In using reinforcement learning to play Othello, we can use at least three different strategies: First, we can have a learning agent train against itself. Its evaluation function will become more and more accurate during training, and there will never be a large difference in level of play between the training agent and its opponent. A second strategy would be to train while playing against a player which is fixed, in the sense that its playing style does not change during training. The agent would learn from both its own moves and the moves its opponent makes. The skill levels of the non-learning players can vary. A third strategy consists of letting an agent train against a fixed opponent, but only have it learn from its own moves. This paper examines the differences between these three strategies. It attempts to answer the following research questions:

- How does the performance of each algorithm after learning through self-play compare to the performance after playing against a fixed opponent, whether paying attention to its opponent’s moves or just its own?
- When each reinforcement learning algorithm is trained using its best strategy, which algorithm will perform best?
- How does the skill level of the fixed training opponent affect the final performance when the learning agent is tested against another opponent?

Earlier research considered similar issues for backgammon [14]. There, it was shown that learning from playing against an expert is the best strategy. However, in that paper only TD-learning and one strong fixed opponent were used. When learning from a fixed opponent’s moves as well, an agent doubles the amount of training data it receives. However, it tries to learn a policy while half of the input it perceives was obtained by following a different policy. The problem may be that the learning agent cannot try out its own preferred moves to learn from, when the fixed opponent selects them. This research will show whether this doubling of training data is able to compensate for the inconsistency of policies. It is not our goal to develop the best Othello playing computer program, but we are interested in these research questions that also occur in other applications of RL.

In our experimental setup, three benchmark players will be used in both the train runs and the test runs. The results will therefore also show possible differences between the effect this
similarity between training and testing will have on the test performance for each of the three algorithms.

Outline. In section II we shortly explain the game of Othello. In section III, we discuss the theory behind the used algorithms. Section IV describes the experiments that we performed and the results obtained. A conclusion will be presented in section V.

II. OTHELLO

Othello is a two-player game played on a board of 8 by 8 squares. Figure 1 shows a screenshot of our application with the starting position of the game. The white and the black player place at alternate turns one disc at a time. A move is only valid if the newly placed disc causes one or more of the opponent’s discs to become enclosed. The enclosed discs are then flipped, meaning that they change color. If and only if a player cannot capture any of the opponent’s discs the player passes. When both players have to pass the game is ended. The player who has the most discs of his own color is declared winner, when the number of discs of each color are equal a draw is declared.

The best known Othello playing program is LOGISTELLO [15]. In 1997, it defeated the then world champion T. Murakami with a score of 6-0. The program was trained in several steps: First, logistic regression was used to map the features of the disc differential at the end of the game. Then, it used 13 different game stages and sparse linear regression to assign values to pattern configurations [16]. Its evaluation function was then trained on several millions of training positions to fit approximately 1.2 million weights [15].

III. REINFORCEMENT LEARNING

In this section we give an introduction to reinforcement learning and sequential decision problems. In reinforcement learning, the learner is a decision making agent that takes actions in an environment and receives a reward (or penalty) for its actions in trying to solve a problem [1], [2]. After a set of trial-and-error runs it should learn the best policy, which is the sequence of actions that maximize the total reward.

We assume an underlying Markov decision process, which is formally defined by (1) A finite set of states \( s \in S \); (2) A finite set of actions \( a \in A \); (3) A transition function \( T(s, a, s') \), specifying the probability of ending in state \( s' \) after taking action \( a \) in state \( s \); (4) A reward function \( R(s, a) \), providing the reward the agent will receive for executing action \( a \) in state \( s \), where \( r_t \) denotes the reward obtained at time \( t \); (5) A discount factor \( 0 \leq \gamma \leq 1 \) which discounts later rewards compared to immediate rewards.

