The Impact of deck size Q-Learning Blackjack

Avish Buramdoyal (BRMAVI002)
Supervisor: A/Prof. Tim Gebbie

University of Cape Town
Honours Project Proposal

Abstract

Blackjack or 21 is a popular card-based game of chance and skill played in many casinos. The objective of the game is to win money by obtaining a point total higher than the dealer's without exceeding 21. Finding an optimal blackjack strategy proves to be a difficult but interesting problem, due the stochastic nature of the game and particularly when it involves playing with an unknown deck size. The ideal blackjack strategy will maximize financial return in the long run while avoiding gamblers ruin. The stochastic environment and inherent reward structure of the game presents an appealing problem to reinforcement learning algorithms. This project explores the problem of a finite deck in order to explore the learning rates for Q-learning. The project will then involve implementing and simulating a Q-learning solution for optimal play and investigate the rate of learning convergence of the algorithm as a function of simulation size and deck size.

Keywords: house edge, betting unit, stochastic

1 Introduction: Blackjack

1.1 History of the game

The game of blackjack started in the 18th Century in France [Ofton, 1998] and was called Vingt-et-Un, translating to 21. The rules of the game back at the time somehow differed from modern game rules but had the same objective of getting a score as close to 21 without exceeding it [Ofton, 1998]. The evolution of the game happened greatly in America and a house-banked blackjack system was first introduced in Nevada in 1931 [Ofton, 1998]. The rise in prevalence of legalised games in Las Vegas casinos inspired a number of players in developing optimal strategies for play [Ofton, 1998]. The game during the 20th century was offering bonus payouts including one that paid an extra if a jack of spades/club i.e. a blacjack was dealt along with an ace of spades [Gaming and Bartending, 2019]. It was that time when the game changed its name to blackjack [Gaming and Bartending, 2019]. Roger Baldwin, Wilbert Cantey, Herbert Maisel and James McDermott were known by blackjack insiders as the "Four Horsemen" who were the first in determining the optimal strategy for blackjack play [Haney, 2008].

Baldwin et al. published, in 1956, the basic strategy to play blackjack and since then, numerous researches have been conducted attempting to improve this strategy [Hellemons, 1996]. The first success was in the early 1960's when mathematics professor Edward Thorp published his book "Beat the dealer a winning strategy for the game of 21" [Hellemons, 1996]. This caught the attention of a lot of blackjack players. Casinos had to increase the deck size for each play in an attempt to overcome the effectiveness of the counting system [Hellemons, 1996]. Edward Thorp came up with a counting system known as the Ten-Count system allowing players to keep track of cards which should inform them how to size their bet to their own advantage. [Thorp, 1966]. Since then, numerous refinements have been made with respect to the Ten-Count system, improving the player's advantage.

2 Playing Blackjack

2.1 Rules of the Game

2.1.1 Game Set-up

The objective of blackjack is to get a hand total higher than the dealer without busting[1]. Blackjack is a casino banked game allowing players to compete against the house rather than each other. The game of blackjack consists of a dealer and from 1 to 7 players. A standard deck of 52 cards was initially used for blackjack. After the announcement of the first winning strategies, casinos implemented countermeasures such as varying the deck size, making card counting harder [Thorp, 1966]. Nowadays, casinos in Nevada, Las Vegas and United States use between 1 to 8 decks [Wong, 1994].

Cards 2 to 10 are worth their face values, Jacks, Queens and Kings are counted as 10 and an ace is worth 1 or 11, whichever the most favourable to the player. A hand with an Ace valued as 11 is called a "soft hand" and all other hands are "hard hands". The distinction between soft and hard hands is important as the strategy for a given total of a soft hand can differ from the same total holding a hard hand. This can be illustrated in the strategy table provided in section 11 of Appendix.The players must

[1] busting: exceeding a score of 21

make their initial bets before any cards are dealt ranging between a minimum and maximum set by the casino.

2.1.2 The deal

At the beginning of each play/round, the dealer shuffles the pack of cards. The player then cuts the pack into half. The deal starts off with each player placing a bet before the start of every hand. The players are next dealt with two cards face up and the dealer also gets two cards, one face up and one face down (called a hole card).

It should be noted that a starting hand of an Ace and a 10-valued card is called a Blackjack or natural and beats any other hands even another blackjack. Assuming blackjack pays 3:2, the player receives 1.5 times his initial bet for having a blackjack. If the player doesn't have a blackjack but the dealer does, the player loses his bet. If both the player and dealer have a blackjack, it's a tie and no money is exchanged.

The draw of cards proceeds in a clockwise fashion starting at the left most player of the dealer. The player will look at his hole cards and take actions requested from the dealer. After all players have completed their action per round, the dealer turns his face down card up and decide which action(s) to take.

2.1.3 Actions

The challenge posed to the player is choosing the optimal action at each hand, given his current total and the dealer's face up cards. The actions available at play include standing, hitting, splitting and doubling down.

A player might choose to stand, i.e. takes no additional cards if doing so is unattractive for his current hand total. Hitting is when the player asks for another card from the dealer. A player is allowed to hit as many times as long as he does not bust. The player might even choose to separate two cards of the same face value and make another bet of equal size as the first bet and play each card as a separate hand. This is known as splitting. Another option available to the player is to double his bet in the middle of a hand. The player then has to draw one and only one card from the dealer. A further option to the player would be to take insurance. Under insurance, an additional wager by the player is allowed on the condition that the dealer has an Ace as up card. The player should generally expect the dealer to have a natural to take insurance.

2.2 Examples

To further understand the game of blackjack, we consider examples adapted from [Thorp, 1966]. The player's key decisions as to whether pair split, double down, stand or even draw and the order in which the player makes them are illustrated in a flowchart shown in the Figure 1. The flowchart essentially helps the player know whether to stand or draw more cards to improve his hand off based on his current total.

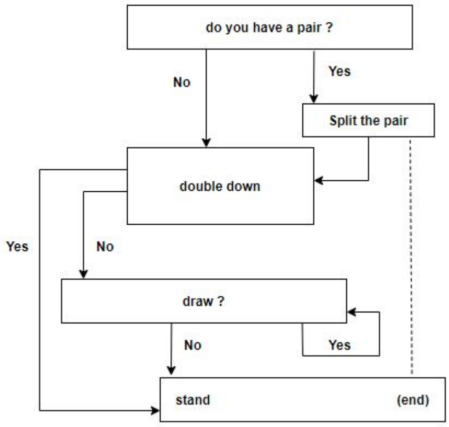
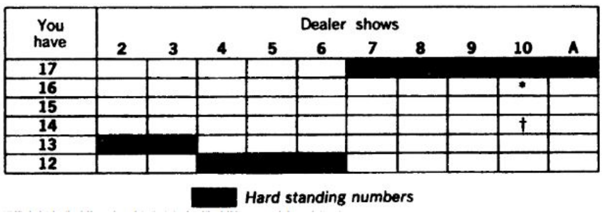


Fig. 1: The Player's Key Decisions

We will now consider examples of possible actions taken by a player based off the basic strategy for soft, hard and split hands. The possible actions will be illustrated through Strategy tables computed by [Thorp, 1966]. The strategy table for hard standing numbers, soft doubling numbers and pair splitting will be considered in this section. The other strategy tables corresponding to other relevant actions will be found in section 11 of Appendix.

