Predictive Hebbian Learning

Terrence J Sejnowski *
Peter Dayan† P Read Montague‡

Abstract

A creature presented with an uncertain and variable environment needs to *anticipate* important future events or risk diminished chances for survival. These events can include the presence of food, destructive stimuli, and potential mates. In short, a nervous system must have means to generate guesses about its most likely next state and the most likely next state of the world. Psychologists have studied conditions under which animals can learn to predict future reward and punishment. In this paper, we review the computational theory that may be relevant for understanding this form of learning. Some of the central mechanisms required for predictive learning have been discovered in both vertebrate Ljungberg *et al*'s (1992) and invertebrate brains (Hammer, 1994).

Prediction

Animals are capable of *predicting* events and the consequences of those events on the basis of the sensory information they receive and directing their actions according to those predictions (see Dickenson, 1980; Mackintosh, 1983; Gallistel, 1990 and Gluck and Thompson, 1987 for reviews). Although these conclusions are well established from the perspective of psychological experiments, the neural mechanisms that underlie this prediction are less well understood. Prediction, and its appropriate use for action, is essentially a computational concept, but this still leads a wide range of possible theories to explain existing data.

One way for an animal to learn to make predictions is for it to have a system that reports on its current best guess, and to have learning be contingent on *errors* in this prediction: learning only happens if the animal becomes surprised based on its prediction. This is the underlying mechanism behind essentially all adaptation rules in engineering (Kalman, 1960; Widrow & Stearns, 1985) and particular learning rules in psychology (Rescorla & Wagner, 1972; Mackintosh, 1983; Pearce & Hall, 1980). We consider below the general requirements for such a signal in the brain.

The construction, delivery and use of an error signal related to predictions about future stimuli would require the following:
i) access to a representation of the phenomenon to be predicted such as the amount of reward or food.
ii) access to the current predictions so that they can be compared to the phenomenon to be predicted.
iii) capacity to influence plasticity (directly or indirectly) in structures responsible for constructing the predictions.
iv) sufficiently wide broadcast of the error signal so that stimuli in different modalities can be used to make and respond to the predictions.

These general requirements are met by a number of diffusely projecting system of axons that are thought to report in part on salient events in the world and within the organism. These axons release neuromodulators that can influence the effectiveness of synapses in these areas. This anatomical motif is a common feature of many nervous systems and is not unique to vertebrate brains. Invertebrates have analogous sets of neurons with extensive axonal arborizations that deliver neuromodulators to widespread target regions (Hawkins and Kandel, 1984; Greenough and Bailey, 1988; Hammer, 1994). Experimental evidence from both behavioral and physiological work suggests that these systems influence

*Howard Hughes Medical Institute, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, CA 92037, and Department of Biology, University of California, San Diego, La Jolla, CA 92093.

†Department of Brain and Cognitive Science, MIT, Cambridge MA 02139.

‡Division of Neuroscience, Baylor College of Medicine, Houston TX 77030.

COLT''95 Santa Cruz, CA USA© 1995 ACM 0-89723-5/95/0007..$3.50

ongoing neural activity as well as a number of critical physiological functions. These systems are also known to be required for normal development of the response properties of cerebral cortical neurons in various regions of sensory cortex

Models of Classical Conditioning

We briefly summarize here the mathematical assumptions that underlie the approach that we have taken to modeling animal learning. At time t, the animal sees various conditioned stimuli (CSs) which are represented in the activity $x(t) = \{x_i(t)\}$ of units x_i, $i \in \{1, N\}$. The units can be considered populations of neurons in the cerebral cortex that represent states of the environment. The animal also receives a scalar reward $r(t)$. The temporal difference (TD) approach (Sutton & Barto, 1987) assumes that the computational task is to use the CSs to fit a function $V(x(t))$ that predicts a *discounted sum of future rewards:*

$$\hat{V}(t) = \sum_{u=t}^{\infty} \gamma^{u-t} r(u) \qquad (1)$$

where $0 \leq \gamma \leq 1$ is a discount factor that models the fact for the animal that future rewards may be worth less than current ones. This equation is key: Predicting the sum over future rewards represents a significant advance over static conditioning models such as the Rescorla-Wagner rule (Rescorla & Wagner, 1972). In non-deterministic problems, where stimuli are not always followed by the same reward consequences, the task is to predict the mean of this quantity.

