Learning to Play Leduc Hold'em Using Reinforcement Learning

UNIVERSITY *of* LIMERICK

OLLSCOIL LUIMNIGH

Department of CSIS

Bachelor of Science in Computer Systems

Author: Jamie Mac Manus

ID: 15147312

Supervisor: J.J Collins

Abstract

In recent years the area of machine learning has gained a lot of ground in a broad range of areas. A particularly interesting problem pertaining to machine learning is how we can develop useful AIs in a more hands off fashion. This problem is addressed by a machine learning paradigm named reinforcement learning. Reinforcement learning allows us to set up an agent in an environment after which the agent can explore the environment and begin to learn which actions that it should take in the different scenarios it can find itself in. This avenue of machine learning is suited to a broad range of problems but in this project we will explore the possibility of applying it to the game of texas hold'em.

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

Since the inception of machine learning, games have been a key problem area that has seen a lot of focus from top academics. For decades games have been used as a platform to test and develop algorithms that have gone on to provide invaluable services that are used in peoples everyday lives. The ability to contribute to this great history was a large motivator when it came to choosing this project.

Although this report will, to an extent, discuss machine learning and how it applies to games in general, the primary focus will be on machine learning techniques when applied to texas hold'em, or variations of it. In the past, methods such as Counterfactual Regret Minimization (CFR) have been used to develop agents that can play texas hold'em to a superhuman level. CFR is an algorithm that allows computation of a strategy through self-play. The metric used to update these strategies is called regret, which measures the difference between the game's outcome and the outcome that could have been achieved if some other action was taken. A large number of simulated games are played, with regret being calculated each time and subsequently being used to compute a strategy that is optimal. One example of CFR being used was the 2018 Annual Computer Poker Competition (ACPC) champion,

slumbot(Jackson 2013).

When applied to large imperfect information games such as texas hold'em CFR generally relies on the creation of abstractions of the game. The CFR algorithm is applied to these abstractions, generating a strategy. The generated strategy is then applied to the full version of the game and, if the original abstraction was accurate, our strategy will work well. This obviously requires a high degree of knowledge of the game in order to create an accurate abstraction.

There have also been attempts to tackle texas hold'em using a machine learning paradigm called reinforcement learning (RL). The great advantage of these approaches is that they do not require game abstractions or the associated domain knowledge.

RL is a way of programming agents by reward and punishment without needing to specify how the task is to be achieved(Kaelbling et al. 1996). RL problems consist of an agent in an environment. The environment consists of a number of states and rewards. The agent is allowed to take certain actions, in certain states. The overall goal is to learn a strategy that will maximise the cumulative reward. On the surface this seems like a perfectly reasonable methodology for solving games such as texas hold'em. However, texas hold'em is an imperfect information game. This means that we do not know the entirety of the state information at any given time i.e we do not know the values of the opponents cards. Thus, from a RL perspective, we do not know the actual state from which we are choosing actions. This makes pure reinforcement learning strategies impractical for these types of games.

However, there has been some success when more custom reinforcement learning methods have been implemented. In one case linear programming techniques and RL were used in order to tackle a simplified version of the game(Dahl 2001). In another RL was combined with techniques inspired by game theory(Heinrich et al. 2015). This latter approach will be the basis for our texas hold'em agent as we attempt to replicate and build upon the results outlined in this paper.

The different approaches will be discussed in greater detail in the background section.

1.2 Objectives

1.2.1 Primary Objectives

Although this project will be largely research based, the primary goal is to create a texas hold'em playing agent. Due to the fact that texas hold'em has an extremely large state space, combined with the fact that it is an imperfect information game, the initial goal will be to tackle a simplified version of the game. Specifically Leduc Hold'em will be used for this version of the project. This version of hold'em consists usually of a six card deck and only one private card, compared to two in texas hold'em.

In(Heinrich et al. 2015) a metric called exploitability was used. This is a measure of how the agent's strategy fares against the best responses to that strategy. For Leduc Hold'em, with a 6 card deck and 300 seconds of training time the initial exploitability was slightly over 1.2 and descended to under .5. When the deck size increased to 60 cards the exploitability began at over 5.5 and converged to roughly 1.2 in the same period of time. As such the success criteria for our initial iteration of the game is to replicate these results, with an allowance for hardware differences that may impact computational speed.

If the success criteria for this simplified version of the game are met, we will then proceed to tackle a more complex version of the game. In this second iteration of the project we will tackle limit texas hold'em with the end goal of recreating the results shown by Heinrich in his second paper on the matter(Heinrich & Silver 2016). In this paper, rather than using exploitability as an evaluation method, the agent was simply compared against the best hold'em agents from the ACPC of the previous year. Thus win rate was used as the evaluation metric. More specifically the measure used was mbb/h or milli big-blinds per hand. This is a measure of the number of big blinds won or lost per thousand hands. Note that the big blind is the larger of the two

mandatory bets required at the beginning of each hand. Against the top three competitors in the ACPC, the fully trained agent achieved a win rate of between -50 and -15 mbb/h. This score is consistent with that of a player that is slightly inferior but still competitive with these superhuman poker agents. As such, if the second iteration of our project is completed the goal will be to have a win rate of -200mbb/h or better against these same agents. Again we give an allowance for the difference in hardware used to train the agent as well our limited time.

It is also my goal to create a product that will be fun and useful for the general public. As such another objective will be to create a website that will allow users to play heads-up against the final product.

1.2.2 Secondary Objectives

As this project is very specific and academic, one of the larger challenges will be to gain a strong knowledge of the domain. This means learning the history of RL, the types of problems that it has been used to solve and the specific details of different RL algorithms.

A successful project will require a high degree of knowledge from the broader domain of RL. However, it is also the case that I must become closely familiar with the existing academic literature in the area of RL with respect to imperfect information games. This will allow me to avoid taking approaches that have previously shown to fail as well as allow me to add value to the existing literature whether that be through literature review or through my own experimental findings.

1.3 Contribution

1.4 Methodology

1.5 Motivation

For the last number of years I have played poker recreationally with friends or online. It became more of an interest of mine as I started to explore the mathematical basis for the game and how players could use their knowledge, intelligence and temperament in order to gain an advantage in a game that, on the face of it, seemed to be largely based on chance. I spent some of my free time researching different aspects of the game. This included gaining some basic knowledge like the probability of making drawing hands as well as learning more technical aspects of the game such as how to calculate expected value of hands, or how to narrow down one's opponent's range of possible hands.

Concurrent with the development of this interest I was also becoming more and more interested in the area of machine learning. Machine learning and the development of artificial intelligence is possibly the most glamourized area of computer science. However, this is probably for good reason because there is something intrinsically interesting about machines that can learn to solve a problem on their own, without direct instructions from a human. The fact that machine learning has made so many strides in recent years was another cause of interest in this area of computer science, especially as the practical viability of ML as means of tackling a wide array of problems in industry continues to increase.

As a result the merging of these two interests as the basis for my final year project seemed like an obvious choice.

