DISTRIBUTED OUTSTAR LEARNING AND THE RULES OF SYNAPTIC TRANSMISSION

Gail A. Carpenter

Center for Adaptive Systems and Department of Cognitive and Neural Systems
Boston University, 111 Cummington Street, Boston, Massachusetts 02215 USA

Abstract

The distributed outstar, a generalization of the outstar neural network for spatial pattern learning, is described. In the outstar, signals from a source node cause weights to learn and recall arbitrary patterns across a target field of nodes. The distributed outstar replaces the source node with a source field whose activity pattern may be arbitrarily distributed. Learning proceeds according to a principle of atrophy due to disuse, whereby a path weight decreases in joint proportion to the transmitted path signal and the degree of disuse of the target node. During learning, the total signal to a node converges toward that node's activity level. Weight changes are apportioned according to the distributed pattern of converging signals. Three synaptic transmission functions, a product rule, a capacity rule, and a threshold rule, are examined for this system. The three rules are computationally equivalent when source field activity is winner-take-all. When source field activity is distributed, catastrophic forgetting may occur. Only the threshold rule solves this problem. Analysis of spatial pattern learning by distributed codes thereby leads to the conjecture that the unit of long-term memory in such a system is an adaptive threshold, rather than the multiplicative path weight widely used in neural models.

Introduction: Outstar learning and distributed codes

The *outstar* is a neural network that can learn and recall arbitrary spatial patterns (Grossberg, 1968). Outstars have played a central role in the theoretical analysis of cognitive phenomena and the corresponding neural models, as well as in applications of these systems (Carpenter and Grossberg, 1991). In particular, all neural network realizations of adaptive resonance theory (ART models) have so far used outstar learning in the top-down adaptive filter (Carpenter and Grossberg, 1987a, 1987b, 1990; Carpenter, Grossberg, and Rosen, 1991). An outstar anatomy consists of a *source node* that sends weighted inputs to a target, or border, field of nodes. We will here consider spatial pattern learning in a more general setting, in which a *distributed outstar network* (Carpenter, 1993) replaces the single outstar source node with an arbitrarily large *source field* (Figure 1).

One possible distributed outstar design is simply to implement outstar learning in each active path. However, such a system is subject to catastrophic forgetting that can quickly render the network useless, unless learning rates are very slow. In particular, if all F_2 nodes

This research was supported in part by British Petroleum (BP 89-A-1204), DARPA (ONR N00014-92-J-4015), the National Science Foundation (IRI-90-00530), and the Office of Naval Research (ONR N00014-91-J-4100).

The author wishes to thank Diana J. Meyers for her valuable assistance in the preparation of the manuscript.

