Chapter 15

Prelude to ART

§ 1. Adaptive Resonance Theory

By most authors' accounts, the birth of adaptive resonance theory (ART) is recognized as being in 1976 with the appearance of [GROS6]. In an important sense this is true, but it diminishes the fact that ART developed over a period of years dating back into the late 1960s. Indeed, many of the key ideas used in [GROS6] will not make sense to the novice unless he has already become familiar with them from Grossberg's earlier publications. By 1976 they had become part of what is often called "the standard argument" used in the first sections of a technical journal paper. The purpose of this chapter is to present the key ideas and findings that are essential for the actual discussion of adaptive resonance theory in chapter 16.

ART networks' undeserved reputation for being very complicated is due to an unfortunate historical accident. The foundational ideas that would lead to ART were discovered and published in the "dark age" of neural network research dating from 1968 until well into the 1980s. In some ways the history of ART can be compared to the Carolingian Renaissance that began with Charlemagne and flourished briefly in the ninth century before vanishing in the tumult of the tenth. Fortunately for ART, Grossberg – unlike Charlemagne – was still alive and active when the darkness lifted. (Had Charlemagne's successors been competent men, the dark ages might have ended 300 years sooner). ART's foundations never did disappear but, like the post-Carolingians, there are many younger theorists who came into the neural network field in the 1980s and 1990s, and who are simply too young to know about the propaedeutic work of the late 1960s and early 1970s [GROS2], [GROS3], [GROS11-18], [ELIA], [GROS4-6].

ART can looked at as the fusion of two major themes: recurrent on-center/off-surround networks and Outstars. There is, of course, more to an ART network than just these two elements, but they are the central elements and everything else exists to support their function. The on-center/off-surround structure is found in abundance in the central nervous system. Its basic form consists of a population of neurons that is tightly coupled and self-excitatory (the on-center) surrounded by other populations with which it has lateral inhibitory connections (the off-surround). Figure 15.1 illustrates the basic on-center/off-surround schema. The designation of a population as on-center or as off-surround is relative. Every population is an on-center to itself and its neighbors are its off-surround. One population is designated as on-center in figure 15.1 for

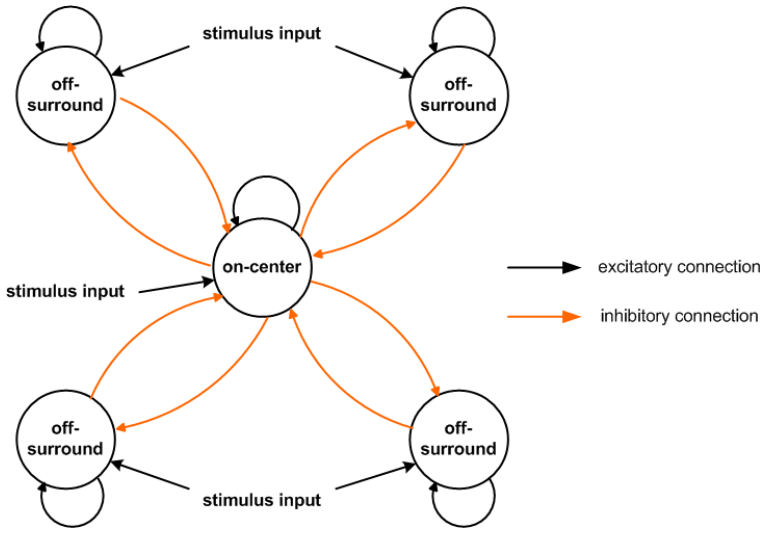


Figure 15.1: Basic on-center/off-surround network anatomy

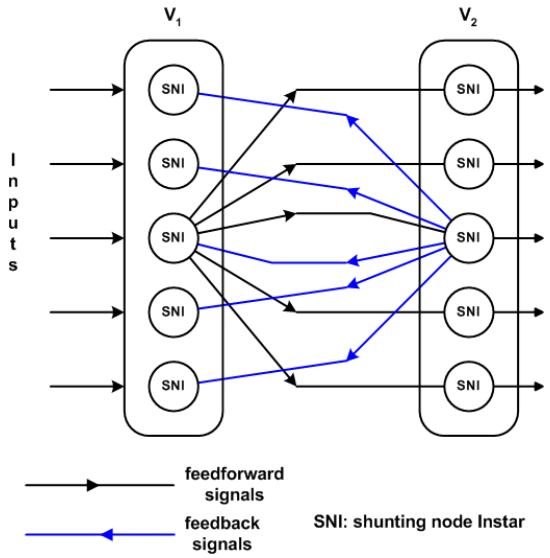


Figure 15.2: Minimal ART anatomy. Feedback projections are made via Outstars. Feedforward projections project into each V_2 node using an Instar anatomy.

purposes of discussion. Lateral inhibitory paths are only shown for the population designated as the on-center population. It is to be understood that the off-center populations all have this same connectivity when they are regarded as an on-center. When the input stimulus to the on-center population is greater than that to the off-surround populations, the on-center node tends to suppress the activities of the off-surround nodes. If the off-surround stimuli are greater, the on-center activity tends to be suppressed. The simplest example of this is seen in the behavior of the MAXNET, which is an on-center/off-surround anatomy.

