Approaching Blackjack using Reinforcement Learning

Chandu Dadi
MAI, Faculty of Computer Science
Technical University of Applied
Science Wurzburg-Schweinfurt
chandu.dadi@study.thws.de

Abstract— **Blackjack is one of the most popular casino games, which makes it ideal for applying reinforcement learning algorithms. The main goal in Blackjack is to obtain a hand less than or equal to 21 in value beating the dealer's hand. This study utilizes the usage of reinforcement learning algorithms to learn the optimal strategies to maximize the chances of the player winning the game. Different strategies of blackjack game namely Basic Strategy and Complete point count system are implemented along with additional rule variations. This study makes use of the simulated environment for blackjack along with the Q-learning algorithm approach in learning the best policy for playing blackjack. The approaches are trained for several episodes to get a clear estimate of the problem-solving ability of these algorithms for the Blackjack game. This paper discusses the winning rates of Basic Strategy and Complete Point Count System approaches.**

Keywords—**Reinforcement Learning, Blackjack, Q-Learning, Complete Point Count System, Basic Strategy**

I. Introduction

Blackjack is one of the famous casino games that involves complex decision-making process to be able to win profits in the long run. This kind of complex decision-making process makes the game suitable to be used in the reinforcement learning approach to learn the optimal decisions to succeed in the game.

The Blackjack game usually consists of a dealer and 1 to 7 players in total. The game uses playing cards of usually a single deck or multiple decks till 8. As the game proceeds, players take turns making decisions based on the cards dealt to them. The dealer initially deals each player two cards. Typically, the dealer deals one card face up and the other face down. When players are dealt two cards, they have the option of hitting, which means receiving extra cards, or standing, which means keeping their existing hand. There are several techniques that a player or dealer may utilise to get an advantage in the game. These many methods make Blackjack a problem that can be solved by using Reinforcement-based algorithms to learn from experience and maximize earnings. The game ends when the dealer or the player gets a card value of 21 for a player to win the game or more than the dealer hand value. If the value of the cards with player is more than 21 then the game is called busted and the game is lost.

Reinforcement Learning is an approach in artificial intelligence that involves an agent to learn the optimal policies to solve the problem at hand. This kind of approach deals with reward mechanism which gets maximized as we continue to make right decisions and the rewards are shortened when the agent tries to make wrong decision from the set of rules that is defined to take action on. When the agent is trained or played for various runs, the agent learns what kind of actions are giving the optimal benefits and thus learns what kinds of decision the agent must make depending on the state of the environment the agent is operating in. This kind of learning technique makes the usage of reinforcement learning algorithm perfect for using in the blackjack based critical decision-making games. An agent defined for this task will be able to understand what kind of decisions it must take when playing with the dealer in the blackjack game and it will get adapted and improved over a period after playing for several episodes. This approach makes possible in defining the complicated rules of the blackjack game for the agent to learn the best strategies.

II. METHODOLOGY

Basic Strategy:

In this technique there is a predefined set of actions—hit, stand, double down, and split—a player can take based on their current hand and the dealer's visible card. It is a basic building block of more complex strategies in the game of blackjack

Card Counting:

This strategy involves keeping track of the number of high and low cards yet to be dealt that are placed on the game board. This strategy gives points namely 1,0, -1 for certain kind of cards to keep a count and using the count to make the action decision later.

Q-Learning:

One of the popular RL methods used in Blackjack is updating value functions (Q-values) of state-action pairs based on rewards received in terms of actions executed in states. This entire approach attempts to learn a policy that maximizes cumulative reward

1. *Basic Strategy:*

The basic strategy in blackjack is a set of actions that players must follow to make the best move based on their hand and dealer's up-card These guidelines help mitigate the house edge and provide the player’s chances of winning in time are high. While some rules may differ depending on the number of decks used and other game variables, the basic principles of basic design remain the same.

The key to basic strategy is deciding when to hit or stop. Players should play when their hand count is low to avoid the risk of a hit. For example, always hit a total of 8 or less. Otherwise, players should stay when their hand levels are large enough to draw additional cards that may exceed 21. Like, always stay a total of 17 or more. Double Down is another important factor. This involves doubling the initial bet in exchange for the promise of one more card. Generally, players must lower the total by double 10 or 11 unless the dealer’s up-cards are 9 or higher

When it comes to splitting pairs, the strategy depends on the value of the cards. Players should always split aces and 8s to increase the chances of a strong hand. However, pairs 5 and 10 should not be divided; Instead, double up on the 5s and stay on the 10s. For pairs of 2s, 3s, 6s, 7s, and 9s, if the dealer's up-card is weak, it's best to split them (2-6 for 2s, 3s, 6s, and 7s; 7 or lower for 9s).

