Book Review

Reinforcement Learning: An Introduction by Richard S. Sutton and Andrew G. Barto, A Bradford Book, The MIT Press, Cambridge, 1998. pp 322, ISBN 0-262-19398-1. USD 42.

The reinforcement learning (RL) problem is the challenge of artificial intelligence in a microcosm; how can we build an agent that can plan, learn, perceive, and act in a complex world? There's a great new book on the market that lays out the conceptual and algorithmic foundations of this exciting area. RL pioneers Rich Sutton and Andy Barto have published *Reinforcement Learning: An Introduction*, providing a highly accessible starting point for interested students, researchers, and practitioners.

In the RL framework, an agent acts in an environment whose state it can sense, and occasionally receives some penalty or reward based on its state and action. Its learning task is to select actions to maximize its reward over the long haul; this requires, not only choosing actions that are associated with high reward in the current state, but "thinking ahead" by choosing actions that will lead the agents to more lucrative parts of the state space. While there are many ways to attack this problem, the paradigm described in the book is to construct a value function that evaluates the "goodness" of different situations. In particular, the *value* of a state is the long-term reward that can be attained starting from the state if actions are chosen optimally. Recent research has produced a flurry of algorithms for learning value functions, theoretical insights into their power and limitations, and a series of fielded applications. The authors have done a wonderful job of boiling down disparate and complex RL algorithms to a set of fundamental components, then showing how these components work together. The differences between Dynamic Programming, Monte Carlo Methods, and Temporal-Difference Learning are teased apart, then tied back together in a new, unified way. Innovations such as "backup diagrams", which decorate the book cover, help convey the power and excitement behind RL methods to both novices and RL veterans like us.

The book consists of three parts, one dedicated to the problem description, and two others to a range of reinforcement learning algorithms, their analysis, and related research issues.

We enthusiastically applaud the authors' decision to articulate the problem addressed in the book before talking in length about its various solutions. After all, a thorough discussion of the problem is necessary to understand the aims and scope of reinforcement learning research, let alone for novices in the field. At 85 pages in length, however, one might wonder what it is about the reinforcement learning problem that its description deserves (or requires?) twice as many pages as the typical journal paper. Is the reinforcement learning problem so complicated that it takes that long to describe and discuss it?

In truth, the Part I does much more than just pose the problem. Chapter 1 contains a highly informal introduction into the broad problem domain: *learning to select actions while interacting with an environment in order to achieve long-term goals*. The example of tic-tac-toe makes concepts such as reward, value functions, and the exploration-exploitation dilemma feel natural—all concepts that find a more mathematical treatment later in the book. The first chapter also provides an invaluable description of the history of reinforcement learning, placing recent research efforts in context. This is an early example of a series of detailed literature reviews, found at the end of each chapter, which could alone justify the expense of purchasing the book.

Next, the book dives into a highly restricted instance of the reinforcement learning problem: the *k-arm bandit problem*. This well-researched problem lacks state transitions—there is only a single state—but, it otherwise possesses the typical characteristics that sets reinforcement learning apart from, say, supervised learning. The placement of the problem is well-chosen, since it illustrates with rigor the key concepts of the algorithms yet to come: the idea of interaction with an environment, reward, value functions, and the exploration-exploitation dilemma. It is followed by what readers knowledgeable in AI might chose as their starting point into the book: the formal, mathematical definition of the reinforcement learning problem. At this point, the authors state clearly the key assumptions that underlie the methods expounded in the book, including the critical Markov assumption that renders the environment's state fully observable. Their crystal-clear description of Bellman optimality leaves little room for misunderstanding. Trust us; even if you are familiar with RL, you will find that Part I is insightful and great fun to read! By the end, we could hardly wait to get to Part II to learn about reinforcement learning algorithms.

At this point it seems appropriate to make a few comments about the general style of the book. The text is extremely accessible, from the first page to the last. The authors have successfully integrated formal algorithmic descriptions and analysis with intuitions, motivations, numerical examples, and exercises (whose solutions can be obtained from the first author). Exercises and examples are merged into the floating text, and we recommend spending some time to think about them! Every exercise has at least one "aha!" insight in it; some have two. Great care was taken in choosing tractable, yet non-trivial, examples of RL problems; optimal blackjack, soap-bubble shape prediction, and cost-effective rental car management were some of our favorites, in addition to 101 variations on the classic gridworld problem. New ideas are introduced, described, discussed, and evaluated with meticulous care, and often theoretical results and/or easily replicable experiments accompany the general discourse on the material. The experiments are nicely done, as they give the reader a hands-on appreciation for the underlying ideas. Because of the conversational writing style, algorithms like TD(λ), which encompass a collection of concepts, are not really described at a single location; instead, the book gradually progresses from basic algorithmic methods to today's state-of-the-art reinforcement learning algorithms. This makes the book a little difficult to use as a desktop reference or for a course that does not follow the same logical train of thoughts. But, there is a lot of wisdom in the book's development of ideas!