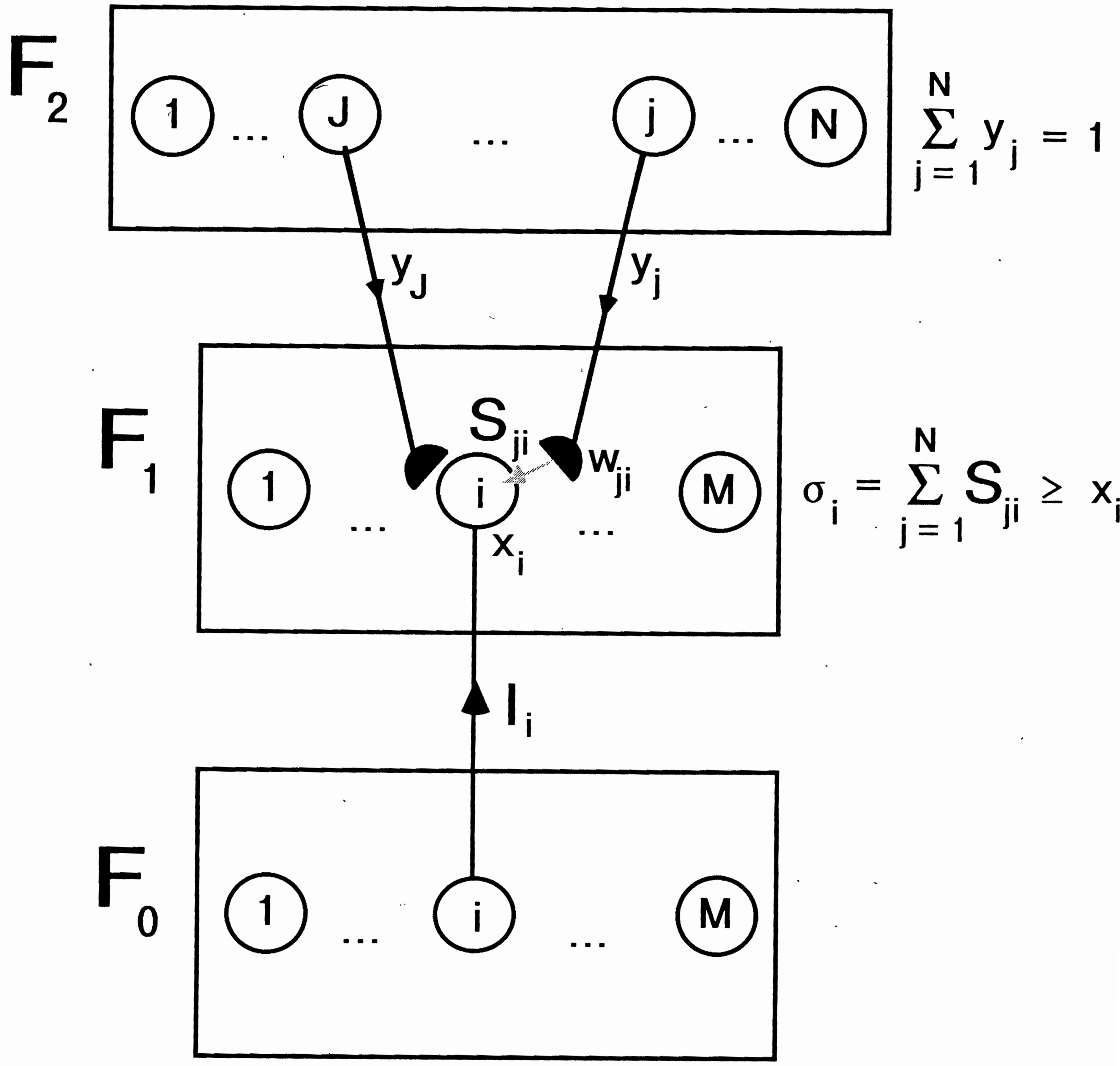

Figure 1. Distributed outstar network for spatial pattern learning. During adaptation a top-down weight w_{ji}, from the j^{th} node of the coding field F_2 to the i^{th} node of the pattern registration field F_1, may decrease or remain constant. An atrophy-due-to-disuse learning law causes the total signal σ_i from F_2 to the i^{th} F_1 node to decay toward that node's activity level x_i, if σ_i is initially greater than x_i. Within this context, three synaptic transmission rules are analyzed.

were active during learning, all $F_2 \rightarrow F_1$ weight vectors would converge toward a common pattern.

A learning principle of *atrophy due to disuse* leads toward a solution of the catastrophic forgetting problem. By this principle, a weight in an active path is assumed to atrophy, or decay, in joint proportion to the size of the transmitted synaptic signal and a suitably defined "degree of disuse" of the target cell. During learning, the total transmitted signal from F_2 converges toward the activity level of the target F_1 node. Unfortunately, this development is, by itself, insufficient. In particular, the network still suffers catastrophic forgetting if signal transmission obeys a *product rule*. This rule, now assumed in nearly all neural models, takes the transmitted synaptic signal from the j^{th} F_2 node to the i^{th} F_1 node to be proportional to the product of the path signal y_j and the path weight w_{ji}. An alternative transmission process is described by a *capacity rule*. However, catastrophic forgetting is even more serious a problem for this rule than for the product rule.

Fortunately, another plausible synaptic transmission rule solves the problem. This *threshold rule* postulates a transmitted signal equal to the amount by which the $F_2 \rightarrow F_1$ signal y_j exceeds an adaptive threshold τ_{ji}. Where weights decrease during atrophy-due-to-disuse learning thresholds increase: formally, τ_{ji} is identified with $(1 - w_{ji})$. When synaptic transmission is implemented by a threshold rule, weight/threshold changes are automatically distributed, with fast learning as well as slow learning. When F_2 makes a choice, the three synaptic transmission rules are computationally identical, and atrophy-due-to-disuse learning is essentially the same as outstar learning. Thus functional differences between the three types of transmission would be experimentally and computationally measurable only in situations where the F_2 code is distributed.

