Game Design Verification using Reinforcement Learning

Eirini Ntoutsi **Dimitris Kalles**

AHEAD Relationship Mediators S.A., 65 Othonos-Amalias St, 262 21 Patras, Greece

and

Department of Computer Engineering and Informatics, University of Patras, Greece

{ntoutsi, kalles}@aheadrm.com

Abstract

Reinforcement learning is considered as one of the most suitable and prominent methods for solving game problems due to its capability to discover good strategies by extended self-training and limited initial knowledge. In this paper we elaborate on using reinforcement learning for verifying game designs and playing strategies. Specifically, we examine a new strategy game that has been trained on self-playing games and analyze the game performance after human interaction. We demonstrate, through selected game instances, the impact of human interference to the learning process, and eventually the game design.

Keywords

Reinforcement learning, machine learning, strategy games, design verification.

1 Background

The game theory domain is been widely regarded as appropriate for understanding the concepts of machine learning. Scientists usually focus on strategic games and make efforts to create "intelligent" programs that efficiently compete with human players. Such games are suitable for further studying because of their complexity and the opportunities they offer to explore winning strategies. Furthermore, evaluation criteria are typically known, whereas the game environment, the moves and the termination conditions can be simulated.

Scientists have long tried to create expert artificial players for strategy games. In 1949, Shannon began to study how computers could play chess and proposed the idea of using a value function to compete with human players. In 1959, Samuel created a checkers program that tried to find "the highest point in multidimensional scoring space". Although the experiments of Samuel's research were impressive they did not exert significant influence (method-wise), until 1988 when Sutton formulated the TD(λ) method for temporal difference learning. Since then, more games such as Tetris, Blackjack, Othello [Leouski 1995], chess [Thrun 1995], backgammon were analysed by applying TD(λ) to improve their performance. During the 1990s, IBM made strenuous efforts to develop (first with Deep Thought, later with Deep Blue) a chess program comparable to the best human player. Whether it succeeded is still a philosophical and technological question.

One of the most successful and hopeful applications of TD(λ) is TD-Gammon [Tesauro 1992, 1995] for the game of backgammon. Using reinforcement learning techniques and after training with 1.5 million self –playing games Tesauro achieved a performance comparable to that demonstrated by backgammon world champions.

The advantage of reinforcement learning domain among other learning methods is that it requires little programming effort for system training. Training is effected by a system's interaction with its environment. RL comprehends changes on the learning environment without having to be re-programmed from scratch.

As far as strategy games are concerned, the most important and critical point of them is to select and implement the computer's strategy during the game. The term *strategy* stands for the selection of the computer's next move considering its current situation, the opponent's situation, consequences of that move and possible next moves of the opponent. RL comes to significant assistance in solving this problem.

In this paper we continue the research of Kalles and Kanellopoulos [2001] on the application of RL to the design of a new strategy game (see section on game description, below, for a detailed game description). The research demonstrated that, when trained with self-playing games, both players had nearly opportunities to win and neither player enjoyed a pole position advantage. In this paper, we aim to explore the extent to which this conclusion continues to stand for the case one of the opponents is human. Specifically, we will try to give answers to questions such as:

- Are games played from a computer against itself enough to accomplish learning?

- Which case is more suitable for learning, a computer playing against itself or a computer playing against a human player?

- Does playing with human players improve the computer performance much more than playing against itself?

The rest of this paper is organised in six sections. The next section presents the details of the game. It includes the basic components of the game, rules for legal pawn movements, special characteristics and playability issues. The third section refers to the game analysis; which methods are used and how they could lead towards learning. The fourth section describes training issues and experimental results. The fifth section refers to the human factor and how this affects the learning procedure. Finally, we put all the details together and discuss lines of future research that have been deemed worthy of following.

2 Game description

The game is played on a square board of size n by two players, called black and white. Two square bases of size a are located on the board. The base at the lower left part of the board belongs to the white player whereas the base at the upper right part of the board belongs to the black player.

At the beginning of the game each player possesses β pawns, but during the game some pawns may be lost.

Each player's goal is to possess the opponent's base; the first that will achieve that is the winner. If some player runs out of pawns the opponent is the winner.

Each pawn can move to an empty square that is vertically or horizontally adjacent, provided that the pawn's maximum distance from its base is not decreased (this mean that backward moves are not allowed).