Chapter 2

Background

The aim of this chapter is to give the reader background information on certain areas of machine learning and game theory in order for them to understand the rest of the report. There will also be a in-depth discussion of existing literature that relates to machine learning in texas hold'em agents.

Machine learning is an area of computer science that tackles how we construct computer programs that improve with experience(Mitchell et al. 1997). The term was coined by Arthur Samuel in a paper in which he discussed machine learning methods using checkers(Samuel 1959). Since then there has been a great deal of advancement in the field. Some of the notable early contributions being the discovery of recurrent neural networks in 1982, the advancement of reinforcement learning by the introduction of Q-Learning in 1989 and the development of a backgammon-playing agent using neural networks and temporal difference learning(Tesauro 1995). Recently we have seen some of this early academic work culminate in more practical achievements such as Facebook's DeepFace system which, in 2014, was shown to be able to recognise faces at a rate of 97.35% accuracy, a rate that is comparable to that of humans. Another example of recent achievement is Google's AlphaGo program which, in 2016, became the first program to beat a professional human player.

It should be becoming clear that machine learning can be a solution to

a wide array of problems and as both hardware and software continue to improve it's reach will only continue to grow. We are starting to see machine learning systems become a key component of many companies business model. Since certain machine learning techniques are great at prediction, machine learning has been widely used for content discovery by companies such as Google and Pinterest. Other business applications include the use of chatbots as a part of customer service, self-driving cars and even in the field of medical diagnostics.

2.1 Reinforcement Learning

The early research for this project yielded reinforcement learning as the most suitable machine learning paradigm for the problem of texas hold'em. However, in order to understand both reinforcement learning and how it would apply to the chosen problem, in depth research was required. This research included a Udemy course(LazyProgrammerInc. 2018) as well as reading in part the book Reinforcement Learning: An Introduction(Sutton et al. 1998).

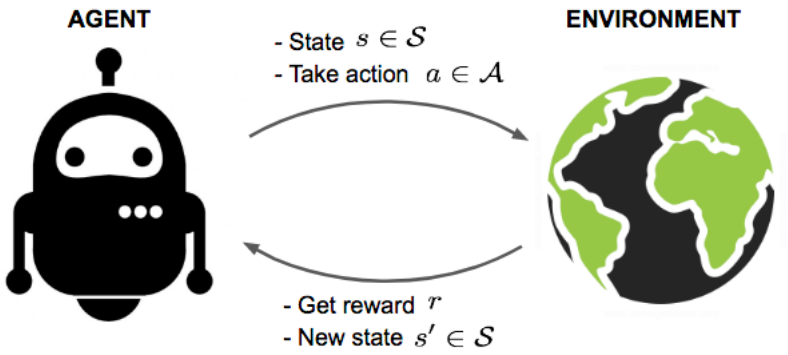

Figure 2.1: Reinforcement Learning

As mentioned in the introduction, reinforcement learning is a method of programming agents by reward and punishment without needing to specify how the task is to be achieved. As such the primary components of a reinforcement learning problem are an agent which exists in an environment.

From a simplified perspective we can think of the environment as a set of states, actions and rewards. The objective for the agent is to maximise cumulative reward. This is done by developing a policy that will dictate which actions should be taken in each state.

2.1.1 Explore-Exploit Dilemma

When it comes to reinforcement learning one of the first questions that we have to ask is how we explore the state space. An example that is often used to conceptualize this problem is the multi armed bandit problem. Let's say an agent is in a room with a number of gambling machines. Each of these machines has an arm that, when pulled will return a reward of 0 or 1 based on some underlying probability(Kaelbling et al. 1996). The agent has a limited number of total pulls. So the question becomes how do we distributed these pulls in order to maximise return? Well, first we have to ensure that we explore enough that we find the machine with the best reward probability and second, we must then exploit this machine to the best of our abilities.

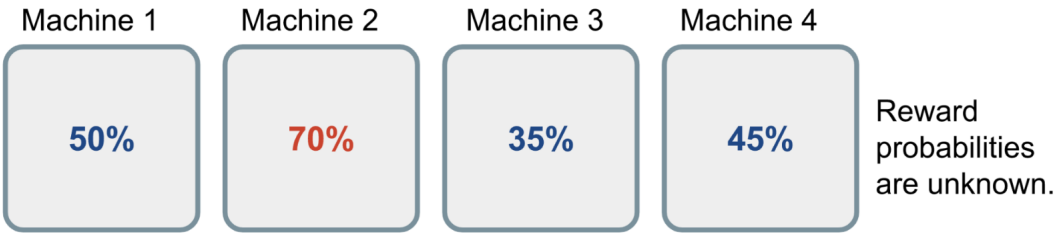

Figure 2.2: Multi Armed Bandit

There are a number of approaches that can be taken to solve this problem, we will now briefly discuss two of these methods.

ϵ-Greedy Solutions

The first approach that we will discuss is the ϵ-greedy strategy. This approach was first proposed in(Watkins 1989) and is a very simple and widely used method. The ϵ-greedy strategy involves choosing a random lever some

proportion ϵ of the time, and choosing the lever that has been established to give the highest reward the rest of the time.

There are a number of variations of this method, the first being the ϵ-first strategy. With this strategy we take all of our random choices first, allowing us to establish the best bandit, after which we exploit this bandit. However, as stated in(Vermorel & Mohri 2005) this simple approach is sub-optimal because asymptotically, the constant factor prevents the strategy from getting arbitrarily close to the optimal lever.

This is where the ϵ-decreasing strategy becomes useful. Here, the proportion of random lever pulls decreases with time. Generally if our initial epsilon value is ϵ_0 then our epsilon value at time t will be $\epsilon_t = \frac{\epsilon_0}{t}$.

Interval Estimation Strategy

Another approach that can be used is called the interval estimation strategy. With this method we initially give an optimistic estimate of the reward to each bandit within a certain confidence interval. Then we simply take a greedy approach to our exploration. Less explored bandits will have a artificially higher reward estimate and thus they will be greedily chosen, thus allowing us to evaluate each of the bandits.

In the context of reinforcement learning, state space exploration through the ϵ-greedy approach is generally sufficient.

2.1.2 Markov Decision Processes

In the last section we have established some methods that can be used to explore environments. We will now discuss in more detail how reinforcement learning environments, and their interaction with reinforcement learning agents, are modelled. Generally finite Markov decision processes (finite MDPs) are used. Markov decision processes provide a formal mathematical framework for sequential decision making, where actions influence immediate rewards as well as subsequent situations(Sutton et al. 1998). MDPs allow us

to create an idealized model of reinforcement learning problems and thus we can make precise theoretical statements.

MDP Dynamics

As mentioned earlier, reinforcement learning problems consist of an agent and an environment interacting. Markov decision processes can be looked at in a similar way. However, there are a number of additional factors that we must consider in order to paint a complete picture.

