Reprinted from the PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES
Vol. 59, No. 2, pp. 368–372. February, 1968.

SOME NONLINEAR NETWORKS CAPABLE OF LEARNING A SPATIAL PATTERN OF ARBITRARY COMPLEXITY*

BY STEPHEN GROSSBERG

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Communicated by C. C. Lin, December 15, 1967

(1) *Introduction.*—This note describes some nonlinear networks which can learn a spatial pattern, in "black and white," of arbitrary size and complexity. These networks are a special case of a collection of learning machines $\mathfrak{M}$ which were introduced in reference 1, where a machine capable of learning a list of "letters" or "events" was described. We list in heuristic terminology some of the properties which arise in the learning of patterns:

(a) *"Practice makes perfect":* Given a "black and white" pattern of arbitrary size and complexity, a nonlinear network $\mathfrak{M}$ can be found which learns this pattern to any prescribed degree of accuracy.

(b) *An isolated machine never forgets:* If the pattern is learned to a fixed degree of accuracy by $\mathfrak{M}$, then $\mathfrak{M}$ will remember the pattern to at least this degree of accuracy until a new pattern is imposed upon $\mathfrak{M}$.

(c) *Overt practice is unnecessary:* $\mathfrak{M}$ remembers the pattern without practicing it overtly.

(d) *Contour enhancement:* If $\mathfrak{M}$ learns the pattern to a "moderate" degree of accuracy, then $\mathfrak{M}$'s memory of the pattern spontaneously improves after practices ceases. As a result, when $\mathfrak{M}$ recalls the pattern, its contours are enhanced in the sense that "darks get darker" and "lights get lighter."

(e) *A new pattern can always be learned:* Even if $\mathfrak{M}$ knows one pattern to an arbitrary degree of accuracy, this pattern can be replaced by any other pattern by a sufficient amount of practice.

(2) *The Machine.*—The nonlinear network which describes $\mathfrak{M}$ is defined as follows for any fixed number $n \geq 1$ of states and any reaction time $\tau \geq 0$.

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

$$y_{jk}(t) = z_{jk}(t)[\Sigma_{m=1}^{n} z_{jm}(t)]^{-1}, \quad (2) \qquad (*)$$

and

$$\dot{z}_{jk}(t) = -u z_{jk}(t) + \beta x_j(t-\tau) x_k(t), \quad (3)$$

where $i,j,k = 1,2,\ldots,n$. (*) describes the following process.

Let G be a graph with vertices $V = \{v_i : i = 1,2,\ldots,n\}$ and directed edges $E = \{e_{jk} : j,k = 1,2,\ldots,n\}$. Each v_i is drawn as a point and e_{jk} is drawn as an arrow facing from v_j to v_k. $x_i(t)$ describes a process going on at v_i, and $y_{jk}(t)$ describes a process going on at the *arrowhead* N_{jk} of e_{jk}. Equation (1) has the following interpretation. At time $t - \tau$, each v_m emits a signal of size $\beta x_m(t-\tau)$ into e_{mi}. This signal travels along e_{mi} at finite velocity until it reaches N_{mi} at time t. The signal thereupon activates the process $y_{mi}(t)$, and a quantity $\beta x_m(t-\tau) y_{mi}(t)$ is instantaneously transmitted from N_{mi} to v_i, and thereby changes the rate of growth $\dot{x}_i(t)$ of $x_i(t)$. Since this is true for every $m = 1,2,\ldots,n$, the total signal received by v_i from all v_m at time t is $\beta\Sigma_{m=1}{}^n x_m(t - \tau)y_{mi}(t)$. $x_i(t)$ also spontaneously decays at the rate $-\alpha x_i(t)$. $I_i(t)$ is the input signal to v_i created by the pattern.

$y_{jk}(t)$ in (2) is the ratio of functions $z_{jm}(t)$ which, as (3) shows, *cross-correlate* the signal $\beta x_j(t - \tau)$ received by N_{jm} from v_j at time t with the value $x_m(t)$ of the contiguous vertex v_m at time t.

These equations can be derived from simple psychological postulates and have a suggestive neural interpretation.[2] They are studied mathematically in reference 3, and are extended to more realistic neural equations in reference 4, which, for example, contain the Hartline-Ratliff equation[5] as a special case. The "contour enhancement" in property (*d*) above will thereupon be seen as an extension of contour enhancement as it is usually discussed in terms of lateral inhibition.

