# Optimisitc initialization of parameterized value functions using a neural network

**Anonymous Authors**[1]

## Abstract

Optimistic initialization of value functions is a popular approach to exploration in tabular reinforcement learning. However, it is rarely analyzed in deep reinforcement learning. We explore this problem through a parameterized value function using linear neural networks and compare our results to an existing popular learning algorithm.

## 1. Introduction

Tradeoff between exploration and exploitation, is an eternal problem that any RL algorithms face. Since this challenge arises everywhere and impacts the overall performance of algorithm, various methods to balance the exploration/exploitation has been proposed. Among those approaches, one of the most fundamental and flexible approach is an optimistic initialization of the value functions. By simply setting the initial values greater than the reward maxima, one can enforce the agents to explore every state(-action) pair at least once at the early stage. Despite of its simplicity, the effect of optimistic initialization is significant, especially on the tabular cases. Although this simple technique is proven to provide the tremendous advantages in tabular context, its effect on deep nonlinear function approximation is yet to be discovered. This project aims to provide 1. a basic method to optimistically initialize the deep network, 2. simple fix on the implementation so that optimistic values remain effective after few gradient steps and 3. empirical analysis.

## 2. Background

As an exploration method in RL, **optimistic initialization** is often used in diverse set of tasks due to its high scalability. Depending on the task, however, the methods for giving this optimism changes slightly. One obvious method is to naively set the initial values to an arbitrary constant for all the states. This method is known to work well in the tabular RL tasks, but it is not guaranteed that it works in other situtations such as function approximation. Particularly for function approximation, it is almost unable to set the optimism naively without knowing the upper bound of return signals. To avoid this, Mochado et al (2015) proposed a method to normalize and shift down the reward signals. This technique allows us to determine an upper bound on any tasks, and helps us to naively initialize the function parameters. However, the method only assumed the linear function approximator, and not yet tested on DRL algorithms.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. **AUTHORERR: Missing \icmlcorrespondingauthor.**

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

## 3. Research Question

In this paper we strive to answer the following 3 part question, Can a parameterized value function using a neural network be optimistically intialized using reward signal shift and normalization? How will the removal of the bias term from the neural network affect optimistic initialization? In a value estimation experiment how will this initialization compare to it's tabular counterpart using SARSA and semi-gradient SARSA with the same reward signal shift?

## 4. Experimental Design

We will be considering the minigrid library to run our experiments in three stationary environments. For the purposes of this paper, we will be focusing on the stationary environments Crossingenv, Distshiftenv and Lavagapenv.

The value estimation algorithm in our tabular setting is SARSA and for the parameterized setting we will be using semi-gradient SARSA, using an $\epsilon$-greedy behavior policy.

The mingrid library API provides a discrete action space, along with encoding states in an image format for the parameterized setting and a coordinate format for the tabular setting.

We will perform SARSA using an epsilon-greedy policy to estimate the state-action value functions using the agent coordinates while initializing the state-action values optimistically. We will extend this idea to the parameterized case by adjusting the weights of the neural network so that

it outputs an optimistic value for all state-action values. We will repeat the parameterized setting experiment by removing the bias term from the neural network and re-adjusting and performing the experiments accordingly. Comparing the results should give us better insight into the process of initializing paramaterized value functions, it's effect on expected return and the siginificance of the bias term in performance.
Furthermore, the mentioned environments were selected to due having simpler environment dynamics and a smaller state-space, making our experiment computationally less expensive.

## 5. Contributions

**Alireza Azimi:**
**Haruto Tanaka:**
**Henry Du:**
**Mashfique Zaman:**

## 6. References

To be added later.