Reinforcement Learning in Motor Control

Andrew G. Barto

November 26, 2003

INTRODUCTION

How do we learn motor skills such as reaching, walking, swimming, or riding a bicycle? There is a large literature on motor skill acquisition which is full of controversies (for an introduction to human motor control, see Schmidt and Lee, 1999), but there is general agreement that motor learning requires the learner, human or not, to receive response-produced feedback through various senses providing information about performance. Careful consideration of the nature of the feedback used in learning is important for understanding the role of reinforcement learning in motor control (see REINFORCEMENT LEARNING). One function of feedback is to guide the performance of movements. This is the kind of feedback with which we are familiar from control theory, where it is the basis of servo control, although its role in guiding animal movement is more complex. Another function of feedback is to provide information useful for improving *subsequent* movement. Feedback having this function has been called *learning feedback*. Note that this functional distinction between feedback for control and for learning does not mean that the signals or channels serving these functions need to be different.

LEARNING FEEDBACK

When motor skills are acquired without the help of an explicit teacher or trainer, learning feedback must consist of information automatically generated by the movement and its consequences on the environment. This has been called *intrinsic feedback* (Schmidt and Lee, 1999). The "feel" of a successfully completed movement and the sight of a basketball going through the hoop are examples of intrinsic learning feedback. A teacher or trainer can augment intrinsic feedback by providing *extrinsic feedback* (Schmidt and Lee, 1999) consisting of extra information added for training purposes, such as a buzzer indicating that a movement was on target, a word of praise or encouragement, or an indication that a certain kind of error was made.

Most research in the fields of machine learning and artificial neural networks has focused on the learning paradigm called *supervised learning*, which emphasizes the role of training information in the form of desired, or 'target', network responses for a set of training inputs (PERCEPTRONS, ADALINES, AND BACKPROPAGATION). However, motor learning is more complex than supervised learning even when it involves extrinsic feedback provided by a trainer. For example, a

trainer can tell or show us what to do, explicitly guide our movements, give us hints on how to deal with difficult parts of a skill, tell us when we have improved or done badly, etc. The aspect of real training that corresponds most closely to the supervised learning paradigm is the trainer's role in telling or showing the learner what to do, or explicitly guiding his or her movements. These activities provide standards of correctness that the learner can try to match as closely as possible by reducing the error between its behavior and the standard. Supervised learning can also be relevant to motor learning when there is no trainer because intrinsic feedback can be used to learn various kinds of *models* that are useful for motor control. Kawato (1999) and Desmurget and Grafton (2000) discuss some of the uses of models in motor control.

In contrast to supervised learning, reinforcement learning emphasizes learning feedback that *evaluates* the learner's performance without providing standards of correctness in the form of behavioral targets (see REINFORCEMENT LEARNING). Although the most obvious evaluative feedback is extrinsic feedback provided by a trainer, most evaluative feedback is probably intrinsic, being derived by the learner from sensations generated by a movement and its consequences on the environment: the kinesthetic and tactile feel of a successful grasp or the swish of a basketball through the hoop. Evaluative feedback is often called *reinforcement* feedback (and it need not involve pleasure or pain). A reinforcement learning has to actively try alternatives, compare the resulting evaluations, and use some kind of selection mechanism to guide behavior toward the better alternatives. This basic idea follows Thorndike's classical "Law of Effect" (Thorndike, 1911) and is commonly called learning by trial-and-error (not to be confused with error-correction, or supervised, learning).

The great Russian physiologist Nikolai Bernstein discussed the role of trial-and-error learning in motor control in his classic 1967 book (Bernstein, 1967). He distinguished his view from the concept of random undirected search which he attributed to the behaviorists. According to Bernstein, the process must be an active search involving "gradient extrapolation" by probabilistic sampling so that each attempt is informed by previously acquired information about "how and where the next step must be taken." This is very much in accord with modern concepts of reinforcement learning, where randomness is often used to generate behavioral variety, but action selections are strongly constrained by evaluations of earlier experience (see REINFORCEMENT LEARNING). To Bernstein, this kind of search was important for motor behavior, especially for movements requiring high levels of co-ordination. Another motor control theorist, Jack Adams, provided an interesting discussion of the role of the law of effect in motor control in a 1978 article (Adams, 1978). Although he called into question some of the details of Thorndike's theories, he affirmed the importance of reinforcement learning in motor control.

