the emissions of atomic particles, among the most random of physical phenomena.

Experiments have documented a relationship between reinforced variability and "voluntary choice" (Neuringer and Jensen 2010). For example, human participants judged the choices of "actors" on a computer screen as approximating a voluntary agent when the responses combined two qualities: "matching" choice proportions to reinforcement proportions, this demonstrating the functional nature of the choices; and varying levels of variability/predictability. Voluntary acts are hypothesized to be *functional* responses that are *more or less predictable* depending upon the demands of the situation, i.e., the contingencies of reinforcement.

Many questions remain. Research is underway attempting to identify the underlying physiological bases of stochastic-response emission, e.g., when male songbirds generate variable songs and when monkeys choose among alternative options (Glimcher 2005). Operant variability procedures have been applied to individuals with psychological disabilities such as attention deficit hyperactivity disorder, autism, and depression. Drugs have been used to distinguish between memory-based and stochastic emission of variable responses. Studies with animal models have shown that reinforced variability facilitates acquisition of difficult-to-learn responses, but attempts to replicate with humans have not been successful. Other areas of initial studies include the influences of reinforced variability on problem solving, motor skills learning, and artistic creativity (see Neuringer 2002 for a review). Each of these areas provides rich questions for additional research.

Cross-References

► Adaptive Learning Through Variation and Selection
► Creativity, Problem Solving, and Feeling
► Learning in Autism
► Matching
► Operant Behavior
► Play, Exploration, and Learning
► Probability Learning
► Reinforcement Learning in Animals
► Schedules of Reinforcement
► Skinner B.F. (1904–1990)
► Variability of Practice

References

Blough, D. S. (1966). The reinforcement of least frequent interresponse times. *Journal of the Experimental Analysis of Behavior, 9*, 581–591.

Glimcher, P. W. (2005). Indeterminacy in brain and behavior. *Annual Review of Psychology, 56*, 25–56.

Neuringer, A. (2002). Operant variability: Evidence, functions, and theory. *Psychonomic Bulletin & Review, 9*, 672–705.

Neuringer, A., & Jensen, G. (2010). Operant variability and voluntary action. *Psychological Review, 117*, 972–993.

Pryor, K. W., Haag, R., & O'Reilly, J. (1969). The creative porpoise: Training for novel behavior. *Journal of the Experimental Analysis of Behavior, 12*, 653–661.

Skinner, B. F. (1974). *About behaviorism*. New York: Knopf.

Reinforcement

► Feedback and Learning

Reinforcement Learning

Luis R. Izquierdo[1], Segismundo S. Izquierdo[2]
[1]University of Burgos, Burgos, Spain
[2]University of Valladolid, Valladolid, Spain

Definition

Reinforcement learners interact with their environment and use their experience to choose or avoid certain actions based on their consequences. Actions that led to high rewards in a certain situation tend to be repeated whenever the same situation recurs, whereas choices that led to comparatively lower rewards tend to be avoided.

Theoretical Background

Reinforcement learning (RL) is a term that is interpreted differently in different disciplines. All interpretations, however, share the same underlying idea that reinforcement learners interact with their environment and use their experience to choose or avoid certain actions based on their consequences. Actions that led to high rewards in a certain situation tend to be repeated whenever the same situation recurs, whereas

choices that led to comparatively lower rewards tend to be avoided. The roots of RL date back to Thorndike's animal experiments on instrumental learning at the end of the nineteenth century. The results of these experiments were formalized in the well-known "Law of Effect," which is nowadays one of the most robust properties of learning in the experimental psychology literature:

▶ *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 to the 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, p. 244)

Nowadays there is little doubt that RL is an important aspect of much learning in most animal species, including many phylogenetically very distant from vertebrates.

This entry gives a brief overview of the two main branches of reinforcement learning in the literature. One of the branches has developed mainly within Economics, especially in Game Theory, and its interpretation of the term RL remains close to its origins in the psychological learning literature. In essence, this branch studies learning rules based on the Law of Effect, which dictates that responses are evaluated on the basis of the satisfaction that *accompanies or closely follows* the response. The other branch has grown within the field of Artificial Intelligence, particularly in the context of Machine Learning, and its interpretation of the term RL is considerably more forward-looking than the interpretation in Economics. To be specific, responses in this branch are not evaluated only on the basis of the immediate reward that follows the response in a certain situation, but special attention is paid to the future consequences of the response in terms of the type of situations the response may lead to, and the magnitude of the rewards that the agent expects to obtain in those future situations. Thus, the aim of RL in Machine learning is to design efficient algorithms to maximize *the flow* of numerical rewards that an agent receives by interacting with its environment, where his decisions not only affect the immediate reward, but also the situation the agent faces next, and, through that, all subsequent rewards too (Sutton and Barto 1998). This branch, which is closely related to dynamic programming, assumes a particular framework – explained below – which can nevertheless be applied to a very wide range of real-world problems.

