Reinforcement Learning Project Proposal Avinash Kori [ED15B006]
CS6700 Harikrishnan R [ED15B021]

Problem Statement

Explore various reinforcement learning (RL) and inverse reinforcement learning (IRL) algorithms on cartpole and inverted pendulum environment, And come up with the algorithm for reward constrained IRL.

Environment Details

For analysing the algorithms we plan to use 'CartPole-v0' [1] environment provided by Openai Gym. The pendulum starts upright, and the goal is to prevent it from falling over. A reward of +1 is provided for every timestep that the pole remains upright. The environment consists of four states, which corresponds to position(x), velocity ($\dot{x}$), angular displacement(θ) and angular velocity($\dot{\theta}$), and two actions (move 'left' or 'right').The episode ends when the pole is more than 15 degrees from vertical, or the cart moves more than 2.4 units from the center. We also plan to test the effect of our algorithms on 'InvertedPendulum-v2'

Investigation/Research

- We are planning to conduct comparative study between value iteration algorithm and various other policy gradient methods (like Monte-Carlo Policy gradient [2] and Proximal Policy Optimization [3])
- We also plan to implement IRL[4] algorithms to learn optimal reward for a given policy, along with the study on effects of constrained reward update for IRL

Details about Algorithm

- **Value Iteration**: In our case we are discretising the continuous states involved in cart-pole / inverted pendulum environment and apply Bellman Operator till convergence
- **Policy Gradient Methods (PG)** In all our experiments with PG we are planning to make use of neural networks for functional approximation.
  - **Monte-Carlo Policy Gradient**: Policy gradient methods are specifically designed for continuous states problems, in case of policy gradient algorithms policies are directly updated without actually calculating return value of a given state.
  - **Proximal Policy Optimization (PPO)**: PPO is an advanced form of policy gradient method in which with each update, policy is constrained in KL divergence sense with it's previous policy.

- **IRL**: IRL is entirely different paradime in reinforcement learning, where optimal reward function is learnt given optimal policies for each and every state. In our experiments we want to explore the effect of constraints on reward function

Analysis

Each algorithm can be analysed based on the iterations it takes to converge to an optimal policy. Moreover by tuning hyperparameters and playing with rewards more insight to the algorithm could be obtained.

In case of IRL, we first plan to use the results obtained using value iteration to train the agent for learning rewards. As this is a fairly unexplored algorithm in cartpole environment, more study has to be done. Rewards obtained using IRL will be compared to that used for Value Iteration to understand more about the same. If time permits, we also plan to explore the effect constrains on rewards during learning phase, and analyse and compare the behaviour of reward function obtained with and without constrains.

References

[1] Barto, A.G., Sutton, R.S. and Anderson, C.W., 1983. Neuronlike adaptive elements that can solve difficult learning control problems. IEEE transactions on systems, man, and cybernetics, (5), pp.834-846.

[2] Sutton, R.S., McAllester, D.A., Singh, S.P. and Mansour, Y., 2000. Policy gradient methods for reinforcement learning with function approximation. In Advances in neural information processing systems (pp. 1057-1063).

[3] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O., 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

[4] Ng, A.Y. and Russell, S.J., 2000, June. Algorithms for inverse reinforcement learning. In Icml (pp. 663-670).