We can think of the problem as consisting of a set of discrete time steps. At each time step the environment supplies the agent some information about the state, S_t. Using this information the agent chooses an action, A_t. Then, as a result of the action, the environment will supply the agent with a reward, R_t, as well as a new state. As such the process of interaction between the agent and the environment can be seen as a trajectory of states, actions and rewards(Sutton et al. 1998):

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2 \dots \tag{2.1}$$

Thus far we understand that states, actions and rewards are related, however questions still exist as to the exact workings of this relationship. The answer is that finite MDPs contain a discrete probability distribution that determines the likelihood that we will reach the state s' and receive reward r at time t based on the previous action a and state s:

$$p(s', r|s, a) \doteq \Pr\{S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a\} \tag{2.2}$$

In simplified terms this means that for a given state-action pair (s, a), the probability of us reaching some new state s' and receiving reward r is determined by the MDPs probability function Pr.

This four argument probability function completely characterizes the dynamics of the MDP and from it anything else we want to know about the environment(Sutton et al. 1998).

MDPs and Learning

The goal of the agent in an MDP is to learn how to maximise the cumulative reward received when traversing the environment. In some cases we will traverse the MDP until we reach some terminal state, T. This type of MDP reflects episodic tasks that will always terminate. In this case we can calculate the cumulative reward,G_t as follows:

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T \tag{2.3}$$

However, in other cases we will model continuous tasks. The problem here is that, if we use the same method of calculating G_t as we do for episodic tasks then in this case G_t will always eventually sum to infinity, regardless of whether we are taking good or bad actions. As such we must introduce the a new concept called discounting. With this approach the aim is to maximise the sum of future discounted rewards. Thus γ, a parameter with a value between 0 and 1, is introduced. As such, for continuous tasks modelled as MDPs our cumulative reward is as follows:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \tag{2.4}$$

Based on our specified value of γ we can alter the weighting of future rewards. For example if we have a low value for γ (eg .5) then the value of rewards more than a few time steps in the future will be very low.

Partially Observable Markov Decision Processes

A partially observable Markov decision Process (POMDP) is used to model environments that are not fully observable(Kaelbling et al. 1996). POMDPs extend MDPs by including a set of observations, O. There is also an observation function, $P(O_t = o|S_t = s)$, that determines the probability of making a certain observation in a certain state. Hold'em can be modelled as a POMDP due to the fact that is an imperfect information game.

However, it is possible that an agent in a POMDP environment can remember the sequence of observations and actions that lead to the current

state. This is a sufficient statistic of it's experience and can thus define a complete information state(Heinrich 2017). As such we can reduce the POMDP to it's underlying MDP by using these full history information states and also extending the relevant transition and reward functions.

2.1.3 Policy Evaluation and Policy Improvement

As mentioned above the primary focus of reinforcement learning is to find a policy (denoted by π) that allows the agent to take actions in states that lead to the maximum possible reward. There are two primary problems that we must solve in order to do so.

The first is called the prediction problem, or policy evaluation. This involves computing the values of states given some arbitrary policy(Sutton et al. 1998). For example a state would have a high value if the reward for reaching that state was high. A state would also have a high value if we were only one action away, according to the supplied policy, from a state that renders a high reward. However a state would have a low value if, according to the policy, there was no path to a state that would return a positive reward in the foreseeable future.

The second problem is known as the control problem, or policy improvement. This involves changing the policy in order to improve our cumulative reward. The policy improvement process can only occur when the we have performed policy evaluation. Let's say, after our evaluation step, we know the value of some state s. Note that this value is calculated with the condition that we take some action a in state s. But, if we take some other action a' would this render a higher value for s? If the answer is yes then we update the policy.

These two operations can be seen as the core of reinforcement learning. In the next section we will discuss different reinforcement learning methodologies. Some of the main differences are in how method each approaches the prediction and control problems.

2.1.4 Dynamic Programming

Dynamic programming (DP) refers to a collection of algorithms that can be used to compute optimal policies given a perfect model of the environment as a Markov decision process(Sutton et al. 1998). DP is not widely used in practical reinforcement learning applications due to its assumption of a perfect MDP and it's high computational requirements. Despite this it is very important from a theoretical standpoint as it serves as an introduction to a number of important reinforcement learning concepts. Furthermore, it provides a basis for many algorithms that are used in practical reinforcement learning applications.

Policy Evaluation in Dynamic Programming

When discussion policy evaluation we talk about a state-value function or a value function. This is simply the mapping of states to their corresponding values and is denoted by v.

Since the environments's dynamics are completely known we can apply an iterative solution to finding the value function. If we consider a series of approximate value functions $v_0, v_1, v_2, ...$ The initial value function, v_0 is chosen arbitrarily and each successive generation is obtained by using the Bellman equation for v_π as an update rule(Sutton et al. 1998):

$$v_{k+1}(s) = \mathbb{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1})|S_t = s] \tag{2.5}$$

In order to produce each successive approximation of v_{k+1} from v_k we apply the operation outlined to each state s. As shown above our new value for s is based on a combination of the expected immediate rewards(R_{t+1}), and the expected values of each of the states that we can transition to(S_{t+1}) given policy π. It can be shown that as $k \to \infty$ v_k will converge to v_π, the correct value function for policy π. This algorithm is called *iterative policy evaluation*(Sutton et al. 1998).

Policy Improvement in Dynamic Programming

Since we have now determined how good it is to follow v_π we can use this information to determine how we should modify this policy in order to improve it's value. If we assume that π is a deterministic policy then $\pi(s)$ will return some action that we must take. Now the question becomes what if we take some other action $a \neq \pi(s)$? Well we must consider whether or not choosing this action, and then continuing to use the existing policy will improve the value of the policy. If it does then we will choose this new action.

The logical extension of this approach is to apply it to each state and each possible action. As such we will select what appears to be the best action at each state. We can thus denote our new greedy policy π' as:

$$\pi'(s) = argmax\, \mathbb{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1}) | S_t = s, A_t = a] \tag{2.6}$$

Essentially here we are determining the value of each available action in the current state, using the same operation as outlined in the policy evaluation phase. Then the argmax function will select the action with the highest value. Finally we assign this action to be the one we will choose in state s, according to the new policy π'.

Note that in this section we have outlined an algorithm with respect to deterministic policies, however a lot of times in reinforcement learning we deal with stochastic policies. This means that we take actions in states according to some probability distribution rather than always choosing the same action in a particular state. This is not a problem as the ideas mentioned apply equally well to stochastic policies.

2.1.5 Monte Carlo

In Monte Carlo, unlike dynamic programming, we do not assume complete knowledge of the environment. Monte Carlo methods require only experience. We sample sequences of states, actions, and rewards from interaction with the environment(Sutton et al. 1998). These sequences are called episodes.

Monte carlo evaluation is an episodic process this means that we only update our action values after an episode has completed.

Monte Carlo Policy Evaluation

In Monte Carlo methods we take a fundamentally different approach to policy evaluation. As mentioned this method is focused on using episodic experience. In order to evaluate a state we can simply average the rewards returned after visiting that state. As we observe more returns the average will converge to the actual expected value of the state.