The basic minimal ART anatomy is shown in figure 15.2. It consists of two on-center/off-surround layers, V_1 and V_2. Each node in V_1 projects to each node in V_2, and the fan-in to each V_2

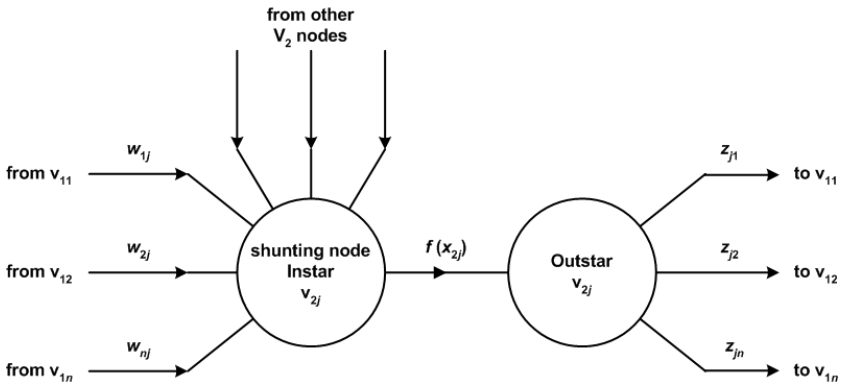


Figure 15.3: Detailed diagram of each v_2 node's input/output anatomy.

node uses an Instar anatomy. Each node in V_2 projects back to each node in V_1, and the feedback from each V_2 node is made by means of an Outstar. Figure 15.3 illustrates the input/output anatomy of a V_2-layer node. In general a V_1-layer node, v_{1i}, is the same except that V_1 nodes do not have an Outstar output. Input weights W_{2j} and output weights Z_{2j} are adaptable.

Each node in figure 15.2 represents a population of B neurons and has a level of excitation x, with $0 \leq x \leq B$, representing how many neurons in the population are active. The quantity $B - x$ therefore represents how many neurons in the population are inactive. In the most general case, each node v_i can represent a different population size B_i, although the most commonly encountered ART networks typically use the same population size B for all nodes. This network anatomy is termed a ***lumped network*** [GROS14]. The meaning of this term is as follows. Each population in each node is regarded as being made up of both excitatory and inhibitory neural subnetworks. If the excitatory and the inhibitory subpopulations have the same parameters and receive the same inputs, then the two subpopulations are indistinguishable with respect to every input source and with respect to their temporal dynamics. In more advanced ART networks it is possible to divide each node into an excitatory population and an inhibitory population, and to give each population differing sets of parameters. Such a network is said to be ***unlumped*** [ELIA], [GROS15].

All parameters and variables in an ART system are non-negative. Although many of the details of the network are very similar to what we have seen in the earlier chapters of this text, ART systems employ two quite different features, and these make all the difference. The first is that ART does not use the classical Instar map model. Rather, it uses a special kind of Instar map developed by Grossberg. We call it the ***shunting node Instar*** or SNI. Second, the activation functions *f(u)* employed in an ART system are different from all the more classical activation functions we have seen so far. In one of his early works, Grossberg studied how the properties of the activation function affect the behavior of a network constructed from SNIs [GROS14]. He

found that the details of the activation function are crucial to how the network performs – much more crucial than is typically the case with the simpler network systems we have already studied. We will explore both these unique aspects of ART in turn.

§ 2. Shunting Node Instars

When used as a general term, the name 'shunting node Instar' refers to a general class of map models within which various special cases are distinguished. This usage is similar to our use of the generic term 'Instar' for a specific general form of map model with names to identify distinct species of the class such as Adaline, perceptron, and RBF Instar. In this section the defining equation for a general SNI will be presented along with some special notation we will find useful. In practice, most ART systems restrict the parameters of this equation in specific ways, and most published ART systems are less general than the equation we present here. We will give different special case SNI maps a designation of the form $\text{SNI}^{(k)}$ and allow the superscript k to be the identifier of which specific case we are talking about.

We will find it useful to define special symbols to designate particular classes of input signals received by an SNI. There are four classes of inputs:

- ξ^e = the sum of all excitatory inputs from specific SNI nodes;
- ξ^i = the sum of all inhibitory inputs from specific SNI nodes;
- ζ = a non-specific and uniformly distributed input applied to all nodes in a layer;
- K = an excitatory input regarded as originating from an external (non-SNI) stimulus.

We let x represent the SNI's excitation variable and define the following general parameters:

- $A > 0$; a general ***relaxation parameter***;
- $B > 0$; the node's ***population*** parameter such that $0 \le x \le B$;
- $C \ge 0$; a node parameter we will call the ***spatial contrast parameter***.

Using these definitions, the general form of the SNI dynamical equation is

$$\dot{x} \overset{def}{=} \frac{dx}{dt} = -A \cdot x + (B - x) \cdot \xi^e - (C + x) \cdot \xi^i + \zeta + K \,. \tag{15.1}$$

Grossberg developed adaptive resonance theory entirely within the framework of differential equations. Thus, t in (15.1) is a continuous time parameter. For computer simulations we will need a discrete-time form of (15.1). We convert (15.1) to difference equation form using Euler's method and obtain

$$x(t + \Delta t) = \left(1 - \Delta t \cdot \left(A + \xi^e(t) + \xi^i(t)\right)\right) \cdot x(t) + (\Delta t \cdot B) \cdot \xi^e(t) - (\Delta t \cdot C) \cdot \xi^i(t) + \Delta t \cdot (\zeta(t) + K(t)) \,. \tag{15.2}$$

Generally speaking, in ART systems the variables ξ^e and ξ^i will be non-negative quantities. A condition sufficient on the time step Δt to ensure the stability of (15.2) and its proper performance within an ART system is given by

$$0 \le 1 - \Delta t \cdot \left(A + \xi^e_{\max} + \xi^i_{\max} \right) < 1 \Rightarrow \Delta t \le \frac{1}{A + \xi^e_{\max} + \xi^i_{\max}} \ . \tag{15.3}$$

The activation variables that give rise to ξ^e and ξ^i in an ART system are usually bounded within the range from 0 to 1 by sigmoid activation functions. If there are N_e excitatory inputs and N_i inhibitory inputs producing ξ^e and ξ^i at the SNI node, and assuming the adaptation dynamics of the system can ensure all the synaptic weights of the SNI similarly remain bounded in the 0 to 1 range, a sufficient condition for the time step is

$$\Delta t = \frac{\alpha}{A + N_e + N_i}, \quad 0 < \alpha \le 1 \, . \tag{15.4}$$

When the system has multiple SNI nodes, as will generally be the case, (15.4) is determined by the SNI node for which the sum in the denominator of (15.4) is the largest.