Hard Standing
In the majority of hands, the player will neither double down nor split a pair and this decision according to [Thorp, 1966] is the most important part of the strategy. Hard standing number for a specific dealer's up card is the smallest total a player stands on against that dealer's up card. Figure 2 below depicts a table of hard standing numbers. The table shows that if the dealer has a 7 as up card, a player should stand on hard totals of 17 or more and draw with hard totals of 16 or less.

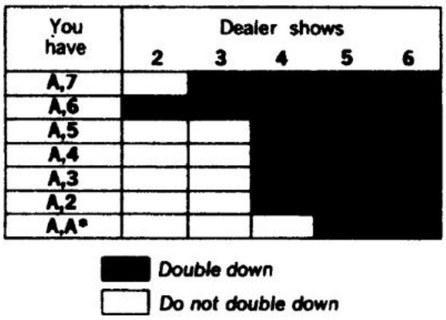


You have	Dealer shows									
	2	3	4	5	6	7	8	9	10	A
17						■	■	■	■	■
16									*	
15										
14									†	
13	■	■								
12			■	■	■					

■ Hard standing numbers

Fig. 2: Pictorial list of hard standing numbers

Soft Doubling
A player might be required to take different actions based on soft hands as illustrated in Figure 3. It the player's soft total does not appear in the table, he should not double down but instead proceed on decision of whether to draw or stand. It should be noted that based on the set-up adapted by [Thorp, 1966], doubling down is only done on totals of 13 or more.

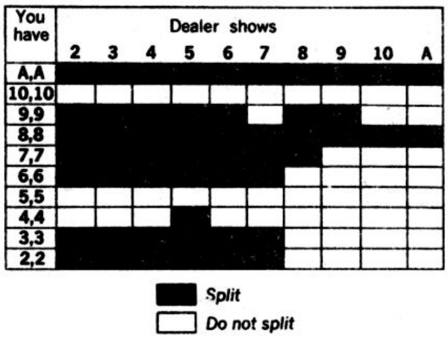


You have	Dealer shows 2	3	4	5	6
A,7	□	■	■	■	■
A,6	■	■	■	■	■
A,5	□	□	■	■	■
A,4	□	□	■	■	■
A,3	□	□	■	■	■
A,2	□	□	■	■	■
A,A*	□	□	□	■	■

■ Double down
□ Do not double down

Fig. 3: Pictorial list of soft doubling numbers

Splitting
Having a pair allows the player to split the 2 cards and play each as a separate hand. Figure 4 is an easy table to recall relative the other strategy tables. The table suggests to always split aces, 8's and never split 5's and 10's Aces should be split as it provides a very good chance of getting a winning hand (score of 21) [Thorp, 1966]. 8's should also be split as a score of 16 is in general, a bad total [Thorp, 1966]. It increases the player's chance of busting on attempt to improve his hand total for drawing more cards [Thorp, 1966]. Splitting 5 is not to the player's advantage as it replaces an initially good total that is good to double down on and splitting of 10's equally replaces a good hand score of 20 by two generally less better than the average [Thorp, 1966].



You have	Dealer shows 2	3	4	5	6	7	8	9	10	A
A,A	■	■	■	■	■	■	■	■	■	■
10,10	□	□	□	□	□	□	□	□	□	□
9,9	■	■	■	■	■	□	■	■	□	□
8,8	■	■	■	■	■	■	■	■	■	■
7,7	■	■	■	■	■	■	■	□	□	□
6,6	■	■	■	■	■	■	□	□	□	□
5,5	□	□	□	□	□	□	□	□	□	□
4,4	□	□	□	■	□	□	□	□	□	□
3,3	■	■	■	■	■	■	□	□	□	□
2,2	■	■	■	■	■	■	□	□	□	□

■ Split
□ Do not split

Fig. 4: Pictorial list of hard standing numbers

The other possible actions (hard doubling, soft standing) are also provided in section 11 of Appendix.

2.2.1 Settlement

If the player and the dealer both have the same total hand value, it is a tie and no money is exchanged. If neither the player nor the dealer busts, the one with the higher total hand value wins the bet. If the player busts and the player's total is below 21, the player wins the bet.

3 Project Assumptions

The project is premised on 5 key assumptions:

(i) We can simulate single and multiple decks games,

(ii) Natural blackjack pays 3:2,

(iii) The dealer stands on soft 17,

(iv) No insurance offered and

(v) No surrender possible

4 Hypotheses

Hypothesis 1 (Primary Hypothesis): Fixed or finite deck size counting rules can be used to learn winning strategies for black-jack using Q-learning.

Hypothesis 2 (Secondary Hypothesis): Learning is less effective as the deck size increases, and there may be a finite deck size at which learning is no-longer feasible.

These hypothese will be investigate by achieving the project objective and aims given in section 7

5 Review

5.1 Learning to Play Blackjack

For purpose of this project, we will follow the blackjack basic strategy which was initially determined by Baldwin et al [Hellemons, 1996]. A basic strategy is simply a proper playing decision for every possible hand against the dealer [Thorp, 1966]. Every possible combination of the player hands' total and dealer's face up card has a mathematically correct play and these can be summarised in a Strategy Table framework shown in section 11 of Appendix. The optimal action per each round therefore depends on the player's current hand and the dealer's face up card [Wong, 1994].

According to Ed Thorp, a good player can take advantage of the basic strategy by keeping track of cards dealt in previous plays [Thorp, 1966]. The track kept is used as information for the next round and is known as card counting. Card counting is based off the basic strategy providing a mathematically proven opportunity to the player's advantage over the house [Grabianowski, 2013]. Generally, a player looks for hands when more high cards are left to be played than a standard deck would have [Pagat, 2019]. This was the key to Edward Thorp Ten Count system which meant to determine the ratio of high to low cards in the deck [Thorp, 1966] and This was done by assigning a value to each card using one deck in a Blackjack game.The values for Ed Thorp Ten count system is given in Table 1.

Card	Value Assigned
A-9	+4
10	-9

Tab. 1: Ten Count System

Thorp's method applies for a single-deck game with the player having a slight advantage over the house, improving the house edge[2] from 6 percent in favor of the house, to about 1 percent in the player's favour [Thorp, 1966]. This Ten Count system is however of little application in a multi-deck game set-up [Wong, 1994]. It provides a slight advantage for a single deck blackjack game, but has little application in the multi-deck games.

