Reprinted from
STUDIES IN APPLIED MATHEMATICS 5

A Global Prediction (or Learning) Theory for Some Nonlinear Functional-Differential Equations

Stephen Grossberg, *Massachusetts Institute of Technology*

1. Introduction. This article surveys some recent global limit and oscillation theorems for some systems of nonlinear difference-differential equations that define cross-correlated flows on probabilistic networks. These equations comprise the first stage of a theory of learning [1] that attempts to unify, at least qualitatively, some data from psychology, neurophysiology and neuroanatomy by finding common mathematical principles that underly these data. The behavior of these networks can be interpreted as a nonstationary and deterministic prediction theory because the networks learn in a way that imitates the following heuristic example chosen from daily life.

An experimenter $\mathscr{E}$ teaches a human subject $\mathscr{S}$ a list of events (say the list AB of letters) by presenting the letters one after the other to $\mathscr{S}$ and then presenting the list several times in this way. To test if $\mathscr{S}$ has learned the list, $\mathscr{E}$ then presents A alone to $\mathscr{S}$ and hopes that $\mathscr{S}$ will reply with B. If $\mathscr{S}$ does so whenever A is presented, $\mathscr{E}$ can safely assume that $\mathscr{S}$ has learned the list. We shall construct machines (the networks) which learn according to a similar procedure. At least three phases exist in this learning paradigm: (i) the learning trial during which the letters are presented, (ii) a remembering period during which no new material is presented, and (iii) a recall trial during which $\mathscr{E}$ tests $\mathscr{M}$'s memory by presenting A alone and noting how well $\mathscr{M}$ can reproduce B. We shall find that varying the geometry of our networks can dramatically change the qualitative properties of each of these phases. Moreover, the networks often exhibit a monotonic response to wildly oscillating inputs; some of them become easier to analyze when loops—and thus an extra source of nonlinear interactions—are added to them, some exhibit interactions that can be interpreted as "time reversals" on $\mathscr{E}$'s time scale, and some of their stability properties become easier to guarantee as the time lag increases.

2. The networks. The networks $\mathscr{M}$ are defined by equations of the form

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

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

and

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

where α, β and u are positive; n is a positive integer; the matrix $P = \|p_{jk}\|$ is semistochastic (i.e., $p_{jk} \geqq 0$ and $\sum_{m=1}^{n} p_{jm} = 0$ or 1); and

$$\theta(p) = \begin{cases} 0 & \text{if} \quad p \leqq 0, \\ 1 & \text{if} \quad p > 0. \end{cases}$$

The initial data is nonnegative and continuous, subject for convenience to the constraint that $z_{jk}(0) > 0$ if and only if $p_{jk} > 0$. The inputs $I_i(t)$ are nonnegative and continuous functions. Each choice of parameters and initial data defines a different machine $\mathscr{M}$, and each choice of inputs in $[0, \infty)$ defines a different experiment performed by $\mathscr{E}$ on $\mathscr{M}$. Instead of presenting letters to $\mathscr{M}$, $\mathscr{E}$ presents abstract symbols r_i, $i = 1, 2, \cdots, n$. The presentation of r_i to $\mathscr{M}$ at time t_i is represented by a momentary increase of $I_i(t)$ for $t \geqq t_i$ whose shape depends on external circumstances.

3. Cross-correlated flows on probabilistic graphs. Every P can be geometrically realized as a directed probabilistic graph with vertices $V = \{v_i : i = 1, 2, \cdots, n\}$ and directed edges $E = \{e_{jk} : j, k = 1, 2, \cdots, n\}$, where the weight p_{jk} is assigned to e_{jk}. Letting $x_i(t)$ describe the state of a process at v_i and $y_{jk}(t)$ be the state of a process at the arrowhead N_{jk} of e_{jk}, then (2) can readily be thought of as a flow of the quantities βx_j over the edges e_{jk} with flow velocity $v = 1/\tau$. The coefficients $y_{jk}(t)$ in (1) control the size of the $\beta x_j(t - \tau)$ flow from v_j along e_{jk} which reaches v_k at time t by cross-correlating past $\beta x_j(w - \tau)$ and $x_k(w)$ values, $w \in [-\tau, t]$, with an exponential weighting factor $e^{-u(t-w)}$ as in $z_{jk}(t)$ in (3), and comparing this weighted cross-correlation in (2) with all other cross-correlations $z_{jm}(t)$ corresponding to any edge leading from v_j, $m = 1, 2, \cdots, n$.