When the external and the non-specific inputs, K and ζ, are held constant, and if the overall system is fixed-point stable such that ξ^e and ξ^i achieve constant final values, both (15.1) and (15.2) yield the same steady-state value for x,

$$x(\infty) = \frac{B \cdot \xi^e(\infty) - C \cdot \xi^i(\infty) + \zeta + K}{A + \xi^e(\infty) + \xi^i(\infty)} \ . \tag{15.5}$$

Note that this result is independent of Δt, as it should be for a proper difference equation representation of (15.1). As we will see, asymptotic behavior is important in adaptive resonance theory.

It is appropriate also at this time to point out that the presence of the $-C$ term in (15.5) makes it mathematically possible for $x(\infty)$ to be negative. In view of Grossberg's interpretation of what the excitation variable x represents – namely, that it represents how many neurons in a population are active – this mathematical possibility is troubling from the viewpoint of a physical interpretation. In point of fact, some ART network subsystems *do* produce negative values for x. These negative values do not result in negative activations, $f(x)$, because of the activation functions f used in ART systems. (Typically $f(x < 0) = 0$ in ART). Grossberg originally introduced the C term into (15.1) by making an analogy between it and the Nernst potential for potassium [GROS6], but this clearly has little plausible application in interpreting the SNI as a

population model. Still, *C* was introduced for an important *functional* reason and it will not do to banish it from (15.1). Mathematical chicanery it might be, but if so it is important chicanery nonetheless.

One way to restore some plausibility to *C*, and to the negative values of *x* it is capable of producing, is to first note that (15.1) is a model for a lumped network. Lying deeper beneath this model is the unlumped model in which we have separate populations of excitatory and inhibitory neurons [ELIA]. If we reinterpret *B* as representing the number of excitatory projection neurons in the population, then we can plausibly interpret a negative value for *x* as denoting that the inhibitory subpopulation is active and holding the excitatory population in an overall average state of hyperpolarization. The activation function *f* then takes care of the rest of the picture.

§ 3. The Grossberg Normalizers

The previous section has pointed out the practical importance of normalized variables in an ART network. Normalization was not new to neural network theory at the time ART first appeared. A number of researchers, including Malsburg, Kohonen and others, had incorporated mathematical normalization into their network models. The WTA rule from chapter 14 is one well known example of this, as is the α-LMS algorithm developed by Widrow. Grossberg was the first important researcher to point out that, if normalized networks were to command plausibility as models of biological neural systems, normalization methods and normalized forms had to be such that a neural network model could *produce* normalization as an inherent part of the way the network functioned, rather than having normalization imposed upon it as an ad hoc bit of mathematical chicanery. In this section we introduce two normalization networks that meet this requirement and which are found – in one form or the other – in many ART systems. We will call these ART subsystems the ***Grossberg normalizers*** $GN^{(1)}$ and $GN^{(2)}$.

$GN^{(1)}$ is the simpler of the two and appeared first. Its anatomy is called a feedforward (non-recurrent) on-center/off-surround anatomy [GROS5]. Figure 15.4 illustrates the structure of the network. The network has *n* inputs I_i and *n* SNI nodes we will designate as type $SNI^{(0)}$. No lateral connections are made among the Instars. Each input I_i projects an excitatory signal to one Instar node v_i and projects an inhibitory signal to all the other nodes. Thus, each SNI node v_i has for its excitatory and inhibitory inputs

$$\begin{aligned} \xi_i^e &= I_i \\ \xi_i^i &= \sum_{k \neq i} I_k \end{aligned} \quad . \tag{15.6}$$

Each node v_i has an excitation variable x_i which also serves as the SNI's output signal. (15.1) for

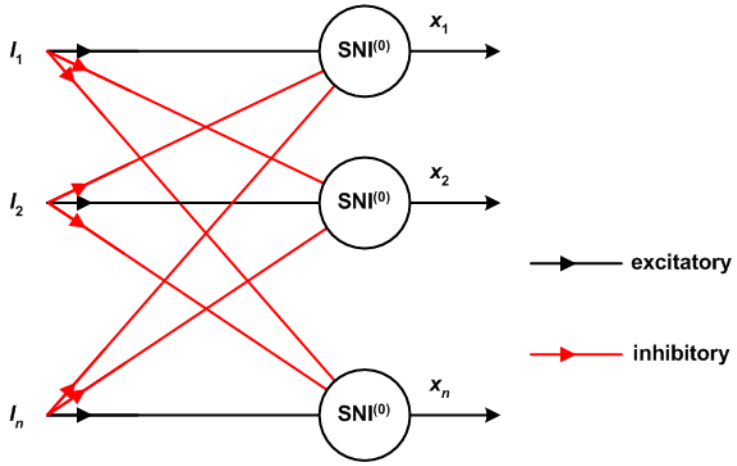


Figure 15.4: Anatomy of $GN^{(1)}$.

the type $SNI^{(0)}$ map is

$$\dot{x}_i = -A \cdot x_i + (B - x_i) \cdot I_i - x_i \cdot \sum_{k \neq i} I_k \ . \tag{15.7}$$

