Tutorial Series on Brain-Inspired Computing
Part 4: Reinforcement Learning: Machine Learning and Natural Learning

Shin ISHII and Wako YOSHIDA
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{ishii,wako-y}@is.naist.jp

Received 29 October 2005
Revised manuscript received 28 February 2006

Abstract The theory of reinforcement learning (RL) was originally motivated by animal learning of sequential behavior, but has been developed and extended in the field of machine learning as an approach to Markov decision processes. Recently, a number of neuroscience studies have suggested a relationship between reward-related activities in the brain and functions necessary for RL. Regarding the history of RL, we introduce in this article the theory of RL and present two engineering applications. Then we discuss possible implementations in the brain.

Keywords: Reinforcement Learning, Temporal Difference, Actor-critic, Reward System, Dopamine.

§1 Introduction
When a rodent is placed in a box, called a Skinner box, the rodent receives food when it happens to press the lever attached to the box. By continually receiving food following lever presses, the rodent associates a cause, the lever press, with an effect, the food, and comes to be motivated to press the lever; that is, a reinforcement occurs. This situation, in which an animal's behavior is modified according to its outcome, is called 'the law of effect', and is considered to be the most primitive aspect of behavioral learning. Although this example was a simple association between a cause and an effect, the case of general motor controls is more complicated. Let us consider, for example, learning to snowboard. When a learner is successful in performing well, on one hand, this outcome is seen as cool and so the behavior seems to be rewarded. Tumbling on the snow, on the other hand, could be painful and thus this outcome could be viewed as a punishment. Intuitively, learning complicated motor controls
in snowboarding would proceed based on such rewards or punishments, but understanding the process requires going beyond simple learning by association. The classical Rescorla-Wagner (R-W) model can explain simple instrumental conditioning exemplified by the rodent case, while it cannot explain sequential motor learning such as occurs in learning to snowboard.

Reinforcement learning (RL) is a natural extension of the R-W model to enable description of sequential learning by means of the conventional delta rule. Over the last decade in the field of neuroscience, the theory of RL has been extended to allow for explanations of not only lower-order neural activities related to rewards but also higher-order brain activities involved in decision making, although their evidence is still controversial. Reinforcement learning has also been recognized as an approach to Markov decision processes which were studied in the 1950s in the field of control theory, and which subsequently have become a major research topic in machine learning. In the field of artificial intelligence, RL is now attracting particular attention, because adaptability to and optimal control in complicated, dynamic and even unknown environments could be described by the RL theory. Paying attention to such a history of RL, in the first part of this article, we introduce the theory of RL and present two engineering applications. In the second part, we discuss possible implementations in the brain.

§2 Theory of Reinforcement Learning

2.1 Value-based Reinforcement Learning

As in Fig. 1, we have a controller (or agent) and a system (or environment) to be controlled (or interacted with). At a discrete time $t$, the controller emits a control signal, often called an action, $a_t$, based on the system's state $x_t$. The controller's function from a state $x_t$ to an action $a_t$ is called a policy, $\pi$, and represented as $a_t = \pi(x_t)$ or $a_t \sim \pi(a_t|x_t)$ for a deterministic or stochastic policy, respectively. The system is assumed to be probabilistic and memory-less, i.e., Markov, so that the state transition probability for reaching a next state $x'$ from a previous state $x$ by an action $a$ is given by $P(x'|x,a)$. A deterministic system is regarded as a special case of the probabilistic system. We assume that