(3) *Spatial Patterns.*—For purposes of learning a spatial pattern, arrange the vertices v_i in a rectangular grid. Not all inputs $I_i(t)$ in (1) represent spatial patterns. For example, the pattern "A" does not depend on the absolute "blackness" of its lines, but only on their relative blackness as compared to the surround. A *pattern* is therefore defined as an input $I_i(t)$ of the form

$$I_i(t) = \theta_i I(t), \quad i = 1,2,\ldots,n, \tag{4}$$

where the θ_i's are arbitrary, but fixed, nonnegative numbers whose sum can be taken equal to 1 without loss of generality. The pattern "A" is the same whether or not we view it in steady light or flickering light. $I(t)$ can therefore oscillate quite wildly without changing the pattern described by the θ_i's. In fact the following theorem holds, which describes the way in which the probabilities $y_{jk}(t) = z_{jk}(t)[\Sigma_{m=1}{}^n z_{jm}(t)]^{-1}$ and the correspondingly defined probabilities $X_k(t) = x_k(t)[\Sigma_{m=1}{}^n x_m(t)]^{-1}$ learn an arbitrary pattern. Other facts and generalizations concerning this learning process are contained in reference 3.

THEOREM 1. *Suppose $u > 2(\alpha - \beta) > 0$ and $\beta > 0$. Let n be any fixed number of states and let τ be any fixed nonnegative reaction time. Let $I_i(t) = \theta_i I(t)$ be any pattern with $I(t)$ nonnegative, continuous, and bounded, and such that positive constants k and T_0 exist for which*

$$\int_0^t e^{\alpha v} I(v)dv \geq ke^{\alpha t}, \quad t \geq T_0. \tag{5}$$

Then for arbitrary nonnegative and continuous initial data in (*), the limits $Q_i = \lim_{t\to\infty} X_i(t)$ and $P_{jk} = \lim_{t\to\infty} y_{jk}(t)$ exist, and obey the equations

$$P_{ji} = Q_i = \theta_i, \quad i,j = 1,2,\ldots,n. \tag{6}$$

Equation (6) says that the probability $X_i(t)$ of v_i and the correlations $y_{ji}(t)$ of all N_{ji} touching v_i learn the relative weight θ_i of the pattern, just so long as the absolute intensity $I(t)$ of the pattern is not "too small" in the sense of (5). $I(t)$ can in fact oscillate very wildly without violating (5). A pattern can therefore be learned to arbitrary accuracy if only it is presented sufficiently often. In order to learn ever more subtle gradations of shading in the pattern, it suffices to take the number n of vertices in the rectangular grid ever larger.

Equation (5) requires that $I(t)$ take on positive values at arbitrarily large values of t. We now describe what happens if a "truncated" pattern $I_i^{(w)}(t) = \theta_i I^{(w)}(t)$ is presented, where $I^{(w)}(t) = I(t)$, $0 \leq t < w$, and $I^{(w)}(t) = 0$, $t \geq w$. That is, $\mathfrak{M}$ is exposed to the pattern only in the time interval $[0,w)$.

THEOREM 2. *Suppose $u > 2(\alpha - \beta) > 0$ and $\beta > 0$. Let $n \geq 2$ (to avoid trivialities) and $\tau \geq 0$. Let $I_i(t) = 0, t \geq w$, for all $i = 1,2,\ldots,n$. Then for arbitrary nonnegative and continuous data in $[w - \tau, w]$, the limits Q_i and P_{ji} exist and lie in the interval $[m_i(w), M_i(w)]$, where*

$$m_i(w) = \min\{X_i(w), y_{ki}(w):\quad k = 1,2,\ldots,n\}$$

$$M_i(w) = \max\{X_i(w), y_{ki}(w):\quad k = 1,2,\ldots n\}.$$

Denoting the functions of (*) which are exposed to $I_i^{(w)}(t)$ by superscripts "(w)" (for example, $X_i(t)$ becomes $X_i^{(w)}(t)$), we find the following corollary.

COROLLARY 1.

$$\lim_{w\to\infty}\lim_{t\to\infty} X_i^{(w)}(t) = \lim_{w\to\infty}\lim_{t\to\infty} y_{ji}^{(w)}(t) = \theta_i, \quad i,j = 1,2,\ldots,n. \tag{7}$$

Proof: By Theorem 1, $\lim_{w\to\infty} m_i(w) = \lim_{t\to\infty} M_i(w) = \theta_i$.