In the Economics literature, reinforcement learning is seen as a form of bounded rationality in decision-making. The decisions that economic agents make lead to different payoffs, and much effort in the field of Economics is devoted to studying different decision-making procedures. In stark contrast with the postulates of perfect rationality, reinforcement learners are assumed to have incomplete knowledge of the environment they are embedded, and act in a simple stimulus–response way: The propensity to repeat a certain decision is positively related to the amount of satisfaction the agent obtained when he made such a decision in the past.

In general, reinforcement learners in Economics decide what action to select stochastically, and use the immediately received payoff to adjust the probability of undertaking each action accordingly. The precise way in which these probabilities are revised in the light of experience is what distinguishes the different RL models in the Economics literature. More specifically, reinforcement models tend to differ in the following, somewhat interrelated, features:

- Whether learning slows down with experience or not, i.e., whether the model accounts for the "Power Law of Practice," which states that the sensitivity of an agent to a new single observation decreases with his cumulative experience. This feature is present in RL models where agents select an action with probability proportional to the *cumulative* payoff that the agent has obtained with that action in the past (e.g., Erev and Roth 1998). In contrast, a popular model where learning does not fade with time is Bush and Mosteller's (1955) linear stochastic model of reinforcement learning.
- Whether the model allows for avoidance behavior in addition to approach behavior. Approach behavior is the tendency to repeat the associated choices after receiving a positive stimulus; avoidance behavior is the tendency to avoid the associated actions after receiving a negative stimulus (one that does

not satisfy the player). Models that allow for negative stimuli tend to define an aspiration level against which achieved payoffs are evaluated (e.g., Bush and Mosteller 1955). This aspiration level may be fixed or vary endogenously.

- Whether "forgetting" is considered, i.e., whether recent experience plays a larger role than past experience in determining behavior.
- Whether the model imposes a positive bias in favor of the most recently selected action (inertia).

Each of the features above can have important implications for the dynamics of the particular model under consideration and for the mathematical methods that are adequate for its analysis. For example, when learning slows down with experience, theoretical results from the theory of stochastic approximation and from the theory of urn models can often be applied, whereas if the learning rate is constant, results from the theory of distance diminishing models (Norman 1972) tend to be more useful.

Besides having been analyzed theoretically, some RL models in Economics have also been empirically tested. In the context of experimental game theory with human subjects, several authors have used simple RL models to try to explain and predict behavior in a wide range of games (e.g., Erev and Roth 1998). Empirical evidence for this backward-looking interpretation of RL seems to be strongest in animals with limited reasoning abilities and in human subjects who have no information beyond the payoff they receive and specifically may be unaware of the strategic nature of the situation.

In the context of machine learning, the field of RL is devoted to solving a particular problem called the RL problem (Sutton and Barto 1998). In the RL problem, there is an agent with a particular goal and an environment with which the agent interacts. In broad terms, the RL problem consists in finding efficient algorithms that enable the agent to achieve his goal. Formally, the problem is represented as a *Markov Decision Process*. The agent–environment interaction occurs in discrete time-steps. In any given time-step t, the agent perceives the environment's *state* s_t and selects an *action* a_t. As a consequence of selecting action a_t in state s_t the agent receives a numerical *reward* $r_t(s_t, a_t)$ and finds himself in another state $s_{t+1}(s_t, a_t)$. Both the state transition function $s_{t+1}(s_t, a_t)$ and the reward function $r_t(s_t, a_t)$

are generally stochastic and satisfy the *Markov Property*, i.e., their distribution depends only on the current state s_t and the current action a_t.