4. Outstars. We now choose the geometry P of (2) to illustrate the learning of a list r_1r_2 of two symbols. Let $p_{1i} = 1/(n-1)$, $i = 2, 3, \cdots, n$, and let all other p_{jk} equal zero. This system is called an *outstar* since all positive weights p_{1i} are directed away from the single source vertex v_1.

Learning in an outstar can be described heuristically in the following way [2] : (i) "practice makes perfect"; (ii) an isolated outstar never forgets what it has been taught; (iii) an isolated outstar remembers without practicing overtly; (iv) the memory of an isolated outstar spontaneously improves if it has previously received a moderate amount of practice; (v) all errors can be corrected, although prior learning and a large number of response alternatives can diminish the learning speed; and (vi) the act of recalling r_2 given r_1 does not destroy $\mathscr{M}$'s memory of prior learning.

Mathematically speaking, these properties are described by a sequence $G^{(1)}$, $G^{(2)}, \cdots, G^{(N)}, \cdots$ of outstars with identical but otherwise arbitrary positive and continuous initial data, whose inputs are formed from the following ingredients:

(a) Let $\{\theta_j : j = 2, \cdots, n\}$ be a fixed but arbitrary probability distribution.

(b) Let f and g be bounded, nonnegative, and continuous functions in $[0, \infty)$ for which there exist positive constants k and T_0 such that

$$\int_0^t e^{-\alpha(t-w)} f(w)\, dw \geqq k, \quad t \geqq T_0,$$

and

$$\int_0^t e^{-\alpha(t-w)} g(w)\, dw \geqq k, \quad t \geqq T_0.$$

(c) Let $U_1(N)$ and $U(N)$ be any positive and monotone increasing functions of $N \geqq 1$ such that

$$\lim_{N\to\infty} U_1(N) = \lim_{N\to\infty} U(N) = \infty.$$

(d) For every $N \geqq 1$, let $h_N(t)$ be any nonnegative and continuous function that is positive only in $(U(N), \infty)$.

The input functions $I_k^{(N)}$ of $G^{(N)}$ are defined in terms of (a)–(d) by

$$I_1^{(N)}(t) = f(t)[1 - \theta(t - U_1(N))] + h_N(t) \tag{4}$$

and

$$I_j^{(N)}(t) = \theta_j g(t)[1 - \theta(t - U(N))], \qquad j = 2, \cdots, n. \tag{5}$$

Letting the functions of $G^{(N)}$ be denoted by superscripts "(N)" (e.g., y_{1j} is written as $y_{1j}^{(N)}$), and defining the ratios $X_j^{(N)} = x_j^{(N)}[\sum_{m=2}^n x_m^{(N)}]^{-1}$ for every $N \geqq 1$ and $j = 2, \cdots, n$, we can state the following theorem.

THEOREM 1. *Let $G^{(1)}, G^{(2)}, \cdots, G^{(N)}, \cdots$ be outstars with identical but otherwise arbitrary positive and continuous initial data, and any inputs chosen as in* (4) *and* (5). *Then*:

(A) *for every $N \geqq 1$, the limits $\lim_{t\to\infty} X_j^{(N)}(t)$ and $\lim_{t\to\infty} y_{1j}^{(N)}(t)$ exist and are equal, $j = 2, \cdots, n$;*

(B) *for every $N \geqq 1$ and all $t \geqq U(N)$, $X_j^{(N)}(t)$ and $y_{1j}^{(N)}(t)$ are monotonic in opposite senses, and*

$$\lim_{N\to\infty} X_j^{(N)}(U(N)) = \lim_{N\to\infty} y_{1j}^{(N)}(U(N)) = \theta_j, \qquad j = 2, \cdots, n.$$

In particular, by (A) *and* (B),

$$\lim_{N\to\infty} \lim_{t\to\infty} X_j^{(N)}(t) = \lim_{N\to\infty} \lim_{t\to\infty} y_{1j}^{(N)}(t) = \theta_j, \qquad j = 2, \cdots, n;$$

