Efficient TD Training

Several successful applications of TD-learning methods in various games, which include Samuel’s checkers player [277, 279], Tesuaro’s TD-Gammon [317, 318, 319], Baxter et al.’s KnightCap [18, 19], Schraudolph et al.’s Go program [294, 295], or Schaeffer et al.’s TDL-Chinook [283], to name only the best-known examples, confirmed wide applicability of this type of learning in game domain. Quite a lot of research papers devoted to TD-learning and games were published, in which particular attempts are thoroughly described focusing on various aspects of TD paradigm. Based on the amount of the literature one might suppose that all relevant questions concerning this type of learning have already been answered and if there exist any ambiguities they must refer to secondary issues or specific implementation details.

However, quite surprisingly, the research reports do not agree even on very basic, fundamental issues as, for example, the effectiveness of self-play TD versus training with external opponents. In the former case the basic questions refer to the exploration-vs.-exploitation dilemma, which consists in a tradeoff between playing along already known, optimal (to the knowledge of the learning program) lines and the exploration of new possibilities at the risk of playing moves weaker that the optimal ones.

In the case of training against external opponents the key issues are related to the choice of the trainers: the number of them and their playing strength. Intuitively, the more training opponents the better, since variety of the trainers leads to variety of possible game situations that the learning program is faced with. On the other hand, at some point the problem of convergence of the training process comes into play. As far as the strength of the trainers is concerned, the intuitive approach is to use opponents of comparable or slightly superior playing level than the learning program, gradually increasing their average playing abilities along with the development of the trainee. Another possibility is to use random or pseudo-random opponents.\footnote{The distinction between \textit{random} and \textit{pseudo-random} players will be cleared in section 9.2.0.}
in the initial phase of training and replace them with stronger players as the training process progresses. Last but not least problem is weighing the pros and cons of training against human players vs. artificial ones.

Finally, the choice of the specific TD-learning method, being either TD(\(\lambda\)) or TDLeaf(\(\lambda\)), as well as the choice of \(\lambda\) decay schedule have significant influence on attained results and in the literature there are different points of view regarding this matter.

This chapter attempts to address these issues basing to a large extent on the author’s experience with TD-GAC - a TD-based program trained to play give-away checkers, introduced briefly in section 6.3.3. The TD-GAC experiment aimed, in particular, at comparison between various training strategies including the variants of learning exclusively on lost and drawn games, learning on all games but with omitting weak moves, or learning with focus on opponents stronger than the learning program by playing several games in a row against the same opponent whenever the program kept losing against that opponent.

Certainly the results, ideas and comments presented in this chapter do not pretend to provide complete and definite knowledge in the above-mentioned areas. The goal is rather to attract readers’ interest to these topics since the more theoretical and experimental results are available the better understanding and better outcomes of TD-learning applications in games.

9.1 External Opponents vs. Self-playing

One of the fundamental questions that need to be answered when planning a TD-learning experiment is the choice between self-play training mode and learning based on games played against external opponents (independent of the learning system). Despite very basic nature of this issue, the experiments published in the literature are inconclusive with regard to whether it is more profitable to favor self-playing or rather to train with external opponents. Hence, one of the interesting and challenging issues is further investigation and formalization of the strengths and weaknesses of both training approaches.

9.1.1 Self-playing

The first widely-known example of a TD-learning process relying on a self-playing scheme was Arthur L. Samuel’s approach to checkers\(^2\) [277, 279]. The program was equipped with \textit{a priori} defined set of expert features potentially relevant to building board evaluation function, and in the course of self-play learning was able to successfully define the meaningful subset of features and their weights in order to form linear evaluation function.

\(^2\) Formally speaking, at that time the term TD-learning was not yet invented, but the learning method used by Samuel was very akin to this type of training.