It is worth noting that a state s could be visited more than once in an episode. As such we can either average the returns following the first visit to s or we could average the returns after each visit to s. These two methods are called *first-visit* and *every-visit* respectively.

Monte Carlo Action Values

In Monte Carlo methods, the lack of a model means that we cannot use only state values in order to obtain a policy. Rather state-action pairs are generally evaluated. This mapping of state-action pairs to values is referred to as the q function or the action value function. In this case the evaluation problem for action values is to estimate $q_\pi(s, a)$, the expected return when starting in state s, taking action a, and thereafter following policy π(Sutton et al. 1998).

The method for policy evaluation using state-action pairs is almost identical to that outlined above. The only difference being that instead of averaging rewards for each state, we average rewards for each action taken when a state is visited. There is one problem with this approach in the context of deterministic policies. The problem being that in following a deterministic policy we will only receive returns for a single action, thus only one action value estimate will be improved. In order to negate this problem we can specify that every episode begin in a state-action pair, with the probability of starting in each state-action pair being non-zero. This is called the *exploring-starts*

method. Another approach would be to ensure that we are using a stochastic policy with the probability of selecting each action being non-zero.

Monte Carlo Policy Improvement

In Monte Carlo methods, the overall policy improvement algorithm is the same as outlined in the dynamic programming section. That is, we alternate between modifying the value function to more closely approximate the current policy, and using the value function to improve the policy.

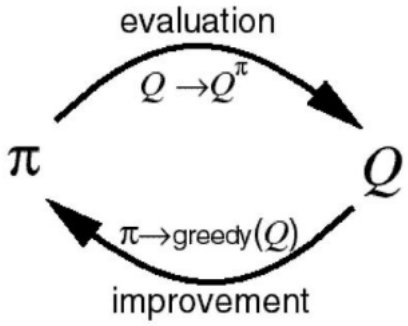

Figure 2.3: Monte Carlo Policy Improvement

2.1.6 Temporal Difference Learning

The final reinforcement learning method we will discuss is the Temporal Difference(TD) learning method. TD learning combines dynamic programming(DP) and and Monte Carlo(MC) ideas(Sutton et al. 1998). Like MC, we can learn directly from experience, without a model of the environment's dynamics. However, like DP we update state value estimates based on other learned estimates, without needing to wait for an episode to complete and the return of some final outcome.

The selective use of different aspects of these reinforcement learning methodologies by TD learning has a number of advantages. Obviously the fact that a model of the environment is not needed makes it easier to implement TD

methods compared to DP . TD methods are also conducive to solving problems with long episodes or even continuous tasks with no episodes at all due to the fully online nature of this learning algorithm.

TD Learning Policy Evaluation

Unlike MC, with TD learning we need only wait until the next time step in order to update the value function. This is exemplified by the *TD(0)* or *one-step TD* method in which we make the update immediately on transition from state S_t to state S_{t+1}. The more general case of this method is the *TD(λ)* or *n-step TD*. With *TD(0)* the update rule is as follows:

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]. \tag{2.7}$$

As such the new value for some state S_t is dependant on the previous value of the state($V(S_t)$), along with the reward(R_{t+1}) gained from transitioning to state S_{t+1} plus the discounted(γ) estimated value of that state($V(S_{t+1})$). The sum of the latter is multiplied by α which is a small positive fraction that influences the learning rate.

This rule is applied for each state visited in an episode and for each episode.

TD Learning Policy Improvement - SARSA

At this point it is worth noting that there are two distinct methods of handling policy improvement. The first is on-policy and the second is off-policy. On-policy reinforcement learning is when the policy being evaluated or improved is the same policy that is used to make decisions. In off-policy reinforcement learning the policy being used to generate behavior is not the same as the policy being evaluated or improved.

SARSA is an example of an on-policy algorithm. The policy improvement mechanism is the same here as outlined in the previous sections. With SARSA, like in Monte Carlo we utilise the action-value function($q_\pi(s, a)$)

rather than the state-value function. Thus the policy evaluation step is slightly modified as follows:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]. \quad (2.8)$$

This update utilises the quintuple of events, $(S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1})$, which is where the name SARSA originates.

TD Learning Policy Improvement - Q-Learning

Q-Learning is an off-policy control algorithm. This was one of the early breakthroughs in reinforcement learning as it allows the direct approximation of the optimal action value function independent of the policy being followed. The update rule is as follows:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma max Q(S_{t+1}, a) - Q(S_t, A_t)]. \quad (2.9)$$

Here our update rule is similar to that of SARSA apart from the fact that we use the value of the best action available ($maxQ(S_{t+1}, a)$) when updating our action value.

2.1.7 Monte Carlo Tree Search

In this section we will discuss Monte Carlo Tree Search(MCTS), a decision time planning algorithm. This algorithm uses tree search combined with value estimations from MC simulations in order to efficiently solve very large MDPs. One application of this algorithm was in DeepMind's AlphaGo(Silver et al. 2016) which became the first computer program to beat a highly ranked Go champion.

2.2 Game Theory

After taking a deep dive on reinforcement learning and the papers surrounding RL in texas hold'em it became clear that a pure reinforcement learning

approach would not be feasible. The main reason for this was the fact that texas hold'em is an imperfect information game. As outlined by(Dahl 2001):

> Note that the concept of game state values, which is the key to solving perfect information games, does not apply to imperfect information games, because the players do not know from which game states they are choosing.

As such it became apparent that some other techniques would have to be incorporated in order to create a competent texas hold'em agent. Since our implementation will be following the method discussed by(Heinrich 2017) the following section will give the requisite background in order to facilitate later discussion of this method.

2.2.1 Modelling Games

When we attempt to solve a game, we must first understand how we are to model that game. We must also carefully consider how we define a form or structure of the model(Myerson 2013). In this section we will discuss three important forms that games can take and the varying utility of each in different scenarios.

Extensive Form Games

Extensive form games are a model of players sequential interaction, with explicit representation of a number of key aspects of the game. They were formally defined in(Kuhn 1953) as consisting of the following components:

- N - a set of players.
- S - a set of states that represent the nodes of a rooted game tree.
- $A(s)$ - a set of actions for each state representing the edges to the following states.
- U - a set of information states (one set for each player).

- *Player Function* - determines who is to act at a given state.
- *Information Function* - determines which states are indistinguishable for the player by mapping them to the same information state.
- *Return Function* - maps terminal states to each player's return/payoff.

Extensive form games allow us to richly describe game situations. This allows us to uncover characteristic differences between games and the structural features which determine these differences(Kuhn 1953).

We can also define behavioral strategies for players, which consist of a probability distribution over actions given information states(Heinrich 2017). This is denoted by $\pi^i(a|u^i)$. If we have a collection of strategies for all players in the game then this is called a strategy profile, π. π^{-i} is the set of strategies in π not including π^i. This will be discussed later when we discuss Heinrich's implementation in more depth.

