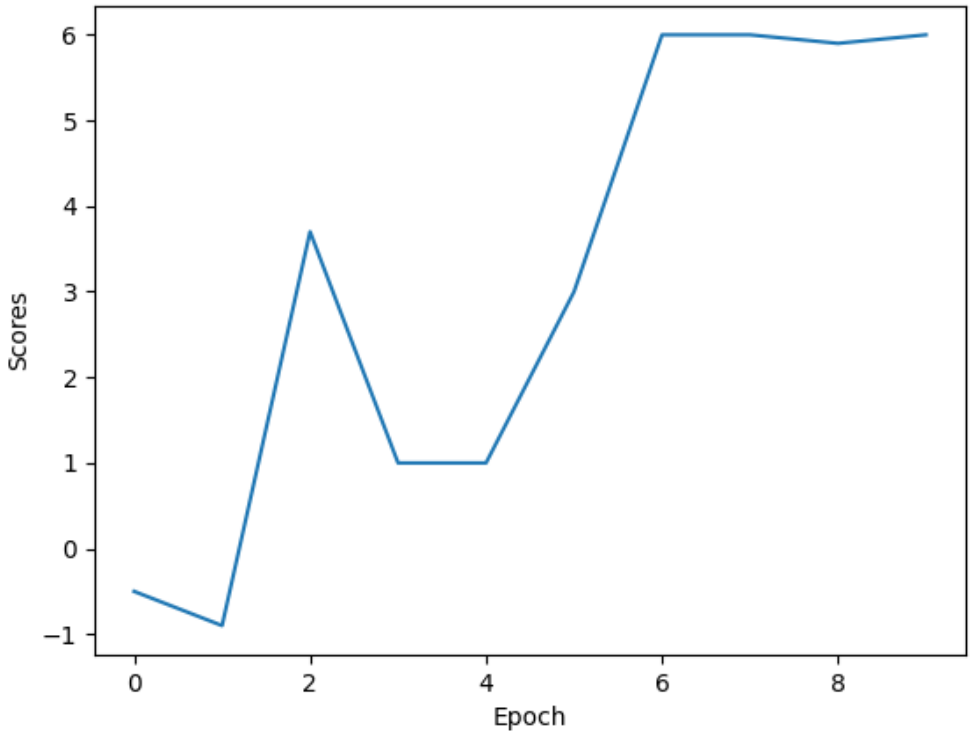
- sample training score vs epoch plot 2



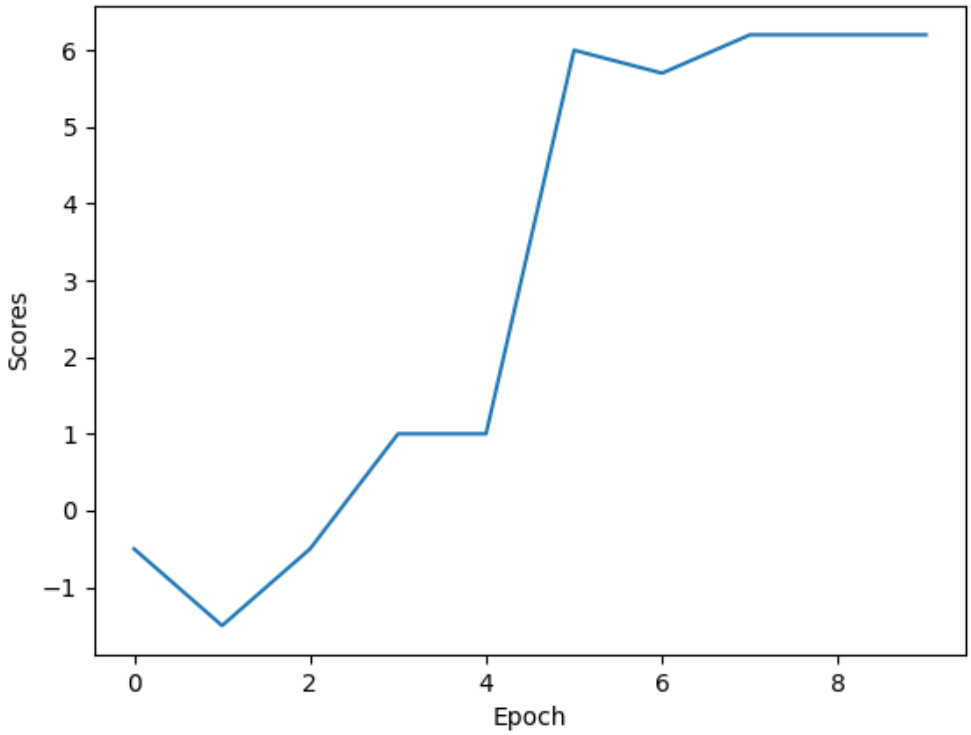
2. **b)** Completed `initialize_models()`, `get_q_values()` functions in `q2_2_nature_torch.py`. After training the agent was able to

achieve the maximum *~6.2 average reward* by training with the provided parameters within *9.37676191329956 ~ 9.57876191329956s s.*

- sample training score vs epoch plot 1



- sample training score vs epoch plot 2



3. Note that some extra lines of codes were added at the end of the three run files to measure the time. (tagged with `# new`)

4. **To Infer: DQN implemented following the NIPS paper trains faster and reaches the maximum average reward .**