Exploratory Project

Exploring Safe and Effective Reward Patterns in Reinforcement Learning

Presented By:
Prashant Kumar Singh Yadav
(22074023)
Prachi
(22075061)
Tejus Diwakar
(22074032)

Under the guidance of
Dr. Lakshmanan Kailasam

Abstract

This project explores the design and evaluation of reward patterns in reinforcement learning (RL) to enhance learning efficiency and safety. The study compares penalized reward systems with traditional goal-based reward systems using Q-learning in RL environments. Our analysis indicates that penalized reward systems exhibit improved learning performance and safety compared to goal-based systems, particularly in scenarios with complex and sparse reward structures. The findings suggest that careful design of reward patterns can significantly impact the learning process in RL, emphasizing the importance of safe and effective reward strategies in autonomous systems.

Table of Content

Introduction

Reinforcement learning (RL) has emerged as a powerful paradigm for training intelligent agents to make sequential decisions in complex environments. A key component of RL is the reward system, which provides feedback to the agent based on its actions. The design of the reward system plays a crucial role in shaping the learning process and ultimately, the agent's behavior.

This project focuses on exploring safe and effective reward patterns in RL, specifically comparing penalized reward systems with traditional goal-based reward systems. The goal is to investigate how different reward structures impact learning efficiency and safety in RL environments. By analyzing the performance of these reward systems using Q-learning, we aim to provide insights into the design of reward patterns that can enhance the performance and safety of RL agents.

Reinforcement Learning

Reinforcement Learning (RL) is a machine learning paradigm where an agent learns to make decisions by interacting with an environment. The agent performs actions and receives feedback in the form of rewards or penalties, guiding its learning process. The goal of RL is for the agent to learn a policy, which is a mapping from states to actions, that maximizes the cumulative reward over time.

One of the key components of RL is the reward signal, which serves as the primary means of communication between the agent and the

environment. The agent's objective is to learn a policy that maximizes the expected cumulative reward. This is typically formulated as a Markov Decision Process (MDP), where the agent aims to find the optimal policy by exploring and exploiting the environment.

Q-learning is a popular RL algorithm that is used to learn the Q-values of state-action pairs. The Q-value represents the expected cumulative reward that an agent will receive starting from a given state and taking a specific action, and following a particular policy thereafter. By iteratively updating the Q-values based on the observed rewards, Q-learning enables the agent to learn an optimal policy.

In this project, we use Q-learning as the basis for our analysis of reward systems in RL. By comparing different reward structures, including penalized and goal-based reward systems, we aim to understand how different reward patterns influence the learning process and the performance of RL agents.

Objectives

- The main objective of our project is to investigate and compare different reward system approaches in Reinforcement Learning (RL).
- Specifically, we aim to:
    - Evaluate the effectiveness of penalized rewards compared to traditional goal-based rewards.
    - Analyze the concept of safe rewards when penalties are included in the reward system.
- Our goal is to provide insights into designing more efficient and robust reward systems for RL agents.

By addressing these objectives, this project contributes to the ongoing research in RL by providing empirical evidence on the impact of reward design on learning efficiency and safety. The findings of this study have implications for the design of reward systems in various applications of RL, including robotics, gaming, and autonomous systems.

Methodology

1. Experimental Setup
We conducted our experiments using the Gymnasium library (also known as Gym), a toolkit for developing and comparing reinforcement learning algorithms. The gymnasium provides a wide range of environments, each with its own reward structure.

2. Reward Wrappers
To modify the default rewards of the environments, we utilized reward wrappers provided by Gymnasium. These wrappers allowed us to customize the reward signals to suit our experimental needs. We implemented different reward structures, including penalized and goal-based reward systems, by modifying the reward wrappers accordingly.

3. Q-Learning Implementation
We implemented Q-learning, a model-free reinforcement learning algorithm, to train agents in the modified environments. Q-learning is well-suited for environments with discrete state and action spaces, making it a suitable choice for our experiments. We used the Q-learning algorithm to learn the optimal policy for each reward structure.

