LEARNING AND ENERGY-ENTROPY DEPENDENCE IN SOME NONLINEAR FUNCTIONAL-DIFFERENTIAL SYSTEMS

BY

STEPHEN GROSSBERG

Reprinted from the
BULLETIN OF THE AMERICAN MATHEMATICAL SOCIETY
November, 1969, Vol. 75, No. 6
Pp. 1238-1242

LEARNING AND ENERGY-ENTROPY DEPENDENCE IN SOME NONLINEAR FUNCTIONAL-DIFFERENTIAL SYSTEMS

BY STEPHEN GROSSBERG[1]

Communicated by Gian-Carlo Rota, July 14, 1969

1. Introduction. This note describes limiting and oscillatory features of some nonlinear functional-differential systems having applications in learning and nonstationary prediction theory. The main results discuss systems of the form

$$\dot{x}_i(t) = A(W_t, t)x_i(t) + \sum_{k\in J} B_k(W_t, t)z_{ki}(t) + C_i(t) \tag{1}$$

and

$$\dot{z}_{ji}(t) = D_j(W_t, t)z_{ji}(t) + E_j(W_t, t)x_i(t), \tag{2}$$

where $i\in I, j\in J$, and I and J are finite sets of indices such that either $I=J$ or $I\cap J=\varnothing$. The coefficients are continuous functions of t, dependent perhaps on the $|I|(1+|J|)$ dimensional vector function $W=(x_i, z_{ji}: i\in I, j\in J)$ evaluated at times no later than t, and on known functions of t. All coefficients B_j and E_j are also nonnegative, and the initial data and inputs C_i are nonnegative and continuous. The main results discuss the probabilities $y_{ji}(t)=z_{ji}(t)\left[\sum_{k\in I}z_{jk}(t)\right]^{-1}$ and $X_i(t)=x_i(t)\left[\sum_{k\in I}x_k(t)\right]^{-1}$ defined for $i\in I$ and $j\in J$, given choices of initial data and coefficient functionals for which (1) and (2) has a unique bounded solution.

These results apply for example to systems of the form

$$\dot{x}_i(t) = -\alpha_i x_i(t) + \sum_{k=1}^{n}\left[x_k(t-\tau_{ki}) - \Gamma_{ki}\right]^+ p_{ki}z_{ki}^{(+)}(t) - \sum_{k=1}^{n}\left[x_k(t-\tau_{ki}) - \Gamma_{ki}\right]^+ q_{ki}z_{ki}^{(-)}(t) + I_i(t), \tag{3}$$

$$\dot{z}_{ji}^{(+)}(t) = -u_{ji}^{(+)}z_{ji}^{(+)}(t) + v_{ji}^{(+)}\left[x_j(t-\tau_{ji}) - \Gamma_{ji}\right]^{(+)}\left[x_i(t)\right]^{(+)}, \tag{4}$$

and

$$\dot{z}_{ji}^{(-)}(t) = -u_{ji}^{(-)}z_{ji}^{(-)}(t) + v_{ji}^{(-)}\left[x_j(t-\tau_{ji}) - \Gamma_{ji}\right]^{+}\left[-x_i(t)\right]^{+}, \tag{5}$$

$i, j=1, 2, \cdots, n$, where $[w]^+=\max(w, 0)$ for any real number w; or alternatively to (3) along with

$$\dot{z}_{ji}^{(+)}(t) = \left\{-u_{ji}^{(+)}z_{ji}^{(+)}(t) + v_{ji}^{(+)}\left[x_i(t)\right]^+\right\}\left[x_j(t-\tau_{ji}) - \Gamma_{ji}\right]^+ \tag{6}$$

and

$$\dot{z}_{ji}^{(-)}(t) = \left\{-u_{ji}^{(-)}z_{ji}^{(-)}(t) + v_{ji}^{(-)}\left[-x_i(t)\right]^+\right\}\left[x_j(t-\tau_{ji}) - \Gamma_{ji}\right]^+. \tag{7}$$

Such systems and generalizations thereof, known generically as *embedding fields*, describe cross-correlated flows on signed networks which are capable of learning and predicting complicated tasks. They are discussed in special cases along with references to pertinent psychological, neurophysiological, anatomical, and biochemical data in [1]–[8].

2. Main results. The limiting and oscillatory behavior of the probabilities $y_{ji}(t)$ and $X_i(t)$ associated with (1) and (2) is quite insensitive to the detailed form of functional coefficients if $C_i(t)$ represents a spatial pattern; i.e., if $C_i(t)=\theta_i C(t)$ with $\sum_{i\in I}\theta_i=1$ and $\theta_i\geqq 0$.