A. Value Functions

We want our agent to learn an optimal policy \( \pi \) for mapping states to actions. The policy defines the action to be taken in any state \( s: a = \pi(s) \). The value of a policy \( \pi \), \( V^\pi(s) \), is the expected cumulative reward that will be received when the agent follows the policy starting at state \( s \). It is defined as:

\[
V^\pi(s) = E \left[ \sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, \pi \right],
\]

where \( E[\cdot] \) denotes the expectancy operator. The optimal policy is the one which has the largest state-value in all states.

Instead of learning values of states \( V(s_t) \) we could also choose to work with values of state-action pairs \( Q(s_t, a_t) \). \( V(s_t) \) denotes how good it is for the agent to be in state \( s_t \) whereas \( Q(s_t, a_t) \) denotes how good it is for the agent to perform action \( a_t \) in state \( s_t \). The Q-value of such a state-action pair \( \{s, a\} \) is given by:

\[
Q^\pi(s, a) = E \left[ \sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a, \pi \right].
\]

B. Reinforcement Learning Algorithms

When playing against an opponent, the results of the agent’s actions are not deterministic. After the agent has made its move, its opponent moves. In such a case, the Q-value of a certain state-action pair is given by:

\[
Q(s_t, a_t) = E [r_t] + \gamma \sum_{s_{t+1}} T(s_t, a_t, s_{t+1}) \max_a Q(s_{t+1}, a)
\]

Here, \( s_{t+1} \) is the state the agent encounters after its opponent has made his move. We cannot do a direct assignment in this case because for the same state and action, we may receive a different reward or move to different next states. What we can do is keep a running average. This is known as the Q-learning algorithm [11]:

\[
\hat{Q}(s_t, a_t) \leftarrow \hat{Q}(s_t, a_t) + \alpha (r_t + \gamma \max_a \hat{Q}(s_{t+1}, a) - \hat{Q}(s_t, a_t))
\]

where \( 0 < \alpha \leq 1 \) is the learning rate. We can think of (4) as reducing the difference between the current Q value and the backed-up estimate. Such algorithms are called temporal difference algorithms [10]. Once the algorithm is finished, the
agent can use the value of state action pairs to select the action with the best expected outcome:

$$\pi(s) = \arg \max_a \hat{Q}(s, a)$$

(5)

If an agent would only follow the strategy it estimates to be optimal, it might never learn better strategies, because the action values can remain highest for the same actions in all different states. To circumvent this, an exploration strategy should be used. In $\varepsilon$-greedy exploration, there is a probability of $\varepsilon$ that the agent executes a random action, and otherwise it selects the action with the highest state-action value. $\varepsilon$ tends to be gradually decreased during training.

Sarsa, the on-policy variant of Q-learning, takes this exploration strategy into account. It differs from Q-learning in that it does not use the discounted Q-value of the subsequent state with the highest Q-value to estimate the Q-value of the current state. Instead, it uses the discounted Q-value of the state-action pair that occurs when using the exploration strategy:

$$\hat{Q}(s_t, a_t) \leftarrow \hat{Q}(s_t, a_t) + \alpha (r_t + \gamma \hat{Q}(s_{t+1}, a_{t+1}) - \hat{Q}(s_t, a_t))$$

(6)

where $a_{t+1}$ is the action prescribed by the exploration strategy.

The idea of temporal differences can also be used to learn $V(s)$ values, instead of $Q(s, a)$. TD learning (or TD(0) [10]) uses the following update rule to update a state value:

$$V(s_t) \leftarrow V(s_t) + \alpha (r_t + \gamma V(s_{t+1}) - V(s_t))$$

(7)

C. Function Approximators

In problems of modest complexity, it might be feasible to actually store the values of all states or state-action pairs in lookup tables. However, Othello’s state space size is approximately $10^{28}$ [12]. This is problematic for at least two reasons. First of all, the space complexity of the problem is much too large to be stored. Furthermore, after training our agent it might be asked to evaluate states or state-action pairs which it has not encountered during training and it would have no clue how to do so. Using a lookup table would cripple the agent’s ability to generalize to unseen input patterns.

