Reprinted from the PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES
Vol. 58, No. 4, pp. 1329–1334. October, 1967.

*NONLINEAR DIFFERENCE-DIFFERENTIAL EQUATIONS IN PREDICTION AND LEARNING THEORY**

BY STEPHEN GROSSBERG

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Communicated by Norman Levinson, July 31, 1967

1. *Introduction.*—This note introduces some nonlinear difference-differential equations which can be interpreted as a learning theory or, alternatively, as a prediction theory whose goal is to discuss the prediction of individual events, in a fixed order, and at prescribed times. The theory provides a mathematical description of the following kind of experiment. An experimenter $\mathcal{E}$, confronted by a machine $\mathfrak{M}$, presents $\mathfrak{M}$ with a list of "letters" or "events" to be learned. Suppose, for example, that $\mathcal{E}$ wishes to teach $\mathfrak{M}$ the list of letters AB, or to predict the event B, given the event A. $\mathcal{E}$ does this by presenting A and then B to $\mathfrak{M}$ several times. To find out if $\mathfrak{M}$ has learned the list as a result of these list presentations, the letter A alone is then presented to $\mathfrak{M}$. If $\mathfrak{M}$ responds with the letter B, and $\mathfrak{M}$ does this whenever A alone is said, then we have good evidence that $\mathfrak{M}$ has indeed learned the list AB. Thus $\mathfrak{M}$ learns to predict the event B whenever the event A occurs as a result of repeated presentations of the list AB.

We will introduce some mathematical machines that learn lists in the above manner. These machines have some properties that have a familiar intuitive interpretation. For example, in our simplest machine, it is possible to make rigorous the heuristic statements that (1) "practice makes perfect," (2) an isolated system suffers no memory loss, (3) an isolated system remembers without practicing overtly, (4) the memory of an isolated system sometimes spontaneously improves without practice, (5) all errors can be corrected, although the rate of correction is sometimes diminished by response interference due to prior learning, (6) increasing the number of response alternatives sometimes diminishes the learning rate, and so on.

All of the machines we will introduce possess the same dynamical laws. Nonetheless, the exact manner in which a given machine learns and remembers depends crucially on the particular way in which the components of the machine are interconnected; i.e., on its "geometry." For example, properties (2) and (3) above do not hold in all the machines we shall describe. In this note, we will sketch some results for our simplest machine. These results, along with results for more complicated machines, are described in greater detail in another place.[1]

2. *The Nonlinear Systems.*—Each of our machines $\mathfrak{M}$ is described in terms of a positive integer n; positive rate constants α, u, and β; a nonnegative time lag τ; and an $n \times n$ matrix $P = ||p_{ij}||$ whose entries satisfy $p_{ij} \geq 0$ and $\sum_{k=1}^{n} p_{ik} = 0$ or 1. Given these quantities, let

$$\dot{x}_i(t) = -\alpha x_i(t) + \beta \sum_{k=1}^{n} x_k(t-\tau) y_{ki}(t) + I_i(t), \tag{1}$$

$$y_{jk}(t) = p_{jk} z_{jk}(t) \Big[\sum_{m=1}^{n} p_{jm} z_{jm}(t) \Big]^{-1}, \tag{2}$$

and

$$\dot{z}_{jk}(t) = [-uz_{jk}(t) + \beta x_j(t - \tau)x_k(t)]\theta(p_{jk}), \tag{3}$$

for all $i,j,k = 1,2,\ldots,n$, where

$$\theta(p) = \begin{cases} 1 \text{ if } p > 0 \\ 0 \text{ if } p \leq 0. \end{cases}$$

The initial data of this system must always be nonnegative. We also require it to be continuous, and for convenience, suppose that $z_{jk}(0) > 0$ iff $p_{jk} > 0$.