Using this table, the player can now count cards. The player will essentially start with a count of 0. Then, based on the above table, he will add or subtract for every single card revealed. This is known as the "Running Count". The next step is to compute the "True Count" given by:

$$\text{True Count} = \frac{\text{Running Count}}{\text{Decks Remaining}} \tag{1}$$

The greater the count, the more the player should bet given his higher advantage. The general idea is to bet little or nothing when player advantage is low and to bet proportionately high when player advantage is high, i.e. in multiples of the player's betting unit[3] [Thorp, 1966]. By continually adjusting his betting unit,the player is not expected to go broke,and his bankroll is also expected to increase quicker over the long run [Thorp, 1966].

To relate to the mathematical theory of the size bet based on the count, Edward Thorp used the Kelly Criterion . Ed Thorp was introduced to Kelly's Paper in the early 1960s and used the Kelly Criterion [Thorp, 1966] defined below. The Kelly Criterion is an intermediate strategy between maximising one's expected return and minimising the probability of ruin [4] given. It suggests that a player who knows his advantage should bet that percentage of his bankroll [Thorp, 2006].

Kelly Criterion

If starting with X_0 capital we bet an amount B_t at decision time increment i, given some fraction f, according to the rule:

$$B_i = fX_{i-1} \text{ where } 0 \leq f \leq 1. \tag{2}$$

After n trials our capital is then:

$$X_n = X_0(1+f)^S(1-f)^F n. \tag{3}$$

Here S and F are the number of successes and failures, respectively, in n trials,where $S + F = n$.

By solving the Kelly equations (Equation 3) on a computer, a strategy for the game of blackjack which is as close to the optimal desired can be found [Thorp, 2006].

Ruin is re-interpreted to mean that for some arbitrary small positive ϵ:

$$\lim_{n\to\infty} \mathbb{P}[X_n \leq \epsilon] = 1. \tag{4}$$

Harvey Dubner introduced a simplified variation of Thorp's strategy, called the "Hi-Lo card counting strategy" (or the "point count system"), at the Fall Joint Computer Conference in Las Vegas in 1963 [Patterson, 2002]. The Hi-Low system framework is represented in Table 2.

Card	Value Assigned
2,3,4,5,6	+1
7,8,9	0
10,J,Q,K,A	-1

Tab. 2: Hi-Lo System

The value of +1 means as low cards are depleted from the game, chances of busting are low and so fewer cards can hurt you in the future while the value of -1 means as high cards are depleted from the game, player advantage falls. The value of 0 represents neutral cards and favours neither the player nor the house. This refinement in counting system done from the Ten Count system was then included in [Thorp, 1966]. The Hi-Low system is the most commonly used card counting system nowadays [Shackleford, 2019b].

5.2 Expectations

As suggested by [Wong, 1994] in his book "Professional blackjack", a player's advantage or disadvantage using the basic strategy varies with the rules and number decks used. Commonly, a player's disadvantage is around 0.5 percent with the basic strategy [Wong, 1994].

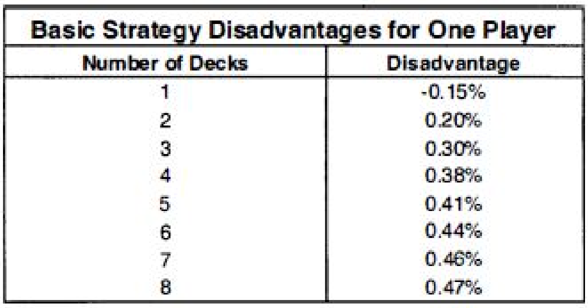
5.2.1 Deck-size variation

It turns out that increasing the number of pack(deck size) slightly cuts the player's advantage [Thorp, 1966]. An analysis of Multiple Deck Games was conducted by [Conrad, 2002]. The same blackjack rules under [Thorp, 1966] used was adapted in their analysis. The only difference in simulation was the absence of a cut card in the deck. The results produced are given in Figure 5 below.

[2] house-edge: Advantage house has over player

[3] betting-unit: Size of player's bet

[4] ruin: agent's surplus level negative



Basic Strategy Disadvantages for One Player	
Number of Decks	Disadvantage
1	-0.15%
2	0.20%
3	0.30%
4	0.38%
5	0.41%
6	0.44%
7	0.46%
8	0.47%

Fig. 5: Impact of Deck size

The chart above shows an increase in player's disadvantage as the number of deck increases. It can be observed that a player using the correct basic strategy in a one-deck game, can have the advantage. This explains why it is so difficult to find a one-deck game with the same standard rules [Conrad, 2002].

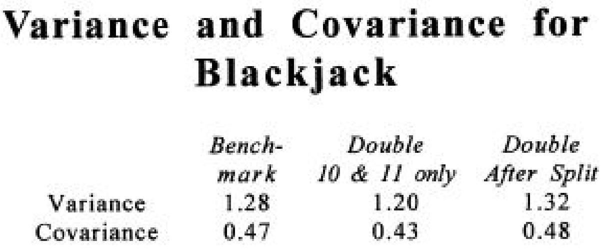
Another important thing to note is the fewer the decks, the more blackjacks you can win relative to a game of more decks [Smith, 2015]. The reason for getting more blackjacks relates to the impact of a card being dealt, i.e a card being removed from deck is higher in a game with fewer overall cards.

5.2.2 Bet-size and strategy variation

[Wong, 1994] suggested that a player's expected win is proportional to his bet size and the amount of time available for play. The more a player bets, the more he will win on lucky hands and the more he loses on unlucky hands. [Wong, 1994] proposed a proportional betting scheme, i.e. betting a fixed proportion of your capital on each hand.

Player and dealer advantage changes from round to round [Wong, 1994]. If the count stays zero or negative, a player should size his bet as small as possible [Wong, 1994]. For a positive count, the player may have an advantage. More specifically, for every unit increase in the true count, the player's advantage goes up by 0.5% [Wong, 1994].

The variance of possible outcomes to the player depends on the specific rules set [Wong, 1994]. A player being able to double after splitting means higher variance (bigger ups and downs). Alternatively, allowing the player to only double down on 10 or 11 means lower variance. Risk or variance of outcomes also depends on the number of simultaneous hands being played by a player [Wong, 1994]. Covariance is used as a measure of how likely the 2 hands are to win or lose together and is shown in Figure 6.

