# **CITS3001 Assignment Report – ThreeChess**

Lachlan D Whang (22466497), Tye Bridgeman (22326712)

There exist a wide variety of research and algorithms developed for the standard, 2-player game of Chess. 3-Player Chess is an interesting variant on this widely popular game which introduces a new degree of entropy and a vast number of new board layouts, and the difficulty involved with training intelligent agents to play a game of N-Player Chess grows exponentially as one observes the state-space explosion with increasing values of N. Thankfully, there exist a number of algorithms which can be used to generalize the agent learning process for any arbitrary number of players in a single game of chess. While a perfect game-playing agent may still be some distance away from our current efforts, this document will explore some of the suitable techniques which we can use to implement a rudimentary agent for 3-Player Chess.

**Greedy Algorithms**

All game-playing algorithms explored in this document work to maximize some utility based on the information available. Basic greedy algorithms constitute the most basic, game-playing agents whose moves are not random nor uninformed. At each turn, greedy agents make a move which maximizes the utility value in the short-term, in hopes that such moves will lead to maximized utility values in the long run. This utility value could be the number of pieces taken by the agent, or the total value of the points of pieces taken, or the total number of pieces remaining on the board that belong to the agent – the list goes on.

**Active Learning & Q-Learning**

The active reinforcement learning problem is characterized by a lack of a fixed policy, unknown reward and unknown transition models, with the latter 2 properties being present in Markov Decision Processes. An intelligent, game-playing agent in an active reinforcement learning environment needs to be able to select the “best” move based on the information it currently has access to at every turn, by calculating the utilities of states (and actions, depending on the algorithm), and selecting the action out of all possible actions in the current state that will produce the maximum known utility value.

Basic greedy algorithms can be applied to the active learning space. One simple greedy algorithm which undergoes the active learning process could involve the following 3 steps in the main game-playing loop.

1. Sense the current state and associated reward .
2. Call the **Adaptive Dynamic Programming** algorithm (a passive learning algorithm) to update the estimates for all previous utilities.
3. Choose the next, optimal action using the same equation for selecting an optimal action in Markov Decision Processes…

… where is the set of actions that can be performed in the current state .

The main problem with this basic, active learning, greedy algorithm is that it **does** **not** **converge on the optimal policy**. The algorithm actively exploits to the maximum possible extent permissible by the bounds of its knowledge, and exploration is minimal.

Q-Learning is the de-facto algorithm for solving the active reinforcement learning problem. The algorithm maintains a set of state-action pairs and their utilities, updating them after every move. Q-Learning is superior to basic greedy algorithms in that it balances **exploitation** with **exploration**, artificially inflating the utility values of state-action pairs that have been visited some number of times less than a predefined hyperparameter, such that when the algorithm attempts to select the action which maximizes this utility, it will explore the results of these actions. Exploration occurs most frequently towards the start of the learning process, and as the algorithm learns the outcome of more actions or moves, the -function responsible for the inflation of these utilities will start to present the actual, calculated utility value to the algorithm, instead of the inflated value, indicating a shift towards exploitation once the algorithm has “matured”.

The biggest challenge facing implementation of a Q-Learning agent is the method by which the algorithm’s knowledge will be **persisted** between training and playing sessions. Once the agent’s state is torn down at the end of a game, the information it has learnt needs to be stored. **I/O methods** are clearly the most straightforward and simplest way to persist this knowledge, but the functions are slow in comparison to the rest of the algorithm, and if the agent is unable to determine whether a position represents the end of a game, it will have to update stored utilities on disk after performing each move – a major detriment to its performance, although it is possible that this situation can be avoided depending on the information available.

**Implementation 1: Q-Learning *(Agent22466497)***

The beauty of the Q-Learning algorithm lies in its simplicity and adaptability. Q-Learning is based on the **Temporal Difference Learning** passive learning algorithm, which seeks to update only the utility value of the most recent state visited at the end of every action performed; Q-Learning similarly updates only the most recent state-action pair utility. While the algorithm itself is indeed comparably dense in comparison to other algorithms such as the **Monte Carlo Tree Search** or the **MiniMax** algorithm, given enough time and the right training data, the performance of a Q-Learning agent can surpass agents implementing either of the aforementioned algorithms, although such performance is not guaranteed for this particular problem instance.