Computational analysis of distributed codes hereby leads unexpectedly to a hypothesis about the mechanism of synaptic transmission in spatial pattern learning systems. That is, the unit of long-term memory in these systems is conjectured to be an adaptive threshold, rather than a multiplicative path weight. The hypothesis is embodied in the *distributed outstar learning law*.

Spatial pattern learning and catastrophic forgetting

The distributed outstar network (Figure 1) features an adaptive filter from a *coding field* F_2 to a *pattern registration field* F_1. During outstar learning, weights in the paths emanating from an F_2 node track F_1 activity. That is, when the j^{th} F_2 node is active, the weight vector $\mathbf{w}_j \equiv (w_{j1}, \dots w_{ji}, \dots w_{jM})$ converges toward the F_1 activity vector $\mathbf{x} \equiv (x_1, \dots x_i, \dots x_M)$ of the target nodes at the outer fringe of the filter. While many variants of outstar learning have been analyzed (Grossberg, 1968, 1972), the essential outstar dynamics are described by the equation:

Basic outstar –

$$\frac{d}{dt} w_{ji} = y_j (x_i - w_{ji}). \tag{1}$$

A special F_2 network called choice, or winner-take-all, is commonly used in ART and competitive learning systems. An F_2 code that chooses the J^{th} node is described by:

$\mathbf{F_2}$ choice –

$$y_j = \begin{cases} 1 & \text{if } j = J \\ 0 & \text{if } j \neq J. \end{cases} \tag{2}$$

In this case, each F_2 node may then be identified with a class, or category, of inputs $\mathbf{I}$. When F_2 makes a choice, outstar learning (1) permits a weight w_{Ji} to change only if the J^{th} F_2 node is active. All other weights to the i^{th} F_1 node remain unchanged when the J^{th} category is selected, so prior learning is preserved. Outstar learning poses a problem, however, when F_2 category representations can be distributed. If a code $\mathbf{y}$ were highly distributed, with all $y_j > 0$, then the outstar learning law (1) would imply that all weight vectors $\mathbf{w}_j$ would converge toward the same F_1 activity vector $\mathbf{x}$. The size of y_j would affect the rate of convergence, but not the asymptotic state of the weights. The severity of this problem can be reduced if learning intervals are required to be extremely short. If, however, the y_j values are nearly uniform or if learning is not always slow, catastrophic forgetting will occur.

A new adaptation rule, called the distributed outstar learning law, solves this problem. Even with fast learning, where weights approach asymptote on each input presentation, the distributed outstar apportions weight changes across active paths without catastrophic forgetting. In the distributed outstar, the rate constant for an individual weight w_{ji} becomes an increasing function of y_j, as in the outstar (1), and also of w_{ji} itself.

When w_{ji} becomes too small, further change is disallowed. Small weights can decrease further only when y_j is close to 1, which occurs when most of the F_2 activity is concentrated at node j. When F_2 activity is highly distributed, only large weights, close to their initial values, are able to change, and the maximum possible weight change in any single path is small. The distributed outstar combines learning by atrophy due to disuse with the adaptive threshold synaptic transmission rule, as follows. Detailed computations and examples are described elsewhere (Carpenter, 1993).

Learning by atrophy due to disuse

The principle of atrophy due to disuse postulates that the strength of an active path will decay when the path is disused. Active "dis-use" is distinct from passive "non-use", where the strength of an inactive path remains constant, as in the outstar (1). To define disuse, a specific class of target fields F_1 is considered. The main hypothesis on F_1 will be that, when F_2 is active, the total top-down input from F_2 to F_1 imposes an upper bound, or limit, on the maximum activity at an F_1 node. In particular, in addition to a bottom-up input I_i, a top-down *priming* input from F_2 is assumed to be necessary for an F_1 node to remain active, once F_2 becomes active. This hypothesis is realized by:

Top–down prime –

$$0 \leq x_i \leq \sigma_i, \tag{3}$$

where σ_i is the sum of all transmitted signals S_{ji} from F_2 to the i^{th} F_1 node:

$$\sigma_i \equiv \sum_{j=1}^{N} S_{ji}. \tag{4}$$

The top-down prime inequality (3) is closely related to the 2/3 Rule of ART (Carpenter and Grossberg, 1987a). One class of F_1 systems that realize σ_i as a top-down prime, or upper bound, on target node activity x_i sets:

$$x_i = I_i \wedge \sigma_i \equiv \min\,(I_i, \sigma_i), \tag{5}$$

where $I_i \in [0,1]$.