Using coordinates the above rule could be defined as follows:

If (x,y) is the current position of the pawn, then it can move to position (x,z), if

$$\max(x-a, y-a) \le \max(x-a, z-a),$$

if the white player moves, or

$$\max(n-a-x, n-a-y) \le \max(n-a-x, n-a-z),$$

if the black player moves.

Legal moves can be categorized in moves of:

- leaving the base (the base is considered as a single square and not as a set of squares, therefore every pawn of the base can move at one step to any of the adjacent to the base free squares), and
- moving from a position to another.

Figure 1 shows examples and counterexamples of moves (the left board demonstrates an illegal move; the centre and right boards demonstrate the loss of pawns).

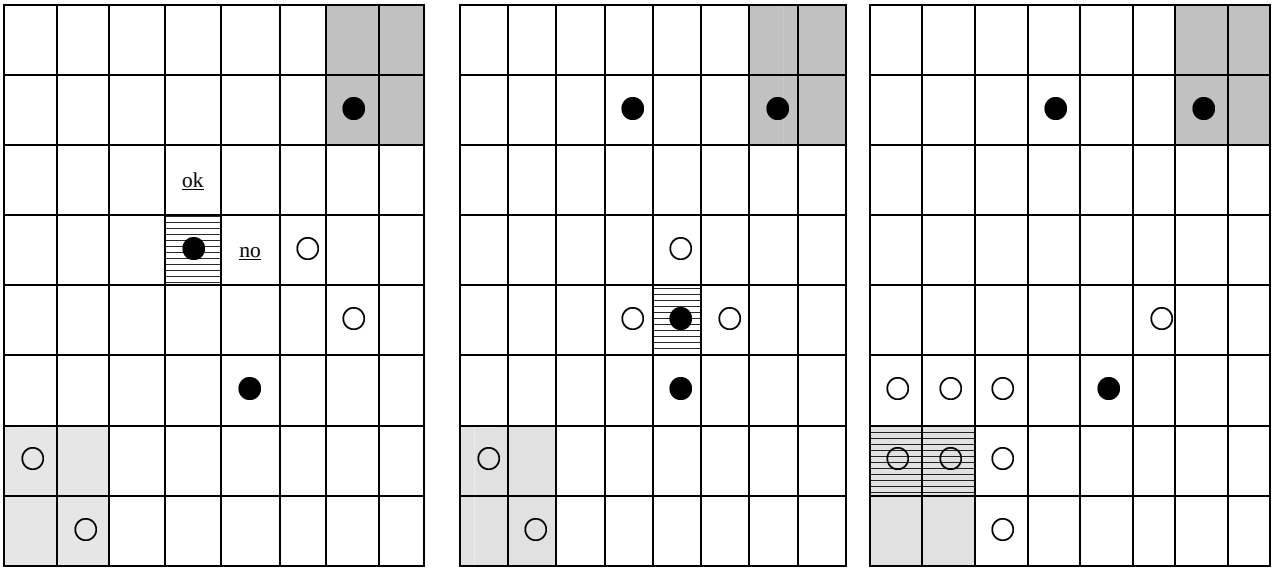

Figure 1: Examples and counterexamples of moves.

Such moves bring about the direct adjustment of the moving pawn with some pawn of the opponent. In such cases the "trapped" pawn automatically draws away from the board game. Alike, in the case that there is no free square next to the base the rest pawns of the base disappear automatically.

3 Game analysis

Since the design of the game the challenge was to design and implement a system that learns how to play through a number of self – playing games. Reinforcement Learning is ideal for this purpose. The basic idea behind RL comes from psychology: the likeliness of repeating an action depends on its consequences. RL is characterized as learning that takes place via continuous interaction of the learning agent with his environment. The agent itself detects which actions to take via trial and error learning with very limited need for human involvement.

The game is a discrete Markov procedure in discrete time, since there are finite states and moves, and since each episode does terminate. The *a priori* knowledge of the system consists of the rules only. The agent's goal is to learn a policy $\pi: S \rightarrow A$ (where *S* being the state space, *A* being the space of legal moves), that will maximize the expected sum of rewards in a specific time; this is called an optimal policy. A policy determines which action should be taken next given the current state of the environment.

The move selection is critical and affects the whole learning procedure. The agent has to decide whether to choose an action that will straightforward maximize its reward or to try a new action for which it does not know anything but it *may* prove to be better (the first case is known as *exploitation*, whereas the second is known as *exploration*). The answer to the above question is (in our case, too) both. Specifically, the system uses an ε-greedy policy, with $\varepsilon=0.9$, which means that in 90% of the cases the system chooses the best-valued action, while in the rest 10% it chooses a random one.

