TD-Gammon, A Self-Teaching Backgammon Program, Achieves Master-Level Play

Gerald Tesauro
IBM Thomas J. Watson Research Center
P. O. Box 704
Yorktown Heights, NY 10598
(tesauro@watson.ibm.com)

Abstract. TD-Gammon is a neural network that is able to teach itself to play backgammon solely by playing against itself and learning from the results, based on the TD(λ) reinforcement learning algorithm (Sutton, 1988). Despite starting from random initial weights (and hence random initial strategy), TD-Gammon achieves a surprisingly strong level of play. With zero knowledge built in at the start of learning (i.e. given only a "raw" description of the board state), the network learns to play at a strong intermediate level. Furthermore, when a set of hand-crafted features is added to the network's input representation, the result is a truly staggering level of performance: the latest version of TD-Gammon is now estimated to play at a strong master level that is extremely close to the world's best human players.

Reinforcement learning is a fascinating and challenging alternative to the more standard approach to training neural networks by supervised learning. Instead of training on a "teacher signal" indicating the correct output for every input, reinforcement learning provides less information to work with: the learner is only given a "reward" or "reinforcement" signal indicating the quality of output. In many cases the reward is also delayed, i.e., is given at the end of a long sequence of inputs and outputs. In contrast to the numerous practical successes of supervised learning, there have been relatively few successful applications of reinforcement learning to complex real-world problems.

This paper presents a case study in which the TD(λ) reinforcement learning algorithm (Sutton, 1988) was applied to training a multilayer neural network on a complex task: learning strategies for the game of backgammon. This is an attractive test problem due to its considerable complexity and stochastic nature. It is also possible to make a detailed comparison of TD learning with the alternative approach of supervised training on human expert examples; this was the approach used in the development of Neurogammon, a program that convincingly won the backgammon championship at the 1989 International Computer Olympiad (Tesauro, 1989).

Details of the TD backgammon learning system are described elsewhere (Tesauro, 1992). In brief, the network observes a sequence of board positions

$x_1, x_2, ..., x_f$ leading to a final reward signal z determined by the outcome of the game. (These games were played without doubling, thus the network did not learn anything about doubling strategy.) The sequences of positions were generated using the network's predictions as an evaluation function. In other words, the move selected at each time step was the move that maximized the network's estimate of expected outcome. Thus the network learned based on the outcome of self-play. This procedure of letting the network learn from its own play was used even at the very start of learning, when the network's initial weights are random, and hence its initial strategy is a random strategy. From an *a priori* point of view, this methodology appeared unlikely to produce any sensible learning, because random strategy is exceedingly bad, and because the games end up taking an incredibly long time: with random play on both sides, games often last several hundred or even several thousand time steps. In contrast, in normal human play games usually last on the order of 50-60 time steps.

Preliminary experiments used an input representation scheme that only encoded the raw board information (the number of White or Black checkers at each location), and did not utilize any additional pre-computed features relevant to good play, such as, e.g., the strength of a blockade or probability of being hit. These experiments were completely knowledge-free in that there was no initial knowledge built in about how to play good backgammon. In subsequent experiments, a set of hand-crafted features was added to the representation, resulting in higher overall performance. This feature set was the same set that was included in Neurogammon.

The rather surprising result, after tens of thousands of training games, was that a significant amount of learning actually took place, even in the zero initial knowledge experiments. These networks achieved a strong intermediate level of play approximately equal to that of Neurogammon. The networks with hand-crafted features have greatly surpassed Neurogammon and all other previous computer programs, and have continued to improve with more and more games of training experience. The best of these networks is now estimated to play at a strong master level that is extremely close to equaling world-class human play. This has been demonstrated in numerous tests of TD-Gammon in play against several world-class human grandmasters, including Bill Robertie and Paul Magriel, both noted authors and highly respected former World Champions. For the tests against humans, a heuristic doubling algorithm was added to the program which took TD-Gammon's equity estimates as input, and tried to apply somewhat classical formulas developed in the 1970's (Zadeh and Kobliska, 1977) to determine proper doubling actions.

Results of testing are summarized in table 1. TD-Gammon 1.0, which had a total training experience of 300,000 games, lost a total of 13 points in 51 games against Robertie, Magriel, and Malcolm Davis, the 11th highest rated player in the world in 1991. TD-Gammon 2.0, which had 800,000 games of training games of experience and was publicly exhibited at the 1992 World Cup of Backgammon tournament, had a net loss of 7 points in 38 exhibition

Program	Training Games	Opponents	Results
TD-Gammon 1.0	300,000	Robertie, Davis, Magriel	-13 pts/ 51 games (-.25 ppg)
TD-Gammon 2.0	800,000	Goulding, Woolsey, Snellings, Russell, Sylvester	-7 pts/ 38 games (-.18 ppg)
TD-Gammon 2.1	1,500,000	Robertie	-1 pt/ 40 games (-.02 ppg)

