In: *Models of Information Processing in the Basal Ganglia,* J. C. Houk, J. Davis and D. Beiser (Eds.), Cambridge, MA: MIT Press, 1995, pp. 215-232.

Adaptive Critics and the Basal Ganglia

Andrew G. Barto
Department of Computer Science
University of Massachusetts, Amherst MA 01003

One of the most active areas of research in artificial intelligence is the study of learning methods by which "embedded agents" can improve performance while acting in complex dynamic environments. An agent, or decision maker, is embedded in an environment when it receives information from, and acts on, that environment in an ongoing closed-loop interaction. An embedded agent has to make decisions under time pressure and uncertainty and has to learn without the help of an ever-present knowledgeable teacher. Although the novelty of this emphasis may be inconspicuous to a biologist, animals being the prototypical embedded agents, this emphasis is a significant departure from the more traditional focus in artificial intelligence on reasoning within circumscribed domains removed from the flow of real-world events. One consequence of the embedded agent view is the increasing interest in the learning paradigm called *reinforcement learning* (RL). Unlike the more widely studied *supervised learning* systems, which learn from a set of examples of correct input/output behavior, RL systems adjust their behavior with the goal of maximizing the frequency and/or magnitude of the reinforcing events they encounter over time.

While the core ideas of modern RL come from theories of animal classical and instrumental conditioning (although the specific term "reinforcement learning" is not used by psychologists), the influence of concepts from artificial intelligence and control theory has produced a collection of computationally powerful learning architectures. Despite similarities between some of these architectures and the structure and function of certain brain regions, relatively little effort has been made to relate these architectures to the nervous system (but see Houk 1992, Klopf 1982, Wickens 1990, and Werbos 1987). In this article I describe the RL system called the *actor-critic* architecture, giving enough detail so that it can be related to basal-ganglionic circuits and dopamine neurons. Specifically, I focus on a component of this architecture called the *adaptive critic*, whose behavior seems remarkably similar to that of the dopamine neurons projecting to the stiatum and frontal cortex (Schultz, this workshop). In a companian article in this volume, Houk, Adams, and Barto (1994) present a hypothesis about how the actor-critic architecture might be implemented by the circuits of the basal ganglia and associated brain structures. My explanation of the

adaptive critic largely follows that of Sutton (1984, 1988).

The adaptive critic is a device that learns to *anticipate reinforcing events* in a way that makes it a useful conjunct to another component, the *actor*, that adjusts behavior to maximize the frequency and/or magnitude of reinforcing events. The adaptive critic also forms the basis of the *temporal difference model* of classical, or Pavlovian, conditioning (Sutton and Barto 1987, 1990) which extends the Rescorla-Wagner model (Rescoral and Wagner 1975) to take into account some of the fine temporal structure of conditioning. The learning rule used by the adaptive critic is due to Sutton, who was developing it as part of his Ph.D. dissertation (Sutton 1984) when it was used in the pole-balancing system of Barto, Sutton, and Anderson (1983). Sutton (1988) developed this class of learning algorithms further, calling them temporal difference (TD) methods.

This line of research work began with the exploration of Klopf's (1972, 1982) idea of *generalized reinforcement* which emphasized the importance of sequentiality in a neuronal model of learning. An earlier precursor, however, is the technique used by Samuel (1959) in his learning program for the game of checkers. Current research on the adaptive critic focuses on its relationship to an optimization technique known as dynamic programming used for solving control problems. This connection follows the research of Werbos (1977, 1987) and Watkins (1989). Barto, Sutton, and Watkins (1990) and Barto, Bradtke, and Singh (1994) provide detailed accounts of this perspective. A remarkable demonstration of the power of the actor-critic architecture is provided by Tesauro's (1992) backgammon playing program, which used an actor-critic architecture to learn how to play world-class backgammon.

Reinforcement Learning

Following the basic idea of Thorndike's "Law of Effect" (Thorndike 1911), the simplest RL algorithms are based on the commonsense idea that if an action is followed by a satisfactory state of affairs, or an improvement in the state of affairs, then the tendency to produce that action is strengthened, i.e., reinforced. Although this is often called "trial-and-error" learning, I prefer to call it learning based on the "generate-and-test" procedure: alternatives are generated, evaluated by testing them, and behavior is directed toward the better alternatives. The reason for my preference is that it is too easy to confuse trial-and-error learning with supervised learning.

For example, an artificial neural network trained using the well-known supervised learning method of error backpropagation (e.g., Rumelhart, Hinton, and Williams 1986) produces an output, receives an error vector, and adjusts the network's weights to reduce the magnitude of the error. This is a kind of trial-and-error learning, but it differs from the kind of learning Thorndike had in mind. Error vectors in supervised learning are derived from standards of correctness: the 'target' responses of supervised learning. In contrast, RL emphasizes response-dependent feedback that *evaluates* the learner's performance by pro-

cesses that do not necessarily have access to standards of correctness. Evaluative feedback tells the learner whether or not, and possibly by how much, its behavior has improved; or it provides a measure of the 'goodness' of the behavior; or it just provides an indication of success or failure. Evaluative feedback does not directly tell the learner what it *should* have done, and although it is sometimes the *magnitude* of an error vector, it does not include *directional* information telling the learner how to change its behavior, as does the error vector of supervised learning. Although evaluative feedback is often called *reinforcement* feedback, it need not involve pleasure or pain.