The agent estimates whether it is good for it to be in a specific position using the $V^{\pi}(s)$ value function. According to $V^{\pi}(s)$ the value of the state *s* of the strategy π equals to the sum of the expected rewards starting from state *s* and following the strategy π. Specifically, the agent is interesting in discovering the optimal strategy (the strategy that will maximize the expected sum of rewards) and for this it uses the optimal value function $V^{\pi}(s)$. Learning comes from the experience in playing or training from samples of positions taken from the game. **Because of the high dimensionality and large state space of this computation we use neural networks as a generalization technique**.

In fact, two neural networks were used, one for each player, because each player has a unique state space, different from its opponent's. Back-propagation was used, setting the RL parameters to $\gamma=0.95$ and $\lambda=0.5$.The input layer nodes are the board positions for the next possible move, totalling n^2-2a^2+10. The hidden layer consists of half as many hidden nodes, whereas the output node has only one node, which can be regarded as the probability of winning beginning from a specific game-board configuration and then taking on a specific move.

At the beginning all states have the same value except for the final states, but after each move the values are updated through the temporal difference learning rule. The algorithm is TD(λ), where λ determines the reduction degree of assigning credit to some action. Using λ only, the eligible states (eligibility traces can be seen as a temporary record of the occurrence of an event, e.g. visiting a state) or actions are assigned credit or blame when a TD error occurs. We replaced eligibility traces instead of accumulating them, because the latter approach has been known to inhibit learningm, when a repeated wrong action generates a large bad trace.

For the experiments we used a game of dimensions 8x2x10 (8: the game board dimension, 2: the base dimension, 10: the number of pawns).

4 Training issues

Initial experiments had suggested that both computer players have nearly equal opportunities to win. However, when we tested the game performance against a human player we realized that the human player was almost always, independently of the moves the black player was following. Obviously the network training was not enough. Tesauro [1992, 1995] reached a high level performance in his TD-Gammon

after playing a huge number (*1,500,000*) of self-playing games. And as Sutton and Barto [1998] point out, in the case of the first *300,000* games, TD-Gammon performance was poor, games lasted hundreds or thousands of moves before one side or the other won, almost by accident.

The above symptoms arose in our game as we noticed that the initial games lasted hundred of moves with the majority of moves being cyclical between two squares. So, we kept on the training procedure and in order to speed up learning we changed the way of assigning reward. In the initial experiments, each action-move is given reward –1, unless the resulting state is a final one; then the reward is +50 for the winner's last move and –50 for the loser's last move. The new reward assignment procedure was more explicit; each action-move is assigned reward not only in final states but also during the learning procedure when it loses some pawn or when it is next to the opponent's base.

The new training results showed a clear improvement in computer playing even in the case it had to compete with a human player.

There were four obvious points of improvement towards the agent's goal to establish an advantage in winning the game.

1. The computer player attempts to protect its base by covering the next-to-base squares in case an opponent's pawn approaches them. This is a clear sign that the computer player has learned to protect itself against attacks.

2. The back-n-forth moves were significantly decreased. Currently, the average number of moves per game has been nearly halved.

3. The area covered by the computer player during the game has been significantly expanded. The computer player does not stop short at squares lying near its base but expands its moves so as to cover distant squares too. This is another sign that the computer player has begun to understand its goal to possess the opponent's base.

4. The computer player protects its pawns. More specifically, it moves carefully so as to avoid adjacency with opponent's pawns, which might cause their loss. Towards this direction the computer player does not hold all next to base squares; note that due to game rules, when all next-to-base squares are occupied, the remaining base pawns are lost. In previous experiments, the computer player played usually with four pawns only, as it lost all the others when it covered all next-to-base squares.

The above were all signs of game performance improvement. Aiming at greater improvement and to speed up learning we decided to examine the impact of human interference to the learning procedure. Our question was: how can a human player improve the game performance and speed up the learning procedure? And, we ask, does this have the same influence as adding handcrafted features (note that the latter has been shown to accelerate learning [Tesauro 1992, 1995]).

5 The human factor in learning acceleration

Training the system by self-playing games restricts the exploration to very narrow portions of the state space, due to the absence of some strong "regularity disturbance" factor. In the case of backgammon, dices play such a role and this is believed to be vital in the success of TD-Gammon. Dices produce a high degree of variability in the positions seen during training and, as a result, the learner explores more of the state space, leading to the discovery of improved evaluations and new strategies.