Motor learning involves feedback carrying many different kinds of information. Consequently, it is incorrect to view motor learning strictly in terms of either supervised, reinforcement, or any other learning paradigms that have been formulated for theoretical study. Aspects of all of these paradigms play interlocking roles, with their relative importance undoubtedly varying with the type of task as well as the developmental stage. However, reinforcement learning may be an essential component of motor learning simply because evaluative feedback is more easily obtained than many other kinds of learning feedback.

Figure 1: Panel A: A Basic Control Loop. A controller provides control signals to a controlled system, whose behavior is influenced by disturbances. Feedback from the controlled system to the controller provides information on which the control signals can depend. Commands to the controller specify aspects of the control task's objective. Panel B: A Control System with Learning Feedback. A *critic* provides the controller with a reinforcement signal evaluating its success in achieving the control objectives.

LEARNING FROM CONSEQUENCES

To illustrate how reinforcement learning applies to motor learning, we first discuss it within the general context of control. Then we describe several special cases related to motor control. Figure **??**, Panel A, is a variation of the classical control system diagram. A controller provides control signals to a controlled system. The behavior of the controlled system is influenced by disturbances, and feedback from the controlled system to the controller provides information on which the control signals can depend. Commands to the controller specify aspects of the control task's objective.

In Figure **??**, Panel B, the control loop is augmented with another feedback loop that provides learning feedback to the controller. In accordance with common practice in reinforcement learning, a *critic* is included that generates evaluative learning feedback on the basis of observing the control signals and their consequences on the behavior of the controlled system. The critic also needs to know the command to the controller because its evaluations must be different depending on what the controller should be trying to do. The critic is an abstraction of whatever process supplies evaluative learning feedback, both intrinsic and extrinsic, to the learning system. It is often said that the critic provides a *reinforcement signal* to the learning system. In most artificial reinforcement learning systems, the critic's output at any time is a number that scores the controller's behavior: the higher the number, the better the behavior. Assume for the moment that the behavior being scored is the some immediately preceding unit of behavior. We discuss what a unit of behavior might be, as well as more complex temporal relationships below. For this process to work, there must be some *variability* in the controller's behavior so that the critic can evaluate many alternatives. A learning mechanism can then adjust the controller's behavior so that it tends toward behavior that is favored by the critic.

A learning rule particularly suited to reinforcement learning control systems implemented as artificial neural networks was developed by Gullapalli (1990) in the form of what he called a Stochastic Real-Valued (SRV) unit. An SRV unit's output is produced by adding a random number to the weighted sum of the components of its input pattern. The random number is drawn from a zero-mean Gaussian distribution. This random component provides the unit with the variability necessary for it to 'explore' its activity space. When the reinforcement signal indicates that something good happened just after the unit emitted a particular output value in the presence of some input pattern, the unit's weights are adjusted to move the activation in the direction in which it was perturbed by the random number. This has the effect of increasing the probability that future outputs generated for that input pattern (and similar input patterns) will be closer to the output value just emitted. If the reinforcement signal indicates that something bad happened, the weights are adjusted to move future output values away from the value just emitted. Another part of the SRV learning rule decreases the variance of the Gaussian distribution as learning proceeds. This decreases the variability of the unit's behavior, with the goal of making it eventually stick (i.e.,

Figure 2: Block Diagram of a Reinforcement Learning Controller of an Arm (after Figure 1 of Lipitkas et al., 1993). Given inputs coding the starting and target positions of the hand, the network controller learns to provide correct parameters to a torque generator which generates, in open-loop mode, time-varying torque signals to the arm. The reinforcement signal evaluates the success of each movement after its completion.

become deterministic) at the best output value for each input pattern. Using this learning rule, an SRV unit learns to produce the best output in response to each input pattern (given appropriate assumptions). Unlike more familiar supervised learning units, it is never given target outputs; it has to discover what outputs are best through an active exploration process.