Normal Form Games

Normal form games are represented by way of a matrix. Although some information is lost in comparison to extensive form games, normal forms are better suited to the derivation of generalized theorems(Kuhn 1953) and thus have their own utility.

An extensive-form game can induce an equivalent normal form of the same game. This can be done through the generation of a set of deterministic behavioral strategies for each player called pure strategies. Each pure strategy is a full game plan that will determine an action for each situation the player may encounter. We can also create mixed strategies which define a probability distribution over the players pure strategies. We can denote a mixed strategy for player i as Π^i. When we restrict the extensive-form return function to normal-form we yield an expected return function. The expected return function for some mixed strategy profile Π is $R^i(\Pi)$(Heinrich 2017).

Sequence Form Games

In order to compute a Nash equilibrium for an extensive form game we can convert the game to a normal form, however normal forms tend to produce very large game trees. The sequence form is an efficient method of representing extensive form games(Koller et al. 1996). This representation is described as a linear-sized strategic description of the game. It decomposes players strategies into sequences of actions and probabilities of realizing those sequences.

In sequence form games, for every player $i \in N$, each of their information states $u^i \in U^i$ uniquely define a **sequence** σ_{u^i} of actions that a player must take in order to reach that information state. These sequences are then mapped to realization probabilities through what is called a realization plan, denoted by x. When two or more strategies have the same realization plan we consider these strategies to be realization equivalent(Von Stengel 1996). This can apply across different types of strategies for example extensive-form, behavioral strategies and normal-form, mixed strategies(Kuhn 1953).

2.2.2 Nash Equilibria

A Nash's equilibrium is a state in which each player in a game has chosen a strategy and none of the players can benefit from changing their strategies, if the other player's strategies remain unchanged. As such, if we reach a strategy that induces a Nash's equilibrium our strategy can no longer be exploited.

In the context of extensive form games the concept of best responses is related to that of Nash's equilibria. If the opponent's strategies are denoted by π^{-i} then the set of best responses are denoted by $BR^i(\pi^{-i})$. Note that if we have a strategy profile π such that $\pi^i \in BR^i(\pi^{-i})$ for every $i \in N$ (i.e for every player) then that game constitutes a Nash equilibrium(Heinrich 2017). Heinrich also discusses the concept of ε-best responses. This is the set of strategies that are within a certain tolerance ε of the best responses. As

such we can define an ε-Nash equilibrium as a strategy profile π such that $\pi^i \in BR^i_\varepsilon(\pi^{-i})$ for all $i \in N$.

2.2.3 Fictitious Play

Fictitious play(FP) is a game-theoretic model of learning through self-play. At each iteration players choose the best responses to their opponents average strategies(Heinrich 2017). These strategies converge to Nash equilibria in certain classes of games, including two-player, zero sum games.

Generalised weakened fictitious self-play (GWFSP) is a method that is built on FP but allows for approximations in players strategies(Leslie & Collins 2006). Thus it is more suitable for machine learning. GWFSP allows for a certain error at each iteration of the algorithm and relies on the fact that this error rate will tend towards zero as time progresses. In the the research done by Leslie and Collins normal form games were studied. There has also been research done into the applicability of FP to extensive form games however, before Heinrich there was no method shown to converge for imperfect information games such as poker. In later sections we will discuss how Heinrich utilised GWFSP as a basis for neural fictitious self-ply, a method that has shown success in games of imperfect information.

2.3 Supervised Learning

Supervised learning was also a common theme in the RL based texas hold'em implementations so once again a brief study of this area was conducted.

Supervised learning involves an agent which observes some example input-output pairs and learns a function that maps from input to output.(Russell & Norvig 2016). This learned function can then be used on new input data, that wasn't used to train the agent and the agent should be able to give an accurate output. The agent must be able to identify general features of the input data and how they map to the output. Common applications of supervised learning include computer vision, speech and pattern recognition

as well as spam detection.

2.4 Texas Hold'em

Texas hold'em is one variant in a family of games called poker. Poker is a group of card games that combine gambling, strategy and skill. All poker variants have three core similarities. There is betting involved, there is imperfect information (ie cards remain hidden until the end of a hand) and the winner is determined by combinations of cards.

2.4.1 Game Structure

Texas hold'em consists of four betting rounds. Initially each player is dealt two private cards. These remain face down and only the person who received these cards may view them. In the next three rounds five public cards are dealt face up on the table. The second round of dealing is called the flop, where three public cards are dealt. The third round is called the turn where one additional public card is dealt. Finally in the fourth round another public card is dealt which is called the river.

At each round, after the cards are dealt, the players are given the opportunity to take a number of betting related actions. We will discuss the permitted actions in the next section.

In order for players to be incentivized to continue playing in a wider array of situations, blinds are required. Blinds are a mandatory bet that must be posted by two of the players present at the game. These two bets are called the big blind and the small blind, the big blind being twice that of the small blind. As hands are played the big and small blinds are posted by different players in order to distribute the cost fairly.

The big and small blind are the first two bets that contribute to what's called the pot. The pot is the collection of all of the current chips bet by the players. When a player wins a hand then what they receive in return is the pot.

The final structural component of the game is player stacks. Each player will start the game with a certain amount of chips. If a player wins a pot then all of the chips in the pot are transferred to the winners stack.

2.4.2 Actions

As mentioned in the previous section, after cards are dealt players are permitted to take a number of actions. If a player is the first to act the they may either check or bet a chosen amount. If a player is not first to act and the previous player has made a bet then they may choose to either fold and forfeit the pot, call the bet by adding the same amount to the pot, or raise by adding the amount previously bet plus some additional chips. Players can go back and forth with bets until they run out of chips in which case they are considered to be "all in".

2.4.3 Hand Values

In poker the best 5 cards available to the player can be played. This means any combination of his own private cards and the public cards can be used. There are 10 major poker hands. These are listed below in ascending order of value:

1. **High Card:** None of the higher hand values achieved, highest card plays.
2. **Pair:** Any two cards of the same rank.
3. **Two Pair:** Two different pairs.
4. **Three of a kind:** Three cards of the same rank.
5. **Straight:** Five cards in a sequence.
6. **Flush:** Five cards of the same suit.
7. **Full House:** Three of a kind with a pair.

8. **Straight Flush:** Five cards in sequence, all of the same suit.

9. **Royal Flush:** A, K, Q, J, 10 - all of the same suit.

2.4.4 Leduc Hold'em

Leduc Hold'em is a simplified version of Hold'em that was first introduced in(Southey et al. 2012). In Leduc Hold'em the deck is reduced to six cards with two suits and three ranks in each suit. Rather than four rounds there are only two. In the first round a single private card is dealt to each player. In the second round a single board card is revealed. In the first round both players have a mandatory bet of one and a raise of two is allowed. In the second round players can raise by four. Each round allows for at most two bets.

2.5 Variations of MCTS applied to poker

As mentioned MCTS has been shown to be a very powerful method for solving large perfect-information games such as Go. However if we wish to apply this algorithm to imperfect information games like poker then a number of modifications must be applied. Such an approach was outlined in (Heinrich 2017) in which a variation of MCTS called smooth UCT was implemented to tackle both leduc hold'em and limit hold'em.