We let I denote the sum of all the I_i terms and define the normalized value $\theta_i = I_i/I$. The steady state solution for (15.7) is

$$x_i = \frac{B \cdot I_i}{A + I} = \frac{I_i}{I} \cdot \frac{B \cdot I}{A + I} = \theta_i \cdot \frac{B \cdot I}{A + I} \ . \tag{15.8}$$

We see from (15.8) that the steady state excitation x_i is bounded by $0 \le x_i \le BI/(A+I)$. This is due to the on-center/off-surround distribution of the input signals. Furthermore, $x_i < B$, which satisfies the constraint on excitation variables introduced earlier. We obtain the difference equation form of (15.7) by applying Euler's method, subject to the constraint on Δt worked out above. The result is

$$x_i(t + \Delta t) = (1 - \Delta t \cdot (A + I)) \cdot x_i(t) + \Delta t \cdot B \cdot I_i \tag{15.9}$$

which has the same steady state solution (15.8).

In ART systems it is generally assumed that inputs I_i change slowly compared to the dynamic action of the various layers of SNI maps in the system. It is therefore common to simply use (15.8) directly for $GN^{(1)}$ rather than to actually go through the computational steps of (15.9). Mathematically this is no different from the ad hoc computation method for other normalization calculations used by other network system models. What is different here is that the particular normalization produced by $GN^{(1)}$ is the steady state value obtained from a network system model. It is, in other words, a *natural* consequence of neurodynamics.

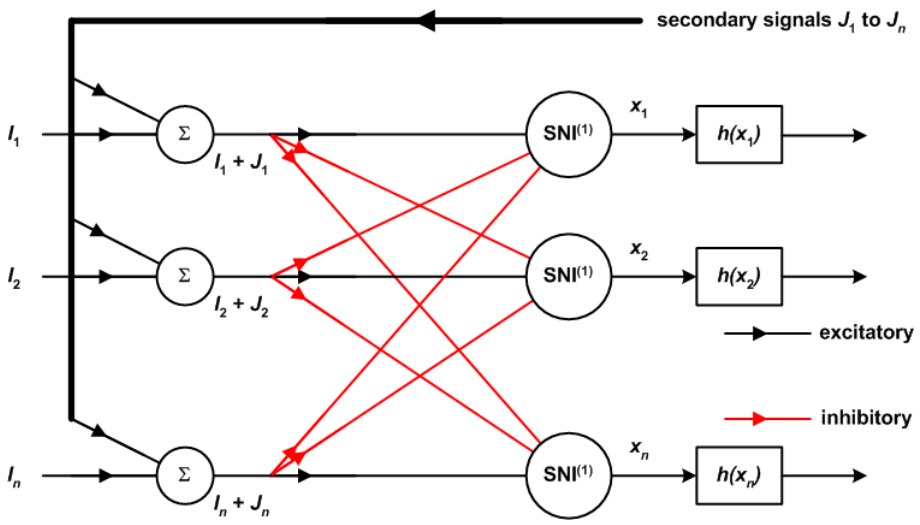


Figure 15.5: Anatomy of $GN^{(2)}$.

$GN^{(2)}$ is a variation on the theme of $GN^{(1)}$. The structure of this normalizer is shown in figure (15.5). Like $GN^{(1)}$, $GN^{(2)}$ is a non-recurrent layer of Instars, but of a class we will call $SNI^{(1)}$. Unlike $GN^{(1)}$, the SNI nodes of $GN^{(2)}$ require a nonlinear activation function $h(x)$ defined by

$$h(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} . \tag{15.10}$$

In the ART literature the notation $[x]^+$ is often used to designate this function. We will call this activation function the ***Heaviside extractor*** because it is equivalent to multiplying x by the Heaviside function of x.

The usefulness of anatomy $GN^{(2)}$ comes into play when the input to the normalizer consists of two converging afferent tracts, $\{I_1 \cdots I_n\}$ and $\{J_1 \cdots J_n\}$. It performs a limited amount of ***contrast enhancement***, as we will explain shortly. Define $G_i = I_i + J_i$. We set $J = \Sigma_{i=1:n} J_i$ and, similarly, $G = \Sigma_{i=1:n} G_i = I + J$. The excitatory and inhibitory inputs are defined by replacing I_i and I_k by G_i and G_k in (15.6). $SNI^{(1)}$ is then defined by the dynamical equation

$$\dot{x}_i = -A \cdot x_i + (B - x) \cdot G_i - (C + x_i) \cdot \sum_{k \neq i} G_k . \tag{15.11}$$

The difference between $SNI^{(1)}$ and $SNI^{(0)}$ is the presence of the spatial contrast parameter C in $SNI^{(1)}$. C is related to B by $B = (n - 1) \cdot C$. Let $\omega_i = G_i/G$. The steady state solution of (15.11) is then

$$x_i(\infty) = \frac{nCG}{A + G} \cdot \left(\omega_i - \frac{1}{n} \right). \tag{15.12}$$

One property of this solution we may immediately note is this. If the G_i pattern is uniform, i.e. if $\omega_i = 1/n$ for every G_i, then $x_i(\infty) = 0$ for each node. More generally, if $G_i = a + b_i$ where a is a constant term and $\Sigma_{i=1:n}\, b_i = 0$, we have $G = n \cdot a$ and $\omega_i = b_i/G + 1/n$. (Electrical engineers refer to the constant term a as the "dc" component[1] in the signal). In this case (15.12) reduces to

$$x_i(\infty) = \frac{nCb_i}{A+G}.$$

In other words, $x_i(\infty)$ for each node is specifically determined by the spatially-varying component of G_i and any constant "background" in the input pattern is suppressed across the output pattern of the normalizer. *Only* the spatially-varying part of the pattern $\{G_1 \cdot \cdot \cdot G_n\}$ is presented at the outputs of the $SNI^{(1)}$ nodes. (The "dc" component a appears only in the gain term G). We may further note that since some of the b_i are negative, $x_i(\infty) < 0$ for those terms, and this is why the Heaviside extractor is needed in this normalizer.