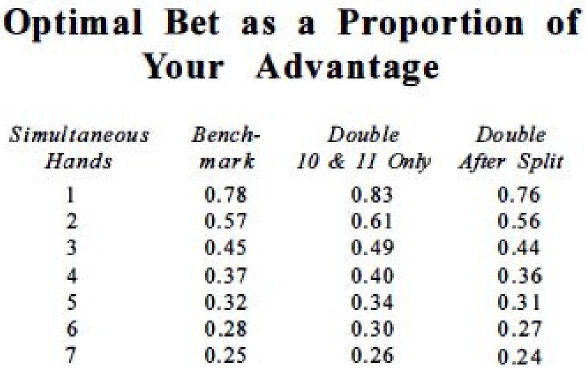


Variance and Covariance for Blackjack

	Bench-mark	Double 10 & 11 only	Double After Split
Variance	1.28	1.20	1.32
Covariance	0.47	0.43	0.48

Fig. 6: Variance and Covariance for Blackjack

The optimal bet size to a player changes when playing more than one hand in one play [Wong, 1994]. This is so as the two hands are not independent. [Wong, 1994] proposes to use the table below indicated in Figure 7 to size one's bet for playing simultaneous hands. The table indicates the different bet players should make as a proportion of their edge.



Optimal Bet as a Proportion of Your Advantage

Simultaneous Hands	Bench-mark	Double 10 & 11 Only	Double After Split
1	0.78	0.83	0.76
2	0.57	0.61	0.56
3	0.45	0.49	0.44
4	0.37	0.40	0.36
5	0.32	0.34	0.31
6	0.28	0.30	0.27
7	0.25	0.26	0.24

Fig. 7: Variance and Covariance for Blackjack

A player will bet 56% of his edge if he follows the strategy of doubling after splitting while a player following a strategy of doubling down on 10 or 11 only will bet 61% of his edge. For purposes of simplicity, Stanford proposes to round the proportions and use 80%, 60% and 45% as proportion of edge bet for 1, 2 and 3 simultaneous hands respectively.

It is also important to note that the formula 5 is used for computing the proportion of edge bet which is based off the Kelly Criterion and is given by:

$$\text{Proportion} = \frac{1}{(v + (n-1)c)}, \tag{5}$$

where v = variance, c = covariance and n = number of simultaneous hands.

The 5 rules suggested by [Wong, 1994] to allow players win at a faster rate are:

(i) Making bigger bets in situations where the player has an edge

(ii) Finding a game with better rules

(iii) Finding a game with fewer decks

(iv) Finding a game with better penetration

(v) Playing more hands per hour

6 Reinforcement and Q-Learning

6.1 Reinforcement Learning Algorithm

Reinforcement learning problem involves relying on responses from the environment to learn and more specifically to map situations to actions [de Granville, 2005]. The responses under reinforcement learning techniques take the form of rewards to guide the agent in developing his policy [de Granville, 2005]. The aim is to maximise a numerical reward signal. Playing blackjack is naturally formulated as an episodic finite Markov Decision Process (MDP) [S.Sutton and Barto, 2014]. The environment is as such defined by a set of states, actions, transition probabilities, and expected rewards modelled as an MDP. Important quantities in a reinforcement learning algorithm also include a value-state function $V^{\pi}(s)$ and an action-value function $Q^{\pi}(s,a)$ of being in a particular state s. Value functions specify what is good for the agent in the long run [S.Sutton and Barto, 2014].

$V^{\pi}(s)$ enables the agent to compute the expected reward of being in state s, following policy π [S.Sutton and Barto, 2014]. $Q^{\pi}(s,a)$ now allows the agent to compute the expected reward of being in state s, taking action a, and thereby following policy π [S.Sutton and Barto, 2014]. The optimal state-value function and the optimal action-value function are denoted by $V^{*}(\mathrm{s})$ and $Q^{*}(\mathrm{s,a})$ respectively [S.Sutton and Barto, 2014]. One of the challenges that arise in reinforcement learning, is the trade-off between exploration and exploitation [S.Sutton and Barto, 2014]. In order to learn, the agent must try to balance between exploration and exploitation of the environment [S.Sutton and Barto, 2014]. During exploitation, the agent selects the action yielding the highest value known as the greedy actions and selects the non-greedy actions during exploration of the environment.

6.2 Implementing Q-learning

6.2.1 One-step Q-Learning

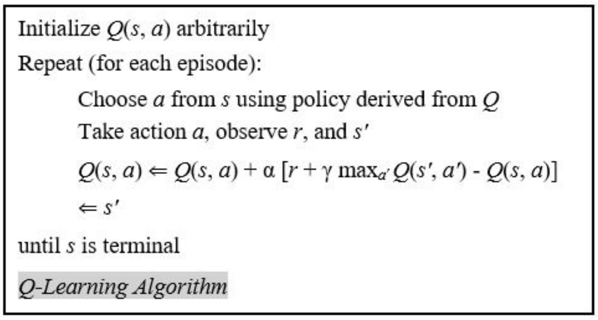
One of the most important breakthroughs in reinforcement learning was the development of an off-policy TD control algorithm known as Q-learning [S.Sutton and Barto, 2014]. The simplest form of q-learning is a one-step Q learning as given by:

$$Q\left(s,a\right) \Leftarrow Q\left(s,a\right) + \alpha\left[r + \gamma \max_{a'} Q\left(s',a'\right) - Q\left(s,a\right)\right] \tag{6}$$

In Equation 6, α is the learning rate, allowing to determine the size of the update made on each time-step and γ is the discount rate, allowing to determines the value of future rewards.

6.2.2 Q-Learning Algorithm

Under the assumption of the one-step Q-learning and a variant of stochastic[5] approximation conditions on the sequence of step-size parameters, it has been shown that Q converges with probability 1 to Q^{*} [S.Sutton and Barto, 2014]. The Q-learning algorithm is an excellent method for approximating an optimal blackjack strategy as it allows learning to take place during play [de Granville, 2005].
From the one-step Q-Learning, we derive the Q-Learning Algorithm as indicated below:



```
Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
    Choose $a$ from $s$ using policy derived from $Q$
    Take action $a$, observe $r$, and $s'$
    $Q(s, a) \Leftarrow Q(s, a) + \alpha\,[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$
    $\Leftarrow s'$
until $s$ is terminal
```

Q-Learning Algorithm

Fig. 8: Q-Algorithm

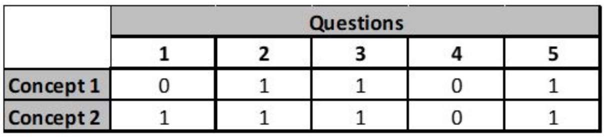
Blackjack can be easily formulated as an episodic task, where the terminal state of an episode corresponds to the end of a hand [S.Sutton and Barto, 2014]. The state representation will consist of the agent's current total, the dealer's face up card value and whether the hand being dealt is a soft or hard one. This will essentially be the strategy table of the player. A reward structure should also be adopted allowing the agent to learn. The idea behind the reward structure will be such that the agent receives a positive feedback for actions resulting in the required terminal state and receives a negative feedback for not reaching a terminal state [S.Sutton and Barto, 2014]. It should also be noted that the reward structure will be a function of the size of the agent's bet [S.Sutton and Barto, 2014].

