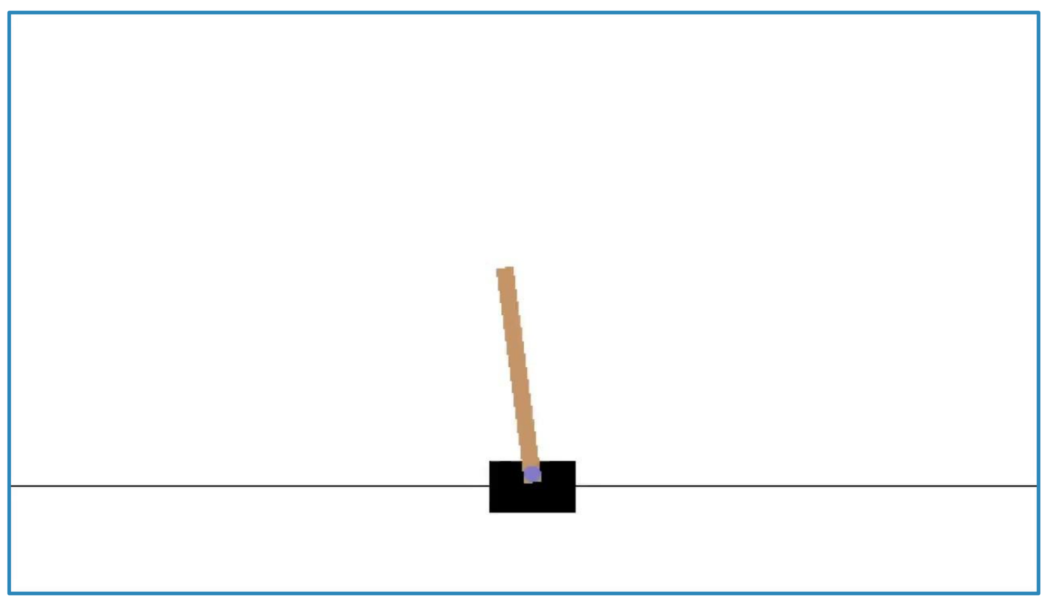
Action-Selection Strategies for Exploration of a Cart Pole

Mostafa Hussein, Eduardo Alvarado

Deep Learning Lab WS 17/18 - Albert-Ludwigs-Universität Freiburg

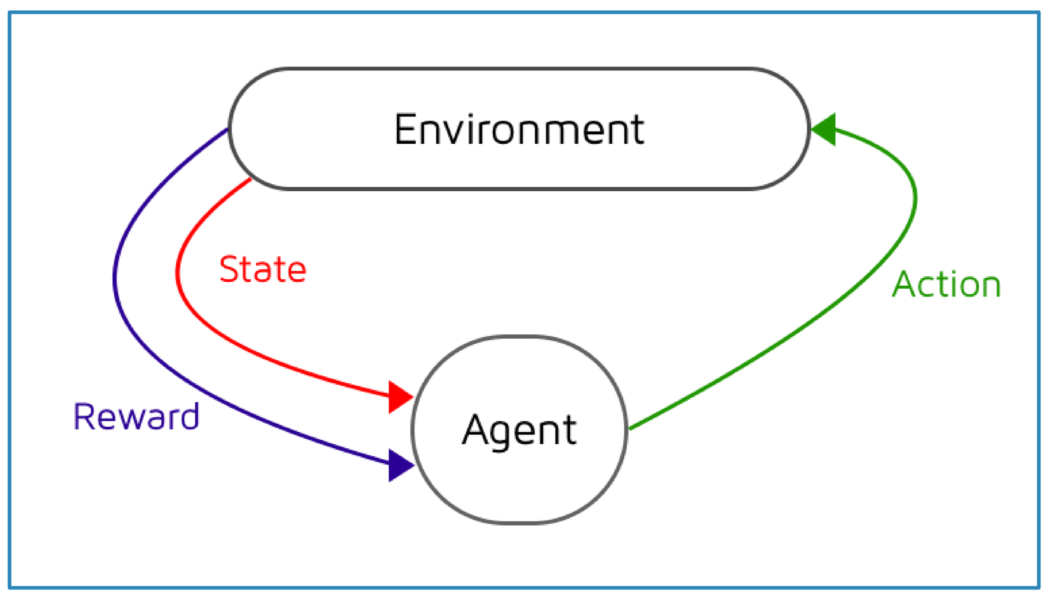
Abstract

- A pole is attached by an un-actuated joint to a cart, which moves along a frictionless track.
- The pendulum starts upright, and the objective is to prevent it from falling over.
- A reward of +1 is provided for every timestep that the pole remains upright.



Project Goals

- To find out the best action selection strategy in order to keep the pendulum upright.
- Experiment different RL Exploration Approaches.
- Compare the metrics of each one (average total reward).



RL and Exploration Principles

- We implement a DQN which outputs the predicted Q-values from given current states.
- To learn an optimal strategy, we need to expose the agent to as many states as possible.
- An agent needs to make the right decision to choose the action which can lead him to the terminal state, with the highest sum of total reward.
- A balanced ratio of exploration/exploitation can significantly affect the total learning time and the quality of learned policies.

Approaches

- **Random Approach**
- **Greedy Approach**
- **Ɛ-Greedy Approach**
- **Decaying Ɛ-Greedy Approach**
- **Boltzmann Approach**
- **UCB1 Approach:**

- An optimistic guess is constructed as to how good the expected payoff of each action is.
- The agent chooses the action with the highest guess, and if it is right, it keeps exploiting it, by incurring little regret.
- If the guess is wrong, the optimistic guess decreases and it switches to another action.
- Balance between exploration/exploitation.

$$a_t = argmax_{a \in A}\left(Q(a) + \sqrt{\frac{2\log t}{N_t(a)}}\right)$$

$$where\ t = n^{\circ}\ of\ steps, N_t(a) = execution\ frequency\ of\ action$$

Conclusion

- UCB1 Approach is acting optimally by giving highest average sum reward after small number of episodes, then it starts converging.
- Decaying Ɛ-Greedy reaches a better performance than normal Ɛ-Greedy for more episodes, due to the combination of changing exploration and exploitation.
- Random and Boltzmann Approaches did not outperform for this problem.