OVERCOMING THE DISTAL ERROR PROBLEM

As a simple illustration of how reinforcement learning can be useful in motor learning, consider the problem of learning to reach to specific points in space starting from a variety of initial hand positions. Lipitkas et al. (1993) proposed a particularly straightforward method (although not as a model of the human learning process, which is much more complex). Their controller is an artificial neural network receiving inputs coding the initial spatial location and the desired, or target, spatial location of the hand (ignoring hand orientation). The six outputs of the network provide parameters to a torque generator that generates time-varying signals for driving the joint actuators of a dynamic arm model (Figure **??**). The time-varying signals are parameterized by six numbers determining characteristics of their wave-like shapes (e.g., giving the magnitudes and relative timing of the half-waves). During each movement, the controller operates in open-loop mode, generating the torque time functions without the aid of sensory feedback. The problem for the network is to learn a function associating each pair of hand starting and target positions to the values of the six torque-generator parameters that will accomplish the movement.

A straightforward application of supervised learning is not possible here because the required training examples are not available: It is not known what parameters will work for any pair of starting and target positions (except possibly the trivial cases in which the starting position is already the target position, but these are not useful as training examples). This is an instance of what has been called the *distal error problem* (Jordan and Rumelhart, 1992) for supervised learning. This problem is present whenever the standard of correctness required for supervised learning is available in a coordinate system that is different from the one in which the learning system's activity must be specified for learning. In the case of learning how to move the hand from a starting position to a target position, the standard of correctness is the target position, but what must be learned are the control signals to the joint actuators, that is, to the muscles. The hand position error is distal to the output of the controller that has to be learned. Although a nonzero distal error vector indicates that the controller made an error, it does not tell the controller how it should change its output in order to reduce the error.

The distal error problem can be solved by using a model of the controller's influence on the arm's movement (possibly learned via supervised learning) to translate distal error vectors into error vectors required for supervised learning (Jordan and Rumelhart, 1992). Another approach is to learn an inverse model of the controller's influence on the arm's movement (Jordan and Rumelhart,

1992; Kawato, 1999). Reinforcement learning offers another way of to overcome the distal error problem because it does not need learning feedback in the form of error vectors. Continuing with the reaching example, Lipitkas et al. (1993) defined a reinforcement signal that attains a maximum value of one if the hand reaches the desired position and stops there. The signal decreases depending on the distance between the hand's final position and the target position and on its tangential velocity as it passes the target position. The reinforcement signal could include other criteria of successful movements as well. With inputs coding starting and target hand positions, the network employs SRV units to generate six parameter values using its current weights. The torque generator generates a movement using these parameter values. When the movement is completed, it is scored by the reinforcement signal, and the network's weights are changed according to Gullapalli's SRV learning rule. After a few thousand movements with different starting and target hand positions, the system could move with reasonable accuracy for new pairs of starting and target positions as well as for the pairs on which it was trained. This amount of practice is required because the system effectively has to search the six-dimensional parameter space for each starting and target position. A more complicated example of reinforcement learning using SRV units is the work on biped walking by Benbrahim and Franklin (1997).

The relative advantages and disadvantages of supervised and reinforcement learning approaches to the distal error problem have been discussed by many researchers. It is clear that reinforcement learning approaches are simpler, but reinforcement learning is usually slower in terms of the amount of experience required for learning. This is true because reinforcement learning methods extract less information from each experience than do the model-based supervised approaches. However, in some problems it is easier to learn the right actions than it is to model their effects on a complicated process. Reinforcement learning methods are also more plausible from the perspective of neuroscience (see below), while the backpropagation process often used by supervised approaches is more difficult to reconcile with what we know about neural mechanisms. In practical terms, which approach is more advantageous will depend on aspects of the specific problem being considered.