Instead of trying to *match* a standard of correctness, an RL system tries to *maximize* the goodness of behavior as indicated by evaluative feedback.[1] To do this, the system has to probe the environment—perform some form of *exploration*—to obtain information about how to change its behavior. It has to actively try alternatives, compare the resulting evaluations, and use some kind of selection mechanism to guide behavior toward the better althernatives. I discuss the distinction between reinforcement and supervised learning in more detail elsewhere (Barto 1991, 1992).

To actually build an RL system, one has to be more precise about the objective of learning. What does evaluative feedback evaluate? If the learning system's life consisted of nothing but a series of discrete trials, each consisting of a discrete action followed by an evaluation of that, and only that, action, the situation would be simple. But actions can have delayed as well as immediate consequences, and evaluative feedback generally evaluates the consequences of all of the system's past behavior. How can an RL system deal with complex tangles of actions and their consequences occuring throughout time? This has been called the *temporal credit assignment problem*. The concept of an adaptive critic is one way to approach this problem: the critic learns to provide useful *immediate* evaluative feedback based on predictions of future reinforcement. According to this approach, RL is not only the process of improving behavior according to given evaluative feedback; it also includes learning to improve evaluative feedback.

The Actor-Critic Architecture

The actor-critic architecture is usually viewed within the framework of control theory. Figure 1A is a variation of the classical control system block 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. The controller inputs labeled 'context' provide information pertinent control task's objective. You might think of the context signals as specifying a 'motivational state' that implies certain control goals.

[1]Klopf's (1972, 1982) theory of *heterostasis*, in contrast to *homeostasis*, emphasizes the significance of the difference between seeking to match and seeking to maximize.

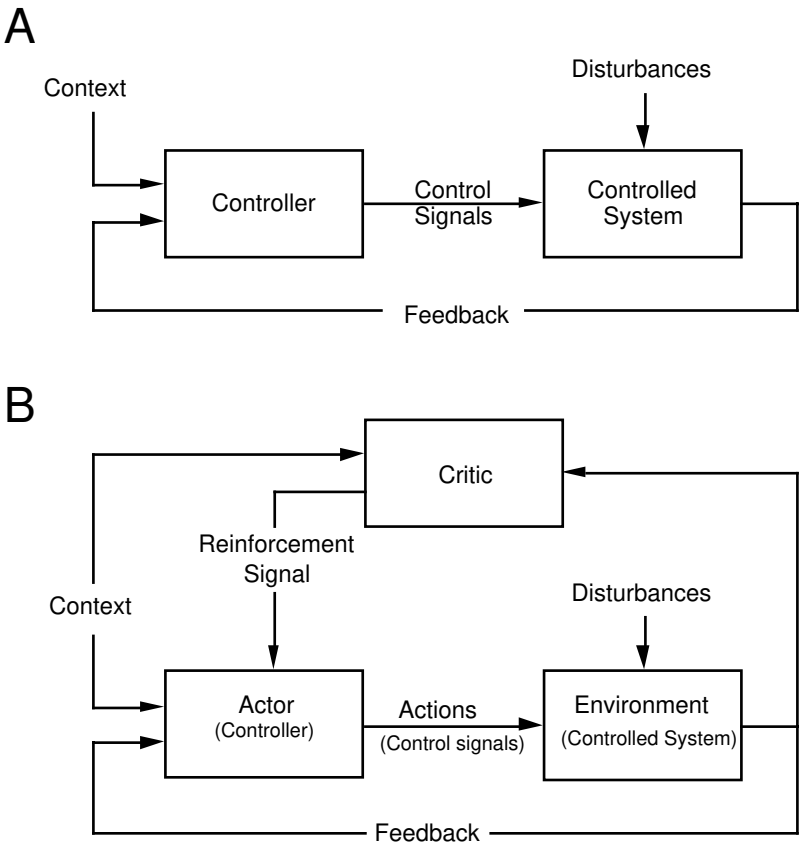

Figure 1: The Actor-Critic Architecture as a Controller. A: A Basic Control Loop. A controller provides control signals to a controlled system, whose behavior is influenced by disturbances. Feedback from the controlled system to the controller provides information on which the control signals can depend. The context inputs provide information pertinent to the control task's objective. B: The actor-critic architecture. A critic provides the controller with a reinforcement signal evaluating its success in achieving the control objectives.

Figure 1B extends the block diagram of Figure 1A to the actor-critic architecture. Another feedback loop is added for providing evaluative feedback to the controller, now called the actor. The critic produces evaluative feedback, or reinforcement feedback, by observing the consequences of the actor's behavior of the controlled system, now called the environment. The critic also needs to know the motivational context of the task because its evaluations will be different depending on what the actor should be trying to do. The critic is an abstraction of the process that supplies evaluative feedback to the learning mechanism responsible for adjusting the actor's behavior. In most artificial RL systems, the critic's output at any time is a number that scores the actor's immediately preceding action: the higher the number, the better the action.

The actor-critic architecture is an abstract learning system, and care must be taken in relating it to animals and their nervous systems. It can help us in thinking about animal reinforcement learning, but it also can be misleading if it is taken too literally. Specifically, it is deceptive to identify the actor with an animal and the environment with the animal's environment. It is better to think of the actor-critic architecture as a model of any reinforcement learning component, or subsystem, of an animal. There are probably many such subsystems, only some of them directly controlling the animal's overt motor behavior.