Table 1: Results of testing TD-Gammon in play against world-class human opponents. Version 1.0 used 1-ply search for move selection; versions 2.0 and 2.1 used 2-ply search. Version 2.0 had 40 hidden units; versions 1.0 and 2.1 had 80 hidden units.

games against top players Kent Goulding, Kit Woolsey, Wilcox Snellings, former World Cup Champion Joe Sylvester, and former World Champion Joe Russell. The latest version of the program, version 2.1, had 1.5 million games of training experience and achieved near-parity to Bill Robertie in a recent 40-game test session: after trailing the entire session, Robertie managed to eke out a narrow one-point victory by the score of 40 to 39.

According to an article by Bill Robertie published in *Inside Backgammon* magazine (Robertie, 1992), TD-Gammon's level of play is significantly better than any previous computer program. Robertie estimates that TD-Gammon 1.0 would lose on average in the range of 0.2 to 0.25 points per game against world-class human play. (This is consistent with the results of the 51-game sample.) This would be about equivalent to a decent advanced level of human play in local and regional Open-division tournaments. In contrast, most commercial programs play at a weak intermediate level that loses well over one point per game against world-class humans. The best previous commercial program scored -0.66 points per game on this scale. The best previous program of any sort was Hans Berliner's BKG program, which in its only public appearance in 1979 won a short match against the World Champion at that time (Berliner, 1980). BKG was about equivalent to a very strong intermediate or weak advanced player and would have scored in the range of -0.3 to -0.4 points per game.

Based on the latest 40-game sample, Robertie's overall assessment is that TD-Gammon 2.1 now plays at a strong master level that is extremely close to equaling the world's best human players. In fact, due to the program's steadiness (it never gets tired or careless, as even the best of humans inevitably do), he thinks it would actually be the favorite against any human player in a long money-game session or in a grueling tournament format such as the World Cup competition.

The only thing which prevents TD-Gammon from genuinely equaling world-class human play is that it still makes minor, practically inconsequential technical errors in its endgame play. One would expect these technical errors to cost the program on the order of .05 points per game against top humans. Robertie thinks that there are probably only two or three dozen players in the entire world who, at the top of their game, could expect to hold their own or have an advantage over the program. This means that TD-Gammon is now probably as good at backgammon as the grandmaster chess machine Deep Thought is at chess. Interestingly enough, it is only in the last 5-10 years that human play has gotten good enough to rival TD-Gammon's current playing ability. If TD-Gammon had been developed 10 years ago, Robertie says, it would have easily been the best player in the world at that time. Even 5 years ago, there would have been only two or three players who could equal it.

The self-teaching reinforcement learning approach used in the development of TD-Gammon has greatly surpassed the supervised learning approach of Neurogammon, and has achieved a level of play considerably beyond any possible prior expectations. It has also demonstrated favorable empirical behavior of TD(λ), such as good scaling behavior, despite the lack of theoretical guarantees.

Prospects for further improvement of TD-Gammon seem promising. Based on the observed scaling, training larger and larger networks with correspondingly more experience would probably result in even higher levels of performance. Additional improvements could come from modifications of the training procedure or the input representation scheme. Some combination of these factors could easily result in a version of TD-Gammon that would be the uncontested world's best backgammon player.

However, instead of merely pushing TD-Gammon to higher and higher levels of play, it now seems more worthwhile to extract the principles underlying the success of this application of TD learning, and to determine what kinds of other applications may also produce similar successes. Other possible applications might include financial trading strategies, military battlefield strategies, and control tasks such as robot motor control, navigation and path planning. At this point we are still largely ignorant as to why TD-Gammon is able to learn so well. One plausible conjecture is that the stochastic nature of the task is critical to the success of TD learning. One possibly very important effect of the stochastic dice rolls in backgammon is that during learning, they enforce a certain minimum amount of exploration of the state space. By stochastically forcing the system into regions of state space that the current evaluation function tries to avoid, it is possible that improved evaluations and new strategies can be discovered.

References

H. Berliner, "Computer backgammon." *Scientific American* **243:1**, 64-72 (1980).

B. Robertie, "Carbon versus silicon: matching wits with TD-Gammon." *Inside Backgammon* **2:2**, 14-22 (1992).

R. S. Sutton, "Learning to predict by the methods of temporal differences." *Machine Learning* **3**, 9-44 (1988).

G. Tesauro, "Neurogammon wins Computer Olympiad." *Neural Computation* **1**, 321-323 (1989).

G. Tesauro, "Practical issues in temporal difference learning." *Machine Learning* **8**, 257-277 (1992).

N. Zadeh and G. Kobliska, "On optimal doubling in backgammon." *Management Science* **23**, 853-858 (1977).