The next major assumption is that $\hat{V}(t)$ can be treated as a function of $x(t)$. This amounts to assuming that the environment has a Markov property (future rewards do not depend on past rewards except through the current stimulus state $x(t)$). Assuming this, $V(x(t))$ should satisfy a consistency condition:

$$V(x(t)) = r(t) + \gamma V(x(t+1)) \text{ or equivalently,} \qquad (2)$$

$$\epsilon(t) \equiv r(t) + \gamma V(x(t+1)) - V(x(t)) = 0 \qquad (3)$$

where $\epsilon(t)$ is called the TD error.

In its simplest form, $V(x(t)) = \sum_{i=1}^{N} x_i(t) w_i(t)$ has parameters $w(t) = \{w_i(t)\}$ that are adjusted to make $\epsilon(t) = 0$. It is natural to make the adjustments using the delta rule (Rescorla & Wagner, 1972; Widrow & Stearns, 1985):

$$\Delta w_i(t) = \alpha_i(t) x_i(t) \epsilon(t) \qquad (4)$$

where $\alpha_i(t)$ is a stimulus specific learning rate. Sutton (1988), Dayan and Sejnowski (1994), and others have shown conditions under which this update makes $V(x(t))$ converge to the optimal $V(t)$.

The weights $w(t)$ in equation 4 are updated according the learning rule suggested by Hebb (1949) based on the correlation between presynaptic activity $x_i(t)$ and a quantity that depends on postsynaptic activity $\epsilon(t)$. However the form of the postsynaptic term changes the computational behavior of the algorithm from the traditional Hebb rule, which performs principal components analysis, to something closer to a predictive delta rule.

If the brain takes advantage of the mathematical framework for learning outlined above, then there may be neurons that compute the prediction error in equation 3. Such neurons have recently been discovered in an area of the brain called the ventral tegmental area (VTA). These neurons are dopaminergic and their axons project diffusely to widespread areas of the cortex and the ventral striatum. The latter brain region is known to be an important center for reward learning and is involved in many addictive behaviors.

A significant fraction of neurons in the VTA tend to fire in response to the delivery of reward to a naive animal performing a behavioral task in which some sensory stimulus, such as a light, consistently predicts the delivery of reward. After the task has been learned, however, few cells respond to the delivery of reward and more cells respond to the onset of the predictive sensory cue (Ljungberg *et al* 1992; Schultz *et al* 1993). We suggest the VTA dopaminergic cells are reporting the prediction error for the discounted reward, $\epsilon(t)$ in equation 3. (Quartz *et al*, 1992; Montague *et al*, submitted).

Neuroscientists have previously suggested that neuromodulators might modulate synaptic plasticity. The global signal could act as a "print now" command to regulate when learning takes place. Predictive learning in equation 4 is computationally more powerful than using neuromodulatory influence merely to gate periods of conventional Hebbian learning (Rauschecker, 1991). However, in order for this scheme to work, the representations in the cortex which are being weighted must predict the time of the reward as well as its magnitude. How this is accomplished by the brain is an open research problem (Montague *et al*, submitted).

Models of Instrumental Conditioning

One facet of this framework for classical conditioning is that a link (Quartz *et al*, 1992; Montague *et al*, submitted, Dayan, 1994) can be made to instrumental conditioning, even though the exact relationship between these two is still the subject of substantial debate. A working hypothesis is that animals use their capacity to learn to predict events of importance in the world to control their actions appropriately—not only blinking just before the delivery of a signaled puff of air to a nictitating membrane, but also learning complicated and even comparatively arbitrary sequences of actions in response to external (and internal) stimuli. There is again computational theory as to appropriate learning rules for the resulting control problem. Many of these learning rules were first introduced in the psychology literature.