Figure 2 elaborates the actor-critic architecture emphasize this point. Think of the shaded box in this figure as an animal. The actor is not the same as the entire animal, and its actions are not necessarily motor commmands. Furthermore, the critic (perhaps one of many) is in the animal. I have split the environment box of Figure 1 into an internal and external environment to emphasize that the critic evaluates the actor's behavior on the basis of both its internal and external consequences. The internal environment can contain other actor-critic architectures, some of which do generate overt external behavior. In suggesting how the actor-critic architecture might be related to the basal ganglia, Houk, Adams, and Barto (1994) suggest that the actions of the relevant actor are the signals influencing frontal cortex. Both the frontal cortex and cerebellum are components of this actor's internal environment.

Figure 3 shows the critic in more detail. It consists of a fixed and an adaptive critic. We think of the fixed critic as assigning a numerical *primary reinforcement value*, r_t, to the sensory input (both internal and external) received by the critic at each time instant t; r_t summarizes the strength of that input's primary reinforcing effect, i.e., the reinforcing effect that is wired in by the evolutionary process, not learned through experience. Although in animals this reinforcing effect depends on motivational state, we simplify things by assuming a fixed motivational state (so Figure 3 does not show the context input to the critic). The adaptive critic assigns a different reinforcement value to the sensory input via an adaptive process. The output of the critic at t is the *effective reinforcement signal* sent to the actor. We label it $\hat{r}_t$.

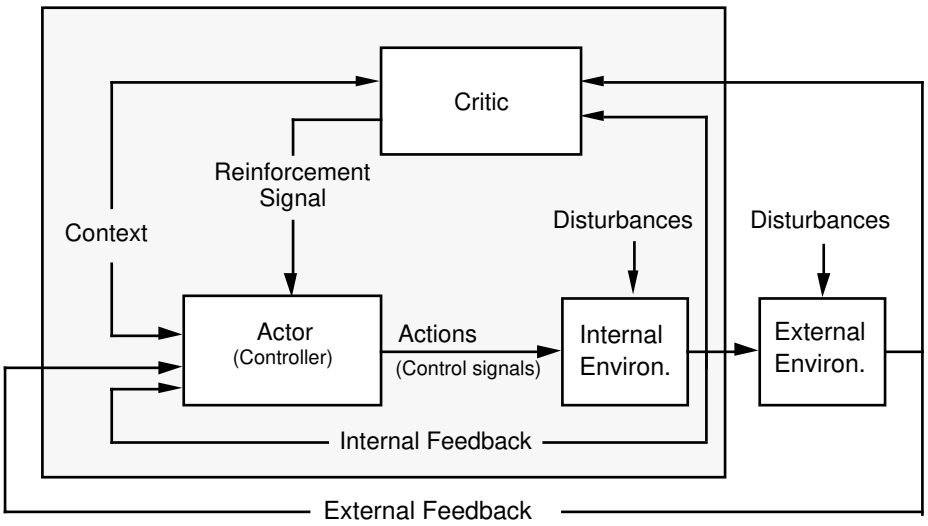

Figure 2: A Hypothetical Animal. The shaded box represents an animal, emphasizing that it is misleading to identify the actor with an entire animal and the critic with an external agent.

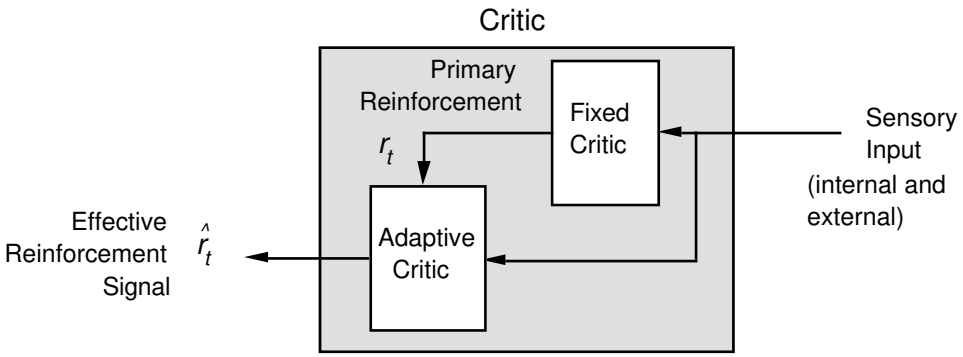

Figure 3: Elaboration of the Critic. It contains fixed and adaptive components.

Imminence Weighting

The basic objective of the actor-critic architecture is to learn to act so as to produce sensory input for which the primary reinforcement value is maximized. But because behavior continues over time, producing sensory input over time, the learning system has to maximize some measure of the entire time course of its input. This measure has to take into account the fact that actions can have long-term as well as short-term consequences on reinforcement, and that sometimes it is better to forgo short-term reward in order to achieve more reward later. Most RL researchers have adopted a measure based on the theory of optimal control. Although mathematical simplicity is its main advantage, this measure has some plausibility for animal learning as demonstrated by the TD model of classical conditioning (Sutton and Barto 1987, 1990). According to this measure, the objective of learning is to act at each time instant so as to maximize *a weighted sum of all future primary reinforcement values*. It is plausible to weight immediate primary reinforcement more strongly than slightly delayed primary reinforcement, which should be more strongly weighted than long-delayed reinforcement. Sutton and Barto (1990) call this *imminence weighting* and suggest that the adaptive critic attempts to predict the imminence-weighted sum of future primary reinforcement.