The reward structure is as follows: For each action that does not result in a transition to a terminal state, a reward of 0 is given [S.Sutton and Barto, 2014]. Once a terminal state has been reached, a reward is given based on the size of the agent's bet.

6.2.3 Q-Matrix

A q-matrix is a matrix describing relations of questions and concepts through states [Barnes, 2015]. It is noted that the q-matrix is a domain-independent model of knowledge [Barnes, 2015]. In other words, this means that the states in the q-matrix are the true information available for each combination of question-concept. The q-learning algorithm is really suitable for this problem as it takes the inherent reward structure of the game which is omitted by supervised learning algorithms [de Granville, 2005]. As an example, we consider a basic example with binary states as shown in the Figure 9 below.

[5] stochastic: subject to random behaviour



	Questions				
	1	2	3	4	5
Concept 1	0	1	1	0	1
Concept 2	1	1	1	0	1

Fig. 9: Binary q-matrix

If we had to read the above table, it would mean that knowledge of concept 1 is required to answer question 2, 3 and 5 while concept 2 is a requirement for answering questions 1, 2, 3 and 5. The description of this problem is analogous to the blackjack problem.

For the base problem [6], we will have 2 q-matrices, one for the case where the dealer stands on soft 17 and a second one where the dealer hits on a soft 17. Given each strategy table is divided into a player's hard, soft and pairs total, we will have $2\ (11+8+8)*10 = 27*10$ q-matrices corresponding to the dimensions of each strategy table for a player's hard total (11∗10) matrix, soft total (8∗10) matrix and pair splitting $(8*10)$ matrix. The strategy table corresponds to the one adapted by [Shackleford, 2019a] found in section 11 of Appendix. The q-matrices will however be updated based upon the significance of the learning paramater α. It should also be noted that 2 reward matrices of the same dimensions of the q-matrices should also be considered with each element being the reward of being in a particular state, i.e. of taking a particular action (hitting/standing/splitting/doubling down).

6.3 Single and Multi-Learning Agent

For the multi-agent set-up, we want to investigate, whether multiple agents outperform independent agents who do not pass information across learning given a fixed number of reinforcement-learning agents in the game. We treat independent agents as a benchmark. We also aim to know the cost attached to cooperative agents. Cooperative agents learn and interact by sharing sensation, episodes and learned policies [Tan, 1993]. To model a Multi-agent interaction, we adopt the concept of Complex Dynamics used in learning complicated games.

Complex Dynamics

Complex dynamics is an area of mathematics focused on the study of dynamical systems defined by iteration of functions on wider and complex state spaces.
[Galla and Farmer, 2013]. Game theory is a standard approach used in modelling strategic interactions but mainly studies the optimal actions of simple games.
[Galla and Farmer, 2013]. Blackjack is a game involving interaction between a player and a dealer. The number of action-state spaces of a blackjack game will be bigger for the case of multi-agents rather than a single agent competing against the house. We therefore consider Complex Dynamics as a potential extension to the existing q-algorithm where more agents now simultaneously compete against the house. We will consider, under the Dynamics approach, the effect of a single agent relative to multiple agents to the blackjack game outcomes.

1. **Single multi-RL[7] agent**
   For the case of a single multi-RL agent, the set-up of the game and methodology used to allow learning is the same as it would be for the case of a single agent. We will use the one-step q-learning given in Equation 6 to update the player's policies.

2. **Two multi-RL agents**
   In the case of two multi-RL agents, the paper by [Galla and Farmer, 2013] considers a 2-person games involving a type of reinforcement learning called Experience-Weighted Attraction (EWA). The EWA assumes a numerical attraction to each strategy. A numerical attraction is simply a postive numerical value attached to a strategy and allows to determine the probability of a player choosing that strategy. [Galla and Farmer, 2013].

   Methodology for two multi-RL agents:

   We consider 2 players A and B such that at each step time t, player $\mu \in \{\text{A}, B\}$ chooses between one of N possible moves. The player picks the ith move with frequency $x_i^\mu(t)$ where $i = 1, \ldots, N$. The frequency vector $\mathbf{x}^*(t) = (x_1^\mu, \ldots, x_N^\mu)$ is the strategy of player μ. The payoffs received by A for playing strategy i and B for playing strategy j is given by Π_{ij}^A and Π_{ji}^B respectively. The players then learn their strategies $\mathbf{x}^\mu$ using that form of reinforcement learning called the EWA [Galla and Farmer, 2013]. Experimental economists have shown that this approach provides a reasonable approximation for how real players learn in games [Galla and Farmer, 2013].

   Under this approach, the probability of a given move is given by Equation:

   $$x_i^\mu(t) = \frac{e^{\beta Q_i^\mu(t)}}{\sum_k e^{\beta Q_k^\mu(t)}} \tag{7}$$

   where Q_i^μ is called the "attraction" for player i to strategy μ

   Using the EWA, players A and B attractions are updated according to the Equations 8 and 9 below:

   $$Q_i^A(t+1) = (1-\alpha)Q_i^A(t) + \sum_j \Pi_{ij}^A x_j^B(t) \tag{8}$$

   $$Q_i^B(t+1) = (1-\alpha)Q_i^B(t) + \sum_j \Pi_{ij}^B x_j^A(t) \tag{9}$$

   Equation 8 shows a situation where both players vary their strategies slowly so that A is able to collect appropriate statistics about B before updating his own strategy [Galla and Farmer, 2013]. The same applies for player B under the equation 9 where he collects information about player A.

[6] base problem: learning to play blackjack with q-learning

[7] Reinforcement-Learning

3. **3 or more Multi RL agent**
   By randomly generating games under EWA, the forces enabling learning are characterized by 3 seperate systems [Galla and Farmer, 2013]:

   (a) convergence to a unique fixed point
   (b) Huge multiplicity of stable fixed points
   (c) Chaotic behaviour[8]

   It is noted that moves, under the ensemble of games studied in this paper are randomly chosen. This means the learning dynamics have a stochastic component. The paper suggests that the analysis of the 2-player game can potentially be extended to multiplier games [Galla and Farmer, 2013]. Games become harder to learn with increased competition, especially if learning algorithms with long memory is used. [Galla and Farmer, 2013] preliminary studies of multiplayer games suggest an increase in the chaotic regime as the number of players increase leading to intermittent bursts of large fluctuations punctuated by relative quiescence of total payoffs of all players [Galla and Farmer, 2013].