For these two reasons, we instead train multi-layer perceptrons to estimate the $V(s)$ and $Q(s, a)$ values. During the learning process, the neural network learns a mapping from state descriptions to either $V(s)$ or $Q(s, a)$ values. This is done by computing a ‘target’ value according to (4) in the case of Q-learning or (7) in the case of TD-learning. The learning rate $\alpha$ in these functions is set to 1, since we already have the learning rate of the neural network to control the effect training examples have on estimations of $V(s)$ or $Q(s, a)$. This means that (4) and (6) respectively simplify to

$$\hat{Q}^{new}(s_t, a_t) \leftarrow r_t + \gamma \max_a \hat{Q}(s_{t+1}, a)$$

(8)

and

$$\hat{Q}^{new}(s_t, a_t) \leftarrow r_t + \gamma \hat{Q}(s_{t+1}, a_{t+1}).$$

(9)

Similarly, (7) simplifies to

$$V^{new}(s_t) \leftarrow r_t + \gamma V(s_{t+1}).$$

(10)

In the case of TD-learning, for example, we use $(s_t, V^{new}(s_t))$ as training example for the neural network trained with the backpropagation algorithm. A Q-learning or Sarsa network consists of one or more input units to represent a state. The output consists of as many units as there are actions that can be chosen. A TD-learning network also has one or more input units to represent a state. It has a single output approximating the value of the state given as input. Figure 2 illustrates the structure of both networks.

D. Application to Othello

In implementing all three learning algorithms in our Othello framework, there is one important factor to account for: The fact that we have to wait for our opponent’s move before we can learn either a $V(s)$ or a $Q(s, a)$ value. Therefore, we learn the value of the previous state or state-action pair at the beginning of each turn – that is, before a move is performed. Every turn except the first, our Q-learning agent goes through the following steps:

1) Observe the current state $s_t$
2) For all possible actions $a'_t$ in $s_t$ use NN to compute $\hat{Q}(s_t, a'_t)$
3) Select an action $a_t$ using a policy $\pi$
4) According to (8) compute the target value of the previous state-action pair $\hat{Q}^{new}(s_{t-1}, a_{t-1})$

Figure 2. Topologies of function approximators. A TD-network (a) tries to approximate the value of the state presented at the input. A Q-learning network (b) tries to approximate the values of all the possible actions in the state presented at the input.
TD-agent performs the following steps:

1) Observe the current state \( s_t \)
2) For all afterstates \( s'_t \) reachable from \( s_t \) use NN to compute \( V(s'_t) \)
3) Select an action leading to afterstate \( s'_t \) using a policy \( \pi \)
4) According to (10) compute the target value of the previous afterstate \( V^{\text{new}}(s'^{a}_{t-1}) \)
5) Use NN to compute the current value of the previous afterstate \( V(s'^{a}_{t-1}) \)
6) Adjust the NN by backpropating the error \( \Delta = V^{\text{new}}(s'^{a}_{t-1}) - V(s'^{a}_{t-1}) \)
7) \( s'^{a}_{t-1} \leftarrow s'^{a}_{t} \)
8) Execute action \( a_t \)

Note that only the output unit belonging to the previously executed action is adapted. For all other output units, the error is set to 0. The Sarsa implementation is very similar, except that in step 3 it uses \((9)\) to compute the target value of the previous state-action pair instead of \((8)\).

In online TD-learning we are learning values of afterstates, that is: the state directly following the execution of an action, before the opponent has made its move. During playing, the agent can then evaluate all accessible afterstates and choose the one with the highest \( V(s'^{a}) \). Each turn except the first, our TD-agent performs the following steps:

3) Adjust the NN by backpropating the difference between the target and the estimate

Similarly, when the TD-agent learns from its opponent it will do the following once an opponent has reached an afterstate \( s'^{a}_{t} \):

1) According to \((10)\) compute the target value of the opponents previous afterstate \( V^{\text{new}}(s'^{a}_{t-1}) \)
2) Use NN to compute the current value of the opponents previous afterstate \( V(s'^{a}_{t-1}) \)
3) Adjust the NN by backpropating the difference between the target and the estimate

3) Learning from Its Own Moves: When an agent plays against a fixed opponent and only learns from its own moves, it simply follows the algorithm described in subsection III-D, without keeping track of the moves its opponent made and the (after-)states its opponent visited.