One of the subtleties of poker is that player information is asymmetric. In other words each player has access to their own private card information but not to their opponents private cards. This means that we cannot represent the search tree as a single, collective entity(Heinrich 2017), rather two search trees must be available to accommodate self-play. The method used by Heinrich to accomplish this goal was the creation of an information function $I^i(s)$. This function will return the information state u^i of the current player (i) given the current state s from the overall game tree. Note that the overall game tree will have separate nodes for any variation in either players

cards. However, player 1's game tree will not have separate nodes where the only differentiating factor is player 2's private cards(Johanson et al. 2011) due to the fact that this information is not available. This means that a number of nodes that are separate in the overall game tree will be grouped in either players individual game trees.

In order to understand smooth UCT we must first explain the meaning of UCT. The abbreviation UCT refers to upper confidence bound(UCB) applied to trees. This means that in the action selection portion of the algorithm a UCB approach is taken, where less explored nodes in the tree are given a positive bias in value. This means that unexplored nodes will be explored which ensures that we discover the best actions at each position in the tree. The value of an action is denoted as follows:

$$Q(u^i, a) + c\sqrt{\log N(u^i)/N(u^i, a)} \tag{2.10}$$

In this expression the Q function denotes the current value estimates for each action a in information state u^i. $N(u^i)$ and $N(u^i, a)$ denote the current visitation count of the information state u^i and the subsequent information state after action a has been taken.

Smooth UCT is a modification of UCT that is inspired by fictitious playHeinrich (2017). As described in the last section the action selection in UCT is purely deterministic. Smooth UCT changes this by utilizing the average strategy a certain proportion of the time. In order to calculate the average strategy for a particular state we create a probability distribution based on the visitation counts of the actions available at that state. For example we could have an information state u^i that has been visited 100 times and has three available actions, a_1, a_2, a_3. Then it would be possible that a_1 has been visited 50 times, a_2 35 times and a_3 15 times. According to the average strategy we would then select a_1 50% of the time, a_2 35% of the time and a_3 15% of the time. In order to determine when the average strategy should be used compared to the UCB approach Heinrich describes a sequence η_k. This sequence decays to 0 as $\lim_{k\to\infty}$ and is expressed as follows:

$$\eta_k = max(\gamma, (1 + d * \sqrt{N_k})^{-}1) \tag{2.11}$$

Here N_k is the total number of plays and γ is a lower limit on η_k and d is a constant that parameterises the rate of decay.

The overall algorithm described by Heinrich for smooth UCT is listed in figure 2.4.

INSERT ALGORITHM FOR SMOOTH UCT HERE

2.6 Other Approaches

Although we have already mentioned that we will be focusing on MCTS it's worth discussing some other notable methods for creating poker agents before we continue. Currently the premier method for tackling full-scale texas hold'em is counterfactual regret minimization. This is the method that has dominated the Annual Computer Poker Competition(ACPC) for the last number of years, however recently some new methods have been outlined which look promising and thus we will discuss these methods as well.

2.6.1 Counterfactual Regret Minimization

Counterfactual regret minimization(CFR) is a method for finding approximate Nash equilibria in imperfect information games and was first outlined in (Zinkevich et al. 2008). Regret is a measure of the difference in utility between following some strategy σ compared to another strategy. Zinkevich introduced counterfactual regret which is regret applied to a single information set ie a single extensive form game state. It was found that by calculating and minimizing regret on an individual state basis that there were performance benefits as well as an improvement in accuracy of the calculated approximate Nash equilibria.

One implementation of CFR is outlined in (Jackson 2013). This paper outlines the implementation used for Slumbot, the 2018 ACPC champion in

no-limit hold'em. Jackson describes his use of CFR in order to generate a static strategy that approximates a Nash's equilibrium. At each iteration a strategy would be computed and when the process had ended an average of these strategies would be used. Jackson also mentions using an abstraction of the game in order to reduce no-limit's massive state space (roughly 10^164 with stack sizes of 100 big blinds)(Johanson 2013). This is done through techniques like grouping similar hand values into buckets and treating them as strategically equal or splitting the game up into it's individual rounds and solving the rounds separately. These techniques allow for a dramatic reduction in state space and if handled correctly, should allow for a strategy that can transfer to the full version of the game successfully.

2.6.2 Neural Fictitious Self-Play

Chapter 3

Application Development (10)

3.1 Requirements

3.2 Design

3.3 Backend API

3.4 Frontend Website

3.5 Testing

3.6 Issues

Chapter 4

Empirical Studies

4.1 Overview

In this chapter we will cover a number of experiments that were conducted in order to investigate the performance of the different algorithms implemented in order to tackle Leduc Hold'em. We will begin with a simplified version of MCTS for POMDPs and will incrementally add to this method in order to see how the performance of our agent evolves. In the table below we have listed the template that we will follow when conduction these experiments.

Section	**Rationale**
Objective	This section will contain an explanation of the purpose of the experiment along with how it was carried out
Algorithm and Coding	This section will go into the details of the algorithm used to produce results
Results	This section will detail the results acquired from the experiments conducted
Analysis	In this section we will examine our results and try to provide insight into the reasoning behind these results

4.2 Experiment 1 - MCTS versus random player

The first experiment conducted involved a simplified version of the algorithm outlined in(Silver & Veness 2010). We set an initial goal of using a random player as our benchmark opponent in order to demonstrate how this algorithm could exploit such a player's strategic inefficiencies.

4.2.1 Objective

The goal of this experiment is to implement MCTS for Leduc Hold'em. The MCTS agent will play against a random player and learn a strategy to exploit this player for maximal reward. Although we are interested in the results gained from playing against the random player we treat the outcome of this experiment as a baseline for our subsequent results. The reasoning for this is that the difficulty in finding a winning strategy against a random player is not high. Thus we expect that the exploitability of the resultant agent will be relatively high due to the fact that it will not have learned all of the strategic intricacies of the game. Rather, it will simply know how to beat a 'dumb' random player. This will give us a platform to build a more sophisticated agent through different mechanisms such as self-play in the subsequent experiments.

4.2.2 Algorithm and Coding

As mentioned we will be following Silver's 2010 implementation. The pseudocode for this algorithm can be seen below. This algorithm ticks the box of being applicable to POMDPs like poker. The one deviation from Silver's implementation is that we do not include the use of belief states. Belief states are used in order to avoid the difficulty of learning a function with unbounded input(Thrun 2000). In our case we are learning values associated with histories of actions and observations that occur in the game. Due to the nature of Leduc Hold'em these histories are limited in their length and thus belief states are not required.