THEOREM 1. *Let* (1) *and* (2) *be given with continuous and nonnegative initial data and inputs, and coefficient functionals continuous in* t *such that*

(1) *all* B_j *and* E_j *are nonnegative,*

(2) $\int_0^\infty B_j(W_v, v)dv=\infty$ *only if* $\int_0^\infty E_j(W_v, v)dv=\infty$,

(3) $\int_0^\infty C(v)dv=\infty$,

(4) *there exist positive constants* K_1 *and* K_2 *such that for every* $T\geqq 0$,

$$\int_T^{T+t} C(v)\exp\left[\int_v^{T+t} A(W_\xi, \xi)d\xi\right]dv \geqq K_1 \quad \text{if } t\geqq K_2, \tag{8}$$

(5) *the solution of* (1) *and* (2) *is bounded.*

Then all the limits $P_{ji}=\lim_{t\to\infty}y_{ji}(t)$ *and* $Q_i=\lim_{t\to\infty}X_i(t)$ *exist and* $Q_i=\theta_i$. *If moreover*

[1] The preparation of this work was supported in part by the National Science Foundation (GP 9003), the Office of Naval Research (N00014-67-A-024-0016), and the A.P. Sloan Foundation.

(9) $$\int_0^\infty E_j(W_t, v)dv = \infty,$$

then $P_{ji}=\theta_i$, *whereas if* $E_j(W_t, t)=0$, *then* $\dot{y}_{ji}(t)=0$.

Speaking heuristically, Theorem 1 says that the probabilities $y_{ji}(t)$ and $X_i(t)$ learn the weights θ_i if they practice them sufficiently often.

Oscillations are described in terms of the functions

$$Y_i(t) = \max\{y_{ji}(t): j \in J\}, \quad y_i(t) = \min\{y_{ji}(t): j \in J\},$$
$$M_i(t) = \max\{Y_i(t), X_i(t)\}, \quad m_i(t) = \min\{y_i(t), X_i(t)\},$$
$$Y_{i,\theta}(t) = \max\{Y_i(t), \theta_i\}, \quad y_{i,\theta}(t) = \min\{y_i(t), \theta_i\},$$
$$M_{i,\theta}(t) = \max\{Y_{i,\theta}(t), X_i(t)\}, \quad \text{and} \quad m_{i,\theta}(t) = \min\{y_{i,\theta}(t), X_i(t)\}.$$

PROPOSITION 1. *Let the conditions of Theorem 1 hold, and suppose also for convenience that* $\sum_{i\in I}x_i(0)>0$ *and* $\sum_{i\in I}z_{ji}(0)>0$ *if* $E_j\not\equiv 0$. *Then for every time* $T\geqq 0$ *and all* $t\geqq T$,

$$m_{i,\theta}(T) \leqq m_{i,\theta}(t) \leqq M_{i,\theta}(t) \leqq M_{i,\theta}(T).$$

If moreover $I(t)=0$ *for* $t\geqq T$, *then*

$$m_i(T) \leqq m_i(t) \leqq M_i(t) \leqq M_i(T).$$

The functions $\dot{Y}_{i,\theta}$, $\dot{y}_{i,\theta}$, $X_i-Y_{i,\theta}$, *and* $X_i-y_{i,\theta}$ *change sign at most once, and not at all if* $y_{i,\theta}(0)\leqq X_i(0)\leqq Y_{i,\theta}(0)$. *If moreover* $I(t)=0$ *for* $t\geqq T$, *then the functions* $\dot{Y}_i$, $\dot{y}_i$, X_i-Y_i, *and* X_i-y_i *change sign at most once for* $t\geqq T$, *and not at all if* $y_i(T)\leqq X_i(T)\leqq Y_i(T)$. *Also sign* $\dot{y}_{ji}(t)=sign\ (X_i(t)-y_{ji}(t))$ *if* $E_j(W_t, t)\not\equiv 0$.

Speaking heuristically, Proposition 1 says that the probabilities $y_{ji}(t)$ and $X_i(t)$ remember the weights θ_i if the inputs $C_i(t)$ cease after a sufficient amount of practice. This does not mean that the weights can always thereafter be reproduced in large outputs $x_i(t)$, since in a system of type (3)–(5), the absolute size of $z_{ji}(t)$ decays exponentially after practice ceases. In (3), (6), and (7), the absolute and relative sizes of $z_{ji}(t)$ can be perfectly remembered after practice ceases in the absence of recall trials.

Conditions which guarantee boundedness of (1) and (2) in cases of applied interest, as well as the following heuristic remarks, will be discussed in [8].

Suppose $I\cap J=\varnothing$, and write $C_i(t)=\theta_i(t)C(t)$ in (1), where $C(t)=\sum_{k\in I}C_k(t)$ and $\theta_i(t)=C_i(t)C^{-1}(t)$. Because $\dot{y}_{ji}(t)=0$ whenever $E_j(W_t, t)=0$, $y_{ji}(t)$ "samples" only the weights $\theta_i(t)$ of a given space-time pattern at times t for which $E_j(W_t, t)>0$. Since the continuous function $\theta_i(t)$ can be arbitrarily well approximated by a sequence $\theta_i(k\xi)$, $k=1, 2, \cdots$, with ξ sufficiently small, $\theta_i(t)$ can be "encoded" in a sequence $y^{(k)}\equiv\{y_{ki}: i\in I\}$, $k=1, 2, \cdots$, of probability distributions which successively sample the pattern briefly every ξ time units, in "avalanche" fashion.