These theorems say that if the pattern is exposed to $\mathfrak{M}$ during $[0,w)$ and if w is taken sufficiently large, then $\mathfrak{M}$ will learn the pattern to an arbitrary degree of accuracy and will *remember* the pattern to at least this degree of accuracy thereafter. $\mathfrak{M}$ does this without "practicing overtly" because the outputs $x_i(t)$ from $\mathfrak{M}$ decay exponentially to 0 for $t \geq w$ whenever $\alpha > \beta > 0$.

Contour enhancement occurs in $\mathfrak{M}$ because of the following corollary, which describes the "envelope"

$$Y_i(t) = \max\{y_{ki}(t):\quad k = 1,2,\ldots,n\}$$

$$y_i(t) = \min\{y_{ki}(t):\quad k = 1,2,\ldots,n\}$$

of correlations whose arrowheads N_{ki} touch v_i.

COROLLARY 2. *For w sufficiently large, one of the following alternatives holds for each $i = 1,2,\ldots,n$:*

(a) *$Y_i^{(w)}(t) \geq X_i^{(w)}(t) \geq \theta_i$, $y_i^{(w)}(t) \geq \theta_i$, and $Y_i^{(w)}(t)$ is monotone decreasing for $t \geq w$; or*

(b) *$\theta_i \geq X_i^{(w)}(t) \geq y_i^{(w)}(t)$, $\theta_i \geq Y_i^{(w)}(t)$, and $y_i^{(w)}(t)$ is monotone increasing for $t \geq w$; or*

(c) *$Y_i^{(w)}(t) \geq \theta_i \geq y_i^{(w)}(t)$, $Y_i^{(w)}(t) \geq X_i^{(w)}(t) \geq y_i^{(w)}(t)$, $Y_i^{(w)}(t)$ is monotone decreasing, and $y_i^{(w)}(t)$ is monotone increasing for $t \geq w$.*

In other words, after a sufficient amount of exposure to the pattern, the envelope of correlations "spontaneously" approaches the pattern probabilities θ_i.

Suppose, for example, that $\theta_i = 0$, which designates a "black" portion of the pattern at state v_i. Then case (a) holds, in which $Y_i^{(w)}(t)$ decreases towards zero. That is, "darks get darker."

To see how $\mathfrak{M}$ recalls a pattern, suppose that a pattern has been practiced over a long time interval $[0,w)$ and that the outputs $x_i(t)$ have decayed nearly to zero in the subsequent interval $[w,W]$. We now show that if even a single speck of light is thereupon shined on the machine at a given vertex (say v_1), then τ time units later the pattern will reappear in all its glory at all the vertices v_i if the reaction time τ is sufficiently large. Since all $x_i(W) \cong 0$, we find by (1) that

$$\dot{x}_1(t) \cong -\alpha x_1(t) + I(t), \quad t \in [W, W+\tau],$$

where $I(t)$ represents the speck of light shined on v_1. Thus a signal is emitted from v_1 to all points v_i. By Theorem 2, $y_{1i}(t) \cong \theta_i$ for $t \in [W, W+\tau]$, and thus

$$\dot{x}_i(t) \cong -\alpha x_i(t) + \beta x_1(t-\tau)\theta_i. \tag{8}$$

Suppose τ is so large that $x_1(t)$ has a chance to decay back toward zero before it receives the signal which it has created in e_{11}. Then by (8),

$$x_i(t) \cong \beta\theta_i e^{-\alpha t}\int_{W+\tau}^{t} e^{\alpha v} x_1(v-\tau)dv,$$

for *all* $i = 1,2,$

interval $[T - \tau, T]$ can be viewed as the initial data for (*) in the interval (T, ∞). Since these values are nonnegative and continuous, and Theorems 1 and 2 hold for all nonnegative and continuous initial data, our contention is proved.

* The preparation of this work was supported in part by the National Science Foundation (NSF GP-7477).

[1] Grossberg, S., "Nonlinear difference-differential equations in prediction and learning theory," these PROCEEDINGS, 58, 1329 (1967).

[2] Grossberg, S., "Embedding fields: A new theory of learning with physiological implications," *J. Math. Psych.*, to appear.

[3] Grossberg, S., "A prediction theory for some nonlinear functional-differential equations, II. Learning of patterns," *J. Math. Anal. Appl.*, to appear.

[4] Grossberg, S., "On learning, information, lateral inhibition, and transmitters."

[5] Ratliff, F., *Mach Bands: Quantitative Studies on Neural Networks in the Retina* (New York: Holden-Day, 1965).