Dealing with a weak hand that includes 11 classic Aces requires specific strategies. Players should play or double on soft hands like soft 13-18, depending on the dealer's up-card, but ride a gentle 19 or higher to maximize your chance.

2. *Complete Point Counting System:*

Card counting one of the blackjack strategies where players look at the high and low cards left in the deck. Deal cards in this way have values: +1 for low cards, -1 for high cards, and 0 for intermediate cards. As cards are played, this number of runs helps players calculate the design of the rest of the deck and adjust their bets and choose their choices accordingly

Higher cards (10s, face cards, and Aces) are better for the player because they increase the chances of playing Blackjack or expensive hands. Inspection lets players know when the deck is "hot" (full of high cards) and they should bet more, or "cold" (full of low cards) and they should bet less.

Although reinforcement learning (RL) is not used to create cards first, it can work well with RL-based strategies. The combination of card counts, and RL allows the agent to make intelligent choices, using the mathematical edge from the account to fine-tune and optimize their strategy throughout the game

3. *Q – Learning*

Q-learning is one of the most widely used algorithms for reinforcement learning and can be successfully applied for games like blackjack. It enables agents to learn optimal strategies through interaction with the environment. A player's position in blackjack is determined by their current hand, the dealer's face-up card, and whether or not they can split or double The agent has a series of actions which are usually "hit", "put"; "double down" , "split." " It boils down to being done. The Q-Learning algorithm tries to find the value of each action in each situation." The Q-Learning algorithm arbitrarily initializes Q-values for all state-event pairs. Then, the agent plays blackjack several times to determine the environment. The new rules for each state-event pair are:

$$\mathbf{Q(\textsc{s, a}) = Q(\textsc{s, a}) + \alpha * [R + \Gamma * \textsc{max}(Q(\textsc{s', a'})) - Q(\textsc{s, a})]}$$

Where the variable **Q(S,A)** represents the current estimate of the value for taking action A in state S. The term **α** (Alpha), denotes the learning rate, which controls the extent to which new information influences the Q-value, with a range between 0 and 1. The reward R is the immediate payoff received after performing action A in state S. The discount factor **γ** (Gamma), also between 0 and 1, adjusts the importance of future rewards compared to immediate rewards. The term **MAX(Q(S', A')),** signifies the maximum Q-value for the next state S′ across all possible actions A′, reflecting the best future value. Finally, **Q(S', A')** is the Q-value for the action taken in the subsequent stage. The algorithm ensures that the agent adjusts its assessment of the optimal course of action considering rewards obtained and expected outcomes in future

III. IMPLEMENTATION

Basic Strategy implementation:

In this study, we employed Q-Learning to create a simple blackjack strategy. The Q-Learning method allows an agent to learn the best choice strategy by repeatedly analysing and exploiting the game circumstances. The implementation details of our Q-Learning approach for blackjack are provided below.

The blackjack condition is represented by the Blackjack class. This class simulates the game and handles state changes, hand analysis, and game results. The major aspects of this situation include: The state is determined by the player's total card value, the dealer's visible card, and whether the player has a usable Ace, which are represented as a tuple (player_total, dealer_upcard, has_usable_ace_player). The agent can select "hit" (ACTION_HIT) or "stand" (ACTION_STAND), with further actions such as "double down" and "split" excluded for simplicity. The reward system is straightforward: winning the hand gives a +1 reward, losing leads in a -1 reward, and drawing offers no reward. The step() function controls game state changes by updating the player's hand, checking for busts, and determining the outcome based on the dealers end hand.

The Q-LearningAgent class implements the Q-Learning algorithm with important parameters, including a learning rate (α) of 0.1 to decide how new knowledge influences the Q-value update, and a discount factor (γ) of 1.0 to indicate the relevance of future rewards. The exploration rate (ε) begins at 1.0 and decreases with time, balancing exploration (random action selection) and exploitation (selection of actions with the greatest Q-values). After each episode, the exploration rate is lowered by 0.9999 to ensure a steady transition from exploration to exploitation.

