Pavlovian Pattern Learning by Nonlinear Neural Networks

STEPHEN GROSSBERG

Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Mass. 02139

Communicated by Norman Levinson, January 11, 1971

ABSTRACT This note describes laws for the anatomy, potentials, spiking rules, and transmitters of some networks of formal neurons that enable them to learn spatial patterns by Pavlovian conditioning. Applications to space-time pattern learning and operant conditioning are then possible, if the conditioning is viewed as multi-channel Pavlovian conditioning in a highly inhomogeneous anatomy. In suitable anatomies, biases in learning because of axon collaterals with nonuniformly distributed diameters can be corrected if one properly couples the action potential to transmitter potentiation, and chooses signal velocity proportional to axon diameter. These anatomies can contain any number of cells. Anatomies exist in which patterns may be learned without their being practiced overtly, whereas persistent recall of old patterns without the learning of newly imposed patterns is impossible. Physiologically, this constraint has the trivial interpretation that signals from one cell to another first pass through the intervening synaptic knob. Mechanisms that control learning rates at times important to the network (e.g., reward and punishment times) can be discussed. Serial behavior like that described by Lashley is possible: this consists of sequential learning and performance of patterns faster than would be allowed by a motor-feedback control, at velocities influenced by arousal level, with the possibility of abrupt termination of performance if conflicting environmental demands arise. Analogs of pattern completion and mass action exist, as do phase transitions in memory (for some rate parameters and anatomies, memory is rigid, for others, it is plastic). The laws limit the ways in which these networks can be interconnected to yield specific discrimination, learning, memory, and recall capabilities.

1. INTRODUCTION

This note summarizes some results on pattern learning by neural networks. The learning mechanism is Pavlovian conditioning [1, 2]. This mechanism is described by systems of nonlinear functional–differential equations that represent cross-correlated flows on signed directed networks, or Embedding Fields [3]. Our theorems discuss the learning behavior of any finite number of formal neurons in suitable anatomies under very weak physiological constraints. They show how these neurons learn arbitrary spatial patterns. Once spatial pattern learning is assured, one can generalize the results to include learning of any number of arbitrary space–time patterns [2, 4], the discrimination of such patterns [5], and various influences of operant conditioning (to be published) on the learning process.

Some conditions are the best possible for the systems under study. They include various unusual mathematical properties having such empirical interpretations as: pattern completion [6]; mass action [7]; recurrent networks that can behave like nonrecurrent networks if the numerical values of spiking thresholds in excitatory recurrent interneurons and the arousal level of the system are properly chosen; mechanisms for rapidly performing complex sequential acts without motor feedback and at velocities depending on the arousal level of the system, and for terminating such performance when more important environmental demands arise [8]; “Now Print” mechanisms for speeding up learning during significant events [9]; cell body ensembles of any size that fire with different time lags, thresholds, and axon path weights without causing long-term biases in learning; phase transitions in memory whereby, for some choices of rate parameters or anatomy, memory is plastic, and for other choices, memory is rigid; and a factorization of system responses into “pattern” variables (“information” variables) and “energy” variables (“power” variables).

2. UNBIASED LEARNING WITH ARBITRARY AXON WEIGHTS GIVEN ACTION POTENTIALS AND CHEMICAL SYNAPSES

We find that two types of anatomy (or network connections) and variants thereof are particularly well suited to pattern learning. Let any finite number of cells (or network vertices) α send axons (or directed edges) to any finite number of cells β. The cases α = β and α ∩ β = φ permit perfect pattern learning even if the strengths of the axon connections from α to β are arbitrary positive numbers. In these anatomies, axon diameters can be chosen with complete freedom, and one can grow axons between cells separated by arbitrary distances without concern about their diameters. More elaborate anatomies are needed in realistic cases. Our theorems delineate some basic principles that can be extended to various such cases. Not all anatomies behave well, however. See refs. 4, 5 and 10 for examples of other anatomies.

Abbreviations: CS, conditioned stimulus; UCS, unconditioned stimulus; UCR, unconditioned response.
Only certain types of signal transmission between cells can compensate for differences in connection strengths and thereby yield unbiased learning. The simplest possibility is the following. Let signals (e.g., the action potential, ref. 11) propagate along the circumference of a cylindrical axon to the axon’s synaptic knob (or arrowhead of the directed edge). Let the signal disperse throughout the cross-sectional area of the synaptic knob (e.g., as ionic fluxes). Let local chemical transmitter production in the knob be proportional to the local signal density. Finally, let the effect of the signal on the postsynaptic cell be proportional to the product of local signal density and local available transmitter density and the cross-sectional area of the knob. By contrast, a mechanism whereby signals propagate throughout the cross-sectional area of the axon could not produce unbiased learning given arbitrary axon connection strengths, or at least such a mechanism is still elusive. Also, even given an action potential, unbiased learning would not occur without the interaction of the signal with the chemical transmitter production step. Electrical synapses alone presumably could not execute the desired transformation. Of course, all these conclusions are based on empirical interpretations of the mathematics, and such interpretations are never infallible.