Algorithm 1 Partially Observable Monte-Carlo Planning

procedure SEARCH(h)
  repeat
    if $h = empty$ **then**
      $s \sim \mathcal{I}$
    else
      $s \sim B(h)$
    end if
    SIMULATE($s, h, 0$)
  until TIMEOUT()
  return $\underset{b}{\text{argmax}}\, V(hb)$
end procedure

procedure ROLLOUT($s, h, depth$)
  if $\gamma^{depth} < \epsilon$ **then**
    return 0
  end if
  $a \sim \pi_{rollout}(h, \cdot)$
  $(s', o, r) \sim \mathcal{G}(s, a)$
  return r + γ.ROLLOUT($s', hao, depth$+1)
end procedure

procedure SIMULATE($s, h, depth$)
  if $\gamma^{depth} < \epsilon$ **then**
    return 0
  end if
  if $h \notin T$ **then**
    for all $a \in \mathcal{A}$ **do**
      $T(ha) \leftarrow (N_{init}(ha), V_{init}(ha), \emptyset)$
    end for
    return ROLLOUT($s, h, depth$)
  end if
  $a \leftarrow \underset{b}{\text{argmax}}\, V(hb) + c\sqrt{\frac{\log N(h)}{N(hb)}}$
  $(s', o, r) \sim \mathcal{G}(s, a)$
  $R \leftarrow r + \gamma$.SIMULATE($s', hao, depth + 1$)
  $B(h) \leftarrow B(h) \cup \{s\}$
  $N(h) \leftarrow N(h) + 1$
  $N(ha) \leftarrow N(ha) + 1$
  $V(ha) \leftarrow V(ha) + \frac{R - V(ha)}{N(ha)}$
  return R
end procedure

Figure 4.1: Partially Observable Monte Carlo Planning

Listing 4.1: Code will go here

```
import numpy as np
```

4.2.3 Results

The first metric that we used in order to analyse the results of this experiment is cumulative reward. This is simply the sum of the output of the reward function over time. In our case the function directly corresponds to the size of the pot won or lost in the game, thus we can have either a positive or negative reward. In figure 4.2 we see the reward over time increasing. In order to obtain these results we ran the algorithm for 10000 iterations and repeated this process 100 times. We then averaged our cumulative rewards at each iteration across these 100 repetitions to give the graph shown. Figure 4.3 shows the rate of increase of cumulative reward. In the case of figure 4.3 we applied the MCTS algorithm for 200000 iterations and repeated this

process 10 times, averaging the results.

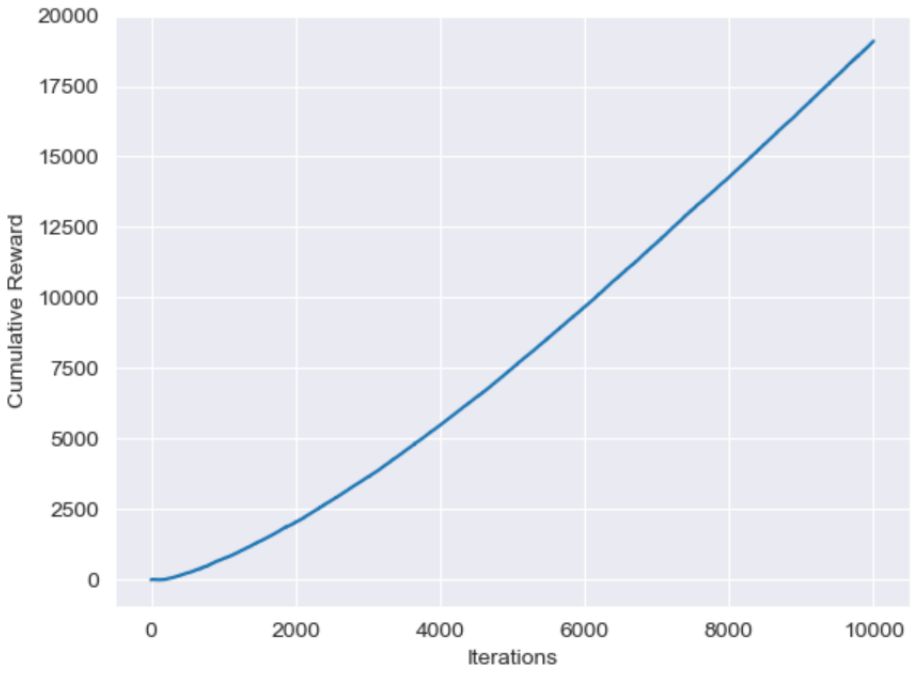

Figure 4.2: MCTS cumulative reward over time vs random player - 10000 Iterations

The second metric used to produce results was exploitability. This is a measure of the reward that can be gained by playing a best response strategy against the agent. In order to calculate the best response strategy, and hence the exploitability we must take our agent's action selections and insert them into the game tree(Heinrich 2017). In other words, wherever the agent must take an action in the game tree, we choose the best action based on our MCTS estimations. The result of doing this is a single-player MDP . We can then solve this MDP using one of our previously defined RL techniques eg Dynamic Programming. This will generate an optimal policy which is equivalent to the best response strategy. Then, based on the reward gained from the applying the best response strategy we can calculate the exploitability of our original MCTS agent.

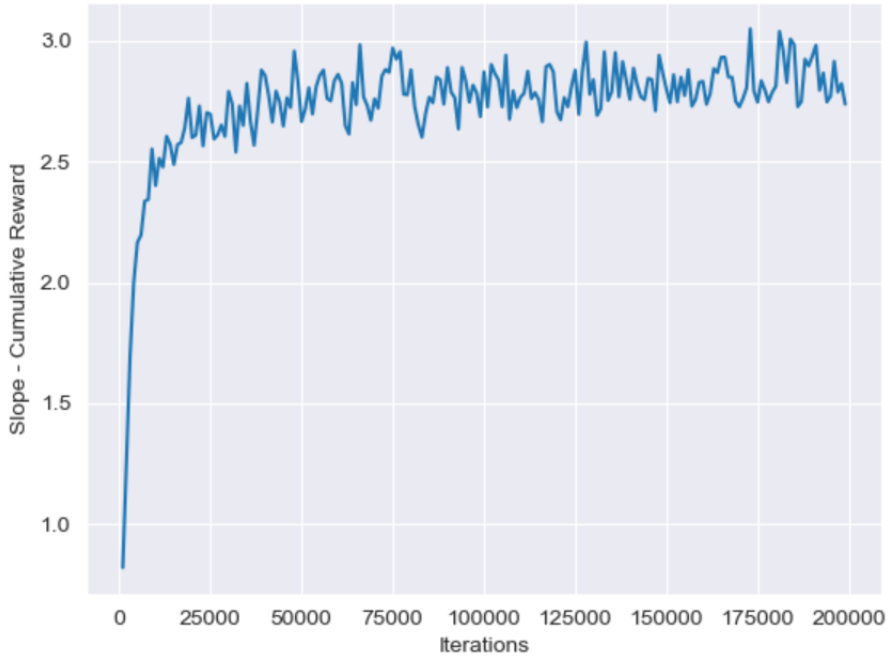

Figure 4.3: MCTS slope of cumulative reward over time vs random player - 200000 Iterations

4.2.4 Analysis

We will first analyse our results from the cumulative reward metric. As shown by figure 4.2 we see that initially there is a gradual increase in cumulative reward, with the slope growing until there is a constant rate of increase. This demonstrates that during the initial phase of the algorithm we have not yet uncovered the most beneficial action selection for all states and the exploration phase is still in effect. However, by visual inspection we can see that over time the rate of increase in cumulative reward begins to stabilise which indicates that a concrete strategy that can exploit the random play has been established. This hypothesis is further supported by figure 4.3 as we see a dramatic upswing in the rate of increase of cumulative reward followed by a levelling of the graph.