Suppose $I=J=\{1, 2, \cdots, n\}$ and consider for specificity the following system of type (3)–(5).

$$\dot{x}_i(t) = -\alpha x_i(t) + \beta\sum_{k=1}^{n}[x_k(t-\tau)-\Gamma_k]^+z_{ki}(t) + I_i(t)$$

and

$$\dot{z}_{ji}(t) = -uz_{ji}(t) + \gamma[x_j(t-\tau)-\Gamma_j]^+x_i(t),$$

$i, j=1, 2, \cdots, n$. If $x_j(t)\leqq\Gamma_j$ for large t, then $P_{ji}\not\equiv\theta_i$ in general, since (9) is violated, and an input pulse to x_j alone at times $t\gg 0$ cannot reproduce the relative weights θ_i in the outputs x_i. Since $Q_j=\theta_j$, $x_j(t)\sim\theta_j x(t)$ where $x=\sum_{i=1}^n x_i$. To guarantee (9), and in particular that $x(t)>\Gamma_j\theta_j^{-1}$ for arbitrarily large t, it suffices by (8) to let $K_1(1-e^{-\alpha\tau})^{-1}>\Gamma_j\theta_j^{-1}$. Thus for any $\theta_j>0$, the equation $P_{ji}=\theta_i$ can be guaranteed by choosing the positive intervals of the total input $I(t)$ with sufficient intensity and/or duration. The bound $K_1(1-e^{-\alpha\tau})^{-1}$ can be enlarged by iterating the equation

(10) $$\dot{x}(t) = -\alpha x(t) + \beta\sum_{k=1}^{n}[x_k(t-\tau)-\Gamma_k]^+z_k(t) + I(t)$$

and

(11) $$\dot{z}_j(t) = -uz_j(t) + \gamma[x_j(t-\tau)-\Gamma_j]^+x(t),$$

where $z_j=\sum_{i=1}^n z_{ji}$, in intervals of length τ, if K_1 is sufficiently large, and the gaps between successive values of $\Gamma_j\theta_j^{-1}$ are sufficiently small. In fact, asymptotic values of $x(t)$ can depend on the "entropy" of the weights θ_i. For example, if all $\Gamma_j=0$, then (10) and (11) asymptotically become

$$\dot{x}(t) \sim -\alpha x(t) + \beta\sum_{k=1}^{n}\theta_k^2 x(t-\tau)z(t) + I(t)$$

and

$$\dot{z}(t) \sim -uz(t) + \gamma x(t-\tau)x(t),$$

suggesting that $x(t)$ is asymptotically an increasing function of $\Omega=\sum_{k=1}^n\theta_k^2$. Ω attains its maximum if some $\theta_i=1$, and its minimum if

all $\theta_i = 1/n$. Then "asymptotically maximal energy transfer," described by the mapping $I(t) \to x(t)$ for $t \gg 0$, occurs if the input pattern is "maximally ordered."

Suppose all $x_i(T) \cong 0$ and $y_{ji}(T) \cong \theta_i$. Let a single input pulse I_k perturb a given x_k for $t \geqq T$. Will this pulse gradually destroy the memory in the y_{ji}? The answer is "yes" if all thresholds $\Gamma_i = 0$. The answer is "no" if the thresholds Γ_i are so large that the signals from x_k to each x_i create outputs x_i proportional to θ_i without exceeding Γ_i. In this sense, signal thresholds localize the system's memory, and localized inputs create context effects in the system.

The fact that learning occurs for so general a choice of coefficients in Theorem 1 means heuristically that the sensory transducers and signal generators of the system can be given by a wide choice of monotonically increasing functionals of peripheral inputs and state functions that maintain system boundedness. A particular choice of B_k or E_k merely determines how quickly a given pattern will be learned by $y^{(k)}$, and thereby determines an index of how important to $y^{(k)}$ the given pattern is. The introduction of thresholds in B_k and E_k guarantees that $y^{(k)}$ can preferentially sample or ignore prescribed input characteristics.

References

1. S. Grossberg, *A prediction theory for some nonlinear functional differential equations.* II, J. Math. Anal. Appl. **22** (1968), 490–522.

2. ———, *Embedding fields: a theory of learning with physiological implications*, J. Mathematical Psychology **6** (1969), 209–239.

3. ———, *On learning, information, lateral inhibition and transmitters*, Math. Biosci. **4** (1969), 255–310.

4. ———, *On the production and release of chemical transmitters and related topics in cellular control*, J. Theoret. Biol. **22** (1969), 327–364.

5. ———, *On the serial learning of lists*, Math. Biosci. **4** (1969), 201–253.

6. ———, *Some networks that can learn, remember, and reproduce any number of complicated space-time patterns.* II, SIAM J. Appl. Math. (to appear).

7. ———, *Some physiological and biochemical consequence of psychological postulates*, Proc. Nat. Acad. Sci. U.S.A. **60** (1968), 758–765.

8. ——— *On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks*, J. Stat. Physics **1** (1969), 319–350.

Massachusetts Institute of Technology, Cambridge, Massachusetts 02139