IV. EXPERIMENTS AND RESULTS

In training our learning agents, we use feedforward multi-layer perceptrons with one hidden layer consisting of 50 hidden nodes as function approximators. All parameters, including the number of hidden units and the learning rates, were optimized during a number of preliminary experiments. A sigmoid function:

\[
f(a) = \frac{1}{1 + e^{-a}}
\]  

(11)
is used on both the hidden and the output layer. The weights of the neural networks are randomly initialized to values between -0.5 and 0.5. States are represented by an input vector of 64 nodes, each corresponding to a square on the Othello board. Values corresponding to squares are 1 when the square is taken by its opponent and 0 when it is empty. The reward associated with a terminal state is 1 for a win, 0 for a loss and 0.5 for a draw. The discount factor \( \gamma \) is set to 1.0. The probability of exploration \( \varepsilon \) is initialized to 0.1 and linearly decreases to 0 over the course of all training episodes. The learning rate for the neural network is set to 0.01 for Q-learning and Sarsa, and for TD-learning a value of 0.001 is used.

<table>
<thead>
<tr>
<th>Value</th>
<th>100</th>
<th>25</th>
<th>10</th>
<th>5</th>
<th>2</th>
<th>1</th>
<th>0.5</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80</td>
<td>-26</td>
<td>-24</td>
<td>-1</td>
<td>-5</td>
<td>28</td>
<td>-18</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>46</td>
<td>-16</td>
<td>-4</td>
<td>1</td>
<td>-3</td>
<td>6</td>
<td>-20</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>-5</td>
<td>-3</td>
<td>-1</td>
<td>4</td>
<td>3</td>
<td>-12</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-6</td>
<td>-1</td>
<td>-2</td>
<td>3</td>
<td>0</td>
<td>-5</td>
<td>-5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>-13</td>
<td>12</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>-24</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>87</td>
<td>-25</td>
<td>27</td>
<td>-1</td>
<td>5</td>
<td>36</td>
<td>-3</td>
<td>100</td>
</tr>
</tbody>
</table>

Figure 3. Positional values used by player HEUR (a) and player BENCH (b, trained using co-evolution [17]).
A. Fixed Players

We created three fixed players: one random player RAND and two positional players, HEUR and BENCH. These players are both used as fixed opponents and benchmark players. The random player always takes a random move based on the available actions. The positional players have a table attributing values to all squares of the game board. They use the following evaluation function:

\[ V = \sum_{i=1}^{64} c_i w_i \]  

(12)

where \( c_i \) is 1 the square \( i \) is occupied by the player’s own disc, -1 when it is occupied by an opponent’s disc and 0 when it is unoccupied, and \( w_i \) is the positional value of a square \( i \). The two positional players differ in the weights \( w_i \) they attribute to squares. Player HEUR uses weights used in multiple other Othello researches [18], [17], [9]. Player BENCH uses an evaluation function created using co-evolution [17] and has been used as a benchmark player before as well [9]. The weights used by HEUR and BENCH are shown in figure 3.

The positional players use (12) to evaluate the state directly following an own possible move, i.e. before the opponent has made a move in response. They choose the action which results in the afterstate with the highest value.

Table I

PERFORMANCES OF THE FIXED STRATEGIES WHEN PLAYING AGAINST EACH OTHER. THE PERFORMANCES OF THE GAMES INVOLVING PLAYER RAND ARE THE AVERAGES OF 472,000 GAMES (1,000 GAMES FROM EACH OF THE 472 DIFFERENT STARTING POSITIONS).