Figure 4 (from Sutton and Barto 1990) illustrates the idea of imminence weighting for the particlar time course of primary reinforcement shown in Panel A. One can think of this as a sequence of unconditioned stimuli (hence its labeling as US/λ, where λ is a normalization factor that we do not discuss here). Figure 4B shows a particular imminence weighting function, which specifies how the weight given to primary reinforcement falls off with delay with respect to a particular time t. Figure 4C shows how the primary reinforcement signal is transformed by the imminence weighting function applied at time t to give reduced weight to delayed primary reinforcement. The quantity the adaptive critic is trying to predict at time t is the area under this curve. To obtain the correct predictions for other times, this weighting function is slid along the time axis to that its base starts at the time in question, the primary reinforcement signal is reweighted according to the new position, and the new area is calculated. An example for another time, t', is shown in Figures 4D and 4E. By repeating this process for every time, one obtains the sequence of correct predictions shown in Figure 4F. If the adaptive critic is correctly predicting the imminence-weighted sum of future primary reinforcement for the primary reinforcement signal of Figure 4A, its predictions should look like Figure 4F.

The simplest way to explain how it is possible to predict an imminence-weighted sum of future primary reinforcement is to adopt a *discrete-time* model of the learning process. Consequently, suppose t takes on only the integer values $0, 1, 2, \ldots$, and think of the time interval from any time step t to $t+1$ as a small interval of real time. I make the additional assumption, again for simplicity, that *at minimum* it takes one time step for an action to influence primary reinforcement. This is the basic delay through the environment and the critic. Hence, by the *immediate primary reinforcement* for an action taken at time t, I mean r_{t+1}, and by the *immediate effective reinforcement*, I mean $\hat{r}_{t+1}$. Of course, this

Figure 4: Imminence Weighting (reprinted from Sutton and Barto 1990). A: A primary reinforcement signal representing a sequence of unconditioned stimuli (USs). B: An imminence weighting function. C: Primary reinforcement weighted by the imminence weighting function. The correct prediction at time t is the area under this curve. D and E: Imminence weighting for time t'. F: The correct predictions at each time. The heights at times t and t' equal the total areas in C and E.

minumum delay can be different for different actor-critic systems in the nervous system.

Using a discrete-time version of imminence weighting, the objective of the actor is to learn to perform the action at each time step t that maximizes a weighted sum of the primary reinforcement values for time step $t+1$ and all future times, where the weights decrease with decreasing imminence of the primary reinforcement value:

$$\alpha_1 r_{t+1} + \alpha_2 r_{t+2} + \alpha_3 r_{t+3} + \cdots, \tag{1}$$

with $\alpha_1 > \alpha_2 > \alpha_3 > \ldots$. Typically, these weights are defined in terms of a *discount factor*, γ, with $0 \leq \gamma < 1$, as follows:

$$\alpha_i = \gamma^{i-1},$$

for $i = 1, 2, \ldots$. Then the imminence-weighted sum of Equation 1 is

$$r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i}.$$

The discount factor determines how strongly future primary reinforcement should influence current actions. When $\gamma = 0$, the imminence-weighted sum is just the immediate primary reinforcement r_{t+1} (because $0^0 = 1$). In this case, the desired actions maximize only the immediate primary reinforcment. As γ increases toward one, future primary reinforcement becomes more significant. Here we think of γ as being fixed close to one, so that the long-term consequences of actions are important and the adaptive critic plays an essential role in learning.

If actions were reinforced by immediate primary reinforcement only, learning would depend only on the short-term consequences of actions. This learning objective, which has been called a *tactical* objective (Werbos 1987), ignores the long-term consequences of actions. Since immediate primary reinforcement is usually lacking entirely (formalized by letting $r_t = 0$ for those times when there is no primary reinforcement), a purely tactical learning system cannot learn how to manipulate its environment in order to bring about future primary reinforcement. Even worse, acting only to attain immediate primary reinforcement can disrupt, or even preclude, attaining better primary reinforcement in the future. A *strategic* objective (Werbos, 1987), on the other hand, takes into account long-term as well as short-term consequences.

How tradeoffs between consequences at different times are handled is determined by exactly how one defines the strategic objective, imminence weighting by means of a discount factor being one definition. With discounting, any amount of primary reinforcement that is delayed by one time step is worth a fraction (γ) of that same amount of undelayed primary reinforcement. As γ increases toward one, the delay makes less and less difference, and the objective of learning becomes more strategic.

The idea of the adaptive critic is that it should learn how to provide an effective reinforcement signal so that when the actor learns according to the tactical objective of maximizing

immediate effective reinforcement, it is actually learning according to the strategic objective of maximizing a long-term measure of behavior. Here, the long-term measure is the imminence-weighted sum of future primary reinforcement. In order to do this, the adaptive critic has to predict the imminence-weighted sum of future primary reinforcement, and these predictions are essential in forming the effective reinforcement, as discussed below. Because effective reinforcement incorporates these predictions, the actor only needs to perform tactical learning with respect to the effective reinforcement signal: it is geared so that the action at time t is always reinforced by the immediate effective reinforcement $\hat{r}_{t+1}$.

An Input's Value

We call the imminence-weighted sum of the primary reinforcement values from $t+1$ into the future the *value* of the sensory input (internal and external) at time t. Let V_t denote this value; that is,

$$V_t = \sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i}. \tag{2}$$

The objective of learning, then, is to learn to influence the environnment so that the sensory inputs received by the learning system have the highest possible values. The job of the adaptive critic is to estimate these values and use them to determine the immediate effective reinforcement.