Objective - 1

Comparing Reward Representations

- **Goal-Reward Representation:**
  - Agent receives a reward of 0 while en route to the goal (Intermediate Reward).
  - Agent receives a reward of +1 (or higher) when the goal is reached.
  -
- **Action-Penalty Representation:**
  - Agent incurs a penalty of -1 (or less) for each step taken (Intermediate Reward).
  - The episode ends when the agent reaches the goal.

We have compared the above two reward systems by applying Q-learning on **The Frozen Lake Problem.**

Frozen Lake Problem

The Frozen Lake problem is a classic grid-world environment often used as a benchmark in reinforcement learning (RL) research.

It's a task that simulates an agent navigating through a grid of frozen tiles (representing a frozen lake) to reach a goal tile (representing safety or a reward) while avoiding holes (which represent danger or penalties).

The goal is to find an optimal policy for the agent to reach the goal tile while minimizing the chances of falling into a hole.

Results :

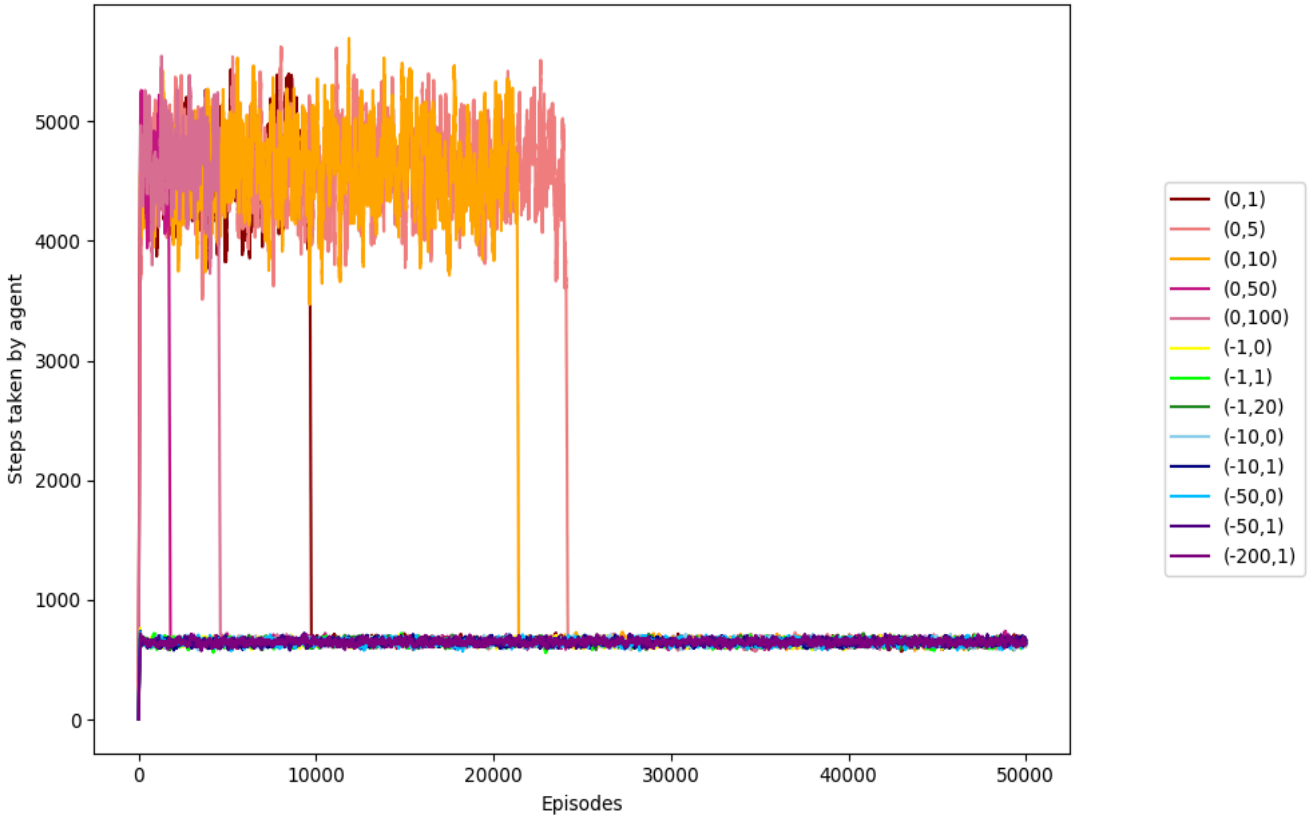


Figure: Graph of the steps taken by the agent plotted against the episodes for different reward systems (Intermediate, Final).

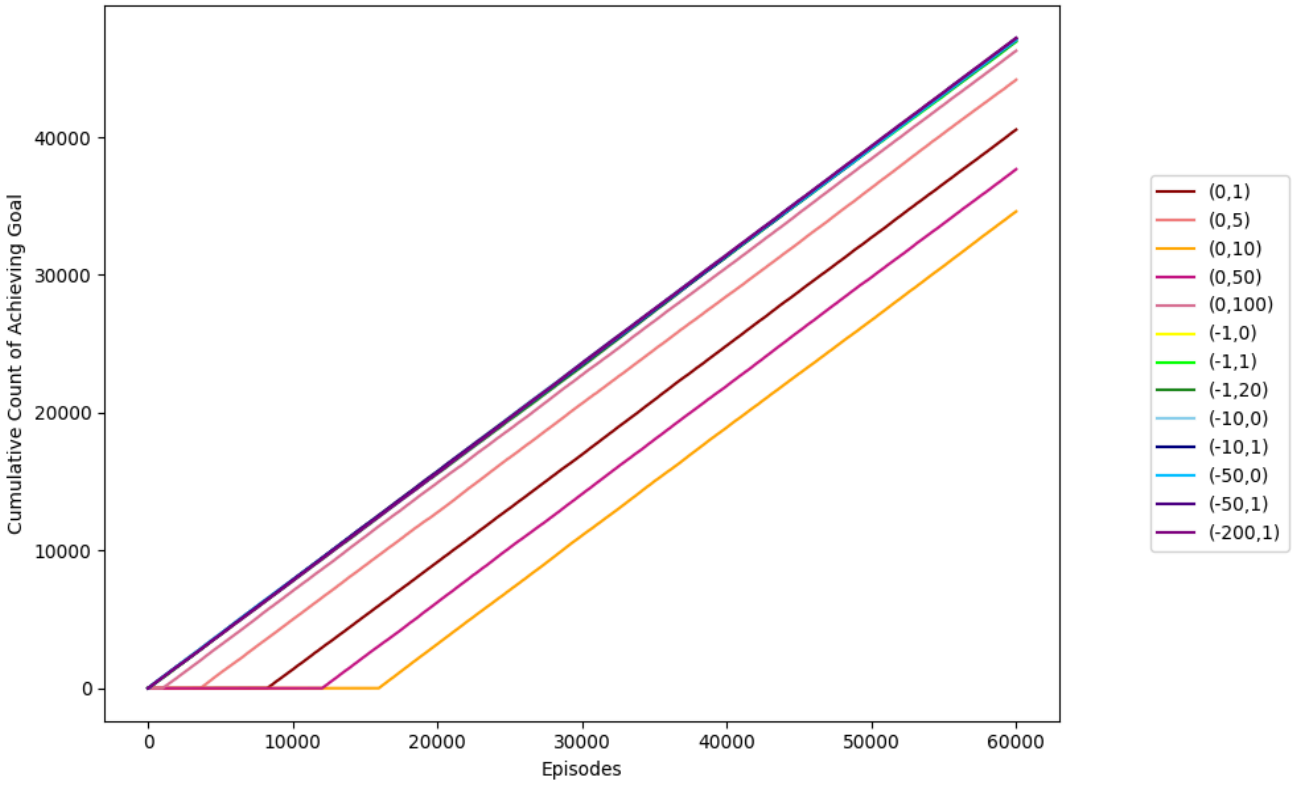


Figure: Graph of the Cumulative count of success achieved plotted against episodes for different reward systems (Intermediate, Final).