Another aspect of the writing style worth mentioning is that it is a bit dogmatic in places. For example, the book argues that *evolutionary methods*—methods that directly search the policy space without constructing a value function—do not fall into the scope of reinforcement learning. Why not? One of the criticisms the book gives of evolutionary methods is that they are imprecise about credit assignment—the whole policy is rewarded or punished. But, can't this same criticism be leveled at methods that use function approximation to represent the value function? This and other examples leave us with the feeling that the authors were a bit dismissive of RL methods that aren't TD. The reader should keep in mind that the material in the book is extremely valuable, but occasionally biased.

Let's move on to the second, algorithmic part of the book. Had one of us written this book, we would have been tempted to start its algorithmic part with a description of Q-learning, pointing out that this is by far the most popular reinforcement learning algorithm to date. Not so these authors. The second part of the book introduces three families of methods, which form the basis of virtually all reinforcement learning algorithms: Dynamic programming (DP), Monte Carlo (MC) methods, and temporal difference (TD) learning. This way the reader not only understands contemporary reinforcement learning algorithms, but also appreciates their roots and connections to related (and often older) methods. If you now wonder what the differences are: DP methods compute value functions by backing up values from successor states to predecessor states; they systematically update one state after another, using a model of the next-state distribution. MC methods don't require such a model and instead sample entire trajectories to update the value functions, based on the episodes' final outcomes. TD methods integrate ideas from both DP and MC: Like MC methods, they learn from sampled trajectories, but unlike them and like DP methods, they backup values from state to state (bootstrap). The celebrated Q-learning algorithm is finally introduced in the middle of Chapter 6, as a special case of TD. The intelligent progression of chapters in Part II makes it a pleasure to read!

The third and final part of the book contains five quite different chapters, in which the authors explore issues that go beyond the basic RL paradigms. It begins with a discussion of eligibility traces and TD(λ). A proof of the equivalence of the forward view and the backward view of TD(λ) provides a refreshing mathematical insight in a book that otherwise contains only very few formal results. A thoughtful discussion on function approximation in reinforcement learning follows. This topic is currently an active research area, and the authors pay tribute to this by carefully describing methods that have been found to work well along with their pitfalls and limitations. This section lacks a description of backpropagation, which plays a key role in several successful applications of neural networks to reinforcement learning, but otherwise provides a range of useful examples. Subsequently, a chapter on reinforcement learning and planning forges ties to more traditional work in artificial intelligence and psychology. After a final, brief summary of the entire book, the authors conclude their monologue by a description of six successful applications of reinforcement learning. Included here is an in-depth description of TD-Gammon, one of the most captivating successes of the entire field because it learns to play the game of backgammon on par with the very best human players.

Now, what is there to be said about the book as a whole? It is a gentle, insightful, and balanced introduction into a topic that over the last decade has been subject to intense research. More than that, it ties together the basic ideas in an astonishingly clear way, with many new insights into the relation of different reinforcement learning algorithms. So, should you buy it? If you are a novice to AI and want to learn something about a highly active field in AI, the answer is: definitely yes. If you are a teacher who wants to design a course on reinforcement learning, then yes, buy it, and also purchase "Neuro-Dynamic Programming" by

Bertsekas and Tsitsiklis to avoid running out of material after the first few weeks. If you are an active reinforcement learning researcher who'd like to know the latest and best in the reinforcement learning: Buy it anyhow—be it just for the bibliographical remarks and clever examples—but, be aware that the book isn't intended as a guide to current research. It does, however, highlight intriguing open problems and point out promising lines of future research—indicators that the field of reinforcement learning is healthy, changing, and on the verge of new and important discoveries!

References

Bertsekas, DP and Tsitsiklis, JN (1996). *Neuro-Dynamic Programming*, Athena Scientific, Belmont, Massachusetts.

Reviewed by Sebastian Thrun, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, and Michael L. Littman, Department of Computer Science, Duke University, Durham, North Carolina.