We now state some results for the simplest machine to illustrate the theory. This system is characterized by the matrix P with entries $p_{12} = p_{13} = \ldots = p_{1n} = 1/(n-1)$ and all other entries equal to zero. This system obeys the equations

$$\dot{x}_1(t) = -\alpha x_1(t) + I_1(t), \tag{4}$$

$$\dot{x}_j(t) = -\alpha x_j(t) + \beta x_1(t - \tau)y_{1j}(t) + I_j(t), \tag{5}$$

$$y_{1j}(t) = z_{1j}(t)[\sum_{k=2}^{n} z_{1k}(t)]^{-1}, \tag{6}$$

and

$$\dot{z}_{ij}(t) = -uz_{1j}(t) + \beta x_1(t - \tau)x_j(t), \tag{7}$$

where $j = 2,3,\ldots,n$. This system has the following geometrical interpretation as a graph G^2 with vertices $V = \{v_i: \ i = 1,2,\ldots,n\}$, and directed edges $E = \{e_{1j}: \ j = 2,3,\ldots,n\}$. $x_i(t)$ is interpreted as the state of a process at v_i, and $y_{1j}(t)$ is interpreted as the state of a process at the arrowhead of e_{1j}. Then (4)–(7) can readily be thought of as a flow of the quantity $x_1(t)$ over the edges e_{1j}, with flow velocity $v = 1/\tau$. That is, at every time $w = t - \tau$, the quantity $\beta x_1(w)$ is transmitted from v_1 along each edge e_{1j} and reaches the arrowhead of e_{1j} at time $w + \tau = t$. This quantity then instantaneously activates the process described by $y_{1j}(t)$, and a total magnitude $\beta x_1(t - \tau)y_{1j}(t)$ reaches v_j from v_1 at time t. $x_j(t)$ changes at a rate equal to this input, as in(5). $x_j(t)$ also decays spontaneously at a rate $-\alpha x_j(t)$, and grows at a rate equal to the input $I_j(t)$, which the experimenter controls.

The quantity $y_{1j}(t)$ appearing in (5) is itself influenced by all the quantities $z_{1j}(t)$, as (6) shows. We interpret $z_{1j}(t)$ as follows. At time t, a quantity $\beta x_1(t - \tau)$ reaches the arrowhead of e_{1j}, and this arrowhead impinges on v_j which has the value $x_j(t)$. $z_{1j}(t)$ *cross-correlates* $\beta x_1(t - \tau)$ and $x_j(t)$, in the sense that it changes at a rate proportional to $\beta x_1(t - \tau)x_j(t)$, as (7) shows. $z_{1j}(t)$ also decays spontaneously at rate $-uz_{1j}(t)$. $y_{1j}(t)$ compares $z_{1j}(t)$ with the sum of all $z_{1k}(t)$ belonging to any edge leading away from v_1, as in (6). It is the relative magnitude $y_{1j}(t)$, rather than $z_{1j}(t)$, that controls the size of the transmission from v_1 to v_j in (5).

In order to teach this machine $\mathfrak{M}$ the list AB, we assign to letter A the vertex v_1, to letter B the vertex v_2, and so on down to letter Z and vertex v_{26}. Suppose that we present A and B to $\mathfrak{M}$ at a periodic rate with A occurring at times $t = 0, w + W, 2(w + W), \ldots, n(w + W), \ldots$, and B occurring at times $t = w, 2w + W, 3w + 2W, \ldots, (n + 1)w + nW, \ldots$. Each presentation of a letter to $\mathfrak{M}$ at a time t_0 is represented by an "input pulse" to the corresponding vertex with "onset time" t_0. An *input*

pulse is a continuous and nonnegative function J which is positive in a finite interval. The *onset time* of J is $\inf\{t: J(t) > 0\}$. The experiment described above can thus be mathematically formulated as inputs

$$I_1(t) = \sum_{k=0}^{\infty} J_1(t - k(w + W)), \tag{8}$$

$$I_2(t) = \sum_{k=0}^{\infty} J_2(t - w - k(w + W)), \tag{9}$$

and

$$I_j(t) \equiv 0, \qquad j \neq 1,2, \tag{10}$$

in (5), where J_1 and J_2 are input pulses with onset time zero.

