Reinforcement Learning using Temporal Difference Learning and Boltzman Distribution

Ronit Kumar Kataria
Department of Computer Science
Habib University
Karachi, Pakistan
rk06451@st.habib.edu.pk

Faraz Ali
Department of Computer Science
Habib University
Karachi, Pakistan
fa06396@st.habib.edu.pk

***Abstract*—This report describes an agent trained using temporal difference (TD) learning and the Boltzmann distribution to solve a grid world task over multiple episodes. The agent learns to navigate a 2D grid world with obstacles and rewards, with the goal of reaching a target location while avoiding hazards. The TD learning algorithm enables the agent to update its value estimates based on the difference between its predicted and observed rewards, while the Boltzmann distribution provides a probabilistic approach to action selection that allows the agent to explore the environment while still exploiting its knowledge. Our experiments show that the trained agent is able to successfully navigate the grid world and achieve high reward across multiple episodes, demonstrating the effectiveness of the TD learning and Boltzmann distribution approach.**

***Index Terms*—Temporal Difference Learning, Reinforcement Learning, Boltzman Distribution, Agent, Grid World**

I. Introduction

Grid world problems are a class of reinforcement learning tasks in which an agent must navigate a 2D grid world with obstacles and rewards, with the goal of reaching a target location while avoiding hazards. These tasks are commonly used in the study of artificial intelligence and machine learning, as they provide a simple yet challenging environment for agents to learn and develop strategies for achieving goals.

One popular approach for training agents to solve grid world tasks is reinforcement learning, which involves learning from feedback in the form of rewards or penalties. Temporal difference (TD) learning is a type of reinforcement learning algorithm that enables agents to update their value estimates based on the difference between predicted and observed rewards. This approach has been shown to be effective in solving a wide range of reinforcement learning problems, including grid world tasks.

In this report, we apply TD learning to train an agent to solve a grid world task over multiple episodes. We also use the Boltzmann distribution as a probabilistic approach to action selection, which allows the agent to explore the environment while still exploiting its knowledge. Our experiments show that the trained agent is able to successfully navigate the grid world and achieve high reward across multiple episodes, demonstrating the effectiveness of the TD learning and Boltzmann distribution approach.

II. Methodology

The section discusses the implementation of Temporal Difference Learning and Boltzman Distribution to solve the 2D Grid World Problem.

A. Problem Formulation

We start by first creating an $n \times n$ grid. The grid was initialised with *states* whereby each cell of the grid was considered as a state. All states were given a value of 0 and a reward of -1 initially. Once initialised, the grid also had two parameters, namely - *redStates* and *greenStates*. These are percentage of red and green states in our grid, whereby a green state is a positive reward terminal state and red is a negative penalty terminal state. Once the grid had been initialised with the terminal states added, we finally assigned a random starting point for our agent. This was done to ensure that the agent is capable of learning given any configuration. The *agent* is then initialised by assigning a starting position and the alpha and gamma values that are to be used to perform Temporal Difference Learning. The explanation of Alpha and Gamma values is further explained in the next section.

B. TD and Boltzman Algorithm

The Boltzmann distribution is a probability distribution that assigns probabilities to a set of possible states based on their energy levels, with lower energy states being more probable than higher energy states. In reinforcement learning, the Boltzmann distribution can be used as a probabilistic approach to action selection, enabling agents to explore the environment while still exploiting their current knowledge. This is given by:

$$P(a|s) = \frac{e^{\frac{Q(s,a)}{k}}}{\sum_j e^{\frac{Q(s,a_j)}{k}}}$$

We used this to decide what action to perform at each state. The $Q(s, a)$ is the *q-value* or the value function for each state based on the state and the action taken. Theoretically, for our case, we could have at most four *q-values* for a single state based on four possible actions - up, down, right and left. These *q-values* are calculated using TD for each state. It is given by:

$$V(s_t) = V(s_t) + \alpha(R_{t+1} + \gamma(V(s_{t+1})) - V(s_t))$$

III. Result Model and Experiments

This section will contain the final model and our experiments for the 2D Grid World Problem using TD and Boltzman Distribution.

A. Experimenting with TD

In temporal difference (TD) learning, the learning rate parameter α and the discount factor parameter γ control the balance between exploration and exploitation, and the weighting of immediate and future rewards, respectively.

Changing the value of alpha can affect the rate at which the agent updates its value estimates. Higher values of alpha cause the agent to place more weight on recent experiences, while lower values cause the agent to place more weight on past experiences. Therefore, a higher alpha value can lead to faster learning, but can also result in the agent being overly influenced by recent experiences, while a lower alpha value can lead to slower learning, but can result in a more stable and reliable estimate of value.

Changing the value of gamma can affect the extent to which the agent values immediate rewards versus future rewards. Higher values of gamma cause the agent to place more weight on future rewards, while lower values cause the agent to place more weight on immediate rewards. Therefore, a higher gamma value can lead to the agent prioritizing long-term rewards and taking a more strategic approach, while a lower gamma value can lead to the agent prioritizing short-term rewards and taking a more myopic approach.

Ultimately, the optimal values of alpha and gamma that we kept for this algorithm was $\alpha = 0.2$ and $\gamma = 0.5$

B. Experimenting with Boltzman Distribution

In the Boltzman Distribution, the temperature parameter affects the "softness" or "hardness" of the distribution, and can be adjusted to control the degree of exploration versus exploitation in the agent's action selection.

Specifically, increasing the temperature parameter in the Boltzmann distribution makes the distribution "softer", meaning that the agent is more likely to explore new actions and take risks. Conversely, decreasing the temperature parameter makes the distribution "harder", meaning that the agent is more likely to exploit its current knowledge and take actions that are already known to lead to high rewards.

Therefore, adjusting the temperature parameter can affect the balance between exploration and exploitation, and may need to be tuned to achieve optimal performance in a given task or environment. A higher temperature parameter may be useful when the agent is still exploring the environment and gathering information, while a lower temperature parameter may be more useful once the agent has gained some level of knowledge and wants to optimize its actions.

Ultimately, we ended up by first initialising the temperature value at 1 and then decreasing it by 95% after each episode. The calculation was causing some trouble for too small values of temperature hence, once it reaches 0.1, we stopped decreasing it further.

IV. Results

Provided with the above values for our RL model using TD and Boltzman Distribution, we were able to affectively train our agent. We tested grids of upto 50×50 and ran 500 episodes. Our agent was able to converge to positive reward terminal states effectively.

References

[1] Dr. Saleha Raza. *CS 451 Notes*. [Online]. Available: https://hulms.instructure.com/courses/2501

[2] OpenAI. *ChatGPT*. [Online]. Available: https://openai.com/chat.

[3] https://web.stanford.edu/group/pdplab/pdphandbook/handbookch10.html

[4] https://reneelin2019.medium.com/the-basic-idea-of-how-boltzmann-distribution-is-used-in-inverse-reinforcement-learning-cf18274e213c

[5] Bartosz Mikulski. [Online]. Available: https://www.mikulskibartosz.name/using-boltzmann-distribution-as-exploration-policy-in-tensorflow-agent/