PERGAMON

Neural Networks 13 (2000) 133–135

Neural Networks

www.elsevier.com/locate/neunet

Book Review

Learning to maximize rewards: A review of R.S. Sutton and A.G. Barto's *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998, 380 pages, ISBN 0-262-19398-1, $42.00

"Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond." (Thorndike, 1911).

The idea of learning to make appropriate responses based on reinforcing events has its roots in early psychological theories such as Thorndike's "law of effect" (quoted above). Although several important contributions were made in the 1950s, 1960s and 1970s by illustrious luminaries such as Bellman, Minsky, Klopf and others (Bellman, 1957; Farley & Clark, 1954; Grossberg, 1975; Klopf, 1982; Michie & Chambers, 1968; Minsky, 1961; Samuel, 1963), the last two decades have witnessed perhaps the strongest advances in the mathematical foundations of reinforcement learning, in addition to several impressive demonstrations of the performance of reinforcement learning algorithms in real-world tasks. The introductory book by Sutton and Barto, two of the most influential and recognized leaders in the field, is therefore both timely and welcome.

The book is divided into three parts. In the first part, the authors introduce and elaborate on the essential characteristics of the reinforcement learning problem, namely, the problem of learning "policies" or mappings from environmental states to actions so as to maximize the amount of "reward" received from interactions with the environment over time. Important concepts such as value functions, action-value functions (or Q-functions), Markov Decision Processes (MDPs) and the Bellman optimality equation are introduced and lucidly explained using well-chosen examples.

The second part of the book focuses on three classes of methods for solving the reinforcement learning problem: Dynamic Programming (DP), Monte Carlo estimation and Temporal Difference (TD) learning. DP methods bootstrap, i.e. learn based on previous learned values without waiting for a final outcome, but they require an accurate model of the environment to learn optimal value functions and policies. Monte Carlo methods, on the other hand, can learn from on-line or simulated experience (sequences of states, actions and rewards) without a prior model of the environment's dynamics, but they do not bootstrap and are not suitable for step-by-step incremental learning. The best of both worlds is harnessed in TD learning algorithms which, although harder to analyze, require no model of the environment and are fully incremental. These algorithms derive their name from the fact that they learn value functions based on the temporal difference between predictions of state- or action-values at successive time steps. The roots of TD learning can be found in the early checkers-playing program of Samuel (1963)and the work of Klopf (1982), but it was undoubtedly the authors themselves who played an instrumental role in establishing TD learning as a viable machine learning technique. The chapter on TD learning begins with a clear explanation of the method's relationship to DP and Monte Carlo estimation and goes on to discuss issues such as optimality and convergence. The tradeoff between exploring different policies and exploiting the current policy is addressed by focusing on four important TD methods for reinforcement learning: "Sarsa," Q-learning, Actor–Critic methods and R-learning.

In the last part of the book, the authors discuss several facets of the reinforcement learning problem that serve to unify the methods and ideas discussed in the previous chapters. The chapter on eligibility traces, which are essentially memory variables that record the eligibility of states for undergoing learning changes, unifies TD and Monte Carlo methods. The chapter on planning and learning explores the extent to which state-space planning methods (such as DP) that are based on explicit models of the environment can be integrated with TD learning and Monte Carlo methods. The example of Dyna agents, originally introduced by Sutton, is used to illustrate how planning based on simulated experience from a learned model can be intermixed with on-line learning from real experience to considerably speed up the search for the optimal policy. An entire chapter is devoted to function approximation and generalization across the state space of a problem given only limited numbers of sample experiences. The techniques that are briefly discussed include least-squares and gradient-descent methods, coarse coding, radial basis functions and Kanerva coding. In the final chapter, in order to emphasize the practical utility of the techniques presented in the book, the authors provide

E-mail address: rao@salk.edu (R.P.N. Rao).

PII: S0893-6080(99)00098-2

in-depth case studies of some of the most successful applications of reinforcement learning to date including Gerry Tesauro's celebrated backgammon program called TD-Gammon that has learned to play at the grandmaster level by using the TD learning algorithm in several thousand games that it played against itself.