For the sake of brevity, I have to disregard a lot of important technical details about how this is even possible, but let me give some hints about these details. Because an estimate of an input's value is a prediction of the future, how can these estimates be made? Doesn't an input's value depend on the course of action the learning system will take in the future? Indeed, doesn't it depend on all kinds of unpredictable aspects of the environment? First, prediction is possible if one assumes that environmental situations tend to recur, so that a prediction is really a kind of recollection of what happened in the same situation, or in similar situations, in the past. The critic's sensory input must be rich enough to allow the detection of situations having the same or similar futures (formalized as *states* of a dynamic system). Second, for much of our discussion, we assume that the learning system's policy of acting stays fixed throughout the prediction process. This does not mean that the actor always produces the same action, but that it always responds the same way whenever the same situation recurs: its response rule, or policy, is fixed. Of course, because the whole point of RL is to change this response rule, this is only a subproblem of the entire RL problem. Finally, when the sensory input cannot resolve all the unpredictable aspects of the environment (i.e., when what *appears* to be a previously sensed situation is followed by a different course of events), probability theory is invoked. By the value of an input we really mean the *expected* imminence-weighted sum of future primary reinforcement values: the average over all the possible future scenarios.

So the adaptive critic is supposed to estimate the values of sensory inputs so it can

compute a suitable effective reinforcement, $\hat{r}_t$. Let the critic's estimate of V_t by denoted P_t; it is a *prediction* of the imminence-weighted sum of future primary reinforcement. Then from Equation 2 we would like the following to be true:

$$P_t \approx V_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots, \tag{3}$$

where $\approx$ means 'approximately equal'.

Learning to Predict

It is relatively easy to devise a supervised learning system for learning to predict the future values of specific signals. For example, suppose we wanted to have a prediction at any time t of the primary reinforcement signal at $t+1$; that is, suppose we want $P_t = r_{t+1}$ for all t. This is a one-step-ahead prediction problem, and the usual kind of error-driven supervised learning system (e.g., Rumelhart et al. 1986) can be used to solve it. This system would need, at each time t during learning, an error between its actual prediction, P_t, and the prediction target (the quantity being predicted), r_{t+1}. It can obtain this error simply by computing P_t, *waiting one time step* while remembering P_t, then observing the actual r_{t+1}. It also has to remember for one time step the sensory input on which the prediction was based in order to update its prediction function.

For example, suppose P_t is the output of a simple linear connectionist unit:

$$P_t = \sum_{i=1}^{m} v_t^i x_t^i,$$

where v_t^i and x_t^i, for $i = 1, \ldots, m$, are respectively the connection weights and input activations at time t. Then the standard delta learning rule for one-step-ahead prediction is

$$v_{t+1}^i = v_t^i + \eta[r_{t+1} - P_t]x_t^i, \tag{4}$$

where $\eta > 0$ is the learning rate. If we think of this update equation being applied at time $t+1$, then P_t is the *remembered* prediction, x_t^i is the *remembered* input activation, and r_{t+1} is the *currently observed* primary reinforcement value.

This is perhaps clearer if we rewrite the learning rule as it would appear if it were applied at time t instead of $t+1$:

$$v_t^i = v_{t-1}^i + \eta[r_t - P_{t-1}]x_{t-1}^i. \tag{5}$$

This form, equivalent to Equation 4, makes it clear that the previous prediction and the previous input activations have to be remembered as well as the previous connection weights. Following Klopf (1972, 1982), we say that input activity at $t-1$ (i.e., $x_{t-1}^i \neq 0$) makes the connection weight v^i *eligible* for modification at t. In neural terms, eligibility would be

a synaptically local memory for storing information about the past activity of the presynaptic fiber. Houk, Adams, and Barto (1994) postulate that this notion of eligibility is implemented by a period of high receptivity of spiny neuron synapses to the reinforcing effects of dopamine.

Of course, for this learning rule to work there must be information in the input stream that is predictively useful. If one wanted to predict more than one time step into the future, the procedure would be essentially the same except that it would need to remember the predictions and the input activations for the entire time interval until the actual predicted value becomes available. Consequently, for a prediction interval of k time steps, the procedure would need to keep in memory k past predictions and k past input activations. Any kind of eligibility mechanism for this situation would have to be much more complicated than the simple period of receptivity mentioned above. One of the advantages of the adaptive critic is that it can learn to predict many time steps into the future without the need for a more complicated eligibility mechanism.

The Adaptive Critic Learning Rule

The adaptive critic learning rule begins with the one-step-ahead prediction method described in the previous section. However, for the critic the prediction targets are the *values* of inputs, which involve all future primary reinforcement, not just the primary reinforcement at the next time step. Extending the one-step-ahead method to this situation in the most obvious way would require an infinite amount of storage and the weights could not be updated until an infinite amount of time had passed.