The run_training function sets up the training schedule, which includes several episodes of blackjack in which the agent interacts with the environment to learn the optimal strategy. Episode execution is a key aspect in which the agent begins a new game, chooses actions depending on current policy, changes Q-values, and receives prizes. During training, performance metrics such as total reward per episode, win percentage, epsilon values, and profit per episode are tracked.

Complete Point Count System Implementation:

The point count system used in this study is designed to track and apply to card value classification in blackjack games, with the aim of improving the decision-making and strategy in this framework in the card values and suits are defaulted: the card is represented by an Ace in numerical value as 1, and with a 10 value card .All face cards are treated as value 10. The suit of the card is 'Hearts', 'Diamonds', 'Clubs' and 'Spades'.

The core of the system is the PointCountSystem class that controls the number of cards. When started, the system is structured with a specific number of decks and a dictionary that assigns point values to cards. The point values are conveniently given to help determine the probability of drawing high and low cards: cards numbered 2 through 6 are given a value of +1, to help determine a better deck for the player Cards numbered 7 through 9 are neutral such that neutrality acquires the value of 0; and 10, face cards, and the value of Aces is assigned a value of -1, indicating a poor deck. This point assignment is based on the standard Hi-Lo card counting strategy, which is widely used in Blackjack.

As cards are drawn from the deck, the PointCountSystem updates the number of runs based on the value of the card. This number of runs is important for estimating the quality of the rest of the deck. Additionally, the system includes options for resetting the count at the start of each game, as well as realistic counting. They are achieved by dividing the run count by the number of decks remaining, which yields a normalized measure that changes the number of decks still in play

In the case of Blackjack, the PointCountSystem is integrated to keep track of the counts throughout the game. Initially the deck is built based on predefined card values and suits, and as cards are drawn during play the point count is updated accordingly, allowing for real-time adjustments and current strategies condition of the deck.

Integrating the point counting system into the blackjack environment supports more informed decision making by following a balance of high and low cards. This is important for strategy games because higher numbers indicate a greater chance of drawing the right card, thus influencing decisions such as hits, stands, doubles, or splits impact The system provides players with an automated means of assessing the potential consequences of their actions and adjusting their strategies accordingly

IV. RESULTS

This section details on the results that are achieved while running the above-mentioned basic strategy and complete point count system environment.

Basic Strategy with Q-Learning:

The Basic strategy approach is now equipped with Q – learning algorithm to learn the optimal policy in playing the game. The game is played for episodes of 100000. The agent is made to take actions from the basic strategy implementation to learn progressively as the agents run through the episodes. The basic strategy approach provided us with an estimate of 42.20% of win percentage when trained. There is lot of uncertainty in the game which makes the agent hard to grasp all the rules and is the result of such an accuracy when compared to the tie and the loss.

The Q-value optimization plot for the basic strategy is in [1] .

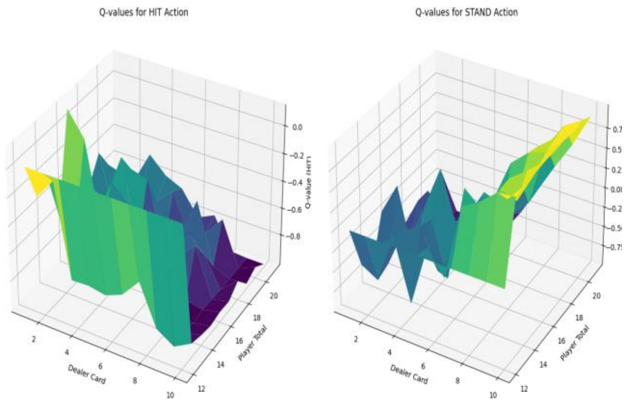

Fig 1: Q-value optimization plot for the basic strategy with Hit and Stand Actions.

The plots show Q values for HIT (left) and STAND (right) actions in blackjack, considering player hand values and dealer cards. HIT Q-values are higher for lower player hand values, indicating that they prefer to play when the player's hand is weak. Conversely, STAND Q-values are higher for higher player hand values, indicating that it is better to stand with a stronger hand. These plots shows the optimal action based on the player's hand and the dealer's upcard, and indicate when to play or pause to maximize expected rewards in the game.