3. *Mathematical Results.*—We announce here some results for the special case of the list AB periodically presented to $\mathfrak{M}$. These results can be substantially generalized.[1] They describe the limiting and oscillatory behavior of the probabilities $y_{1j}(t) = z_{1j}(t)[\sum_{k=2}^{n} z_{1k}(t)]^{-1}$, and $X_j(t) = x_j(t)[\sum_{k=2}^{n} x_k(t)]^{-1}$, as $t \to \infty$.

THEOREM 1. *The limits* $Q_j = \lim_{t\to\infty} X_j(t)$ *and* $P_{1j} = \lim_{t\to\infty} y_{1j}(t)$ *exist and equal*

$$Q_2 = P_{12} = 1, \tag{11}$$

and

$$Q_k = P_{1k} = 0, \qquad k = 3,\ldots,n. \tag{12}$$

Moreover, the functions $\dot{y}_{1j}$ and $f_j = y_{1j} - X_j$ change sign at most once. When $j = 2$, they do not change sign at all if $X_2(0) \geq y_{12}(0)$, whereas they change sign once if $X_2(0) < y_{12}(0)$. When $j \neq 2$, they do not change sign at all if $X_j(0) \leq y_{1j}(0)$, whereas they change sign once if $X_j(0) > y_{1j}(0)$.

That is, perturbing vertices v_1 and v_2 periodically forces all the mass $\sum_{k=2}^{n} X_k(t)$ in the vertices and all the mass $\sum_{k=2}^{n} y_{1k}(t)$ in the edges to be concentrated in v_2 and e_{12}, respectively, as $t \to \infty$.

Theorem 1 describes what happens when A and B are each presented infinitely often to $\mathfrak{M}$. No experiment lasts more than a finite amount of time, however. We therefore study what happens when v_1 and v_2 are each perturbed exactly N times, followed by one perturbation of v_1. This corresponds to an experiment in which AB is presented N times and then only A is presented in the hope that the output B will be produced in reply.

Our first result describes what happens *after* AB has been presented N times, and thus when only A alone is presented. This case corresponds to a machine in which there is a t_0 such that $I_j(t) = 0$, for all $t \geq t_0$ and $j = 2,3,\ldots,n$. We will assume for simplicity that t_0 is chosen after v_1 has received at least one input.

THEOREM 2. *If* $I_j(t) = 0$ *for all* $t \geq t_0$ *and* $j = 2,3,\ldots,n$, *then* $X_j(t)$ *and* $y_{1j}(t)$ *are monotonic in opposite senses and*

$$\lim_{t\to\infty} X_j(t) = \lim_{t\to\infty} y_{1j}(t).$$

In particular, if $X_j(t_0) = y_{1j}(t_0)$, then $X_j(t) = y_{1j}(t) =$ constant, $t \geq t_0$. In all cases, $X_j(t)$ and $y_{1j}(t)$ are contained in the interval $[m_j, M_j]$, where $m_j = \min\{X_j(t_0), y_{1j}(t_0)\}$ and $M_j = \max\{X_j(t_0), y_{1j}(t_0)\}$. If we can guarantee by presenting AB sufficiently often that m_2 and M_2 are close to 1 at some time $t = t_0$, then we can replace the infinite strings of input pulses in (S) and (9) by finite strings of N pulses with an arbitrarily small change of the limits in (11) and (12), if N is taken sufficiently large. This we now do.