An important constraint in our theorems is that the time lag from a given cell for signal transfer to all the cells in a functionally coordinated unit depend only on the source cell. How can different axons from the given cell have the same time lag if they have different lengths? Clearly, then, signal velocity is proportional to axon length. But signal velocity is a local property of signal transmission, whereas axon length is a global feature of the anatomy. How can this global property be converted into a locally discernible one? A simple way is to let axon length be proportional to axon diameter, and then to let signal velocity be proportional to axon diameter. The latter is often the case [12]. The former is qualitatively true: longer axons of a given cell type are usually thicker. Intuitively, this condition means that one idealized cell of a given type can be converted into another of the same type simply by blowing up spatial and temporal scales by a common factor; that is, “form” is invariant under size changes. We call this property spatiotemporal self-similarity [13]. Actually the theorems extend to the case when time lags from a given cell differ, but then the learned pattern is often more complex than a spatial pattern.

Our results are dramatically altered if, for example, in the case \( \alpha = \beta \), the cells do not send axons to themselves. The only known mathematical results in this case discuss three cells interacting with zero spiking thresholds and instantaneous signal transmission between cells [14, 15]. In fact, these results suggest that this is a “bad” anatomy for pattern learning. The existence of self-excitatory axons in a recurrent network is made plausible by the idea that even randomly growing axon collaterals that succeed in reaching all other cell bodies of a network can also with high probability reach the mother cell body.

A famous example of Pavlovian conditioning is the following (1). A hungry dog, presented with food (the unconditioned stimulus, or UCS) will salivate (the unconditioned response, or UCR). A bell (the conditioned stimulus, or CS) does not initially elicit salivation, but will do so after pairing CS and UCS several times. We will discuss the interaction (“pairing”) of cells \( C = \{ v_i : j \in J \} \) activated by any finite number \( (\leq |J|) \) of CS’s with the cells \( \Theta = \{ u_i : i \in I \} \) activated by any UCS spatial pattern. Ref. 16 discusses the empirical interpretations of the following equations for cell body potentials \( z_i \) and synaptic knob transmitters \( z_{ji} \).

### 3. MATHEMATICAL RESULTS

Consider the system

\[
\dot{z}_i = A z_i + \sum_{k \in J} B_{kj} z_{ki} + C_i, \quad (1)
\]

and

\[
\dot{z}_{ji} = D_{ji} z_{ji} + E_{ji} z_{ki}, \quad (2)
\]

where \( i \in I, j \in J, \) and \( I \) and \( J \) are sets of indices of any finite size. The symbols \( A, B_{ij}, D_{ij}, \) and \( E_{ij} \) are continuous functionals, not necessarily linear, with all \( B_{ij} \) and \( E_{ij} \) nonnegative (they represent spiking frequency terms). The input function \( C_i \) is nonnegative and continuous in \( t \), and all initial data are nonnegative. The behavior of this system depends crucially on its anatomy. As remarked in Section 1, we will choose \( I = J \) or \( I \cap J = \phi \). Our method is readily extended to cases in which each cell \( v_j, j \in J \), sends axons to all \( v_i, i \in I \), and other cells: simply relativize all computations to cells \( v_i, i \in I \).

Our theorems discuss the response to any spatial pattern \( C_i(t) = \theta_i C(\theta) \), where \( \theta_i \geq 0 \) and \( \Sigma \theta_k = 1 \). Such an input is called a spatial pattern since, in daily life, the identification of a picture is invariant under fluctuations in total input intensity \( C(t) \) over a broad physiological range. The relative intensity \( \theta_i \) at each spatial point characterizes the picture. Thus we study the limiting behavior of “pattern” variables: the relative potentials \( X_i = x_i [\Sigma \xi_k x_k]^{-1} \) and the relative transmitters \( Z_{ji} = z_{ji} [\Sigma \xi_k x_k]^{-1} \). All their oscillations can also be classified [16]. Once behavior of these variables is established in general, analysis of the “total energy” variables \( x = \Sigma \xi_k x_k \) and \( z = \Sigma \xi_k Z_{jk} \) can be carried out for particular choices of functionals. Then behavior of \( x \) and \( z \) is also known; compare ref. 16.

Our first theorem will be expressed in terms of the function \( f(S, T) = \int C \exp (\int T A d\theta) d\theta \); the functions \( M(t) : [0, \infty) \to J \) such that \( Z_{\{m(0)\}J}(t) = \max \{ Z_{ji}(t) : j \in J \} \) and \( m(t) : [0, \infty) \to J \) such that \( Z_{\{m(0)\}J}(t) = \]}
of $E_p$. Proposition 1 will show that the local flow condition is not superfluous.