<table>
<thead>
<tr>
<th></th>
<th>HEUR - BENCH</th>
<th>BENCH - RAND</th>
<th>RAND - HEUR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.55 - 0.45</td>
<td>0.80 - 0.20</td>
<td>0.17 - 0.83</td>
</tr>
</tbody>
</table>

B. Testing the Algorithms

To gain a good understanding of the performances of both the learning and the fixed players, we let them play multiple games, both players playing black and white. All players except RAND have a deterministic strategy during testing. To prevent having one player win all training games, we initialize the board as one of 236 possible starting positions after four turns1. During both training and testing, we cycle through all the possible positions, ensuring that all positions are used the same number of times. Each position is used twice: the agent plays both as white and black. Table I shows the average performance per game of the fixed strategies when tested against each other in this way. We are interested in whether the relative performances might be reflected in the learning player’s performance when training against the three fixed players.

1In other literature, 244 possible board configurations after four turns are mentioned. We found there to be 244 different sequences of legal moves from the starting board to the fifth turn, but that they result in 236 unique positions.

Table II

PERFORMANCES OF THE LEARNING ALGORITHMS WHEN TESTED VERSUS PLAYER BENCH. EACH COLUMN SHOWS THE PERFORMANCE IN THE TEST SESSION WHERE THE LEARNING PLAYER PLAYED BEST, AVERAGED OVER A TOTAL OF TEN EXPERIMENTS. THE STANDARD ERROR (\( \sigma/\sqrt{n} \)) IS SHOWN AS WELL.

<table>
<thead>
<tr>
<th></th>
<th>Train vs. Q-learning</th>
<th>Sarsa</th>
<th>TD-Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>BENCH</td>
<td>0.871 ± 0.009</td>
<td>0.859 ± 0.006</td>
<td>0.799 ± 0.007</td>
</tr>
<tr>
<td>BENCH-LRN</td>
<td>0.780 ± 0.008</td>
<td>0.816 ± 0.006</td>
<td>0.629 ± 0.009</td>
</tr>
<tr>
<td>Itself</td>
<td>0.721 ± 0.011</td>
<td>0.609 ± 0.011</td>
<td>0.723 ± 0.017</td>
</tr>
<tr>
<td>HEUR</td>
<td>0.582 ± 0.008</td>
<td>0.574 ± 0.018</td>
<td>0.522 ± 0.009</td>
</tr>
<tr>
<td>HEUR-LRN</td>
<td>0.563 ± 0.008</td>
<td>0.427 ± 0.015</td>
<td>0.355 ± 0.010</td>
</tr>
<tr>
<td>RAND</td>
<td>0.330 ± 0.010</td>
<td>0.307 ± 0.009</td>
<td>0.356 ± 0.011</td>
</tr>
<tr>
<td>RAND-LRN</td>
<td>0.418 ± 0.018</td>
<td>0.300 ± 0.012</td>
<td>0.332 ± 0.008</td>
</tr>
</tbody>
</table>

C. Comparison

We use the fixed players both to train the algorithms and to test them. In the experiments in which players HEUR and BENCH were used as opponents in the test games a total of 2,000,000 games were played during training. After each 20,000 games of training, the algorithms played 472 games versus respectively BENCH or HEUR without exploration. Tables II and III show the averages of the best performances of each algorithm when testing against players BENCH and HEUR after having trained against the various opponents through the different strategies: Itself, HEUR, HEUR when learning from its opponent’s moves (HEUR-LRN), BENCH, BENCH when learning from its opponent’s moves (BENCH-LRN), RAND and RAND when learning from its opponent’s moves (RAND-LRN).

Table III

PERFORMANCES OF THE LEARNING ALGORITHMS WHEN TESTED VERSUS PLAYER HEUR. EACH COLUMN SHOWS THE PERFORMANCE IN THE TEST SESSION WHERE THE LEARNING PLAYER PLAYED BEST, AVERAGED OVER A TOTAL OF TEN EXPERIMENTS. THE STANDARD ERROR (\( \sigma/\sqrt{n} \)) IS SHOWN AS WELL.