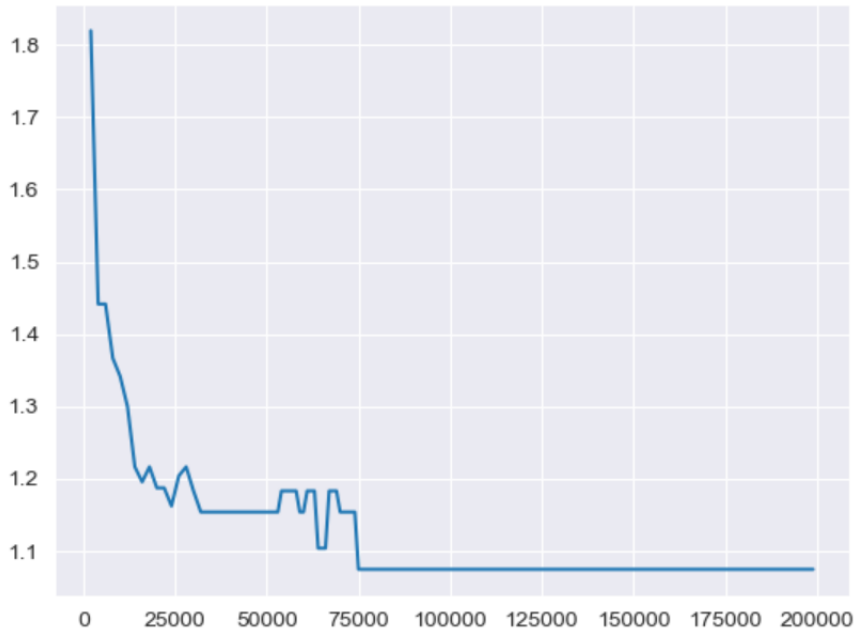

Figure 4.4: MCTS cumulative reward over time vs random player - 10000 Iterations

4.3 Experiment 2

4.3.1 Objective

4.3.2 Algorithm and Coding

4.3.3 Results

4.3.4 Analysis

4.4 Experiment 3

4.4.1 Objective

4.4.2 Algorithm and Coding

4.4.3 Results

4.4.4 Analysis

4.5 Experiment 4

4.6 Experiment 5

4.7 Experiment 6

4.8 Experiment 7

Chapter 5

Conclusions

5.1 Summary

5.2 Reflections

5.3 Future Work

Bibliography

Dahl, F. A. (2001), A reinforcement learning algorithm applied to simplified two-player texas hold'em poker, *in* 'European Conference on Machine Learning', Springer, pp. 85–96.

Heinrich, J. (2017), Reinforcement Learning from Self-Play in Imperfect-Information Games, PhD thesis, UCL (University College London).

Heinrich, J., Lanctot, M. & Silver, D. (2015), Fictitious self-play in extensive-form games, *in* 'International Conference on Machine Learning', pp. 805–813.

Heinrich, J. & Silver, D. (2016), 'Deep reinforcement learning from self-play in imperfect-information games', *arXiv preprint arXiv:1603.01121* .

Jackson, E. (2013), Slumbot nl: Solving large games with counterfactual regret minimization using sampling and distributed processing, *in* 'AAAI Workshop on Computer Poker and Incomplete Information'.

Johanson, M. (2013), 'Measuring the size of large no-limit poker games', *arXiv preprint arXiv:1302.7008* .

Johanson, M., Waugh, K., Bowling, M. & Zinkevich, M. (2011), Accelerating best response calculation in large extensive games, *in* 'Twenty-Second International Joint Conference on Artificial Intelligence'.

Kaelbling, L. P., Littman, M. L. & Moore, A. W. (1996), 'Reinforcement learning: A survey', *Journal of artificial intelligence research* **4**, 237–285.

Koller, D., Megiddo, N. & Von Stengel, B. (1996), 'Efficient computation of equilibria for extensive two-person games', *Games and economic behavior* **14**(2), 247–259.

Kuhn, H. (1953), 'Extensive games and the problem of information', *InH. KuhnandA. Tucker, editors, Contributions to the Theory of Games* pp. 193–216.

LazyProgrammerInc. (2018), 'Artificial intelligence: Reinforcement learning in python'.
URL: *https://www.udemy.com/artificial-intelligence-reinforcement-learning-in-python/*

Leslie, D. S. & Collins, E. J. (2006), 'Generalised weakened fictitious play', *Games and Economic Behavior* **56**(2), 285–298.

Mitchell, T. M. et al. (1997), 'Machine learning. 1997', *Burr Ridge, IL: McGraw Hill* **45**(37), 870–877.

Myerson, R. B. (2013), *Game theory*, Harvard university press.

Russell, S. J. & Norvig, P. (2016), *Artificial intelligence: a modern approach*, Malaysia; Pearson Education Limited,.

Samuel, A. L. (1959), 'Some studies in machine learning using the game of checkers', *IBM Journal of research and development* **3**(3), 210–229.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al. (2016), 'Mastering the game of go with deep neural networks and tree search', *nature* **529**(7587), 484.

Silver, D. & Veness, J. (2010), Monte-carlo planning in large pomdps, *in* 'Advances in neural information processing systems', pp. 2164–2172.

Southey, F., Bowling, M. P., Larson, B., Piccione, C., Burch, N., Billings, D. & Rayner, C. (2012), 'Bayes' bluff: Opponent modelling in poker', *arXiv preprint arXiv:1207.1411* .

Sutton, R. S., Barto, A. G., Bach, F. et al. (1998), *Reinforcement learning: An introduction*, MIT press.

Tesauro, G. (1995), Td-gammon: A self-teaching backgammon program, *in* 'Applications of Neural Networks', Springer, pp. 267–285.

Thrun, S. (2000), Monte carlo pomdps, *in* 'Advances in neural information processing systems', pp. 1064–1070.

Vermorel, J. & Mohri, M. (2005), Multi-armed bandit algorithms and empirical evaluation, *in* 'European conference on machine learning', Springer, pp. 437–448.

Von Stengel, B. (1996), 'Efficient computation of behavior strategies', *Games and Economic Behavior* **14**(2), 220–246.

Watkins, C. J. C. H. (1989), Learning from delayed rewards, PhD thesis, King's College, Cambridge.

Zinkevich, M., Johanson, M., Bowling, M. & Piccione, C. (2008), Regret minimization in games with incomplete information, *in* 'Advances in neural information processing systems', pp. 1729–1736.