When F_2 primes F_1, by (3), the *degree of disuse* D_i of the i^{th} F_1 node is defined to be:

$$D_i = (\sigma_i - x_i) \geq 0. \tag{6}$$

A learning principle of atrophy due to disuse postulates that a path weight decays in proportion to the degree of disuse of its target node. We here consider a class of learning equations that realize this principle in the form:

$$\frac{d}{dt} w_{ji} = -S_{ji} D_i. \tag{7}$$

Weights can then decay or stay constant, but never grow, when $S_{ji} \geq 0$ and $D_i \geq 0$. With the degree of disuse D_i defined by (6), the learning law (7) becomes:

Atrophy due to disuse –

$$\frac{d}{dt} w_{ji} = -S_{ji}(\sigma_i - x_i). \tag{8}$$

Initially,

$$w_{ji}(0) = 1 \tag{9}$$

for $i = 1, \ldots, M$ and $j = 1, \ldots, N$. When F_2 makes a choice (2), the atrophy-due-to-disuse law (8) reduces to:

$$\frac{d}{dt} w_{ji} = \begin{cases} -w_{Ji}(w_{Ji} - x_i) & \text{if } j = J \\ 0 & \text{if } j \neq J \end{cases} \tag{10}$$

for all three synaptic transmission rules defined below. With fast learning, the dynamics of (10) are equivalent to those of the outstar (1).

Synaptic transmission functions

We will here consider three rules for synaptic transmission. The F_2 path signal vector $\mathrm{y} = (y_1, \ldots y_j, \ldots y_N)$ is assumed to be normalized:

$$\sum_{j=1}^{N} y_j = 1, \tag{11}$$

but can otherwise be arbitrary.

The first rule postulates that the $F_2 \rightarrow F_1$ transmitted signal is jointly proportional to the path signal y_j and the weight w_{ji}:

Product rule –

$$S_{ji} = y_j w_{ji}. \tag{12}$$

A different synaptic transmission rule assumes that the path signal y_j is itself transmitted directly to the i^{th} F_1 node, until an upper bound on the path's capacity is reached. With this upper bound equal to the path weight w_{ji}, the net signal obeys the:

Capacity rule –

$$S_{ji} = y_j \wedge w_{ji} \equiv \min\,(y_j, w_{ji}). \tag{13}$$

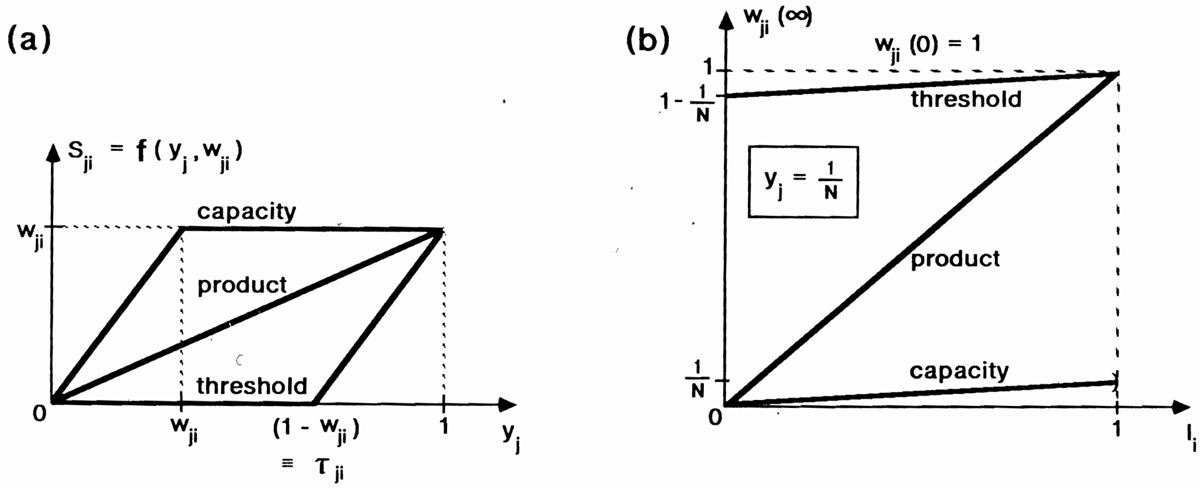

Figure 2. (a) A synaptic transmission parallelogram. S_{ji} is the transmitted signal from the j^{th} F_2 node to the i^{th} F_1 node. By the product rule, $S_{ji} = y_j w_{ji}$. By the capacity rule, $S_{ji} = y_j \wedge w_{ji}$. By the threshold rule, $S_{ji} = [y_j - (1 - w_{ji})]^+ = [y_j - \tau_{ji}]^+$. The three rules agree when **y** is a binary code. (b) Asymptotic weight values for a fully distributed code, where $y_j = \frac{1}{N}$. As a function of I_i, the dynamic range of $w_{ji}(\infty)$ depends critically upon the choice of synaptic transmission rule. During learning, weights decrease, from an initial value of $w_{ji}(0) = 1$, except when $I_i = 1$.

The geometry of the graph in Figure 2a suggests consideration of a third signal function, to complete a transmission rule parallelogram. The third signal describes a:

Threshold rule –

$$S_{ji} = [y_j - (1 - w_{ji})]^+. \tag{14}$$

It is awkward to try to interpret (14) in terms of the weight w_{ji}. However, a natural interpretation can be made if the unit of long-term memory is taken to be an adaptive signal threshold τ_{ji} rather than the path weight w_{ji}. Namely, by setting:

$$\tau_{ji} \equiv 1 - w_{ji}, \tag{15}$$

the threshold rule (14) becomes:

$$S_{ji} = [y_j - \tau_{ji}]^+. \tag{16}$$

Path weights vs. signal thresholds as the unit of long-term memory

An F_2 code is maximally compressed when the system makes a choice. Consider now the opposite extreme, when an F_2 code is maximally distributed. That is, let:

$$y_j = \frac{1}{N}. \tag{17}$$

for $j = 1, \ldots, N$. All weights $w_{1i}, \ldots, w_{Ni}$ obey equation (8) and all are initially equal, by (9). Therefore the weights w_{ji} ($j = 1, \ldots, N$) to a given F_1 node will remain equal to one another during learning, for any transmission function S_{ji}. For all three synaptic transmission rules the total top-down signal σ_i converges to the bottom up-signal I_i at each F_1 node i when the F_2 code (17) is maximally distributed. However, the total weight change varies dramatically (Figure 2b). When F_2 makes a choice the maximum total weight change at a given node equals $(1 - I_i) \in [0,1]$ for all three rules. With distributed F_2 activity and a product rule, all weights w_{ji} converge to I_i and the maximum total weight change is $N(1 - I_i) \in [0, N]$. Within a few input presentations, all weights w_{ji} would, in all likelihood, decay irreversibly to zero. Similar problems occur for other distributed codes $\mathbf{y}$. In this sense, the product rule leads to catastrophic forgetting.

The situation with the capacity rule is even worse (Figure 2b). When the F_2 code is fully distributed, all weights w_{ji} decay to $\frac{I_i}{N} \in [0, \frac{1}{N}]$, unless $I_i = 1$; and the maximum total weight change at the i^{th} node is $N(1 - I_i)$. Thus, unless $\mathbf{I}$ is a binary vector, the full dynamic range of weight values is nearly exhausted upon the first input presentation.