In our game we use the human factor. The human player gives the computer opportunities to explore a large state space different from what it has seen until now by playing against itself. A human opponent can create long-term viewed playing sequences that help a computer player to follow a **loosely guided unexplored path**.

Experimental results presented below prove the above assertions. After training the network with 119,000 self-playing games, we trained it by playing alternatively self-playing games and human-computer games. More specifically, we followed the training sequence shown in Figure 2 (where light-shaded squares correspond to human-computer games).

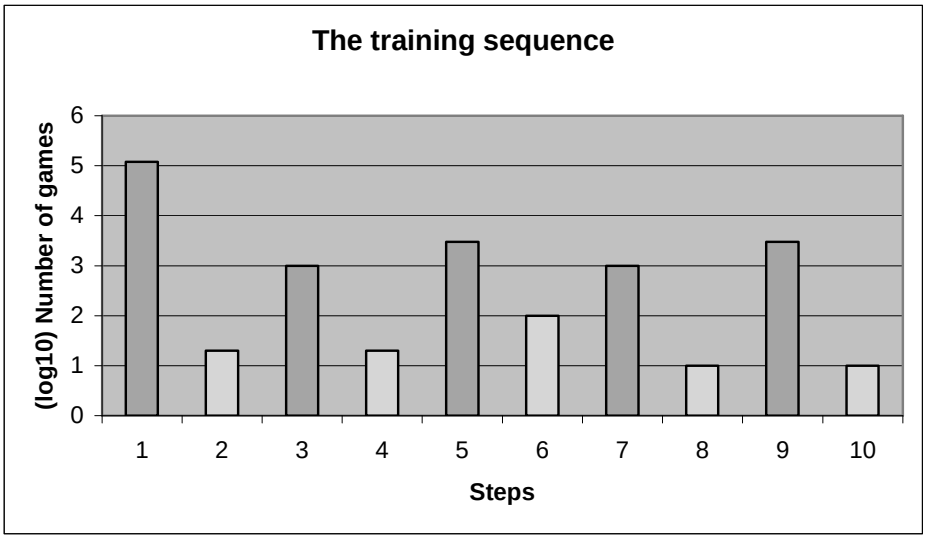

Figure 2: The training sequence.

In all these human-computer games the human had a specific goal: to possess the opponent's base by capturing a particular next-to-base square (see Figure 3). Our aim was to check whether the computer could learn from human attacks and how this would affect the learning procedure. The number of computer-human games was comparably smaller than the number of computer games. We intended to give the computer the opportunity to face states different from those it had explored. After playing 160 human-computer games in combination with 18,160 self-playing games (totalling 137,160 games) the computer's performance has been rapidly improved, as it almost never allowed the human to enter its base through that particular square (see Figure 3 for such a game instance).

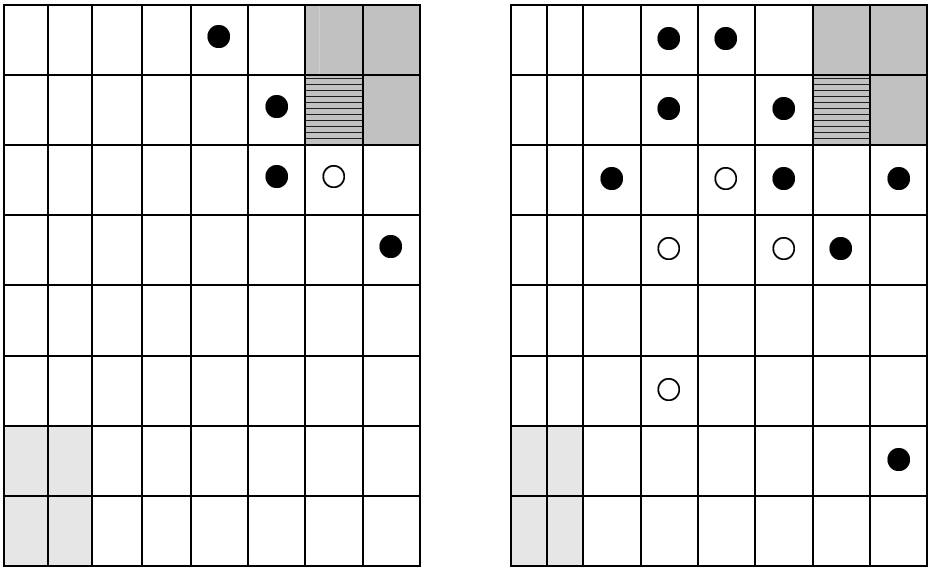

Figure 3: Original (left) and improved (right) game performance.