Observation :

- The step count for a goal-based reward system is higher initially for a large number of episodes that indicates it has not encountered the goal yet and still exploring.

- On the other hand, the Action-Penalty reward system has more number of success count which indicates its faster learning.

Conclusion :

- Penalizing the agent for each step taken (action-penalty representation) results in more efficient learning compared to providing a reward only when the goal is reached (goal-reward representation).

- Essentially, the zero-initialization of the Q table and the zero rewards means the agent has no guidance and explores randomly until the goal is encountered by accident. With action penalties, the agent tries to avoid actions it has already taken, resulting in much more directed exploration.

Objective-2

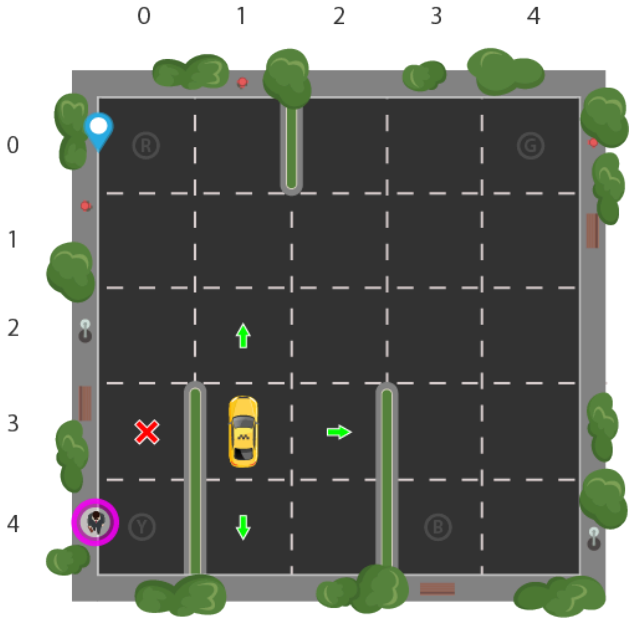
Searching for Safe Reward System

• In environments with wrong moves or illegal actions, we need to appropriately penalize these actions.

• We fixed penalties for illegal actions to discourage the agent from taking them. The challenge was to determine the appropriate reward combination (Intermediate Steps, Success, Penalty) that would deter the agent without overly discouraging exploration.

We have used Q-learning on the **Taxi Problem** to see the better combination of a safe reward system.

Taxi Problem



The Taxi Problem is a classic reinforcement learning (RL) task often used as a simple benchmark for testing and evaluating RL algorithms.

