32
Introduction

(1983)
Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson

Neuronlike adaptive elements that can solve difficult learning control problems
IEEE Transactions on Systems, Man, and Cybernetics SMC-13:834–846

One of the hardest and most important problems for both brains and learning systems to solve is called the "credit assignment" problem, that is, knowing how to change connection strengths in complicated networks so as to produce a network that can do what is wanted. There are a number of forms of this problem. In multilayer networks it is hard to give rules for specifying appropriate connections for the inner layers, that is, layers that contain "hidden units" that are neither input nor output units. The Neocognitron (paper 31) learned in a multilayer network, but by a sequential directed learning procedure, where only a single layer at a time is plastic, and where the types of features that the layer should respond to were known. It was not until the development of the generalized error correcting rule now called back propagation (papers 41 and 42) that it was possible to give efficient general learning rules for multilayer networks.

Barto, Sutton, and Anderson discuss a different aspect of the credit assignment problem. Error correction techniques such as the Widrow-Hoff rule and back propagation require detailed knowledge about the nature of the error. In general, the system must know exactly what the appropriate response was and what the system response was, so that a detailed error signal can be formed and used to make corrections in the network. There are many places where detailed information about the error is not available—only knowledge that an error was made. An example of a complex system of this type would be a chess game, where a loss might have been caused by an error at any one of many earlier moves. Trying to find where the mistake was and when it was made gives substance to many chess arguments.

The example used by Barto, Sutton, and Anderson in their paper analyzes and suggests a solution to another class of credit assignment problems. They use a cart, which holds a pole that is free to pivot. The cart can move on a track between two stops. The object of the system is to keep the pole from falling over by pushing the cart back and forth between the stops. An error is made when the pole falls over or the cart hits the stop. The physics of this task is familiar to anyone who has balanced a broom or a baseball bat on the hand. It is quite easy with a little practice.

The problem here is that the error feedback is not very informative. When the pole falls over or the cart hits the stop, all it means is that a mistake was made some time in the past. Even with perfect control after the initial mistake, the error could not be prevented. There is no idea of when the mistake was made or how large the mistake was.

The strategy used by Barto, Sutton, and Anderson is to assume that there are two adaptive devices involved: first is the *associative search element*, which takes the current state of the physical system (i.e., the position and velocity of the cart) and, by an associative learning rule, gives an output specifying a control action—the force to be

applied to the cart; second is what they call the *adaptive critic element*. The critic is looking ahead and predicting the expected reinforcement from the environment, given a particular action from the associative search element. The learning rule of the associative search element incorporates a prediction of reinforcement from the critic, so the two elements are tightly coupled.

Since the critic element is constantly trying to predict the expected reinforcement associated with particular input states, and modifying itself appropriately with experience, the associative search element is constantly modifying its weights, even when an explicit error has not been made. In human terms, the critic element is acting something like a parent predicting direction of future reinforcement—either negative ("If you eat that, you will get a stomach ache.") or positive ("If you clean up your room, Santa Claus will remember you."). The associative search element is taking the predicted reinforcement into account when it learns what to do.

For convenience in simulation and analysis, Barto, Sutton, and Anderson only use a single adaptive element critic and a single associative search element. However, they point out that their system could easily be turned into a more traditional multineuron network with only minor modifications.

An interesting aspect of the approach taken in this paper is that it is consistent with a large body of psychological literature on animal learning. In fact, use of the adaptive critic element is close to the Rescorla-Wagner model of classical conditioning, probably the most successful current model of classical conditioning. A highly recommended paper by Sutton and Barto in *Psychological Review* (1981) discusses the connections between adaptive system theory and psychological theory in great detail. This paper by Sutton and Barto is also notable for bringing the Widrow-Hoff error correction technique and related work in adaptive control theory to the attention of psychologically oriented neural modelers, where it had a major impact.

Reference

R. S. Sutton and A. G. Barto (1981), "Toward a modern theory of adaptive networks: expectation and prediction," *Psychological Review* 88:135–171.