(C) *for every $N \geqq 1$ and $j = 2, \cdots, n$, the functions $\dot{y}_{1j}^{(N)}$, $F_j^{(N)} = y_{1j}^{(N)} - X_j^{(N)}$, and $G_j^{(N)} = X_j^{(N)} - \theta_j$ change sign at most once and not at all if $F_j^{(N)}(0)G_j^{(N)}(0) \geqq 0$. Moreover, $F_j^{(N)}(0)G_j^{(N)}(0) > 0$ implies $F_j^{(N)}(t)G_j^{(N)}(t) > 0$ for all $t \geqq 0$.*

Part (C) shows in particular that $y_{1j}^{(N)}$ is quite insensitive to fluctuations in f and g.

COROLLARY *The theorem is true if*

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

and

$$I_j^{(N)}(t) = \theta_j \sum_{k=0}^{N-1} J_2(t - w - k(w + W)), \qquad j = 2, \quad , n,$$

where J_i is a continuous and nonnegative function that is positive in an interval of the form $(0, \lambda_i)$, $i = 1, 2$; *w and W are nonnegative numbers whose sum is positive*; *and*

$$\Lambda(N) > w + (N - 1)(w + W) + \lambda_2$$

The case of learning a list r_1r_2 requires the further specialization that $\theta_j = \delta_{j2}$ (see [2] and [3] for further details).

5. Learning of spatial patterns by a complete graph with loops. Suppose $P = (1/n)E_n$, where all entries in E_n equal 1. Then every vertex v_i is connected to every vertex v_j by an equal weight, so $\mathscr{M}$ is called a *complete graph with loops.* We shall show that such a graph can learn a spatial pattern of arbitrary complexity that is presented even with a rapidly oscillating input just so long as the input represents sufficiently many presentations of the pattern. Again (i) "practice makes perfect"; (ii) an isolated machine never forgets; (iii) an isolated machine remembers without overtly practicing; (iv) "contour enhancement" occurs, in the sense that after a moderate amount of practice "darks get darker" and "lights get lighter" in $\mathscr{M}$'s memory; and (v) a new pattern can always be learned to replace an old pattern. Moreover, (vi) "pattern completion" occurs, in the sense that even a speck of light shone at one vertex can reproduce the entire pattern at all vertices after learning has occurred. However, (vii) the very act of perturbing the graph with any recall input other than the pattern that was learned gradually destroys $\mathscr{M}$'s memory of the original pattern.

A *spatial pattern* is, in the present context, a collection of inputs $I_i(t) = \theta_i I(t)$, where $\{\theta_i : i = 1, 2, \cdots, n\}$ is a fixed but arbitrary probability distribution, and $I(t)$ is a bounded, nonnegative, and continuous function. As in Theorem 1, we define a sequence $G^{(1)}, G^{(2)}, \cdots, G^{(N)}, \cdots$ of complete graphs with loops, now subjected to inputs of the form

$$I_j^{(N)}(t) = \theta_j J(t)[1 - \theta(t - U(N))], \tag{6}$$

where

$$\int_\tau^t e^{-\alpha(t-v)} J(v)\, dv \geqq k, \quad t \geqq T_0. \tag{7}$$

We must also properly constrain the parameters α, β, u and τ. For example, let $\sigma(\tau) \equiv u + 2s(\tau)$, where $s(\tau)$ is the largest real part of the zeros of $R_\tau(s) = s + \alpha - \beta e^{-\tau s}$.

THEOREM 2. *Consider any sequence $G^{(1)}, G^{(2)}, \cdots, G^{(N)}, \cdots$ of complete graphs with loops possessing equal but otherwise arbitrary nonnegative and continuous initial data. Let $\alpha > \beta$ and $\sigma(\tau) > 0$, and suppose that the inputs satisfy* (6) *and* (7). *Then letting $X_i^{(N)} = x_i^{(N)}[\sum_{m=1}^n x_m^{(N)}]^{-1}$,*

(A) *for every $N \geqq 1$, the limits $\lim_{t \to \infty} X_i^{(N)}(t)$ and $\lim_{t \to \infty} y_{ki}^{(N)}(t)$ exist and are equal;*