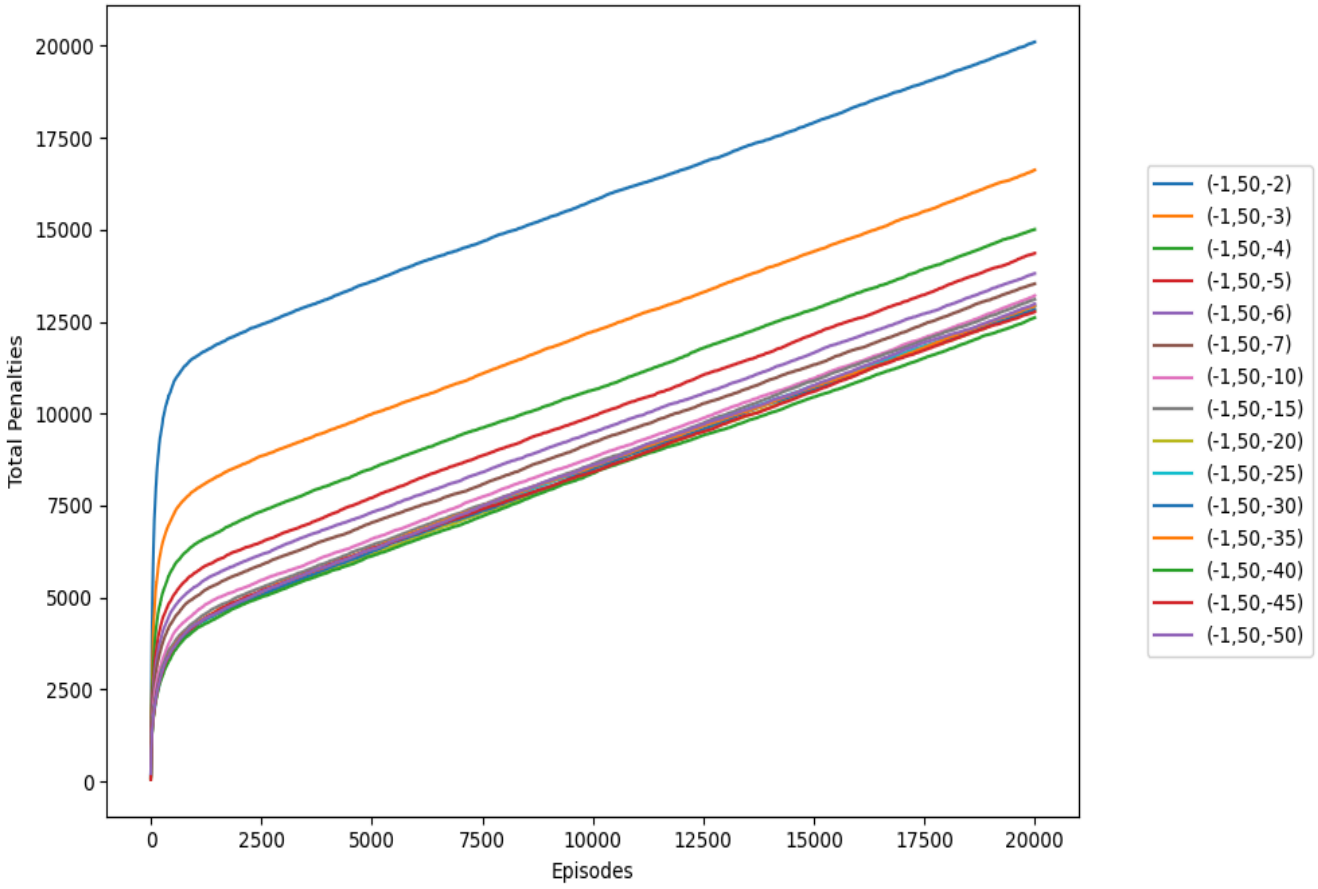
In this problem, an agent (taxi) navigates through a grid world to pick up passengers and drop them off at their destinations.

