CS6700: Reinforcement Learning Project Report

By

AVINASH KORI [ED15B006]
HARIKRISHNAN R [ED15B021]

Study and analysis of various reinforcement and inverse reinforcement learning algorithms on opensource Gym environement

NOVEMBER 2018

TABLE OF CONTENTS

Page

1 Introduction **1**
- 1.1 Problem Statement . . . 1
- 1.2 Environment Details . . . 2
- 1.3 Code avalibilty and Structure . . . 3

2 Expirements with Reinforcement Learning Algorithms **5**
- 2.1 Value Iteration . . . 5
  - 2.1.1 Theory . . . 5
  - 2.1.2 Environment Modification . . . 6
  - 2.1.3 Results . . . 6
  - 2.1.4 Observation . . . 7
  - 2.1.5 Effect of Exploration . . . 7
  - 2.1.6 Effect of Number of Bins . . . 7
- 2.2 Linear Approximation of Value Iteration . . . 8
- 2.3 Monte-Carlo Policy Gradient . . . 9
- 2.4 Proximal Policy Optimization . . . 10

3 Expirements with Inverse Reinforcement Learning Algorithms **11**
- 3.1 Section . . . 11
- 3.2 Subsection . . . 11
- 3.3 Subsubsection . . . 11

4 Comparision of Various RL and IRL Algorithms **13**
- 4.1 Section . . . 13
- 4.2 Subsection . . . 13

Bibliography **15**

CHAPTER 1

INTRODUCTION

Begins a chapter. Example: When the beloved cellist (Christopher Walken - outstanding) of a world-renowned string quartet receives a life-changing diagnosis, the group's future suddenly hangs in the balance: suppressed emotions, competing egos and uncontrollable passions threaten to derail years of friendship and collaboration. Featuring a brilliant ensemble cast (including Philip Seymour Hoffman,

1.1 Problem Statement

Explore various reinforcement learning (RL) and inverse reinforcement learning (IRL) algorithms on cartpole and inverted pendulum environment. We also propose constrained optimization algorithm in case of linear programming IRL problem. Algorithms explored in case of RL:

- Value Iteration
- Linear Approximation for Value Iteration
- Monte-Carlo Policy Gradient
- Proximal Policy Optimization

Algorithms explored in case of IRL:

- Linear Programming
- TBD

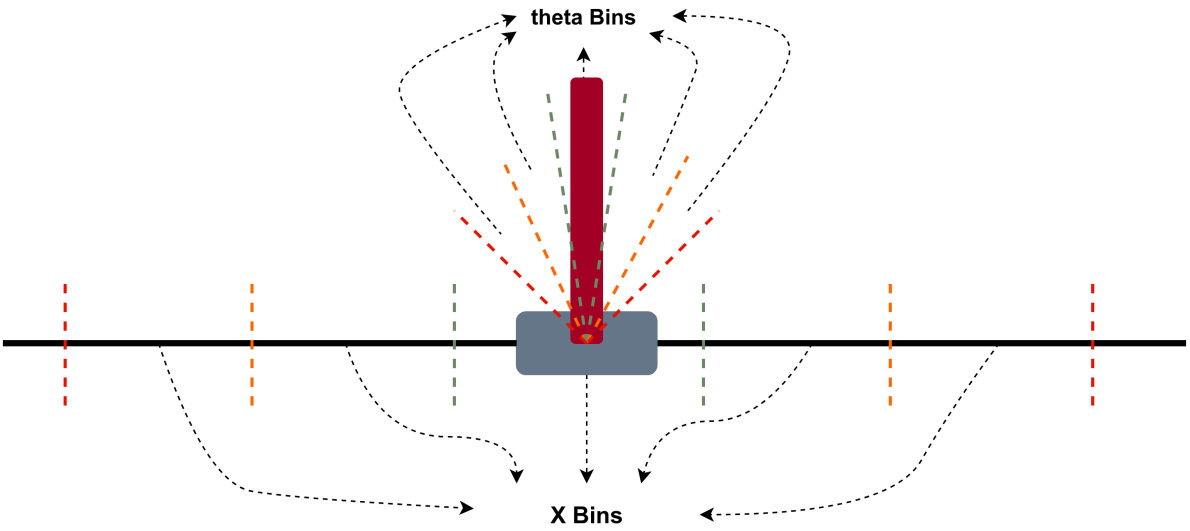


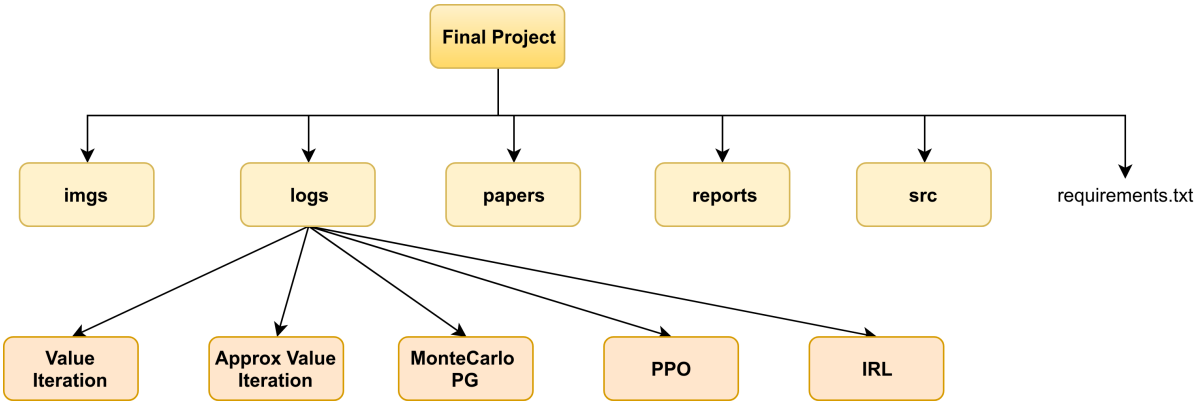
Figure 1.1: Caption

1.2 Environment Details

For all our experiments in this report we make use of publically available open-source environments, 'CartPole-v0' and 'InvertedPendulum-v2' [1]. Figure 1.1 shows CartPole environment, with few location and angle markers (denoted by dashed lines, These are used in case on discretising the setup, i.e converting infinite stages to finite stages).

- **States**: Continuous States, (x, $\dot{x}$, θ, $\dot{\theta}$)
  - x: Location of the cart
  - $\dot{x}$: Velocity of the cart
  - θ: Angle made by pole with the axis perpendicular to the cart
  - $\dot{\theta}$: Angular velocity of the pole about the axis perpendicular to the plain containing pole and cart
- **Reward**: +1 for every iteration, simulation (number of iteration) extends till pole crosses certain threshold angle (critical angle θ_c)
- **Actions**: Discrete actions, 0, 1
  - **0**: Apply force in left direction
  - **1**: Apply force in right direction

The problem in continuous state, discrete action system, Infinite horizon problem. Which can be treated as discounted schotastic shortest path (SSP[3]) problem.



1.3 Code avalibilty and Structure

This Report comes with a dedicated GitHub repository where all codes, animations and pre-trained models will be uploaded.
(https://github.com/koriavinash1/Dynamic-Programming-and-Reinforcement-Learning/tree/master/FinalProject). Figure 1.3 shows the code structure followed for this project.

CHAPTER 2

EXPIREMENTS WITH REINFORCEMENT LEARNING ALGORITHMS

In this section we would like to experiment with different Reinforcement Learning algorithms while analyzing and comparing them. For this purpose we have chosen 'CartPole-v0' environment explained in "***". In the first section we use Value Iteration where we have to discretize the given environment. While it is well known that discretization might lead to some of information, the main aim is to reduce this loss of information and provide a platform for fair comparison of algorithms. We then use Q-Learning with linear approximation, which bridges the gap between continuous and discrete sate space models. This is followed by Policy Gradient algorithm and Proximal Policy Optimization(PPO) which is an improved version of Policy graident.
"***"

2.1 Value Iteration

2.1.1 Theory

VI [2] [proposed by Bellman, 1957] is an iterative procedure that calculates the expected utility of each state using the utilities of the neighboring states until the utilities calculated on two successive steps are close enough, i.e.

$$\max_{s_i} |U(s_i) - U'(s_i)| < \epsilon$$

where ϵ is a predefined threshold value. The smaller the threshold is, the higher is the precision of the algorithm. Given certain conditions on the environment, the utility values are guaranteed to converge. Given a utility matrix we can calculate a corresponding policy according to the

maximum expected utility principle, i.e. choosing

$$\pi^*(s_i) = argmax_a \sum_j P_{ij}^a U(s_j)$$

as an optimal policy.