We consider an infinite collection of experiments $\{S_N: N = 1,2,\ldots\}$. Each experiment S_N has the same initial data as the system of Theorem 1. Moreover, in S_N, AB occurs N times at a periodic rate and is followed by a single presentation of A. We denote the functions of S_N by superscripts "(N)." For example, I_1 is written as $I_1^{(N)}$. Thus

$$I_1^{(N)}(t) = \sum_{k=0}^{N-1} J_1(t - k(w + W)) + J_1(t - \Lambda(N)),$$

$$I_2^{(N)}(t) = \sum_{k=0}^{N-1} J_2(t - w - k(w + W)),$$

and

$$I_j^{(N)}(t) \equiv 0, \qquad j = 3,\ldots,n,$$

where $\Lambda(N) \gg w + (N-1)(w + W)$. Our goal is to see how presenting AB an ever larger number N of times influences the guessing of B given A on the test trial $J_1(t - \Lambda(N))$. This goal is achieved, essentially, by thinking of the Nth experiment as the case of Theorem 1 for small times and as the case of Theorem 2 for large times.

THEOREM 3. *Given any sequence* $\{S_N: N = 1,2,\ldots\}$ *of experiments, then,*

(a) *for every* $N \geq 1$, *the limits* $Q_j^{(N)} = \lim_{t\to\infty} X_j^{(N)}(t)$ *and* $P_{1j}^{(N)} = \lim_{t\to\infty} y_{1j}^{(N)}(t)$ *exist and* $Q_j^{(N)} = P_{1j}^{(N)}$,

(b) *for every* $N \geq 1$ *and all* $t \geq w + (N-1)(w + W) + \sup\{v: J_2(v) > 0\}$, $X_2^{(N)}(t)$ *and* $y_{12}^{(N)}(t)$ *are contained in an interval* $[m_2^{(N)}, M_2^{(N)}]$ *such that* $\lim_{N\to\infty} m_2^{(N)} = \lim_{N\to\infty} M_2^{(N)} = 1$. *In particular,*

$$\lim_{N\to\infty} Q_2^{(N)} = \lim_{N\to\infty} P_{12}^{(N)} = 1 \tag{13}$$

and

$$\lim_{N\to\infty} Q_j^{(N)} = \lim_{N\to\infty} P_{1j}^{(N)} = 0, \tag{14}$$

$j = 3,\ldots,n$.

Equations (13) and (14) are the finite analogues of the limits (11) and (12) for the infinite experiment with inputs (8)–(10).

(c) *For every* $N \geq 1$ *and* $j = 2,\ldots,n$, *the functions* $y_{1j}^{(N)}$ *and* $f_j^{(N)} = y_{1j}^{(N)} - X_j^{(N)}$ *change sign at most once. When* $j = 2$, *they do not change sign at all if* $X_2^{(N)}(0) \geq y_{12}^{(N)}(0)$, *whereas they change sign once if* $X_2^{(N)}(0) < y_{12}^{(N)}(0)$. *When* $j \neq 2$, *they do not change sign at all if* $X_j^{(N)}(0) \leq y_{1j}^{(N)}(0)$, *whereas they change sign once if* $X_j^{(N)}(0) > y_{1j}^{(N)}(0)$.

4. *Learning.*—We illustrate the effects of the probabilities $y_{1j}^{(N)}(t)$ on the out-

puts $x_j^{(N)}(t)$ in the Nth experiment S_N when N is taken large and t is so large that the inputs corresponding to the presentations of AB are zero. The test trial $J_1(t - \Lambda(N))$ is also assumed to occur after the inputs corresponding to AB are zero.

For $N \gg 1$ and $t \geq \Lambda(N)$, $y_{12}^{(N)}(t) \cong 1$ and $y_{1j}^{(N)}(t) \cong 0$, $j \neq 1,2$, by Theorem 3. By contrast, we readily see that each output $x_i^{(N)}(t)$ converges exponentially to zero when no inputs are present, and thus $x_i^{(N)}(\Lambda(N)) \cong 0$ for all $i = 1,2,\ldots,n$. To study the effects of $J_1(t - \Lambda(N))$, we integrate (5) and find that $x_j^{(N)}(t) \cong 0$ for all $t \geq \Lambda(N)$ and all $j = 3,4,\ldots,n$, whereas