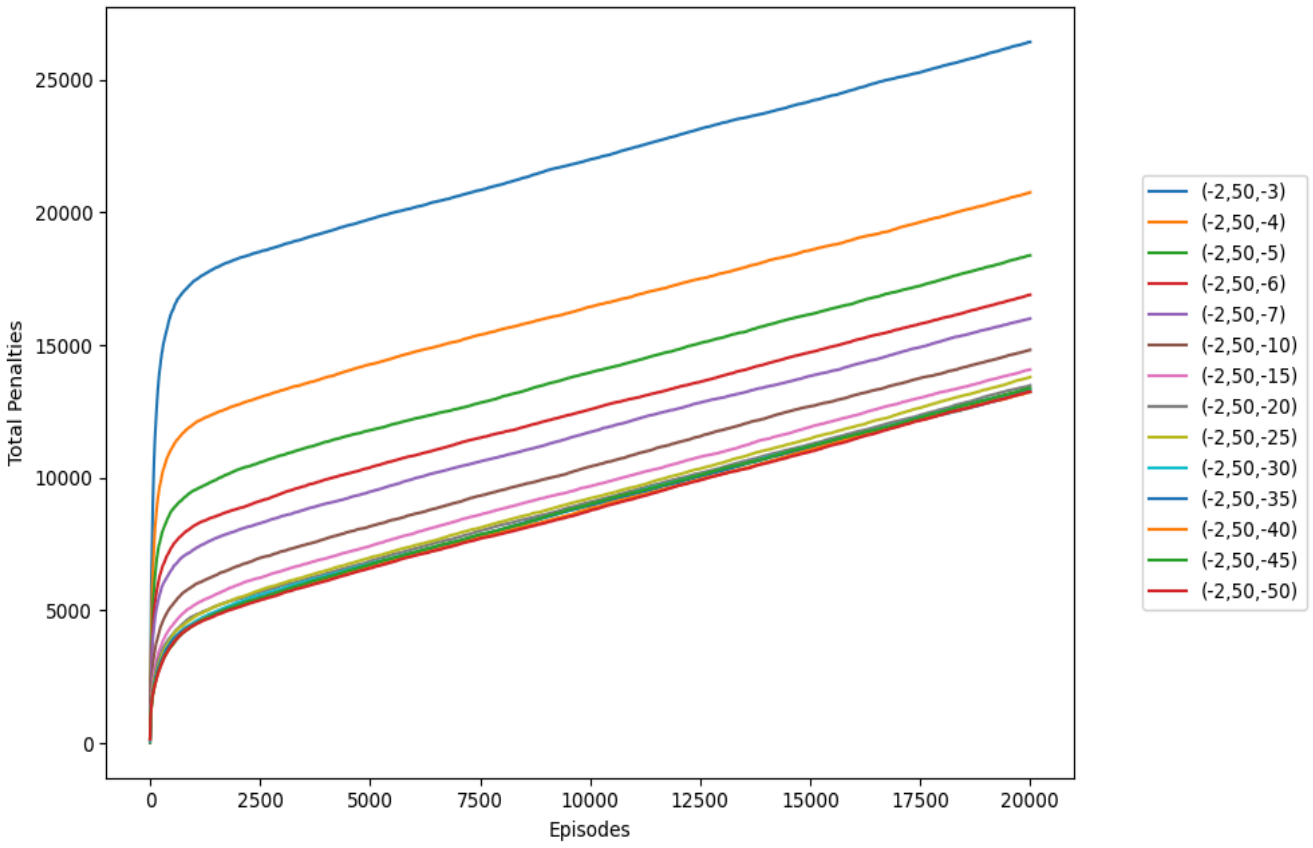
The goal of the agent is to learn an optimal policy for efficiently navigating the grid world and completing tasks while maximizing cumulative rewards.

Here, a Penalty will be given for illegal actions of incorrect attempts to pick up/drop off passengers.

Results :

Some graphs of the Cumulative Count of Illegal actions taken (Penalty Received) are plotted against episodes.





Observation :

- The Count of Penalties for Reward Systems that have an Illegal Action Penalty close to the Intermediate or on-route steps penalty is higher than those with relatively higher Illegal Action Penalty (or more negative reward).

- After a certain relative factor of them, it achieves a saturation in the count of Illegal Actions.

- Also, the Final Success step reward must be higher absolutely with respect to the enroute step rewards for successful learning.

Conclusion :

- The Illegal Action Penalty must be higher than the Intermediate Steps Penalty for safer learning.

- Additionally, there should be a relative factor between the two penalties, which in our case is 5. This factor depends on the state size of the MDP of the problem.

- This will ensure that the agent marks the illegal action with such a low reward that it never selects it again while exploiting the Q-table.

Source Code

You can find the source code of our experiments on the below GitHub repository.
https://github.com/Prashant-ksy/Enhancing-reward-system-in-RL.git

References

- Reinforcement Learning: An Introduction
  By Richard S. Sutton and Andrew G. Barto
  http://incompleteideas.net/book/the-book-2nd.html

- Designing Rewards for Fast Learning
  https://arxiv.org/pdf/2205.15400

- Gymnasium Documentation
  https://gymnasium.farama.org/index.html