In the special case where one of the afferents, say $\{J_1 \cdot \cdot \cdot J_n\}$, is uniform, that afferent tract will be suppressed. For example, let $J_i/J = 1/n$ and let $I_i = \theta_i \cdot I$. Then (15.12) reduces to

$$x_i(\infty) = \frac{nCI}{A+I+J}\left(\theta_i - \frac{1}{n}\right)$$

and the steady-state outputs depend only on the I_i except for the reduction in total excitation due to J. A final special case of interest is the one for which the J_i are each proportional to their corresponding I_i, i.e. $J_i = \theta_i \cdot J$ and $I_i = \theta_i \cdot I$. In this case, (15.12) becomes

$$x_i(\infty) = \frac{nC \cdot (I+J)}{A+I+J}\left(\theta_i - \frac{1}{n}\right)$$

which differs from the previous expression only by an amplification due to J.

We obtain the difference equation form of (15.11) as usual by applying Euler's method. The result is

$$x_i(t+\Delta t) = (1 - \Delta t \cdot (A+G)) \cdot x_i(t) + \Delta t \cdot \left(nCG \cdot \left(\omega_i - \frac{1}{n}\right)\right) \qquad (15.13)$$

where we have used $B = (n-1) \cdot C$ to obtain this expression. The steady state solution of (15.13) is given again by (15.12), as it must be. Similarly to $GN^{(1)}$, it is common practice to simply use (15.12) in computing the response of $GN^{(2)}$.

[1] "dc" stands for "direct current." It is old electrical engineer jargon dating from the days of Thomas Edison and the development of electric power generators. Today it refers to any signal that does not vary.

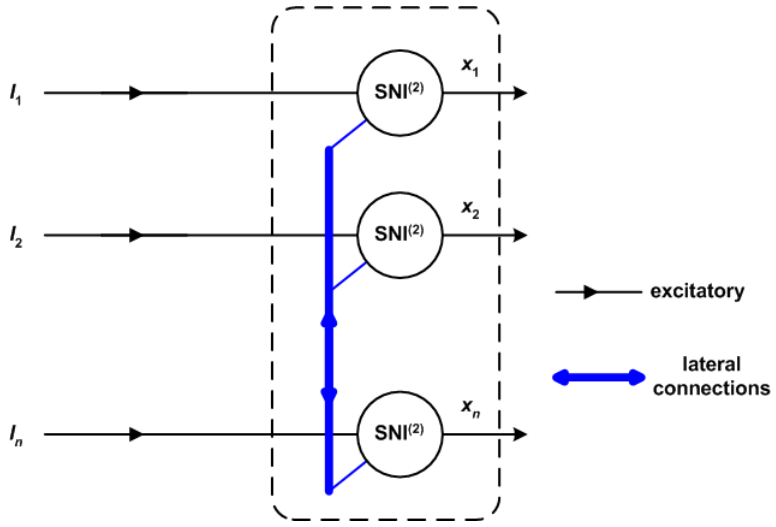


Figure 15.6: Anatomy of contrast enhancer 1.

§ 4. Contrast Enhancer 1

The heart of all ART systems is the contrast enhancer function carried out by recurrent on-center/off-surround networks. The simplest and most basic of these networks is depicted in figure 15.6. We will call this network ***contrast enhancer 1*** or $CE^{(1)}$ for short.

$CE^{(1)}$ was just called the "simplest" of Grossberg's contrast enhancers. In a way this is a bit misleading because no nonlinear recurrent neural network is "simple" in comparison to, say, linear/time-invariant networks. But simplicity is relative, and $CE^{(1)}$ *is* the simplest representative of its class and has been given the most complete mathematical treatment [GROS14].

$CE^{(1)}$ uses a different species of Instar, $SNI^{(2)}$, than do the Grossberg normalizers. Its general variable definitions for the *i*-th node are

- $\xi_i^e = f(x_i) + I_i$, where f is the activation function;
- $\xi_i^i = \sum_{k \neq i} f(x_k)$;
- $K_i = 0$; $\zeta = 0$;
- $C = 0$.

The dynamical equation for $SNI^{(2)}$ is

$$\dot{x}_i = -A \cdot x_i + (B - x_i) \cdot (f(x_i) + I_i) - x_i \cdot \sum_{k \neq i} f(x_k), \quad i = 1, \cdots, n. \tag{15.14}$$

Each node receives an excitatory input from itself and inhibitory inputs from all the other nodes in its layer. In some ways the behavior of an $SNI^{(2)}$ layer resembles that of the MAXNET; however, it also differs in a number of important ways due to the shunting action in (15.14). The

layer carries out a normalization of its total activity

$$x \overset{def}{=} \sum_{k=1}^{n} x_k \tag{15.15}$$

and is capable of performing contrast enhancement in a variety of ways that is determined by the specific activation function f employed in SNI$^{(2)}$. It also ***stores*** an equilibrium output pattern, which consists of the n excitations $\{x_1, \cdots, x_n\}$, after the input excitations I_i cease. This is referred to in the literature as "short term memory" or STM. Unlike the MAXNET, the inputs I_i need not be turned off, although for purpose of analysis we will initially pretend the inputs are applied over the time interval from $-T$ to 0 and then turned off at $t = 0$. The initial conditions $x_i(0)$ established by this are the initial conditions for the reverberation dynamics we will analyze. Later we will remove this restriction and see what happens in the case where the inputs are persistent during the reverberations of CE$^{(1)}$. What we will find is that CE$^{(1)}$ uses negative feedback to stabilize itself and prevent persistent inputs from saturating its nodes, unlike what happens in the MAXNET.