<table>
<thead>
<tr>
<th></th>
<th>Train vs. Q-learning</th>
<th>Sarsa</th>
<th>TD-Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEUR</td>
<td>0.910 ± 0.009</td>
<td>0.809 ± 0.005</td>
<td>0.775 ± 0.005</td>
</tr>
<tr>
<td>HEUR-LRN</td>
<td>0.651 ± 0.006</td>
<td>0.728 ± 0.013</td>
<td>0.666 ± 0.007</td>
</tr>
<tr>
<td>Itself</td>
<td>0.641 ± 0.016</td>
<td>0.631 ± 0.015</td>
<td>0.767 ± 0.005</td>
</tr>
<tr>
<td>BENCH-LRN</td>
<td>0.470 ± 0.012</td>
<td>0.361 ± 0.011</td>
<td>0.725 ± 0.009</td>
</tr>
<tr>
<td>BENCH</td>
<td>0.397 ± 0.016</td>
<td>0.440 ± 0.012</td>
<td>0.610 ± 0.010</td>
</tr>
<tr>
<td>RAND-LRN</td>
<td>0.356 ± 0.014</td>
<td>0.498 ± 0.009</td>
<td>0.644 ± 0.015</td>
</tr>
<tr>
<td>RAND</td>
<td>0.426 ± 0.007</td>
<td>0.428 ± 0.016</td>
<td>0.644 ± 0.015</td>
</tr>
</tbody>
</table>

Table IV

PERFORMANCES OF THE LEARNING ALGORITHMS WHEN TESTED VERSUS PLAYER RAND. EACH COLUMN SHOWS THE PERFORMANCE IN THE TEST SESSION WHERE THE LEARNING PLAYER PLAYED BEST, AVERAGED OVER A TOTAL OF TEN EXPERIMENTS. THE STANDARD ERROR (\( \sigma/\sqrt{n} \)) IS SHOWN AS WELL.

<table>
<thead>
<tr>
<th></th>
<th>Train vs. Q-learning</th>
<th>Sarsa</th>
<th>TD-Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itself</td>
<td>0.949 ± 0.003</td>
<td>0.946 ± 0.002</td>
<td>0.975 ± 0.002</td>
</tr>
<tr>
<td>BENCH-LRN</td>
<td>0.893 ± 0.006</td>
<td>0.906 ± 0.005</td>
<td>0.928 ± 0.003</td>
</tr>
<tr>
<td>RAND-LRN</td>
<td>0.893 ± 0.004</td>
<td>0.896 ± 0.003</td>
<td>0.924 ± 0.007</td>
</tr>
<tr>
<td>RAND</td>
<td>0.892 ± 0.007</td>
<td>0.885 ± 0.004</td>
<td>0.917 ± 0.003</td>
</tr>
<tr>
<td>HEUR-LRN</td>
<td>0.914 ± 0.004</td>
<td>0.837 ± 0.007</td>
<td>0.814 ± 0.008</td>
</tr>
<tr>
<td>HEUR</td>
<td>0.850 ± 0.007</td>
<td>0.827 ± 0.007</td>
<td>0.912 ± 0.004</td>
</tr>
<tr>
<td>BENCH</td>
<td>0.792 ± 0.017</td>
<td>0.783 ± 0.018</td>
<td>0.879 ± 0.007</td>
</tr>
</tbody>
</table>

For each test session, the results were averaged over a total of ten experiments. The tables show the averaged results in the session in which the algorithms, on average, performed best.
Figures 4 and 5 show how the performance develops during training when tested versus players BENCH and HEUR. The performances in the figures are a bit lower than in the tables, because in the tables the best performance during an epoch is used to compute the final results.

In the experiments in which the algorithms are tested versus player RAND, a total of 500,000 training games were played. Table IV shows the best performance when training against each of the various opponents through the different strategies. Figure 6 shows how the performance develops during training when testing versus player RAND.