COLLECTIVE BEHAVIOR

Another property of reinforcement learning that might be relevant to motor control is the ability of a "team" of reinforcement learning systems to learn to cooperate so that the team as a whole improves performance. Here is an example presented in a 1965 lecture by the cybernetician Mikhail Tsetlin (Tsetlin, 1973), a pioneer in the study of simple reinforcement learning systems called "learning automata." He presented the basic idea as follows in terms of human players (the so-called Goore game). Suppose there is a referee and some number of players. The referee can see the players but the players cannot see one another. At the sound of a buzzer, each player is to raise one or two fingers. The referee determines what percentage of players raised one finger, then pays each player a fixed amount with a probability that depends only on this percentage (and is the same for each player). The process repeats each time the buzzer sounds. It turns out that for any number of players each implementing a sufficiently competent reinforcement learning rule, eventually each player will settle on raising either one or two fingers so that the percentage of those raising one finger is (with probability close to 1) a local maximum of whatever payoff function the referee uses. This occurs with no direct communication among the players and no agreements of any kind among them.

It is possible to extend this result to one in which the referee provides payments based not just on the percentage of players raising one finger, but on *any function whatsoever* of the pattern of players' fingers. One can see how this is an instance of the problem of learning with a distal teacher, with the added complication that the payoff, or reinforcement signal, to each player is extremely noisy due to the noise introduced by the actions of the other players (in addition to the referee's probabilistic payoff method).

Tsetlin speculated that the recruitment of motor units can be reduced to this type of problem. Here, the problem would be to activate the right number of motor units to obtain a pull of a given force. The referee corresponds to a process that evaluates the results of the collective behavior of the entire pool of motor units on the resulting force. The collective behavior of reinforcement learning systems has been studied by many researchers (e.g., Narendra and Thathachar, 1974; Barto, 1985), although no modern work following up Tsetlin's suggestion about motor unit recruitment appears to exist.

CREDIT ASSIGNMENT PROBLEMS

The challenge of reinforcement learning is often summed up as various kinds of *credit assignment* problems. A scalar evaluation of a complex mechanism's behavior does not indicate which of its many action components, both internal and external, were responsible for the evaluation. This makes it difficult to determine which of these components deserve the credit (or the blame) for the evaluation. This problem is sometimes referred to as the *structural* credit assignment problem: How is credit assigned to the internal workings of a complex structure? One approach is to assign credit equally to *all* the components so that through a process of averaging over many variations of the behavior, the components that are key in producing laudable behavior end up gaining the most strength, while inappropriate components are weakened. This is the general approach illustrated above by the Goore game.

The fact that reinforcement learning can work under these circumstances makes neural implementation quite plausible. A single reinforcement signal uniformly *broadcasted* to all the sites of learning, either neurons or individual synapses, is consistent with anatomical and physiological evidence showing the existence of diffusely projecting neural pathways by which neuromodulatory chemicals can be widely and nonspecifically distributed. It has been suggested that some of these pathways may play a rule in reward-mediated learning. A specific hypothesis is that dopamine mediates synaptic enhancement in the corticostriatal pathway in the manner of a broadcasted reinforcement signal (see DOPAMINE, ROLE OF). This may be one of the ways in which reinforcement learning is implemented for motor control.

Another aspect of the credit assignment problem occurs when the temporal relationship between a system's behavior and evaluations of that behavior is not as simple as assumed above. How can reinforcement learning work when the learner's behavior is temporally extended and evaluations occur at varying and unpredictable times? Under these more realistic conditions, it is not always clear what elements of behavior are being evaluated. This has been called the *temporal* credit assignment problem. It is especially relevant in motor control because movements extend over time and evaluative feedback may become available only after the end of a movement. An approach to this problem that is receiving considerable attention is the use of methods by which the critic itself

can learn to provide useful evaluative feedback immediately after the evaluated event. According to this approach, reinforcement learning is not only the process of improving behavior according to given evaluative feedback; it also includes learning how to improve the evaluative feedback itself. The strong parallels between algorithms for adapting evaluative feedback (temporal difference methods; see REINFORCEMENT LEARNING) and the properties of dopamine producing neurons in the brain (see DOPAMINE, ROLE OF) make it plausible that the brain uses similar methods for dealing with the temporal credit assignment problem.