It is the adaptive threshold rule alone that limits the total weight change to $(1 - I_i) \in [0,1]$ for maximally distributed as well as maximally compressed codes $\mathbf{y}$. In fact, if $\mathbf{y}$ is *any* F_2 code that becomes active when all w_{ji} are initially equal to 1, then:

$$w_{ji} \to 1 - y_j(1 - I_i). \tag{18}$$

Equivalently:

$$\tau_{ji} \to y_j(1 - I_i), \tag{19}$$

by (15). Thus the total weight/threshold change at each F_1 node i is bounded by $(1 - I_i)$ for any code, provided only that $\mathbf{y}$ is normalized (11). An F_2 code $\mathbf{y}$ would typically be highly distributed, with all y_j close to $\frac{1}{N}$, when a recognition system has no strong evidence to choose one category j over another. In this case, the change of each threshold τ_{ji} is automatically limited to the narrow interval $[0, y_j]$, reserving most of the dynamic range for subsequent encoding. Only when evidence strongly supports selection of the F_2 category node J over all others, with y_J therefore close to 1, would weights be allowed to vary across most of their dynamic range. In particular, it is only when y_J is close to 1 that a weight w_{Ji} is able to drop, irreversibly, toward 0, if I_i is small. Even with fast learning and with all $y_j > 0$, other weights w_{ji} to the i^{th} node would change little.

Conclusion: Distributed outstar learning

The analysis of distributed spatial pattern learning leads to the selection of a synaptic transmission rule with an adaptive threshold. In terms of the threshold τ_{ji} in the path from the j^{th} F_2 node to the i^{th} F_1 node, a stable learning law for distributed codes is defined as the:

Distributed outstar -

$$\frac{d\tau_{ji}}{dt} = S_{ji}(\sigma_i - x_i), \tag{20}$$

where S_{ji} is the thresholded path signal $[y_j - \tau_{ji}]^+$ transmitted from the j^{th} F_2 node to the i^{th} F_1 node and where σ_i is the sum:

$$\sigma_i \equiv \sum_{j=1}^{N} S_{ji} = \sum_{j=1}^{N} [y_j - \tau_{ji}]^+. \tag{21}$$

Initially,

$$\tau_{ji}(0) = 0. \tag{22}$$

In a system such as ART 1 (Carpenter and Grossberg, 1987a) or fuzzy ART (Carpenter, Grossberg, and Rosen, 1991), where F_1 dynamics are defined so that the total top-down signal σ_i is always greater than or equal to x_i, the distributed outstar allows thresholds τ_{ji} to grow but never shrink. The principle of atrophy due to disuse implies that a threshold τ_{ji} is unable to change at all unless (i) the path signal y_j exceeds the previously learned value of τ_{ji}; and (ii) the total top-down signal σ_i to the i^{th} node exceeds that node's activity x_i. In particular, if τ_{ji} grows large when the node j represents part of a compressed F_2 code, then τ_{ji} cannot be changed at all when node j is later part of a more distributed code, since threshold changes are disabled if $y_j \leq \tau_{ji}$. The adaptive threshold τ_{ji} thereby replaces strong F_2 competition as the guardian, or stabilizer, of previously learned codes.

References

Carpenter, G.A. (1993). A distributed outstar network for spatial pattern learning. Technical Report CAS/CNS TR-93-036, Boston, MA: Boston University. Submitted for publication.

Carpenter, G.A. and Grossberg, S. (1987a). A massively parallel architecture for a self-organizing neural pattern re cognition machine. *Computer Vision, Graphics, and Image Processing*, **37**, 54-115.

Carpenter, G.A. and Grossberg, S. (1987b). ART 2: Stable self-organization of pattern recognition codes for analog input patterns. *Applied Optics*, **26**, 4919–4930.

Carpenter, G.A. and Grossberg, S. (1990). ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural Networks*, **3**, 129-152.

Carpenter, G.A. and Grossberg, S. (Eds.) (1991). **Pattern Recognition by Self-Organizing Neural Networks**. Cambridge, MA: MIT Press.

Carpenter, G.A., Grossberg, S., and Rosen, D.B. (1991). Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, **4**, 759-771.

Grossberg, S. (1968). Some nonlinear networks capable of learning a spatial pattern of arbitrary complexity. *Proceedings of the National Academy of Sciences*, **59**, 368–372.

Grossberg, S. (1972). Pattern learning by functional-differential neural networks with arbitrary path weights. In K. Schmitt (Ed.), **Delay and Functional-Differential Equations and Their Applications**. New York: Academic Press, pp. 121-160.