Although CE$^{(1)}$ is capable of winner-take-all operation, it is also capable of other operations including ties, with some subset of the x_i variables taking on non-zero and equal final values (called a ***locally uniform*** output distribution), and contrast enhancement (multiple non-zero final values with unequal final x_i results; this is called a ***fair*** output distribution of the non-zero survivors of the tournament competition). What result is obtained is a complicated function of both the activation function used and the intensity of the initial total excitation $x(0)$. It is worth noting at this point that these features of CE$^{(1)}$, although desirable in many competitive layer functions, are generally *incompatible* with ART network performance (figure 15.2). It is for this reason that we will meet other types of CE systems in chapter 16. Still, the theory of CE$^{(1)}$ is the starting point for the development of adaptive resonance systems and so an understanding of this contrast enhancer is a crucial prerequisite for understanding the systems to come later.

Understanding the dynamics of CE$^{(1)}$ is a two-part operation. First, one must understand the general dynamics of (15.14). Second, one must understand how the activation function f affects the outcomes of these dynamics. Thus, our treatment in this section will come in two parts.

§4.1 The Dynamical Characteristics of CE$^{(1)}$

Three general considerations drove the development of the nonlinear system of n coupled first order equations (15.14). First, Grossberg was and is fundamentally dedicated to producing a theory remaining faithful to both biological and psychology reality. The concept of the shunting node Instar, in all its various forms, evolved from an examination of mathematical forms belonging to classes of equations to which other model equations, such as the Hodgkin-Huxley

membrane equation, also belong. This is not because an SNI is or was ever intended to be a neuron model; it is not and Grossberg has always been very clear in stating it is not (although he often points out that his mathematical expressions might also have useful interpretations at the membrane level). But, as the previous chapters of this text have pointed out, a crystalline neural network model is built out of the concept of "average" neurons, tightly coupled netlet populations of such average neurons, and larger-scale proxy population models of neural networks that, at each step in our ascent of the model order reduction hierarchy, maintain the accurate representation of the main signal processing effects found at each level. In a unified framework such as this, it is a natural and eminently plausible hypothesis that the reduced model structures should remain tied to the *class* of mathematical functions to which the lower-level models also belong. This thinking comes through quite evidently in all of Grossberg's early papers, and one who has chosen computational neuroscience for his research field will find the prehistory of ART archived in these papers quite interesting. They paint a backdrop bringing a bit of intellectual relief to counteract the tone of Platonism that often saturates today's overly formal, overly brief, and unnecessarily abstract presentations of ART systems.

Getting down to detail, the second consideration is that the collective behavior of recurrent on-center/off-surround networks is determined by ***normalized variables***, $X_i = x_i \cdot x^{-1}$. [GROS14] lays out a number of important propositions and theorems for the nonlinear system defined by equations (15.14), and these are for the most part theorems expressed in terms of X_i variables and functions of X_i variables. The system of equations (15.14) is a nonlinear system with no known closed form general solutions. Therefore, these property and existence theorems provide our main tools for understanding how the system works.

Third, it is the ***distribution*** $\{X_1(0), X_2(0), \cdots X_n(0)\}$ rather than the individual $x_i(0)$ that determines the time evolution of $CE^{(1)}$. Quenching effects, final output distributions, etc. are understood in terms of the normalized distribution and cannot be reliably pegged to individual inputs I_i. Recurrent SNI layers are in a sense distributed systems despite the spatial quantization brought about by lumped model elements.

We begin the analysis by assuming at $t = 0$ a set of initial excitations $\{x_1(0), x_2(0), \cdots x_n(0)\}$ has been established by the I_i and that these inputs have now been turned off. We rewrite (15.14) as

$$\dot{x}_i = -\left(A + \sum_{k=1}^{n} f(x_k) \right) \cdot x_i + B \cdot f(x_i). \tag{15.16}$$

Summing equations (15.16) over i from 1 to n gives us the dynamical equation for the total

excitation x,

$$\dot{x} = -\left(A + \sum_{k=1}^{n} f(x_k) \right) \cdot x + \sum_{i=1}^{n} B \cdot f(x_i). \tag{15.17}$$

(15.16) and (15.17) are the two base equations from which the system dynamics are derived. It will prove convenient to introduce the following short-hand notations:

- $F = \sum_{k=1}^{n} f(x_k)$;
- $g(u) = u^{-1} \cdot f(u)$, where u is the argument of the activation function. We will call g the ***activation shape function***.

The variables in all these functions are variables in continuous time. We will also require an expression for the time derivative of the normalized variables X_i. Grossberg develops this in his proposition 1 of [GROS14]. The result is

$$\dot{X}_i = B \cdot X_i \cdot \sum_{k=1}^{n} X_k \cdot (g(x_i) - g(x_k)). \tag{15.18}$$

The reader is referred to [GROS14] for a proof of (15.18). (15.16)-(15.18) are the central equations needed for our analysis of the dynamics of $CE^{(1)}$.