(B) *for every $N \geqq 1$ and $t \geqq U(N)$ the functions $X_i^{(N)}(t)$ and $y_{ki}^{(N)}(t)$ lie in the interval $[m_i^{(N)}, M_i^{(N)}]$, where*

$$m_i^{(N)} = \min\{X_i^{(N)}(U(N)), y_{ji}^{(N)}(U(N)) : j = \ ,2, \quad \cdot, n\},$$

$$M_i^{(N)} = \max\{X_i^{(N)}(U(N)), y_{ji}^{(N)}(U(N)) : j = 1,2, \cdot \quad , n\}$$

and

$$\lim_{N \to \infty} m_i^{(N)} = \lim_{N \to \infty} M_i^{(N)} = \theta_i, \qquad i,k = 1,2, \quad , n$$

In particular,

$$\lim_{N \to \infty} \lim_{t \to \infty} X_i^{(N)}(t) = \lim_{N \to \infty} \lim_{t \to \infty} y_{ki}^{(N)}(t) = \theta_i, \quad i,k = 1,2,\cdots,n$$

(C) *for every $N \geqq 1$ and $t \geqq 0$, the functions $\dot{Y}_{i,\theta}^{(N)}$, $\dot{y}_{i,\theta}^{(N)}$, $X_i^{(N)} - Y_{i,\theta}^{(N)}$ and $X_i^{(N)} - y_{i,\theta}^{(N)}$ change sign at most once and not at all if $y_i^{(N)}(0) \leqq X_i^{(N)}(0) \leqq Y_i^{(N)}(0)$, where $y_i^{(N)} = \min\{y_{ki}^{(N)}: k = 1,2,\cdots,n\}$, $Y_i^{(N)} = \max\{v_{ki}^{(N)}: k = 1,2,\cdots,n\}$, $y_{i,\theta}^{(N)} = \min\{y_i^{(N)}, \theta_i\}$, and $Y_{i,\theta}^{(N)} = \max\{Y_i^{(N)}, \theta_i\}$. For $t \geqq U(N)$, $\dot{Y}_i^{(N)}$, $\dot{y}_i^{(N)}$, $X_i^{(N)} - Y_i^{(N)}$ and $X_i^{(N)} - y_i^{(N)}$ change sign at most once and not at all if $y_i^{(N)}(U(N)) \leqq X_i^{(N)}(U(N)) \leqq Y_i^{(N)}(U(N))$.*

COROLLARY 2 (Stability is graded in the time lag). *If $\alpha > \beta$ and $\sigma(\tau_0) > 0$, then* (A)–(C) *hold for all $\tau \geqq \tau_0$ since $\sigma(\tau)$ is monotone increasing in $\tau \geqq 0$.* (See [4] and [5] for further details.)

6. Dependence of memory on geometry. Removing the loops from the graph has a dramatic effect on its memory. For example, let $\tau = 0, n = 3$, and $p_{ij} = \frac{1}{2}(1 - \delta_{ij})$, $i, j = 1, 2, 3$. Then [6] for arbitrary positive initial data satisfying $z_{ij}(0) = z_{ji}(0)$, $\lim_{t \to \infty} X_i(t) = \frac{1}{3}$ and $\lim_{t \to \infty} y_{jk}(t) = \frac{1}{2}(1 - \delta_{ij})$ if $\sigma(0) > 0$ and the inputs are positive only in a finite interval. In other words, $\mathscr{M}$ "forgets" everything it has learned. By contrast, for $\sigma(0) < 0$, the functions $y_{jk}(t)$ can be kept in an interval of prescribed smallness by choosing $|\sigma(0)|$ sufficiently large; i.e., $\mathscr{M}$ "remembers" arbitrarily well. In the complete graph with loops, Theorem 2 shows that $\mathscr{M}$ can remember only spatial patterns if $\sigma(\tau) > 0$. The constraint $\sigma(\tau) < 0$ is necessary in both cases for $\mathscr{M}$ to be able to remember an arbitrary pattern in space-time, such as a list.

7. Serial learning of long lists. The learning of long lists by human subjects differs significantly from the learning of short lists [7], [8]. For example, the beginning and end of a long list are often learned before the middle is learned ("bowing"), whereas a short list, such as AB, can often be learned on a single trial and such that a significant association from B to A is created ("backward learning"). Moreover, the bowing effect is sensitive to the speed with which the list is presented and to the rest interval between list presentations.