The adaptive critic learning rule rests on noting that correct predictions must satisfy a certain consistency condition which relates the predictions at adjacent time steps. Moreover, it is true that any predictions that satisfy this consistency condition for all time steps must be correct. (This is a result from the theory of optimal control that is not particularly obvious.) Suppose that the predictions at any two adjacent time steps, say steps $t-1$ and t, are correct. This means that

$$P_{t-1} = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \quad (6)$$

$$P_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots. \quad (7)$$

Now notice that we can rewrite P_{t-1} as follows:

$$P_{t-1} = r_t + \gamma(r_{t+1} + \gamma r_{t+2} + \cdots).$$

But this is exactly the same as

$$P_{t-1} = r_t + \gamma P_t.$$

This is the consistency condition that is satisfied by the correct predictions. The error by which any two adjacent predictions fail to satisfy this condition is called the *temporal*

difference error (or TD error) by Sutton (1988):

$$r_t + \gamma P_t - P_{t-1}. \tag{8}$$

The adaptive critic uses the TD error to update its weights. The term temporal difference comes from the fact that this error essentially depends on the difference between the critic's predictions at adjacent time steps.

The adaptive critic therefore adjusts its weights according to the following modification of the the one-step-ahead learning rule of Equation 5:

$$v_t^i = v_{t-1}^i + \eta[r_t + \gamma P_t - P_{t-1}]x_{t-1}^i. \tag{9}$$

This rule adjusts the weights to decrease the magnitude of the TD error. Note that if $\gamma = 0$ this is equal to the one-step-ahead learning rule (Equation 5).

In analogy with the one-step-ahead learning rule (Equation 5), we can think of $r_t + \gamma P_t$ as the prediction target: it is the quantity that each P_{t-1} should match. The adaptive critic is therefore trying to predict the next primary reinforcement, r_t, *plus its own next prediction* (discounted), γP_t. On the surface it is not clear that this would work: it is like the blind leading the blind. How can an incorrect prediction be improved by moving it toward another incorrect prediction? The key observation, however, is that the target $r_t + \gamma P_t$ tends to be more accurate than the prediction P_{t-1} because it includes the additional data provided by r_t. It is more like the blind being led by the slightly less blind. Although this method is very simple computationally, it actually converges to the correct predictions under fairly general conditions.

Effective Reinforcement

The output of the adaptive critic at time t is the effective reinforcement $\hat{r}_t$, which reinforces the action made at $t-1$. For the actor-critic architectures with which we have the most experience, the effective reinforcement is same as the TD error:

$$\hat{r}_t = r_t + \gamma P_t - P_{t-1}. \tag{10}$$

Effective reinforcement is therefore the sum of the primary reinforcement, r_t, and the term $\gamma P_t - P_{t-1}$, which corresponds to *secondary reinforcement*. To understand why this makes sense, one has to consider how the learning rule of the actor works.

The Actor Learning Rule

The basic idea of the actor learning rule is that if an action produced in response to a sensory input has the *expected* consequences, then that response tendency remains unchanged. On

the other hand, if its consequences are better (worse) than expected, the response tendency is strengthened (weakened) (cf. the Rescorla-Wagner model; Rescorla and Wagner 1972). When the TD error equals zero, the consequences are as they were predicted by the critic, and no learning should occur. When the TD error is positive (negative), consequences are better (worse) than predicted so that the response tendency should be strengthened (weakened).

Suppose the actor makes decisions by comparing the activities of a collection of linear connectionist units, where there is one unit for each possible action. The action selected is the one whose unit has the most vigorous activity. Let a_t denote the activity at time t of the unit corresponding to action a and suppose that

$$a_t = \sum_{i=1}^{m} w_t^i x_t^i,$$

where w_t^i and x_t^i, for $i = 1, \ldots, m$, are respectively the weights and input activations at time t. The following learning rule for this unit is applied *only if action a was selected for execution at time* $t - 1$:

$$w_t^i = w_{t-1}^i + \zeta \hat{r}_t x_{t-1}^i. \tag{11}$$

where $\zeta > 0$ is the learning rate. The weights of the units for the unselected actions remain unchanged.

Due to the definition of the effective reiforcement $\hat{r}_t$ (Equation 10), this rule it is almost indentical to the adaptive critic learning rule (Equation 9). It possibly has a different learning rate and, more importantly, it applies only to the weights for the action a that was selected at $t - 1$.

Neural Implementation

Figure 5 illustrates how the actor-critic architecture could be implemented by a neural network. Both the actor units and the predictor unit use the same learning rule and the same modulatory signal, $\hat{r}_t$, as a factor in updating their synaptic weights. All the modifiable synapses require local memory to implement the necessary eligibility mechanism. The only difference between these units is that the actor units compete with each other so that only one unit wins the competition, and the learning rule applies only to the winning unit. This could be implemented by suitable lateral inhibition and a slightly different eligibility mechanism for the actor units. Whereas the eligibility mechnanism of the predictor unit remembers only past presynaptic activity, the eligibility mechanism of an actor unit would have to remember past *conjunctions of pre- and postsynaptic activity* in such a way that if it were not selected, none of its synapses could become eligible for modification. Consequently, the modifiable synapses of the prediction unit must use a two-factor learning rule, whereas those of the actor units must use a three-factor learning rule. Finally, some mechanism is required to compute the secondary reinforcement signal