4. **Cooperative Agent**
   Another study performed by [Tan, 1993] suggests that cooperative agents learn faster and converge sooner to optimal actions than independent agents. The paper also noted that extra sensory information can interfere with learning. Sharing of information therefore comes at a communication cost [Tan, 1993].
   This extension of cooperative agents is inspired from the real-life story of 6 M.I.T. students who were expert trained card counters with a scheme to hit Vegas every weekend to make really big gains [Tamburin, 2006]. The 6 M.I.T. members did not however play independently like most card counters do. They went one step further by using spotters, big players and a team bank [Tamburin, 2006] .
   Their approach to beat the house was to use covert signalling to indicate other players how "good" the table they were sitting on was and a system of mnemonic devices to indicate the player who just joined the table how much the count was [Tamburin, 2006]. These 6 students were able to win millions from Vegas showing that teamwork pays off big in a blackjack game on the condition you play the system well and do not give in your emotions [Tamburin, 2006]. This story has then been shaped to fit a book "Bringing Down the House: The Inside Story of Six MIT Students Who Took Vegas for Millions" and a movie "21" in 2008.
   We will, for purpose of this project also consider cooperative players in a blackjack game and investigate the significance of change in total expected payoff of a whole team relative to independent agents.

7 Aims and Objectives

The project aims to demonstrate the different learning rates for different deck size blackjack games. This requires that the following objectives are met:

1. We implement a Q-learning algorithm, and reinforcement learn the simulated blackjack game.
2. We visualise and investigate the rate-of-learning convergence for the algorithms as a function of simulation size and deck size.
3. We will consider possible extensions to base Q-learning algorithm that can enhance convergence for a large number of decks.
4. We aim to quantifying the deck-size threshold for Q-learning.
5. We aim to quantifying the deck-size threshold for a Deep Q-learning algorithm [Choudhary, 2019] so that the learning rates and deck-size thresholds can be compared to vanilla Q-learning [Yoon, 2019].
6. Comparing the performance of the different Q-learning methods to know whether we learn to win when we do not know the size of the deck
7. Comparing the performance of a single RL agent to the multi-RL agents on house edge

8 Data Requirements Specification

There is no specific data requirement as the project data is generated by the Blackjack simulator.

9 Systems Requirements Specification

1. **Hardware Requirements**
   For purpose of this analysis, I will make use of my personal computer to complete all the coding and simulation requirements.

2. **Software Requirements**
   The software to be used for purpose of the project proposal are:

(i) Python Wing 101 v. 7.2 to fulfill coding requirement (blackjack simulator)
(ii) Github for configuration and version control
(iii) Latex using Texworks and Overleaf for versioning and techincal documentation

All software and test-cases will be made available via avb1597 or similar to ensure that the research is really reproducible and that the work can be replicated using test-data, test-code and the described and derived theory.

[8] Chaotic behaviour: small changes within a closed system leading to drastic changes

10 Project Milestone Deliverables

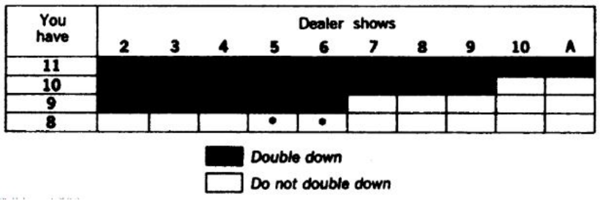
Date	Description	Deliverable
21-Apr	Supervisors' Topics	Supervisor 1
22-Apr	Topics released	Convenor 1
11-May	Project Allocation	Convenor 2
19 May , 1st June	Project Proposal	Student 1
03-Aug	Progress Report	Student 2
20 Oct to 22 Oct	Presentations	Student 3
09-Nov	Final hand-in	Student 4
TBD	Projects marked	Supervisor 2
TBD	Oral Defense	Student 5

Tab. 3: Key dates and deliverables for research project

11 Appendix

Figure 10, 11, 12 below relates to pictorial figures of hard doubling, soft standing and standing numbers adapted from [Thorp, 1966]:

Hard Doubling



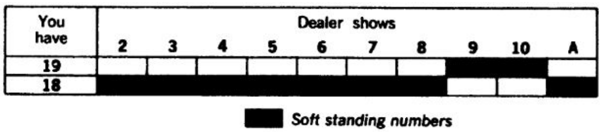
You have	Dealer shows									
	2	3	4	5	6	7	8	9	10	A
11	■	■	■	■	■	■	■	■	■	■
10	■	■	■	■	■	■	■	■	□	□
9	■	■	■	■	■	□	□	□	□	□
8	□	□	□	•	•	□	□	□	□	□

■ Double down

□ Do not double down

Fig. 10: Pictorial list of hard doubling numbers

Soft Standing

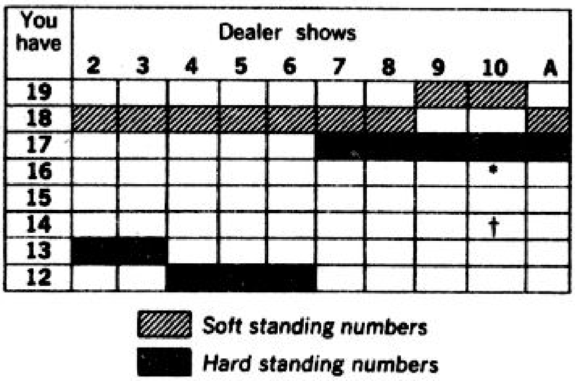


You have	Dealer shows									
	2	3	4	5	6	7	8	9	10	A
19	□	□	□	□	□	□	□	■	■	□
18	■	■	■	■	■	■	■	□	□	■