The capacity for learning predictions over time is used

to solve the temporal credit assignment problem, which comes from the distance in time between making an action and seeing its consequence in terms of getting to the goal. It turns out that temporal difference algorithms learn exactly the predictions of proximities that are required for dynamic programming (Bellman, 1957). In fact, the same error signal that these algorithms use to learn how far states are away from the goal can also be used to criticize the choice of actions (Barto, Sutton & Anderson, 1983; Barto, Sutton & Watkins, 1989).

In instrumental conditioning, an animal has various actions available to it (characterized, say, as coming from set $\mathcal{A}$), and its choice affects its rewards. In general, in tasks such as mazes with multiple choice points, actions can affect rewards in complicated ways. Worse, the animal may face the *temporal credit assignment* problem of working out which action out of a whole sequence was critical for the rewards it received. Markov decision problems provide a general theoretical framework for these tasks (Barto *et al*, 1989), and the techniques of dynamic programming (Bellman, 1957) provide a general theoretical framework for their solution.

The key concept is a *policy* $\pi(x(t))$, which describes how actions are assigned to states (and is something like a stimulus-response function). This assignment will usually be stochastic so that different actions will be tried at the same state. If the animal were to follow a fixed policy, then the TD methods described above would *evaluate* it, in the sense of learning how much reward $\hat{V}^{\pi}(x)$ would be expected in the future, if the animal is at state x and chooses actions according to π. The method of policy iteration in dynamic programming (Howard, 1963) uses this evaluation to improve the policy. Barto *et al* (1983) described a simple version of policy iteration in which each action $a \in \mathcal{A}$ is associated with a set of adjustable parameters $v^a(t) = \{v_i(t)\}$. At time t, the system calculates an action choice 'value' for each action as:

$$u^a(t) = \sum_{i=1}^{N} v_i^a(t) x_i(t) + \eta^a(t) \qquad (5)$$

where $\eta^a(t)$ are random noise values. $u^a(t)$ is used as an estimate of the appropriateness of performing action a relative to doing other actions, so the role of $\eta^a(t)$ is to ensure that each action is tried often enough in each state so that good overall policies can be identified. The ultimate action selected is the largest of these:

$$a^* = \text{argmax}_a u^a(t). \qquad (6)$$

In this formulation, $\epsilon(t)$ in equation 3 is used to criticize the choice of action, indicating whether the one selected was better or worse than the average. $v^{a^*}(t)$ is updated as:

$$\Delta v_i^{a^*}(t) = \beta_i(t) x_i(t) \epsilon(t), \qquad (7)$$

where $\beta_i(t)$ is another stimulus specific learning rate. The $v_i^a(t)$ for the non-selected actions $a \neq a^*$ are left fixed. As learning proceeds, this tends to improve the policy.

The key attraction of this form of policy iteration is that the signal, $\epsilon(t)$, that we postulated the VTA cells to be reporting, has two roles: training the prediction parameters $w(t)$, *and* training the action parameters $v^a(t)$. Evidence that drug self-administration and electrical self-stimulation of the dopaminergic inputs that the projection from the VTA to the ventral striatum suggest that this system ios particularly important for reward learning.

Note that according to these models, classical and instrumental conditioning are very closely related (Mackintosh, 1983). Learning to predict the values of states is classical conditioning; using this information to improve policies or choose actions is instrumental conditioning.

Conclusions

In summary, the TD model of the diffuse systems outlined here allows expectations about the future to influence synaptic change that occur in the present. Certain aspects of decision making can also be understood within this framework (Montague *et al*, submitted). A better developed connection between this model, physiological data, and behavioral decisions may open the way for interesting experiments to uncover the neural mechanisms that influence learning and decision-making in more complicated situations. There are many aspects of learning that have not yet been addressed: How are appropriate representations of sensory stimuli formed by populations of cortical neurons? How is time represented by neural populations? How does the brain reject spurious correelations that are not causal? The advantage of having an overall framework like the one outlined here is that these questons become grounded in ways that are subject to experimental test. The prospect of a rigorous mathematical framework for animal learning will allow sharper questions to be asked, and perhaps more definitive answers to be found.