**The Q-Learning Update Algorithm**

There are 2 main functions associated with a Q-Learning agent. The Q-Learning update algorithm is called after each action to update the Q-utilities for the impacted state-action pair.

function Q_Learning_Update(s, r, a, s', r' , y, n, Q, Nsa) 
if s' is a terminal state: 
Q[s', None] = r' 
// s is null if no states have been visited before 
if s is not null: 
if Nsa[s, a] exists: Nsa[s, a] 1 
else: Nsa[s, a] = 1 
c = a]) // we call function with input Nsa[s, a] 
= (1 — c)Q[s,a] + c (r + ) 

The following input arguments to the Q-learning update algorithm are fixed.

- The previous state-reward-action tuple.  
 - The current state and reward.  
 - The discount factor.  
 - A function which specifies a **learning rate** which **decreases over time**.

The following arguments to the Q-Learning update function have values that may **change** over the course of a single iteration, and correspond to persistent data structures which store the agent’s knowledge.

- The table of Q-values, i.e. the table of state-action pair utilities.

- A 2D table, where keeps track of the number of times action was performed while in state .

**The Q-Learning Agent Model**

The Q-Learning update step is just 1 function which fits into the wider agent model as a whole. On its own, the function above is just another passive learning function - the overall agent code is what makes this an agent which is actively learning. The Q-Learning agent model captures the overall behavior of the agent in a general setting; in the context of the threeChess environment, the agent’s playMove function corresponds directly to the innermost while(true) loop.

function AgentModel_Q_Learning(y, 71) 
// Initialization 
Initialize Q, NS a to empty tables, 
// Main loop: Execute mission after mission 
while (true) 
// no previous state at start 
Initialize variables s, r, a to null 
s' = initial state (chosen randomly, if multiple initial states exist) 
// Execute one game, from start to end 
while (true): 
(s' , r') = SenseStateAndReward() 
Q_Learning_Update(s, r, a, s' , r', y, n, Q, NS a) 
if s' is terminal: break // Reached a terminal state 
a = argmaxf(Q [s', a'], NS a [s', a']) // Choose next action 
ExecuteAction(a) 

*Agent22466497* demonstrates an implementation of the Q-Learning algorithm above.

**The Q-Learning Reward Function**

Certain environment situations possess innate rewards that the agent is provided with at the end of every action. This is not true regarding the threeChess environment, and as such, designing a suitable reward function for this agent is both a necessity, and extremely challenging to perfect. Simple reward functions could include the difference between the number of, or value of, the pieces from the previous state. More complex reward functions could consider the positions of pieces and the number of attacking lines from the agent’s pieces to their opponents, as well as possibly the quantity of moves in each board state and how many of these lead to winning situations (possibly a job for a simple look-ahead algorithm). An immense combination of different potential factors makes designing such a reward function relatively challenging.

*Agent22466497* employs a simple reward function for non-endgame states. At each stage, we consider the current state, and the previous state **directly** **before** our agent’s last action (not the last move taken by any agent). We calculate the reward as the difference between the total value of the agent’s own pieces and the value of the captured pieces in the current state, minus the same calculation for the previous state. If training on endgame positions were possible, potential improvements to the reward function could include assigning higher, positive rewards for reaching winning states, and lower, negative rewards for reaching losing states.

**Tests, Validation Metrics, and Training**

The Random and Manual agents provided are of minimal use for the training process. The Random agent does not provide a good example of intelligent play, and the Manual agent both slows down the training process through the need for user input, and is reliant on the presence of a good player to provide adequate training information for our agent. We need agent(s) which are fast and accurate, and for this, we turn to simple, greedy agents to provide us with both fast and moderately intelligent situations of play.

The speed and efficiency of training are also heavily reliant on the accuracy of the Q-Learning reward function. As mentioned above, the various factors at play in designing a reward function are numerous. A reward function for ThreeChess does not have to capture every aspect of the game, butt failure to cover the most important aspects can lead to undesired play, where certain components of the utility are not reflected or being maximized. Another issue which compounds training difficulty is the amount of information exposed to our agent; the playMove function is the only function which interfaces directly with the game, and this, and this function will not be called when the board is in a Game Over state, meaning that end positions will not be trainable for our agent, and we are unable to assign reward values for these positions.

Agent metrics fall informally into 2 distinct categories.

* **Time** – How long does the agent take to perform moves?
* **Intelligence** – How often does the agent win games? How often does the agent draw games?

Validation tests determine the degree to which the agent meets certain requirements. Given the time and scope of this assignment, we’ll set the acceptable bar for performance of *Agent22466497* at a flat proportion of 0 across 20 games, which corresponds to equal numbers of wins and losses with no counts on draws. We will explore this in more detail in our discussion of agent performance below.