The below plot[2] shows the average win rate of a basic strategy over 100,000 episodes. Initially, as can be seen the win rate rises sharply, indicating rapid improvement, and stabilizes around 40% to 45% after approximately 20,000 episodes. The win rate fluctuates slightly within this range, reflecting the inherent natural randomness of the game. Overall, the strategy quickly becomes effective and maintains a relatively stable win rate over many games. Overall, the plot tells that the win percentage has not been stabilized even after running the game for above mentioned episodes because of the uncertainty present in the game inherently.

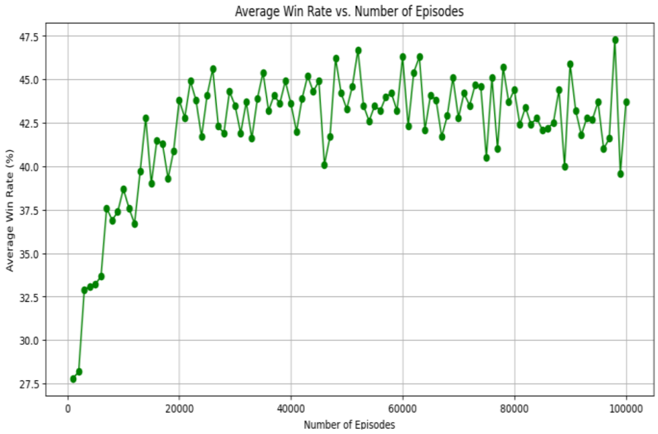

Fig 2 : Average win percentage vs Number of episodes of Basic Strategy

The below plot defines the epsilon decay with respect to how the average win rate changes in [3]. As epsilon grows from 0 to 0.8, the average victory rate falls considerably, from roughly 45% to less than 30%. Lower epsilon values, which indicate a greater willingness to utilise established techniques, result in better victory rates. In contrast, greater epsilon values, which indicate more exploration and less dependence on the existing approach, result in lower victory rates. The plot emphasises the need of striking a balance between exploration and exploitation to maintain an ideal Blackjack win rate.

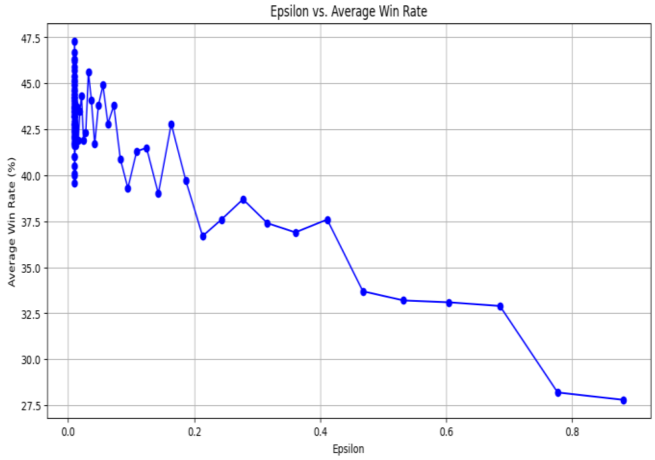

Fig 3 : Epsilon vs the Average Win Rate of Basic Strategy

Complete Point Count System :

In the card counting system approach, the Q learning algorithm is utilised to optimise the action of obtaining the highest rewards. In this method, we created a simple card counting system in which we add the value to smaller card values and deduct the value from larger card values. The agent is trained for 100000 episodes with the actions stand, hit , double-down and split. The overall win percentage that I achieved after running the game for 100000 episodes is 40.94% The agent is equipped with the complete point count system to take extra rule information into account to play the game to learn strategies that achieves the bigger rewards.

