# CS525 Reinforcement Learning - Project Proposal

Team Members: Aniket Patil, Prathamesh Bhamare, Rutwik Bonde
Team Number: 10
Guide: Prof. Yanhua Li

## Problem Statement

Implementation of Reinforcement Learning algorithms on the Highway-Environment

## Description

We plan to implement and showcase comparative study for the following algorithms in the racetrack-v0 environment:

- Proximal Policy Optimization (PPO): This algorithm performs comparably or better than state-of-the-art approaches while being much simpler to implement and tune. State of the art methods like Policy Gradient methods have convergence problems which are addressed by the natural policy gradient. However, in practice, natural policy gradient involves a second-order derivative matrix which makes it not scalable for large scale problems. The computational complexity is too high for real tasks. PPO uses a slightly different approach. Instead of imposing a hard constraint, it formalizes the constraint as a penalty in the objective function. By not avoiding the constraint at all costs, we can use a first-order optimizer like the Gradient Descent method to optimize the objective. Even if we may violate the constraint occasionally, the damage is far less, and the computation is much simpler.
- Deep Deterministic Policy Gradient (DDPG): In this algorithm, the actor directly maps states to actions (the output of the network directly the output) instead of outputting the probability distribution across a discrete action space. The target networks are time-delayed copies of their original networks that slowly track the learned networks. Using these target value networks greatly improve stability in learning.
- Asynchronous Advantage Actor-Critic (A3C): Actor-Critic stands for two neural networks - Actor and Critic. The goal of the actor is in optimizing the policy ("How to act?"), and the critic aims at optimizing the value ("How good action is?"). Thus, it creates a complementary situation for an agent to gain the best experience of fast learning.

## Environment Tools

We will be using the highway-env environment available in the OpenAI Gym documentation. It contains various tasks such as: "highway-v0" (obstacle avoidance and lane keeping), "parking-v0" (car parking), "intersection-v0" (intersection negotiation task with dense traffic). We would like to focus on the racetrack task and implement the proposed algorithms and compare their results.

As extended goal, once we have the results of the algorithms, we would like to validate those results in one of the other tasks in the same highway-env environment.

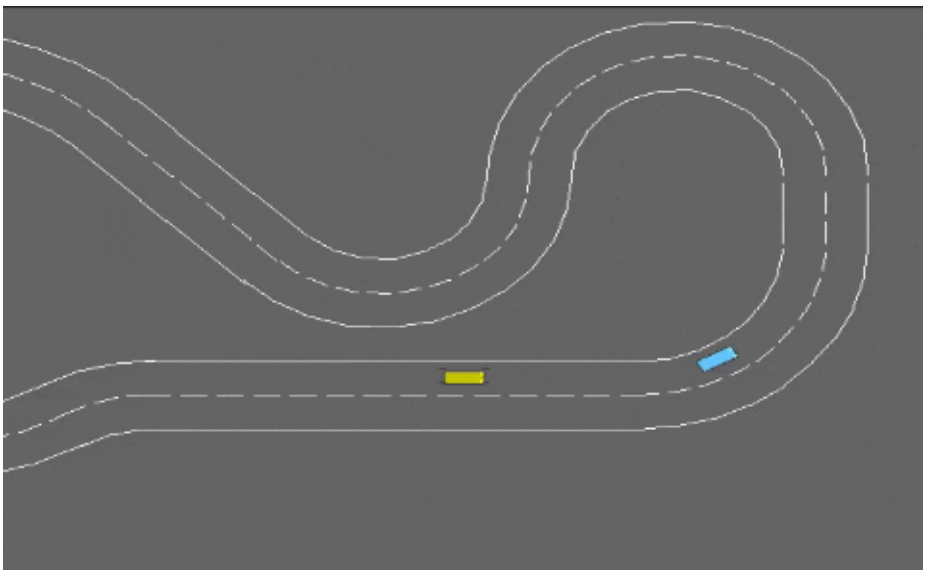

Figure 1: Car Racing Environment

## Project Schedule

| Week | Activity |
|---|---|
| Week 1 - (11/2 - 11/5) | Team formation and project topic brainstorming |
| Week 2 - (11/6 - 11/12) | Project Proposal report, Setup simulation environment |
| Week 3 - (11/13 - 11/19) | PPO, DDPG and A3C implementation, start training |
| Week 4 - (11/20 - 11/26) | Training continuation, results collection and drawing conclusions |
| Week 5 - (11/27 - 12/3) | Finalizing project report and presentation |
| Week 6 - (12/4 - 12/6) | Poster session preparation |

## References

[1] Highway-env Environment - https://github.com/eleurent/highway-env

[2] Which Reinforcement learning-RL algorithm to use where, when and in what scenario? - Ujwal Tewari (Medium)

[3] Schulman, John and Wolski, Filip and Dhariwal, Prafulla and Radford, Alec and Klimov, Oleg (2017) arXiv - Proximal Policy Optimization Algorithms

[4] Mnih, Volodymyr and Badia, Adrià Puigdomènech and Mirza, Mehdi and Graves, Alex and Lillicrap, Timothy P. and Harley, Tim and Silver, David and Kavukcuoglu, Koray (2016) arXiv - Asynchronous Methods for Deep Reinforcement Learning

[5] Lillicrap, Timothy P. and Hunt, Jonathan J. and Pritzel, Alexander and Heess, Nicolas and Erez, Tom and Tassa, Yuval and Silver, David and Wierstra, Daan (2015) arXiv - Continuous control with deep reinforcement learning

[6] RL — Proximal Policy Optimization (PPO) Explained – Jonathan Hui (Medium)

[7] Implementation of Decision-Making Algorithms in OpenAI Parking Environment – Pradeep Gopal (Medium)

[8] Deep Reinforcement learning using Proximal Policy Optimization – Surajit Saikia (Medium)