The modern view of reinforcement learning developed by machine learning researchers uses the framework of stochastic optimal control to study the temporal credit assignment problem (see REINFORCEMENT LEARNING). From this perspective, reinforcement learning algorithms are methods for approximating solutions to complex stochastic optimal control problems via relatively simple mechanistic learning rules. Because optimality principles have played significant roles in theories of motor control (Engelbrecht, 2001), and because stochasticity may be an important element of motor control (Harris, 1998), the modern theory of reinforcement may prove to be of great utility in extending our understanding of motor learning.

DISCUSSION

As emphasized above, motor learning is too complex to view strictly in terms of either supervised learning or reinforcement learning. Feedback used in motor learning ranges from specific standards of correctness to nonspecific evaluative information, and many learning mechanisms with differing characteristics probably interact to produce the motor learning capabilities of animals. However, reinforcement learning principles may be indispensable for motor learning because they seem necessary for improving motor performance when the standards of correctness required by supervised learning are not available.

REFERENCES

Adams, J. A., 1978, Theoretical issues for knowledge of results, in Information Processing in Motor Control and L (G. E. Stelmach, Ed.), New York: Academic Press, pp. 229-240.

Barto, A.G., 1985, Learning by statistical cooperation of self-interested neuron-like adaptive elements, Human Neurobiology, 4: 229-256.

Benbrahim, H. and Franklin, J. A., 1997, Biped dynamic walking using reinforcement learning, Robotics and Autonomous Systems, 22: 283-302.

Bernstein, N., 1967, The Co-ordination and Regulation of Movements, Oxford, Pergamon Press.

* Desmurget, M. and Grafton, S., 2000, Forward modeling allows feedback control for fast reaching movements, Trends in Cognitive Sciences, 4:423-431.

* Engelbrecht, S. E., 2001, Minimum principles in motor control, Journal of Mathematical Psychology, 45:497-542.

Gullapalli, V., 1990, A stochastic reinforcement algorithm for learning real-valued functions, Neural Networks, 3:671-692.

Harris, C. M., 1998, On the optimal control of behavior: a stochastic perspective, Journal of Neuroscience Methods 83: 73-88.

Jordan, M.I. and Rumelhart, D.E., 1992, Supervised learning with a distal teacher, Cognitive Science, 16:307-354.

Kawato, M., 1999, Internal Models for motor control and trajectory planning, Current Opinion in Neurobiology, 9:718-727.

Lipitkas, J., D'Eleuterio, G.M.T., Bock, O., and Grodski, J.J., 1993, Reinforcement learning and the parametric motor control hypothesis applied to robotic arm movements, in Proceedings of the DND Work Ottawa, CDN, 1993.

Narendra, K. and Thathachar, M. A. L., 1989, Learning Automata: An Introduction, Englewood Cliffs, NJ: Prentice Hall.

* Schmidt, R.A. and Lee, T. D., 1999, Motor Control and Learning: A Behavioral Emphasis, Third Edition, Champaign, IL: Human Kinetics Publishers.

Thorndike, E.L., 1911, Animal Intelligence, Darien, CT: Hafner.

Tsetlin, M. L., 1973, Automata Theory and Modeling of Biological Systems, New York: Academic Press.

FIGURE CAPTIONS

Figure 1. Panel A: A Basic Control Loop. A controller provides control signals to a controlled system, whose behavior is influenced by disturbances. Feedback from the controlled system to the controller provides information on which the control signals can depend. Commands to the controller specify aspects of the control task's objective. Panel B: A Control System with Learning Feedback. A *critic* provides the controller with a reinforcement signal evaluating its success in achieving the control objectives.

Figure 2. Block Diagram of a Reinforcement Learning Controller of an Arm (after Figure 1 of Lipitkas et al., 1993). Given inputs coding the starting and target positions of the hand, the network controller learns to provide correct parameters to a torque generator which generates, in open-loop mode, time-varying torque signals to the arm. The reinforcement signal evaluates the success of each movement after its completion.