The next theorem also uses the following functions. Let $N(i): [0, \infty) \to J(1)$ be defined by $Z_{n(i),(i),l}(t) = \max \{Z_{n(i)}(t): j \in J(1)\}$, and let $n(i): [0, \infty) \to J(1)$ be defined by $Z_{n(i),(i),l}(t) = \min \{Z_{n(i)}(t): j \in J(1)\}$, where $J(1) = \{j \in J: \int_0^\infty B_{jj} x_j^{-1} dt = \infty\}$.

Theorem 2

Again suppose that the system is bounded, the UCS is presented sufficiently often,

(vii) the local flow condition holds; that is, for every $j \in J$,

$$\int_0^\infty B_{jj} x_j^{-1} dt = \infty \quad \text{only if} \quad \int_0^\infty E_{j} x \exp\left(-\int_0^\infty D_j dt\right) dt = \infty;$$

and

(viii) those CS’s which are performed continually are also practiced with the UCS sufficiently often; that is, if $J(1) \neq \emptyset$, then condition (iv) holds with $M(i)$ and $m(i)$ replaced by $N(i)$ and $n(i)$. Then the potentials pick up the pattern weights and all transmitters learn the pattern at least partially; that is, all the limits $Q_i$ and $P_{j,t}$ exist with $Q_i = \theta_i$. If, moreover, a CS is practiced with the UCS sufficiently often, then it learns the pattern perfectly; that is, if (ii) holds for some $j \in J$, then $P_{j,t} = \theta_i$.

Corollary 2. Conditions (iii), (vii), and (viii) are implied by conditions (i), (vi), and (ix) for every $j \in J$, $\int_0^\infty B_{j} dt = \infty$ only if $\int_0^\infty E_{j} dt = \infty$. Under these circumstances, if $\int_0^\infty E_{j} dt = \infty$, then $P_{j,t} = \theta_i$. Corollary 2 removes a condition imposed in Theorem 1 of ref. 16.

Proposition 1. Suppose (ix) does not hold. Partition $J$ into subsets $J(2)$ and $J(3)$ such that

$$J(2) = \{j: \int_0^\infty B_{j} dt = \infty \text{ and } \int_0^\infty E_{j} dt < \infty\} \neq \emptyset.$$

Suppose that the system is bounded, that (vi) holds, that

(x) there is perfect memory until recall in $J(2)$; that is, $D_j \geq -\gamma_j E_j$ for some constant $\gamma_j > 0$, $j \in J(2)$; and that

(xi) average performance energy in $J(2)$ does not converge to zero; that is, for every $T \geq 0$,

$$\limsup_{t \to \infty} \sum_{k \in J(2)} \int_0^T B_{kk} \exp\left[\int_0^T A d\xi\right] dv > 0.$$

Then even if $Q_i$ exists, $Q_i \neq \theta_i$, so that even if $P_{j,t}$ exists and $\int_0^\infty E_{j} dt = \infty$, $P_{j,t} \neq \theta_i$.

The extension to arbitrary positive connection weights is achieved by the system

$$x_i = A x_i + \sum_{k \in J} B_{ik} x_k + C_i + \frac{d}{dt} x_i$$

(3)
and

\[ \dot{x}_t = D_t \dot{x}_t + E_t \beta_t^{-1} x_t, \tag{4} \]

where the \( \beta_t \)'s are positive numbers. To achieve performance of the pattern weights \( \theta_t \), it is necessary by (3) that the probabilities \( P^{(2)}_t = \beta_t \dot{x}_t \) converge to \( \theta_t \) as \( t \to \infty \). This will occur under the conditions of Theorem 1 and 2 applied to the variables \( x_t \) and \( \dot{x}_t = \beta_t \dot{x}_t \). The \( \beta_t \)'s in (3) and (4) can be interpreted as follows. Let the radius of the axon from \( v_t \) to \( v_i \) be \( R_{ji} \) and let signal strength be proportional to the axon circumference \( |\theta_t| \) (\( \equiv R_{ji}^2 \)). This accounts for \( \beta_t \) in (3). Let the signal disperse throughout the cross-sectional area \( (\equiv R_{ji}^2) \) of the synaptic knob, yielding a density proportional to \( R_{ji}^{-1} \). This accounts for \( \beta_t^{-1} \) in (4). Thus the definition \( \beta_t = \beta_t R_{ji} \) yields (3) and (4).

Theorems have also been proved for the general non-negative systems

\[ \dot{x}_t = A_t x_t + \sum_{k \in J} B_{kt} \dot{x}_k + C_t \]

and

\[ \dot{\alpha}_t = D_t \dot{\alpha}_t + E_t \alpha_t \]

under conditions which guarantee that they approximate systems of type (1)–(2) sufficiently well as \( t \to \infty \) to yield perfect pattern learning.

This work was supported in part by the Office of Naval Research (N00014-67-A-0204-0016) and by the Alfred P. Sloan Foundation (71609).