■ Soft standing numbers

Fig. 11: Pictorial list of standing numbers

Standing Numbers



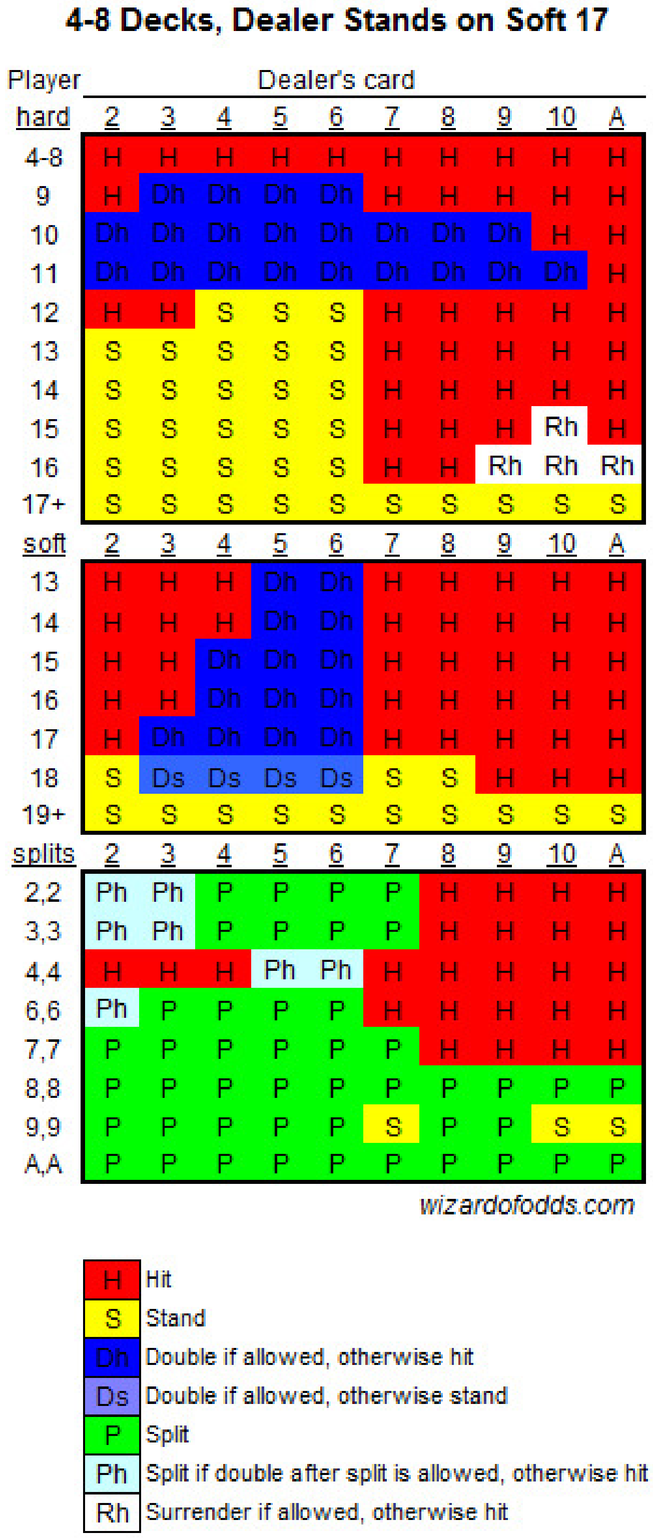
You have	Dealer shows									
	2	3	4	5	6	7	8	9	10	A
19	□	□	□	□	□	□	□	▨	▨	□
18	▨	▨	▨	▨	▨	▨	▨	□	□	▨
17	□	□	□	□	□	■	■	■	■	■
16	□	□	□	□	□	□	□	□	*	□
15	□	□	□	□	□	□	□	□	□	□
14	□	□	□	□	□	□	□	□	†	□
13	■	■	□	□	□	□	□	□	□	□
12	□	□	■	■	■	□	□	□	□	□

▨ Soft standing numbers

■ Hard standing numbers

Fig. 12: Pictorial list of soft standing numbers

The 2 following tables are the strategy tables adapted by [Shackleford, 2019a]. To use the basic strategy, a player looks up his hand along the left vertical edge and the dealer's up card along the top. In both cases an A stands for ace. From top to bottom are the hard totals, soft totals, and splittable hands. There are two charts depending on whether the dealer hits or stands on soft 17



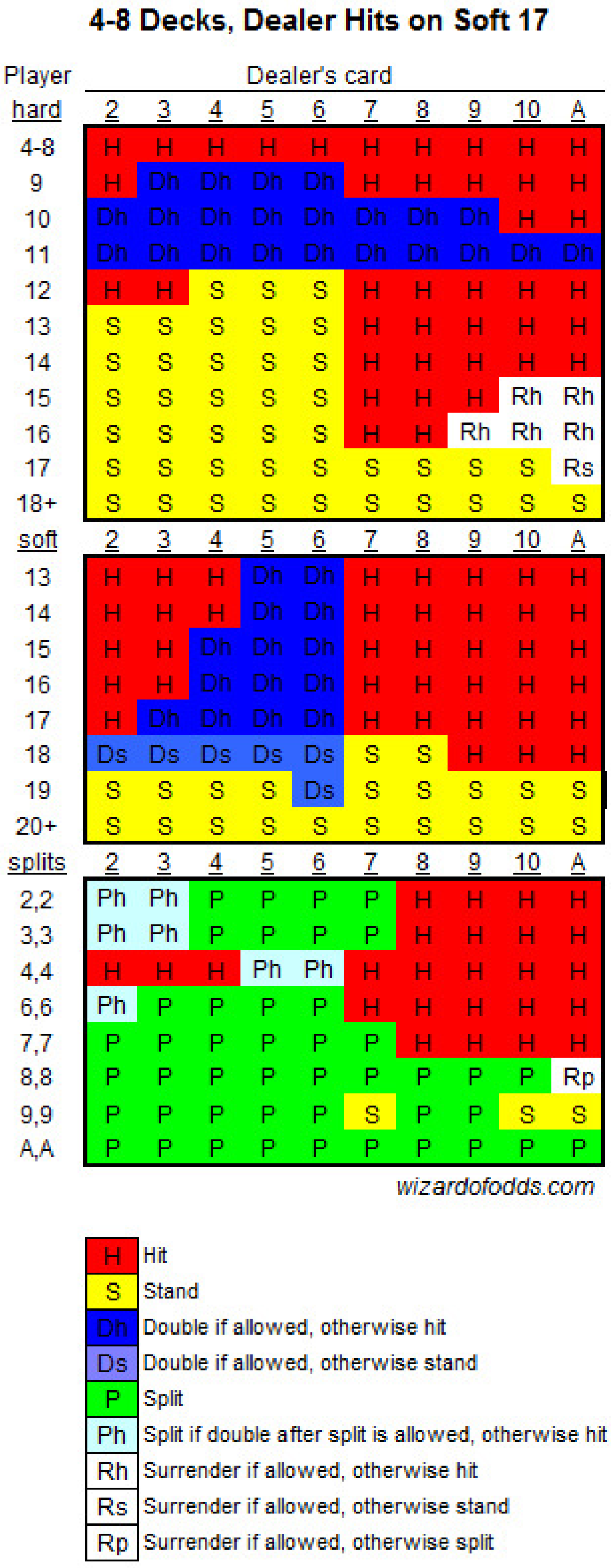
4-8 Decks, Dealer Stands on Soft 17

Player	Dealer's card									
hard	2	3	4	5	6	7	8	9	10	A
4-8	H	H	H	H	H	H	H	H	H	H
9	H	Dh	Dh	Dh	Dh	H	H	H	H	H
10	Dh	Dh	Dh	Dh	Dh	Dh	Dh	Dh	H	H
11	Dh	Dh	Dh	Dh	Dh	Dh	Dh	Dh	Dh	H
12	H	H	S	S	S	H	H	H	H	H
13	S	S	S	S	S	H	H	H	H	H
14	S	S	S	S	S	H	H	H	H	H
15	S	S	S	S	S	H	H	H	Rh	H
16	S	S	S	S	S	H	H	Rh	Rh	Rh
17+	S	S	S	S	S	S	S	S	S	S
soft	2	3	4	5	6	7	8	9	10	A
13	H	H	H	Dh	Dh	H	H	H	H	H
14	H	H	H	Dh	Dh	H	H	H	H	H
15	H	H	Dh	Dh	Dh	H	H	H	H	H
16	H	H	Dh	Dh	Dh	H	H	H	H	H
17	H	Dh	Dh	Dh	Dh	H	H	H	H	H
18	S	Ds	Ds	Ds	Ds	S	S	H	H	H
19+	S	S	S	S	S	S	S	S	S	S
splits	2	3	4	5	6	7	8	9	10	A
2,2	Ph	Ph	P	P	P	P	H	H	H	H
3,3	Ph	Ph	P	P	P	P	H	H	H	H
4,4	H	H	H	Ph	Ph	H	H	H	H	H
6,6	Ph	P	P	P	P	H	H	H	H	H
7,7	P	P	P	P	P	P	H	H	H	H
8,8	P	P	P	P	P	P	P	P	P	P
9,9	P	P	P	P	P	S	P	P	S	S
A,A	P	P	P	P	P	P	P	P	P	P