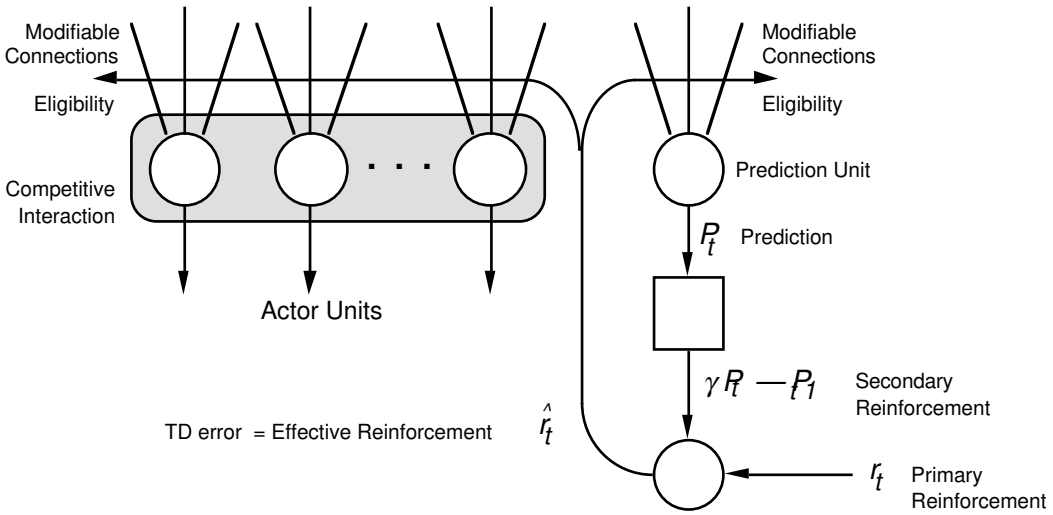

Figure 5: Network Implementation of the Actor-Critic Architecture. Both the actor units and the predictor unit use the same learning rule and the same modulatory signal, $\hat{r}_t$, as a factor in updating their synaptic weights. All the modifiable synapes require local memory to implement the necesary eligibility mechanism. The actor units compete with each other to determine the action, and the learning rule is applied only to the winning actor unit.

$\gamma P_t - P_{t-1}$. This could be accomplished by a kind of neural differentiator that is shown by the box in Figure 5. Houk, Adams, and Barto (1994) elaborate this basic network in relation to the circuitry and intracellular chemistry of the basal ganglia and dopamine neurons.

The Case of Terminal Primary Reinforcement

Most relevant to animal learning experiments are cases in which a sequence of actions has to be accomplished before a primary reinforcing event occurs (e.g., a monkey reaching, picking up a food morsel, and transferring it to its mouth as in Schultz, et al. 1993 and Schultz 1993). In this case, during each trial $r_t = 0$ for all t except when the food is actually tasted, at which time, say T, it is some positive number, say one; so that $r_T = 1$. Suppose that the discount factor γ is very nearly one so that we can effectively ignore it, and further suppose that the adaptive critic starts out by producing $P_t = 0$ throughout the first trial.

Then until the first occurrence of the terminal primary reinforcing event at time T of some trial (due to the accidental execution of the right action sequence), all the TD errors, and hence all the effective reinforcements, are zero. At time T of this first successful trial, the TD error and the effective reinforcement are 1. This positive effective reinforcement causes the actor to increase its tendency to produce the immediately preceding response,

and the positive TD error causes the adaptive critic to adjust its weights so that when the stimulus at time $T-1$ of this successful trial recurs in a later trial, the critic will predict that the immediately following stimulus will have positive value. That is, P_{T-1} will be greater than zero at time $T-1$ of a trial in which the stimulus at $T-1$ is the same (or similar) to the stimulus at $T-1$ of the first successful trial.

Now things become more complicated but also more closely related to the observed responses of dopamine neurons. Consider the next successful trial. In addition to the events of the first successful trial happening again, so that the actor's response tendency and the critic's prediction become stronger, the fact that P_{T-1} is positive has two additional consequences:

1. The TD error and the effective reinforcement at time $T-1$ will now be positive. This quantity is
$$r_{T-1} + \gamma P_{T-1} - P_{T-2},$$
and since both r_{T-1} and P_{T-2} are zero,[2] this equals $\gamma P_{T-1} > 0$. Just as the positive TD error at time T of the first successful trial caused the critic to make a positive prediction at time $T-1$ of later trials, this positive TD error at $T-2$ will cause the critic to make a positive prediction at time $T-2$ of later trials. Also, as effective reinforcement, this positive quantity causes the actor to increase its tendency to produce the response it made at $T-2$ of this successful trial.

2. The TD error and the effective reinforcement at time T will *decrease*. This quantity is
$$r_T + \gamma P_T - P_{T-1}.$$
r_T is still 1 since this is a successful trial; P_T is still zero because it is predicting that zero primary reinforcement occurs *after* the trial;[3] and P_{T-1}, which is positive, is being subtracted. Thus, the TD error and effective reinforcement at time T will be smaller than in earlier trials.

With continued successful trials, which become increasingly likely due to the actor's changing response rule, the TD errors and effective reinforcements propagate backward in time: the activity transfers from later to earlier times within trials. Learning stops when these quantities all become zero, which happens only when the adaptive critic correctly predicts the values of all stimuli, i.e., when all expectations are met by actual events (which

[2]Actually, P_{T-2} might not be zero because it might have increased in intervening unsuccessful trials in which the animal made a mistake only on the last move. But it will be small enough so that the TD error at $T-1$ will still be positive.

[3]Actually, P_T might be nonzero because it is really predicting the imminence-weighted sum of future primary reinforcement, which includes primary reinforcement obtained in later successful trials. However, it is only with considerable experience that P_T takes on significant positive value due to the presumably long duration of the inter-trial interval.

requires certain assumptions about the regularity of the environment and the richness of the sensory input), and the actor always produces the correct actions.