A *policy* is a (potentially stochastic) decision rule that dictates what action to take at each possible state. Thus, given any particular starting state, the use of a certain policy determines a probability distribution for the agent's future rewards. The goal of the agent is to find a policy that maximizes the total expected flow of (potentially discounted) rewards he receives over time, known as his expected return:

$$E\left[\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t)\right]$$

where γ is a discount factor that controls the relative importance of future rewards in comparison with more recent ones. Most often the problem has a natural end state, so the infinite sum is actually truncated. The expected return that an agent implementing policy π will obtain when he starts at state s is called the *value* of state s under policy π, and is denoted $V^\pi(s)$. Thus, the aim of the agent is to find an optimal policy, i.e., a policy that yields the highest possible value at every possible state.

As an example, consider an agent on a 3×3 grid whose goal is to reach the upper-right corner of the grid as soon as possible. The nine states are the different cells in the grid. Assuming the agent has not already reached the upper-right corner, he moves one step in any of the four cardinal directions within the grid at each time-step (see Fig. 1a). The formalization of the agent's goal is done by defining the numerical rewards $r_t(s_t, a_t)$. In this case, the goal can be formalized assigning a reward of -1 to every possible pair (s_t, a_t) except for those involving the upper-right corner (i.e., the natural end state), which can be set to 0. There is no need for discounting, i.e., $\gamma = 1$. Figure 1 shows (a) the set of possible actions in each state, (b) the value $V^{\text{random}}(s)$ of each state under a policy that dictates unbiased random selection of possible actions, (c) the value $V^{\text{optimal}}(s)$ of each state under an optimal policy, and (d) an optimal policy.

Much of the interest in this branch of RL comes from the fact that, in general, the agent starts off without knowing the state transition distribution $s_{t+1}(s_t, a_t)$ and/or the reward distribution $r_t(s_t, a_t)$. In these cases, there is a clear trade-off between *exploration* and
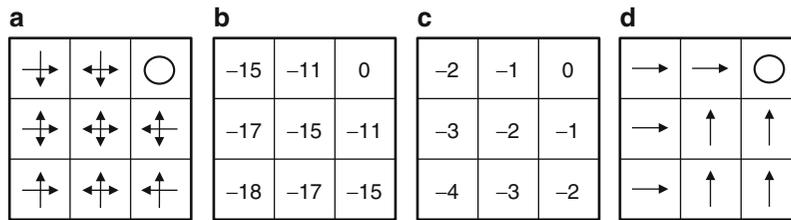
Reinforcement Learning. Fig. 1 For an agent whose goal is to reach the upper-right corner of the grid: (**a**) The set of possible actions in each state, (**b**) the value of each state under a policy that dictates unbiased random selection of possible actions, (**c**) the value of each state under an optimal policy, and (**d**) an optimal policy

exploitation: To secure high rewards the agent must repeat the actions that have proved to lead to high rewards in the past (i.e., *exploit* what he already knows) but, on the other hand, the agent must also try out new actions since these may lead to even higher rewards (i.e., *explore* its environment). Thus, RL algorithms such as Q-learning (Watkins and Dayat 1992) are designed to progressively refine an initial policy maintaining a balance between exploration and exploitation.

Important Scientific Research and Open Questions

In the context of Economics, one of the most important questions to be developed further is the empirical validation of the proposed models, from the double perspective of (a) empirically justifying the hypotheses or behavioral assumptions in the model, and (b) empirically testing the analytical conclusions or predictions that may be extracted from these models (e.g., the expected equilibria in games).

In the context of machine learning, one active line of research is the design and study of RL algorithms that make an efficient use of the available information and have good computational scalability. Another interesting topic is the design of algorithms with modified value functions or objectives, e.g., instead of considering only the expected value of the flow of rewards, one can look for a policy that also considers the variance of the rewards.

Cross-References

References

Bush, R., & Mosteller, F. (1955). *Stochastic models of learning*. New York: Wiley.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review, 88*(4), 848–881.

Norman, M. F. (1972). *Markov processes and learning models*. New York: Academic Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Thorndike, E. L. (1911). *Animal intelligence*. New York: The Macmillan Company.

Watkins, C. J. C. H., & Dayat, P. (1992). Technical note: Q-learning. *Machine Learning, 8*, 279–292.

R

Reinforcement Learning in Animals

ELLIOT A. LUDVIG
Princeton Neuroscience Institute, 3-N-12 Green Hall, Princeton University, Princeton, NJ, USA

Synonyms

Approximate dynamic programming; Neuro-dynamic programming; Reward-related learning; Trial-and-error learning