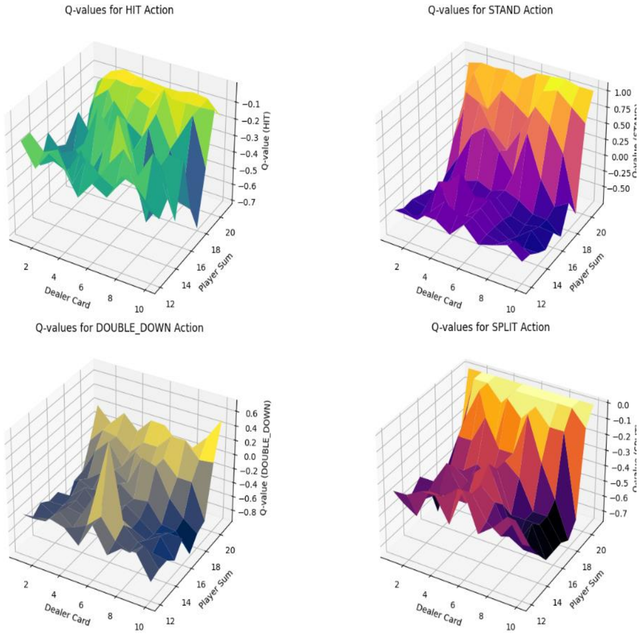

Fig 4 : Q – value optimization plot for Hit , Stand , Double Down and Split of complete Point count system.

The above plot defines the Q values learned by the Complete Point count system agent when trained for 100000 episodes. The Hit action has Lower Q Value especially for the larger player hands which says that playing dominant in the game usually cause more harm than good and has no advantage overall on the game. The Stand action has High Q value for larger player hands which says that it is more useful to stay without hitting when the player hands are strong . The Double Down action has a medium Q value which says that the most of the time this action is beneficial according to situational conditions typically when the player hand value is in mid-range of blackjack game. The Split action shows high Q Values for certain combinations namely when the player has a pair and the dealer's card is low.

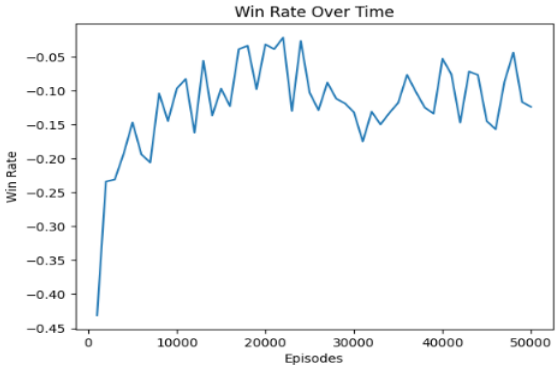

Fig 5 : Win Rate over time vs Episodes for Complete Point Count System

The graph shows that when more episodes are played, the full point count system in Blackjack improves, suggesting that strategy is learned and refined with time. Initially, the system performs badly with a high loss rate, but as the number of episodes grows, the win rate improves, stabilising at -0.15 and -0.05. This shows that, even as the system improves and losses decrease, it still fails to attain a positive win rate. Thus, while the strategy shows potential, more tweaks or complementing tactics are required to attain consistent success in Blackjack.

V. CONCLUSION

This paper studied the usage of artificial intelligence approach namely Reinforcement learning (RL) in playing blackjack to prove that RL can efficiently learn the policies or actions of the blackjack game, The complex rule variations like complete point count system has been implemented to make the agent run and learn from experience. The Study shows that the agent is able to learn as the game progresses and learns the best suit of actions that enables the player to be on the winning side of the game.

The Basic strategy provided a win percentage of 42.2 % when ran with the Q-learning approach while the complete point count system has got 40.94% overall . This shows that the agent is able to make informed decision of the actions better when compared to random selection of actions which gives us only around 20-30% winning rate. That shows a substantial increase in the winning rate which defines the importance of using the reinforcement learning algorithms in the complex decision-making games or tasks like Blackjack.

This study helps the individual to decide what are the best actions an individual can consider to use in the play. The agent can be tested with different set of parameters to control the exploration and exploitation learning capacity of the agent to gain a deeper understanding of the game. Further work includes usage of more complex rule variations and the usage of the other RL learning algorithms such as SARSA , Monte Carlo , Policy Gradient Methods.

REFERENCES

[1] Thorp, E. O. (1966). Beat the Dealer: A Winning Strategy for the Game of Twenty-One

[2] [Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction.

[3] Kakvi, Saqib A. "Reinforcement learning for blackjack." Entertainment Computing–ICEC 2009: 8th International Conference, Paris, France, September 3-5, 2009. Proceedings 8. Springer Berlin Heidelberg, 2009.

[4] MAO, Clifford. Reinforcement Learning with Blackjack. 2019. Doktorarbeit. California State University, Northridge.

[5] Wu, Allen. "Playing Blackjack with Deep Q-Learning."

[6] De Granville, Charles. “Applying Reinforcement Learning to Blackjack Using Q-Learning”, University of Oklahoma.