These facts imply that various learning effects propagate "backwards in time" relative to $\mathscr{E}$'s time scale [8]. For example, if the alphabet $ABC \cdots Z$ is presented just once to $\mathscr{M}$ with a time interval of w units between successive letter presentations, then $\mathscr{M}$ cannot possibly know that Z is the end of the alphabet until at least w units after Z is presented to $\mathscr{M}$, since only then does $\mathscr{M}$ know that Z will not be followed by another letter presented with the same time spacing. Similarly, the short list AB forms part of the alphabet $ABC \cdots Z$, yet the presentation of $CD \cdots Z$ after AB influences $\mathscr{M}$'s recall of AB. These effects are, for example, qualitatively found when in the following system of equations the input represents a serial presentation of a long list:

$$\dot{x}_i(t) = -\alpha x_i(t) + I_i(t), \tag{8}$$

$$y_{jk}(t) = z_{jk}(t)\left[\sum_{m=1}^{n} z_{jk}(t)\right] \tag{9}$$

$$\dot{z}_{jk}(t) = \beta x_j(t - \tau)x_k(t), \quad j \neq k, \tag{10}$$

and

$$z_{jj}(t) = 0, \tag{11}$$

for $i, j, k = 1, 2, \cdots, n$. The system (8)–(11) is a complete n-graph without loops modified by removing the interaction term $\sum_{m=1}^{n} x_m(t - \tau)y_{mi}(t)$ in order to show the primary effect of the serial input ordering on the "associations" $y_{jk}(t)$.

We present the list $r_1 r_2 \cdots r_L$ to $\mathscr{M}$ once at a speed τ, so that

$$I_j(t) = I_{j-1}(t - \tau), \qquad j = 2, 3, \cdots, L,$$

and $I_j(t) \equiv 0, j = L + 1, \cdots, n$, where $I_1(t)$ is assumed to equal zero outside the interval $[0, \tau)$. A computation of $y_{j,j+1}(t)$ at discrete time steps $t = m\tau$ yields the following theorem.

THEOREM 3. *$y_{j,j+1}((j + 1)\tau)$ is a negatively accelerated, monotone decreasing function of $j = 1, 2, \cdots, L - 1$; $y_{L-1,L}(m\tau)$ is monotone increasing in $m \geqq L + 1$; and the function*

$$B(j) = \lim_{m \to \infty} y_{j,j+1}(m\tau)$$

first decreases monotonically to a minimum attained for $j = J$ or $J + 1$, where $J = \max\{j : j \leqq \frac{1}{2}(L - 1)\}$, and then increases monotonically.

In other words, the "correct" associations $y_{j,j+1}(t)$ decrease as a function of j for t chosen right after r_j and r_{j+1} are presented. After r_L is presented, a facilitation effect at the end of the list appears, and this ultimately propagates backwards in j until the middle of the list is reached. Reference [8] describes effects such as these in detail.

REFERENCES

[1] S. GROSSBERG, *Embedding fields: A theory of learning with physiological implications*, J. Math Psych., to appear.

[2] ———, *Nonlinear difference-differential equations in prediction and learning theory*, Proc. Nat Acad. Sci. U.S.A., 58 (1967), pp. 1329–1334.

[3] ———, *A prediction theory for some nonlinear functional-differential equations, I: Learning of lists*, J. Math. Anal. Appl., 21 (1968), pp. 643–694.

[4] ———, *Some nonlinear networks capable of learning a spatial pattern of arbitrary complexity*, Proc. Nat. Acad. Sci. U.S.A., 59 (1968), pp. 368–372.

[5] ———, *A prediction theory for some nonlinear functional-differential equations, II: Learning of patterns*, J. Math. Anal. Appl., 22 (1968), pp. 490–522.

[6] ———, *On the global limits and oscillations of a system of nonlinear differential equations describing a flow on a probabilistic network*, J. Differential Equations, to appear.

[7] C. E. OSGOOD, *Method and Theory in Experimental Psychology*, Oxford University Press, London, 1953, Chapter 12.

[8] S. GROSSBERG, *On the serial learning of lists*, Math. Biosci., to appear.