**Agent Performance**

ThreeChess has an extremely large branching factor, and the number of possible board layouts far exceeds the number in standard 2-player chess, which after only 4 moves, already has over 288 billion distinct possible board states. It is impossible, from a training standpoint, for a Q-Learning agent to capture every possible board position during training, let alone accurately calculate the utility value for all those states. After only roughly 80 to 90 games of training against the aforementioned greedy agents, our Q-Utility persistence file is well in excess of 3 MB – quite a large file for storing the utility of state-action pairs. Thus, the importance of training against intelligent agents is emphasizes – we need to rely on only being able to capture the board positions which are likely to occur, assuming players are intelligent, whereas a random agent is not intelligent and will therefore be more likely to encounter board states reached by taking actions that serve no benefit to most utilities.

In terms of time, as the size of the persistence files increases, the time it takes to read and write these files also increases. As mentioned before, because game over states are not passed to the playMove function (and thus we can block a file-write until the game is over), an agent persisting knowledge will be forced to incur a detriment to its performance in writing its knowledge store to file after every call to this function, before returning a suitable action. While this process is not as much of an issue earlier on in the training process, as the agent’s knowledge store grows, these I/O process times begin to dominate. Fortunately, this issue may be exclusive to training; in a competitive situation, once the agent loads its knowledge from file, we can remove its write-to-file processes, reducing the playMove execution times by a factor of 20 or more. We can also keep the overall execution time low during training by briefly modifying the tournament code to invoke our agent’s write-to-file functions after the conclusion of all games played – again, this does not affect the agent’s performance in tournaments.

We will explore our agent’s performance across several samples of 20 tournament games occurring consecutively to each other with a time limit of 15 seconds per game, disabling agent write-to-file functions when playing against random agents. Note that the agent proportion is significantly more complex than just the proportion of wins to games played.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample Number | Agent opponents | Games Played | Number of wins | Agent Proportion |
| 1 | Greedy, Greedy | 21 | 2 | -0.6666 |
| 2 | Greedy, Greedy | 19 | 0 | -0.5789 |
| 3 | Greedy, Greedy | 16 | 0 | -0.5 |
| 4 | Greedy, Random | 22 | 1 | -0.4091 |
| 5 | Greedy, Random | 26 | 2 | -0.2692 |
| 6 | Random, Random | 23 | 2 | -0.4348 |
| 7 | Random, Random | 10 | 3 | -0.1 |

In terms of the overall game performance, during the start of training, as expected, our Q-Learning agent performs extremely poorly. It fails to win the majority of its games against more intelligent players (at the time), as it is currently in the exploration phase, and its knowledge stores are minimal. However, after roughly 500 to 600 games, the agent begins to encounter state-action pairs which it has commonly visited, and it exploits these states to move itself into more favorable positions; while it still loses the majority of its games, we notice a tangible improvement to the performance of the agent – it captures more frequently and favors higher capture values as these actions produce greater reward values than others, and it will occasionally move its pieces out of dangerous attacking lines to avoid incurring reward losses.

We see a steady trend of improvement in *Agent22466497*’s performance when playing against greedy agents, while that same trend is not observed when playing against random agents. Random agents tend to reach states which serve no benefit to our agent’s utility, meaning that while random agents possess no degree of intelligence, the sheer breadth of states reached could simply overwhelm our agent’s breath of knowledge stores, perhaps an unintended advantage that random agents possess, although this is clearly not on equal footing to greedy agents.

Extrapolating our agent’s performance, we can predict that in the future, our agent may continue along this trend of improvement, although exceptional agent performance may be limited by both the f-function hyperparameter, the quality of the reward function, the quality of the training data/environments, or a combination of all these factors. With careful tuning, we will certainly be able to reach an acceptable 0-proportion, although this may require a vast amount of training situations before our agent’s performance can be generalized against multiple types of agents.

**Implementation 2: XXX *(Agent22326712)***

**References**

Govindan, S. and Wilson, R., 2007. *Faculty & Research › Working Papers › A Decomposition Algorithm For N-Player Games A Decomposition Algorithm For N-Player Games*. PhD. Stanford Graduate School of Business.

Schrittwieser, J. et al., 2019. *Mastering Atari, Go, Chess And Shogi By Planning With A Learned Model*. PhD. Cornell University.