Conclusion

The actor-critic architecture implements one approach to learning when actions have delayed consequences. It has a well-developed theoretical basis, works well in practice, and makes strong contact with animal learning through the TD model of classical conditioning. The adaptive critic computes an effective reinforcement signal such that the action-selection subsystem achieves long-term goals while learning only on the basis of immediate effective reinforcement. The TD error used by the adaptive critic's learning mechanism is the same as the effective reinforcement supplied to the action-selection subsystem. When primary reinforcement occurs only after a sequence of correct actions, the adaptive critic's activity parallels that observed in dopamine neurons during similar animal learning experiments. This suggests the hypothesis that the activity of dopamine neurons plays the dual roles of TD error and effective reinforcment in a neural implementation of the actor-critic architecture. Houk, Adams, and Barto (1994) explore this hypothesis in more detail.

References

Barto, A.G., 1991, Some Learning Tasks from a Control Perspective, in *1990 Lectures in Complex Systems*, L. Nadel and D.L. Stein, eds., pp. 195–223, Redwood City, CA: Addison-Wesley Publishing Company, The Advanced Book Program.

Barto, A.G., 1992, Reinforcement Learning and Adaptive Critic Methods, in *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, D.A. White and D.A. Sofge, eds., pp. 469–491, New York: Van Nostrand Reinhold.

Barto, A.G., Sutton, R.S., and Anderson, C.W., 1983, Neuronlike Elements That Can Solve Difficult Learning Control Problems, IEEE Transactions on Systems, Man and Cybernetics, 13, pp. 835–846, Reprinted in J. A. Anderson and E. Rosenfeld, eds., 1988, *Neurocomputing: Foundations of Research*, Cambridge, MA: The MIT Press.

Barto, A.G., Sutton, R.S., and Watkins, C.J.C.H., 1990, Learning and Sequential Decision Making, in *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, M. Gabriel and J. Moore, eds., pp. 539–602, Cambridge, MA: The MIT Press.

Barto, A.G., Bradtke, S.J., and Singh, S.P., 1994, Learning to Act using Real-Time Dynamic Programming, *Artificial Intelligence Journal,* to appear. [Also available as Computer Science Technical Report 93-02, University of Massachusetts, Amherst, MA.]

Houk, J.C., 1992, Learning in Modular Networks, NPB Technical Report 7, Northwestern University Medical School, Department of Physiology, Ward Building 5-342, 303 E. Chicago Ave., Chicago IL 60611-3008

Houk, J.C., Adams, J.L., and Barto, A.G., 1994, A Model of How the Basal Ganglia Might Generate and Use Neural Signals that Predict Reinforcement, Workshop paper.

Klopf, A.H., 1972, Brain Function and Adaptive Systems—A Heterostatic theory, Air Force Cambridge Research Laboratories technical report AFCRL-72-0164, Bedford, MA.

Klopf, A.H., 1982, *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence*, Washington, D.C.: Hemishere.

Rescorla, R.A. and Wagner, A.R., 1972, A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement, in Black, A.H. and Prokasy, W.F., eds., pp. 64–99, *Classical Conditioning II*, New York: Appleton-Century-Crofts.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J., 1986, Learning Internal Representations by Error Propagation, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol.1: Foundations*, D.E. Rumelhart and J.L. McClelland, eds., Cambridge MA: Bradford Books/MIT Press.

Samuel, A.L., 1959, Some Studies in Machine Learning Using the Game of Checkers, IBM Journal on Research and Development, pp. 210–229. Reprinted in E.A. Feigenbaum and J. Feldman, eds., 1963, *Computers and Thought*, New York: McGraw-Hill.

Schultz, 1993, Workshop paper

Sutton, R.S., 1984, Temporal Credit Assignment in Reinforcement Learning, Ph.D. Dissertation, University of Massachusetts, Amherst, MA.

Sutton, R.S., 1988 Learning to Predict by the Method of Temporal Differences, *Machine Learning*, 3, pp. 9–44.

Sutton, R.S., and Barto, A.G., 1981, Toward a Modern Theory of Adaptive Networks: Expectation and Prediction, *Psychological Review*, 88, pp. 35–170.

Sutton, R.S. and Barto, A.G., 1987, A Temporal-Difference Model of Classical Conditioning, in *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum.

Sutton, R.S., and Barto, A.G., 1990, Time-Derivative Models of Pavlovian Reinforcement, in *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, M. Gabriel and J. Moore, eds., pp. 497–537, Cambridge MA: The MIT Press.

Tesauro, G.J., 1992, Practical Issues in Temporal Difference Learning, *Machine Learning*, 8, pp. 257–277.

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

Watkins, C.J.C.H., 1989, Learning from Delayed Rewards, Ph.D. Dissertation, Cambridge University, Cambridge, England.

Werbos, P.J., 1977, Advanced Forecasting Methods for Global Crisis Warning and Models of Intelligence, *General Systems Yearbook*, 22, pp. 25–38.

Werbos, P.J., 1987, Building and Understanding Adaptive Systems: A Statistical/Numerical Approach to Factory Automation and Brain Research, *IEEE Transactions on Systems, Man and Cybernetics*.

Wickens, J., 1990, Striatal Dopamine in Motor Activation and Reward-Mediated Learning: Steps Toward a Unifying Model, *J. Neural Transm*, 80, pp. 9–31.