Now, derivatives are not difference equations and we must confront the fact that computer implementation requires the conversion of (15.16)-(15.18) into difference equation form. The key concern in this conversion is making sure the dynamics of our discrete-time difference-equation-based system conform to the continuous time results for (15.16)-(15.18). In particular, we must be concerned with how the sampling interval Δt will affect our results. In keeping with the pedagogical level of this text, Euler's method will be used to perform this discretization. In actual practice, more powerful numerical integration methods are usually employed, but their employment requires more background in the mathematics of numerical integration than is given in this book. Applying Euler's method, we obtain for (15.16) and (15.17)

$$x_i(t + \Delta t) - x_i(t) \overset{def}{=} \Delta(x_i(t)) = \Delta t \cdot \{ -(A + F(t)) \cdot x_i(t) + B \cdot f(x_i(t)) \} \tag{d15.16}$$

$$x(t + \Delta t) - x(t) \overset{def}{=} \Delta(x(t)) = \Delta t \cdot \{ -(A + F(t)) \cdot x(t) + B \cdot F(t) \}. \tag{d15.17}$$

Derivation of the discrete time counterpart to (15.18) is a bit more involved but at root involves nothing more complicated than a bit of appropriate algebraic manipulation. The derivation closely follows that of (15.18) in [GROS14]. We define the variable

$$P = \frac{B \cdot F(t)}{x(t)}$$

and use this to obtain the discrete-time counterpart of (15.18) as

$$X_i(t+\Delta t) - X_i(t) \overset{def}{=} \Delta(X_i(t)) = \frac{\Delta t}{1 + \Delta t \cdot [P - (A + F(t))]} \cdot B \cdot X_i(t) \cdot \sum_{k=1}^{n} X_k(t) \cdot (g(x_i(t)) - g(x_k(t))) \tag{d15.18}$$

Note that (d15.18) is the same as (15.18) except for the appearance of the leading multiplicative factor involving P. Now, what is crucial for the limiting behaviors of the system is the algebraic *sign* of (15.18). Therefore, the crucial factor in converting the system over into difference equation form is that (d15.18) always maintains the same sign as (15.18). It is not difficult to deduce that if $f(x_i) \le B$ and $x_i \le B$, a bound for Δt given by

$$\Delta t < \frac{1}{A + B \cdot (n-1)}$$

is a sufficient condition to ensure this. In most cases this is a generally sufficient bound. There are, however, other conditions presented in [GROS14] we would want the discrete time system to meet. In some cases it is possible these may be more stringent than merely our need to maintain the proper sign in (d15.18). Through some limiting arguments that are rather too involved to go into here, a more general sufficient bounding condition for Δt can be set down, namely,

$$\Delta t < \min\left\{ \frac{1}{A + B \cdot (n-1)}, \quad \frac{1}{A + B \cdot (A + B \cdot f(B))} \right\}. \tag{d15.19}$$

With Δt appropriately chosen, we can now safely use the continuous-time propositions and principles derived by Grossberg to analyze the system. Lacking exact closed-form general solutions for (15.16)-(15.18), we must content ourselves with asymptotic and steady-state characteristics of the system. It has been proved [GROS14, proposition 1] that the following properties hold for this system:

$$\dot{X}_i = B \cdot X_i \cdot \sum_{k=1}^{n} X_k \cdot (g(x_i) - g(x_k)), \tag{15.19}$$

$$\begin{aligned} &\dot{x} = x \cdot (B - x) \cdot \left(\sum_{k=1}^{n} X_k \cdot g(x_k) - \frac{A}{B - x} \right) \equiv x \cdot (B - x) \cdot \left(G - \frac{A}{B - x} \right), \\ &G = \sum_{k=1}^{n} X_k \cdot g(x_k) \end{aligned} \tag{15.20}$$

These two equations are key to deriving many of the other properties of the system and to

understanding the effect the choice of activation function will have.

One of the first things we learn from (15.19) is that the nodes in $CE^{(1)}$ exert reciprocal effects on each other. Suppose for nodes *i* and *j* we find

$$g(x_i) - g(x_j) > 0 .$$

This term in (15.19) then contributes to causing $\dot{X}_i$ to tend toward a positive value, thus implying that $x_i = X_i \cdot x$ will increase over time. But by the same token, in the expression for $\dot{X}_j$ there will be a sign reversal in the difference between the two *g* function, thus contributing to a tendency for x_j to decrease over time. In this way the activation function $f = x_i \cdot g(x_i)$ enters into the process of normalization carried out by the network.

In ART and ART-related systems, the activation shape function *g* is always a continuous function. This leads to an important ***order preserving property*** [GROS14, proposition 2] found in these systems. Because we can label the SNI nodes in any way we wish, let us adopt the ***Grossberg enumeration*** schema and number them so that $X_1(0) \le X_2(0) \le \cdot \cdot \cdot \le X_n(0)$. (15.19) then leads to the following important result:

Theorem 1: If $X_1(0) \le X_2(0) \le \cdot \cdot \cdot \le X_n(0)$ then $X_1(t) \le X_2(t) \le \cdot \cdot \cdot \le X_n(t)$ for all $t > 0$.

This theorem does not hold in general for arbitrary network anatomies but it does hold for $CE^{(1)}$ and other types of ART structures. For instance, we saw earlier in this text an example where the classic Mexican Hat competitive network violates theorem 1. The order-preserving property is one of the key useful features of $CE^{(1)}$.