$$x_2^{(N)}(t) \cong \beta e^{-\alpha(t-\Lambda(N)-\tau)} \int_0^{t-\Lambda(N)-\tau} dv \int_0^v e^{\alpha w} J_1(w)dw,$$

for $t \geq \Lambda(N) + \tau$. Thus the output from B, namely $x_2^{(N)}$—and *only* $x_2^{(N)}$ among all the outputs $x_j^{(N)}$ representing letters $B,C,\ldots,Z$—becomes large after the test input to A. That is, periodically presenting AB and thereby periodically perturbing v_1 and v_2 a large number N of times has the consequence that only v_2 produces a large output when v_1 is perturbed at a later time. That is, S_N "learns" that B follows A as N becomes large. Indeed, as N is taken increasingly large, an increasingly large fraction of the output comes from v_2, so that "practice makes perfect."

In an isolated system (i.e., one which is receiving no inputs), all outputs converge exponentially to zero. That is, no "overt practice" occurs. Nonetheless, for sufficiently many trials N and sufficiently large t, Theorem 3 guarantees that the "associations" $y_{1j}^{(N)}(t)$ remain essentially constant. Thus, no "forgetting" occurs even in the absence of overt practice.

Moreover, for sufficiently large N and all large t, it is easily seen that $y_{12}^{(N)}(t)$ is monotone *increasing*. That is, the association from A to B undergoes "spontaneous facilitation" or "reminiscence"[3] even in the absence of overt practice. This effect is most evident when N is large but not so large that the gap $X_2^{(N)}(t) - y_{12}^{(N)}(t)$ is small when the presentations of AB cease; i.e., during experiments providing "moderate practice." It is also most evident immediately after the outputs $x_j^{(N)}$ undergo their rapid exponential decay when the presentations of AB cease; i.e., shortly after moderate practice ceases.

Increasing the number of response alternatives $B,C,\ldots,Z$ (or the number of vertices $v_2,v_3,\ldots v_n$) can decrease the rate of learning by decreasing the effect of z_{12} on the "association" $y_{12} = z_{12}\left[\sum_{k=2}^{n} z_{1k}\right]^{-1}$. Nonetheless, after considerable practice, when $y_{12} \cong 1$, these alternatives have little effect, since $z_{12} \gg z_{1j}$, $j \neq 1,2$.

All errors can be corrected, because Theorems 2 and 3 hold for *all* positive initial data. Suppose, for example, that AB has been taught to $\mathfrak{M}$ for a finite amount of time, until time $t = T_0$, say. Then $y_{12}(T_0) \cong 1$. We can thereafter present AC periodically to guaranteee that $y_{13}(t) \cong 1$ for t sufficiently large. The "error" B given A has hereby been corrected. Nonetheless, it will take longer to bring $y_{13}(t)$ close to 1 if $y_{12}(T_0) \cong 1$ than it would if $y_{1j}(T_0) \cong 1/(n-1)$, $j = 2,\ldots,n$; i.e., "response interference" due to the association $A \rightarrow B$ has occurred.

The above remarks illustrate that the interaction of the outputs (or "stimulus

traces") $x_i(t)$ and the "associations" $y_{1j}(t)$ with the inputs $I_k(t)$ can be subjected to many thought experiments which often have a heuristic interpretation with a familiar psychological ring to them.

5. *Summary.*—Some nonlinear difference-differential equations are introduced which can be interpreted as a learning theory or prediction theory. The simplest case of learning to predict an event B given an event A is briefly discussed.

* This work was supported in part by NONR contract 1841/38.

[1] Grossberg, S., "Global ratio limit theorems for some nonlinear functional-differential equations," *Bull. Am. Math. Soc.*

[2] Busacker, R. G., and T. L. Saaty, *Finite Graphs and Networks* (New York: McGraw-Hill, 1953).

[3] Osgood, C. E., *Method and Theory in Experimental Psychology* (New York: Oxford University Press, 1953).