D. Discussion

These results allow for the following observations:

- **Mixed policies** There is not a clear benefit to paying attention to the opponent’s moves when learning against a fixed player. Tables II, III and IV seem to indicate that the doubling of perceived training moves does not improve performance as much as getting input from different policies decreases it.

- **Generalization** Q-learning and Sarsa perform best when having trained with the same player against which they are tested. When training against that player, the performance is best when the learning player does not pay attention to its opponent’s moves. For both Q-learning and Sarsa, training against itself comes in at a third place in the experiments where the algorithms are tested versus HEUR and BENCH. For TD-learning, however, the performance when training against itself is similar or even better than the performance after training against the same player used in testing. This seems to indicate that the TD-learner achieves a higher level of generalization. This is due to the fact that the TD-learner learns values of states while the other two algorithms learn values of actions in states.

- **Symmetry** The TD-learner achieves a low performance against BENCH when having trained against HEUR-LRN, RAND and RAND-LRN. However, the results of the TD-learner when tested against HEUR lack a similar result. We speculate that this can be attributed to the lack of symmetry in BENCH’s positional values.

Using our results, we can now return to the research questions posed in the introduction:

- **Question** How does the performance of each algorithm after learning through self-play compare to the performance after playing against a fixed opponent, whether paying attention to its opponent’s moves or just its own? **Answer** Q-learning and Sarsa learn best when they train against the same opponent against which they are tested. TD-learning seems to learn best when training against itself. None of the algorithms benefit from paying attention to its opponent’s moves when training against a fixed strategy. We believe this is because the RL agent is not free to choose its own moves when the opponent selects a move, leading to a biased policy.

- **Question** When each reinforcement learning algorithm is trained using its best strategy, which algorithm will perform best? **Answer** When Q-learning and Sarsa train against BENCH and HEUR without learning from their opponent’s moves while tested against the same players, they clearly outperform TD after it has trained against itself. This is a surprising result, since we expected TD-learning to perform better. However, if we compare the performance for each of the three algorithms after training against itself, TD significantly outperforms Q-learning and Sarsa when
tested against HEUR and RAND. When tested against BENCH after training against itself, the difference between TD-learning and Q-learning is insignificant. The obtained performances of Q-learning and Sarsa are very similar.

- **Question** How does the skill level of the fixed training opponent affect the final performance when the learning agent is tested against another fixed opponent?
- **Answer** From table I we see that player HEUR performs better against RAND than BENCH. This is also reflected in the performances of the algorithms versus RAND after having trained with HEUR and BENCH respectively. From table I we see as well that HEUR has a better performance than BENCH when the two players play against each other. This difference in performance also seems to be partly reflected in our results: When Q-learning and Sarsa train against player HEUR they obtain a higher performance when tested against BENCH than vice versa. However, we don’t find a similar result for TD-learning. That might be attributed to the fact that BENCH’s weights values are not symmetric and therefore BENCH might pose a greater challenge to TD-learning than to Q-learning and Sarsa. We believe that BENCH can be better exploited.
using different action networks, as used by Q-learning and Sarsa, since particular action sequences follow other action sequences in a more predictable way when playing against BENCH. Because TD-learning only uses one state network, it cannot easily exploit particular action sequences.

V. Conclusion

In this paper we have compared three strategies in using reinforcement learning algorithms to learn to play Othello: learning by self-play, learning by playing against a fixed opponent and learning by playing against a fixed opponent while learning from the opponent’s moves as well. We found that it differs per algorithm what the best strategy is to train: Q-learning and Sarsa obtain the highest performance when training against the same opponent as which they are tested against (while not learning from the opponent’s moves) while TD-learning learns best from self-play. Differences in the level of the training opponent seem to be reflected in the eventual performance of the training algorithms.

Future work might take a closer look at the influence of the training opponent’s play style on the learned play style of the reinforcement learning agent. In our research, the differences in eventual performance were only analyzed in terms of a score. It would be interesting to experiment with fixed opponents with more diverse strategies and analyze the way these strategies influence the eventual play style of the learning agent in a more qualitative fashion.

References