wizardofodds.com

H	Hit
S	Stand
Dh	Double if allowed, otherwise hit
Ds	Double if allowed, otherwise stand
P	Split
Ph	Split if double after split is allowed, otherwise hit
Rh	Surrender if allowed, otherwise hit

Fig. 13: Strategy Table 1



4-8 Decks, Dealer Hits on Soft 17

Player	Dealer's card									
hard	2	3	4	5	6	7	8	9	10	A
4-8	H	H	H	H	H	H	H	H	H	H
9	H	Dh	Dh	Dh	Dh	H	H	H	H	H
10	Dh	Dh	Dh	Dh	Dh	Dh	Dh	Dh	H	H
11	Dh	Dh	Dh	Dh	Dh	Dh	Dh	Dh	Dh	Dh
12	H	H	S	S	S	H	H	H	H	H
13	S	S	S	S	S	H	H	H	H	H
14	S	S	S	S	S	H	H	H	H	H
15	S	S	S	S	S	H	H	H	Rh	Rh
16	S	S	S	S	S	H	H	Rh	Rh	Rh
17	S	S	S	S	S	S	S	S	S	Rs
18+	S	S	S	S	S	S	S	S	S	S
soft	2	3	4	5	6	7	8	9	10	A
13	H	H	H	Dh	Dh	H	H	H	H	H
14	H	H	H	Dh	Dh	H	H	H	H	H
15	H	H	Dh	Dh	Dh	H	H	H	H	H
16	H	H	Dh	Dh	Dh	H	H	H	H	H
17	H	Dh	Dh	Dh	Dh	H	H	H	H	H
18	Ds	Ds	Ds	Ds	Ds	S	S	H	H	H
19	S	S	S	S	Ds	S	S	S	S	S
20+	S	S	S	S	S	S	S	S	S	S
splits	2	3	4	5	6	7	8	9	10	A
2,2	Ph	Ph	P	P	P	P	H	H	H	H
3,3	Ph	Ph	P	P	P	P	H	H	H	H
4,4	H	H	H	Ph	Ph	H	H	H	H	H
6,6	Ph	P	P	P	P	H	H	H	H	H
7,7	P	P	P	P	P	P	H	H	H	H
8,8	P	P	P	P	P	P	P	P	P	Rp
9,9	P	P	P	P	P	S	P	P	S	S
A,A	P	P	P	P	P	P	P	P	P	P

wizardofodds.com

H	Hit
S	Stand
Dh	Double if allowed, otherwise hit
Ds	Double if allowed, otherwise stand
P	Split
Ph	Split if double after split is allowed, otherwise hit
Rh	Surrender if allowed, otherwise hit
Rs	Surrender if allowed, otherwise stand
Rp	Surrender if allowed, otherwise split

Fig. 14: Strategy Table 1

12 References

[Barnes, 2015] Barnes, T. (2015). The q-matrix method: Mining student response data for knowledge.

[Choudhary, 2019] Choudhary, A. (2019). A hands-on introduction to deep q-learning using openai gym in python.

[Conrad, 2002] Conrad, Kirk A., S. B. (2002). Blackjack betting systems and strategies : the mathematics behind the game. Master's thesis, Muncie, Indiana.

[de Granville, 2005] de Granville, C. (2005). Applying reinforcement learning to blackjack using q-learning.

[Galla and Farmer, 2013] Galla, T. and Farmer, J. D. (2013). Complex dynamics in learning complicated games.

[Gaming and Bartending, 2019] Gaming, C. S. and Bartending (2019). Why does deck size matter in blackjack? https://crescent.edu/post/the-history-of-blackjack.

[Grabianowski, 2013] Grabianowski, E. (2013). How blackjack works? https://entertainment.howstuffworks.com/blackjack7.htm.

[Haney, 2008] Haney, J. (2008). They invented basic strategy. https://lasvegassun.com/news/2008/jan/04/the-inside-straight-they-invented-basic-strategy-j/.

[Hellemons, 1996] Hellemons, H. (1996). Can you still beat the dealer? Master's thesis, Amsterdam.

[Ofton, 1998] Ofton, L. (1998). The history of blackjack and card counting. https://www.blackjackapprenticeship.com/the-history-of-blackjack-and-card-counting/.

[Pagat, 2019] Pagat (2019). Blackjack. https://www.pagat.com/banking/blackjack.html.

[Patterson, 2002] Patterson, L. (2002). Harvey dubner: the forgotten man of blackjack. gambling times,. http://www.gamblingtimes.com/writers/jpatterson/jpattersonsummer2002.htm.

[Shackleford, 2019a] Shackleford, M. (2019a). 4-deck to 8-deck blackjack strategy. http://wizardofodds.com/games/blackjack/strategy/4-decks/.

[Shackleford, 2019b] Shackleford, M. (2019b). Introduction to the high-low card counting strategy. https://wizardofodds.com/games/blackjack/card-counting/high-low/.

[Smith, 2015] Smith, K. (2015). Why does the number of decks matter in blackjack? https://www.blackjackinfo.com/why-does-the-number-of-decks-matter-in-blackjack/.

[S.Sutton and Barto, 2014] S.Sutton, R. and Barto, A. G. (2014). *Reinforcement Learning: An Introduction.* MIT Press.

[Tamburin, 2006] Tamburin, H. (2006). How the mit students beat the casinos at blackjack.

[Tan, 1993] Tan, M. (1993). Multi-agent reinforcement learning: Independent versus cooperative agents.

[Thorp, 1966] Thorp, E. O. (1966). *BEAT THE DEALER. A WINNING STRATEGY FOR THE GAME OF TWENTY ONE.* Vintage; Revised edition.

[Thorp, 2006] Thorp, E. O. (2006). The kelly criterion in blackjack sports betting, and the stock market. *Handbook of Asset and Liability Management*, pages 385–428.

[Wong, 1994] Wong, S. (1994). *Professiona Blackjack.* Pi Yi Press.

[Yoon, 2019] Yoon, C. (2019). Vanilla deep q networks.