To disambiguate the human impact in learning we also ran 137,160 **separate** self-playing games and we compared them with the above experiments. Experiments showed that in the second case the computer had not learned something specific. There was little improvement in its way of playing but this improvement was general and does not correspond to any specific strategy. This happens due to the slow speed of learning; the computer learns through self-playing games but such kind of learning can only be useful after a long number of self-playing games.

The above results are encouraging referring to learning acceleration. But does human interference contribute to the long-term game performance improvement or do we risk degrading the generality of computer playing? The latter would be surely achieved through self-playing games although the number of games required is extremely large. To explore this question we ran more experiments using the neural network weights. Specifically, we ran four sets of experiments, each set consisting of 1,000 computer-vs-computer games. Each set was based on a different training configuration though; see Table 1 for a list of configurations and related performance results.

White player (with):	**Black player (with):**
computer training	computer training
54.2%	45.8%
computer and human training	computer and human training
55%	45%
computer training	computer and human training
50.3%	49.7%
computer and human training	computer training
52.5%	47.5%

Table 1: Cross-testing of learning strategies and percentage of games won.

The term “white player with computer and human training” means that the white computer player bases its play on the knowledge received from the 137,160 compound human-computer games mentioned above, whereas the term “white player with computer training” means that the white computer player bases its play on the knowledge received from the 137,160 self-playing games mentioned above.

The above experiments show that human involvement should be carefully exercised to add value to computer performance. The human (white player) experience proved to be significantly helpful in the case of the black player; the percentage of the black player winning games has been increased from 45.8% to 49.7%. The opposite happens with the white player, whose initial goal was to train the black player with a particular defending strategy. Towards this aim, the white player was rather risky by not exploring new states, and, instead, following the minimal path that would ensure it the black’s base possession. Performance percentage was decreased from 54.2% of winning games to 52.5%.

Another interesting point of the above experimental results is the performance percentage for the case where the training of both computer players contains games against a human opponent. We would expect a reduction in the white player’s performance, but we were surprised to observe its performance increasing from 54.2% of winning games to 55%, which contradicts our intuition. A reason could be the (comparatively) small amount of experiments, so that a decrease of 0.8% may be actually misguiding.

6 Conclusion

Experimental results presented in this paper show that computer performance can take advantage of human knowledge.

We expect to speed up learning by exploring Explanation Based Learning techniques. A combination of RL and EBL could benefit the game providing it with faster learning and the ability to scale to large state spaces in a more structured manner [Dietterich and Flann, 1997].

A parallel improvement of practical value would be to develop a benchmark computer player, however, this is best viewed as a by-product of the game design improvement.

We are confident, however, that this is a most promising research direction with widespread application implications, especially so in simulation of educational environments.

References

1. T. Dietterich, N. Flann. “Explanation-Based Learning and Reinforcement Learning: A Unified View ”, Machine Learning, Vol. 28, 1997.

2. D. Kalles and P. Kanellopoulos. “*On Verifying Game Design and Playing Strategies using Reinforcement Learning”*, ACM Symposium on Applied Computing, special track on Artificial Intelligence and Computation Logic, Las Vegas, March 2001.

3. A. Leouski. “*Learning of Position Evaluation in the Game of Othelo”,* Master’s project: University of Massachusetts, Amherst, 1995.

4. A. Samuel. *“Some Studies in Machine Learning Using the Game of Checkers”*, IBM Journal of Research and Development 3, 1959.

5. C. Shannon. *“Programming a computer for playing chess”*, Philosophical Magazine, Vol. 41 (4), 1950.

6. R. Sutton and A. Barto. *”Reinforcement Learning - An Introduction”*, MIT Press, Cambridge, Massachusetts, 1998.

7. G. Tesauro. *“Practical issues in temporal difference learning”*, Machine Learning, Vol. 8, No. 3-4, 1992.

8. G. Tesauro. *“Temporal Difference Learning and TD-Gammon”*, Communications of the ACM, Vol. 38, No 3, 1995.

9. S. Thrun. *“Learning to Play the Game of Chess”*. Advances in Neural Information Processing Systems 7, 1995.