References

Barto, AG, Sutton, RS and Anderson, CW. *IEEE Transactions on Systems, Man, and Cybernetics,* **13**, 834 (1983).

Barto, AG, Sutton, RS, Watkins, CJCH, *Technical Report 89-95,* (Computer and Information Science, University of Massachusetts, Amherst, MA, 1989

Bellman, R (1957) *Dynamic Programming.* Princeton: Princeton U. Press.

Dayan, P (1994). Computational modelling. *Current Opinion in Neurobiology,* **4**, 212-217.

Dayan, P. and Sejnowski, T. J., TD (λ) converges with probability 1, Machine Learning **14**, 295-301 (1994).

Dickenson, A (1980). *Contemporary Animal Learning Theory.* Cambridge, England: Cambridge University Press.

Gallistel, CR (1990) *The organization of learning.* Cambridge, Mass: MIT Press.

Gluck, MA, Thompson, RF (1987) Modeling the neural substrates of associative learning and memory: a computational approach. *Psychological Rev.* **94**, 176-191.

Greenough, WT, and Bailey, CH (1988). The anatomy of a memory: Convergence of results across a diversity of tests. *Trends in Neuroscience,* **11**, 142-147.

Hammer, M (1994) An identified neuron mediates the unconditioned stimulus in associative olfactory learning in honeybees. *Nature* 366:59-63.

Hawkins RD, Kandel ER (1984) Is there a cell-biological alphabet for simple forms of learning? *Psychological Rev.* **91(3)**:375-91.

Hebb, DO (1949) *The organization of behavior.* New York: Wiley.

Kalman, RE (1960) A new approach to linear filtering and prediction problems. *J. Basic Eng., Trans ASME, Series D* 82(1):35-45.

Ljungberg, T, Apicella, P & Schultz, W (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology,* **67(1)**, 145-163.

MacKintosh, NJ (1983) *Conditioning and Associative Learning.* Oxford University Press: Oxford, UK.

Montague, P. R., Dayan, P. and Sejnowski, T. J., A framework for mesolimbic dopamine systems based on predictive Hebbian learning, *Journal of Neuroscience* (submitted for publication).

Quartz, S, Dayan, P, Montague, PR, Sejnowski, TJ (1992) Expectation learning in the brain using diffuse ascending connections. *Soc. Neurosci. Abstr.* **18**:1210

Pearce JM, Hall G (1980) A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review* 87: 532-52.

Rauschecker JP (1991) Mechanisms of visual plasticity: Hebb synapses, NMDA receptors, and beyond. *Physiological Reviews* 71(2):587-615.

Rescorla, RA & Wagner, AR (1972). A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. In AH Black & WF Prokasy, editors, *Classical Conditioning II: Current Research and Theory,* pp 64-69. New York, NY: Appleton-Century-Crofts.

Schultz, W, Apicella, P, Ljungberg, T (1993) Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J. Neuroscience* 13(3):900-13.

Sutton, RS, Barto, AG (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review,* **88** 2, pp 135-170.

Sutton, RS (1988). Learning to predict by the methods of temporal difference. *Machine Learning,* **3**, pp 9-44.

Sutton, RS, Barto, AG (1987). A temporal-difference model of classical conditioning. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society.* Seattle, WA.

Sutton, RS, Barto, AG (1989). Time-derivative models of Pavlovian reinforcement. In M Gabriel & J Moore, editors, *Learning and Computational Neuroscience.* Cambridge, MA: MIT Press.

Widrow, B, Stearns, SD (1985) *Adaptive signal processing.* Englewood Cliffs, NJ: Prentice-Hall.