What makes this book especially easy to read and its contents easy to digest is the rather liberal use of examples and figures to clarify the technical points raised in each chapter. In addition to several traditional reinforcement learning examples such as the pole-balancing task and varieties of Gridworlds, one also finds unconventional and sometimes thought-provoking examples such as TD learning for the task of "Driving Home" and computing value functions for using a putter versus a driver in playing a hole of golf. Most of the sections in each chapter include exercises and programming assignments that help test one's grasp of the subject matter discussed therein. Also helpful are the boxes with pseudocode that are provided for most of the important learning algorithms derived in text. In many cases, different algorithms are compared to each other based on their performance on a given task, thereby allowing readers to gain a better understanding of the properties of the examined algorithms. The summary at the end of each chapter provides a succinct account of the main topics brought up in the previous pages, with key concepts and terms italicized for emphasis. The historical remarks at the end of each chapter are especially interesting to read since they often trace the evolution of the ideas discussed in the chapter to other fields such as optimal control, animal learning, artificial intelligence and psychology, thereby making explicit the close links between these otherwise disparate fields.

Mathematically sophisticated readers and experts may feel a bit disappointed at the omission of rigorous mathematical proofs of optimality and convergence, but this is understandable given that the book is meant to be an introduction and is targeted primarily at the graduate and advanced undergraduate level rather than the expert level. Such readers and experts may find solace in the book by Bertsekas and Tsitsiklis (1996) which delves more deeply into the mathematical intricacies of reinforcement learning. Another topic that might have been worth discussing but which is mentioned only briefly as a frontier dimension is the ubiquitous problem of hidden state or perceptual aliasing in real-world environments, wherein the current perceptions alone are insufficient to disambiguate the current state. In recent years, this problem has generated great interest in partially observable or hidden Markov decision processes (POMDPs or HMDPs) which form the basis for several different novel approaches to solving this general reinforcement learning problem, some of which are cited in the text.

Several recent experiments have suggested a tantalizing link between reinforcement learning (particularly TD learning) algorithms and dopaminergic reward systems in the brain (Houk, Davis & Beiser, 1995; Montague, Dayan & Sejnowski, 1996; Schultz, Dayan & Montague, 1997). This relationship goes back to early work by Sutton and Barto themselves on psychological models of classical conditioning based on TD learning (for example, Sutton & Barto, 1990). It is therefore surprising that the authors did not devote any significant portion of their text to discussing the successful application of reinforcement learning theory to psychology and neuroscience. The inclusion of a chapter on this topic would probably have made the book even more appealing to psychologists and neuroscientists than it already is, but there is some consolation in the fact that the book does point readers interested in this line of research to the relevant papers.

In summary, this book provides an excellent and easily accessible introduction to the core concepts of reinforcement learning. It should be of interest to students and researchers in computational neuroscience, psychology, machine learning, control engineering, and operations research. Being self-contained, the book is especially ideal for beginners seeking to teach themselves the foundations of reinforcement learning before embarking on research in more advanced topics. The book should also prove useful in teaching reinforcement learning as part of a more general course on artificial intelligence, neural networks or optimal control. Given its broad and in-depth coverage of all the important issues in reinforcement learning, this book appears destined to become the standard text in the field in the years to come.

R.P.N. Rao
Computational Neurobiology Laboratory and Sloan Center for Theoretical Neurobiology, The Salk Institute for Biological Studies, 10010 N. Torrey Pines Road, La Jolla, CA 92037, USA

References

Bellman, R. E. (1957). *Dynamic programming*, Princeton, NJ: Princeton University Press.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*, Belmont, MA: Athena Scientific.

Farley, B. G., & Clark, W. A. (1954). Simulation of self-organizing systems by digital computer. *IRE Transactions on Information Theory*, *4*, 76–84.

Grossberg, S. (1975). A neural model of attention, reinforcement and discrimination learning. *International Review of Neurobiology*, *18*, 263–327.

Houk, J. C. & Davis, J. L. & Beiser, D. G. (Eds.). (1995). Models of information processing in the basal ganglia Cambridge, MA: MIT Press.

Klopf, A. H. (1982). *The hedonistic neuron: a theory of memory, learning and intelligence*, Washington, DC: Hemisphere.

Michie, D., & Chambers, R. (1968). Boxes: an experiment in adaptive control. In E. Dale & D. Michie (Eds.) (pp. 137–152). *Machine intelligence*, 2. Edinburgh: Oliver and Boyd.

Minsky, M. L. (1961). Steps toward artificial intelligence. *Proceedings IRE*, *49*, 8–30.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for

mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16* (5), 1936–1947.

Samuel, A. L. (1963). Some studies in machine learning using the game of checkers. In E. Feigenbaum & J. Feldman (Eds.), *Computers and thought* (pp. 71–105). Malabar, FL: Krieger.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1598.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. W. Moore (Eds.), *Learning and computational neuroscience: foundations of adaptive networks*, Cambridge, MA: MIT Press.

Thorndike, E. L. (1911). *Animal intelligence*, Darien, CT: Hafner.