2.1.2 Environment Modification

For applying Value Iteration to a continuous state-space environment like CartPole, we have to first discretize it. For this purpose the observation space of CartPole was divided into 12 bins with each observation being bounded between -5 and 5. However, down the lane it came to notice that the first observation(x-position) doesn't play a major role while agent learns optimal policy. In short, the entire observation space was discretized as follows:

- x - position was not taken into account
- $\dot{x}$ - velocity was discretized to 12 bins (-5,5)
- θ - velocity was discretized to 12 bins (-5,5)
- $\dot{\theta}$ - velocity was discretized to 12 bins (-5,5)

2.1.3 Results

Below attached is a plot of rewards obtained by agent during training:

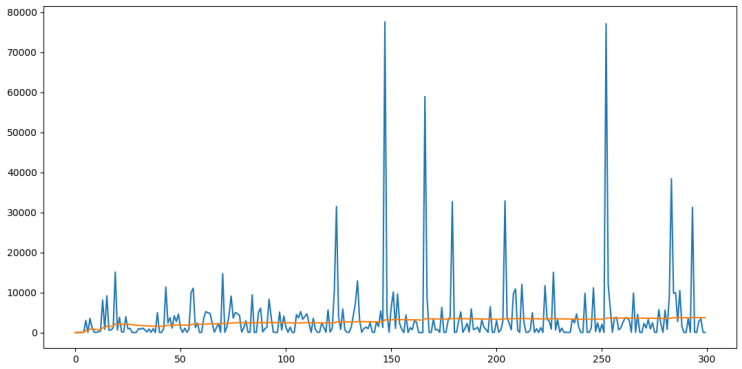


Figure 2.1: Rewards obtained during training (Orange line represents mean of rewards obtained)

It can be seen that initially there is an increase in the mean rewards obtained., which indicates that the agent is learning optimal policy during this time. However, after some points, the mean curve flattens out, indicating no improvement of rewards happen after this point. Also

the reward obtained varies largely for each episode. In this case the mean reward obtained is around 3750, which is a good number compared to the classic requirement of agent requiring to survive at least a reward of 195 when upper cap of steps is 200.

2.1.4 Observation

2.1.5 Effect of Exploration

2.1.6 Effect of Number of Bins

2.2 Linear Approximation of Value Iteration

Begins a subsection.

2.3 Monte-Carlo Policy Gradient

Begins a subsection.

2.4 Proximal Policy Optimization

Begins a subsection.

CHAPTER 3

EXPIREMENTS WITH INVERSE REINFORCEMENT LEARNING ALGORITHMS

Begins a chapter. Example: When the beloved cellist (Christopher Walken - outstanding) of a world-renowned string quartet receives a life-changing diagnosis, the group's future suddenly hangs in the balance: suppressed emotions, competing egos and uncontrollable passions threaten to derail years of friendship and collaboration. Featuring a brilliant ensemble cast (including Philip Seymour Hoffman, Catherine Keener and Mark Ivanir as the three other quartet members), it is a fascinating look into the world of working musicians, and an elegant homage to chamber music and the cultural world of New York. The music, of course, is ravishing (the score is the work of regular David Lynch collaborator Angelo Badalamenti): A Late Quartet hits all the right notes.

3.1 Section

Begins a section.

3.2 Subsection

Begins a subsection.

3.3 Subsubsection

Begins a subsection.

CHAPTER 4

COMPARISION OF VARIOUS RL AND IRL ALGORITHMS

Begins a chapter. Example: When the beloved cellist (Christopher Walken - outstanding) of a world-renowned string quartet receives a life-changing diagnosis, the group's future suddenly hangs in the balance: suppressed emotions, competing egos and uncontrollable passions threaten to derail years of friendship and collaboration. Featuring a brilliant ensemble cast (including Philip Seymour Hoffman, Catherine Keener and Mark Ivanir as the three other quartet members), it is a fascinating look into the world of working musicians, and an elegant homage to chamber music and the cultural world of New York. The music, of course, is ravishing (the score is the work of regular David Lynch collaborator Angelo Badalamenti): A Late Quartet hits all the right notes.

4.1 Section

Begins a section.

4.2 Subsection

Begins a subsection.

4.2.0.1 Subsubsection

Begins a subsection.

BIBLIOGRAPHY

[1] A. G. BARTO, R. S. SUTTON, AND C. W. ANDERSON, *Neuronlike adaptive elements that can solve difficult learning control problems*, IEEE transactions on systems, man, and cybernetics, (1983), pp. 834–846.

[2] R. BELLMAN, *A markovian decision process*, Journal of Mathematics and Mechanics, (1957), pp. 679–684.

[3] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *An analysis of stochastic shortest path problems*, Mathematics of Operations Research, 16 (1991), pp. 580–595.