Next we turn to the limiting properties of $x(t)$ given $x(0)$. These properties depend on the shape of the activation function $f(u) = u \cdot g(u)$, or equivalently on the activation shape function, and on where along $g(u)$ the initial distribution $\{x_1(0), x_2(0), \cdot \cdot \cdot x_n(0)\}$ falls. For practical activation functions, $g(u) = 0$ for $u < 0$. We will further restrict our examination to non-decreasing activation functions, i.e. $f(u_2) \ge f(u_1)$ for $u_2 > u_1$.[2] There are many possible activation shape functions. Two generic practical forms are illustrated in figure 15.7. These are two examples of activation shape functions that produce ***generalized sigmoid activation functions***. Functions such as those depicted in this illustration are described in terms of three activation function regions: (1) *f(u)* is linear; (2) *f(u)* is faster-than-linear; and (3) *f(u)* is constant (flat). Grossberg also analyzed slower-than-linear activation functions in [GROS14] and found they result in undesirable performance. Of the two generic forms in figure 15.7, figure 15.7B has the superior performance.

[2] Radial basis functions are not used in ART systems.

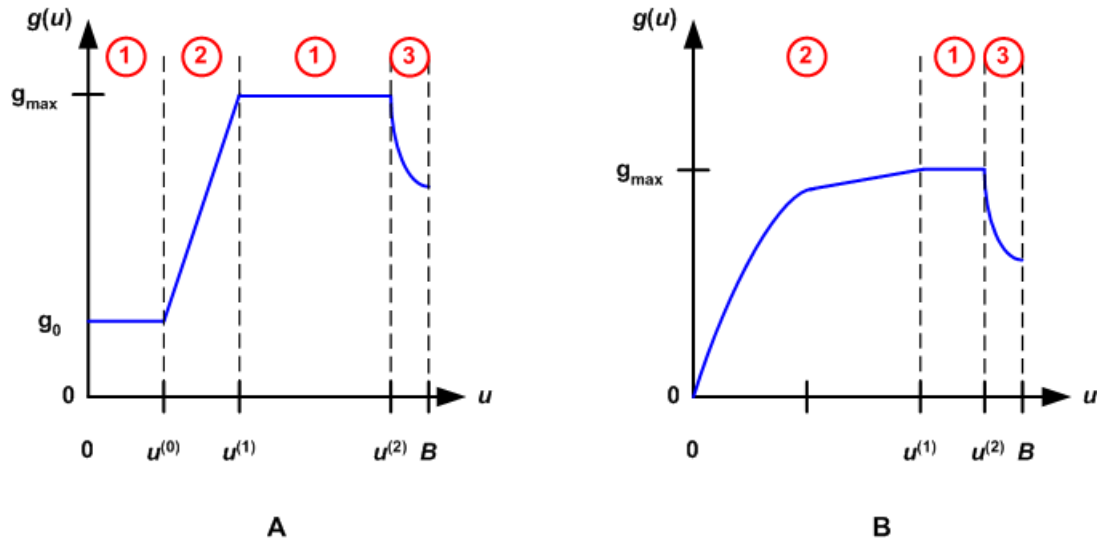


Figure 15.7: Two representative activation shape functions. (A) *g*(0) non-zero. (B) g(0) = 0. There are three regions for *f(u)* for these *g(u)* functions. Region 1: *f(u)* is linear; Region 2: *f(u)* is faster-than-linear; Region 3: *f(u)* is constant. Both activation shape functions depicted here produce generalized sigmoid activations.

We will look here at some general properties of the asymptotic values attained by $x(t)$. In the next subsection we will look at how the design of the activation function affects these results. Our starting point is (15.20) in the form

$$\dot{x} = x \cdot (B - x) \cdot \left(G - \frac{A}{B - x} \right).$$

In [GROS14], Grossberg shows that convex activation functions, such as sigmoids, have the property that limiting values always exist for the x_i variables and for x. $\dot{x} = 0$ defines the steady state condition for the system. Applying this to (15.20) above, we find two possible solutions:

- $x = 0$. This is the trivial solution and corresponds to the case where the reverberations set up in the SNI layer are transient, i.e. all x_i eventually decay to zero.
- $x = B - A/G$. Because G is a function of the X_k distribution and the activation shape function, this solution must typically be analyzed numerically. As it turns out, this solution is the one we desire for the system. It is easily shown that for sigmoid activation functions for which *g(u)* has a flat region where $g = g_{max}$, the upper value for G is bounded by $G \le g_{max}$. Therefore we obtain for an upper bound $x \le B - A/g_{max}$ for these systems. Because x must always be non-zero, this bound is possible only if $g_{max} > A/B$. Otherwise the $x = 0$ solution results. We can also apply this result to the $g = g_0$ region of figure 15.7A. In this case, if the initial X_k distribution is such that *all* the $x_k(0)$ values fall into this region, a $g_0 \le A/B$ will result in the $x = 0$ final solution.
- Because $x = 0$ is a possible solution, implying all the x_i are zero, it is natural to ask (given what has just been said above) what happens if $G(t = 0)$ is also zero. This is clearly a possibility if $g(0) = 0$ as in figure 15.7B. For this consideration, let us assume all $x_i(0) = 0$ and inputs I_i are applied at $t = 0$. In this case, (15.14) gives us $\dot{x}_i(0) = B \cdot I_i > 0$. This tells

us the not-surprising fact that a stimulus will always nudge $CE^{(1)}$ out of an initially relaxed state. The solution x we have discussed above applies only after some time has elapsed, not for all times t. It is what happens *after* this nudge has been applied that concerns us, and for this we must come to grips with the effects of the activation function on the reverberation dynamics.

§4.2 The Effects of the Activation Function on $CE^{(1)}$

We give an abbreviated treatment of the effects of the activation function in this section. Specifically, the discussion here is restricted to cases of $g(u)$ of the general shape illustrated in figure 15.7B. A more general presentation is given in [GROS14]. The reason for this restriction is to simplify the presentation and to focus on properties pertaining to practical ART and ART-like systems. Thus, we consider the generalized sigmoid activation functions here and no others. We will use figure 15.7B to provide specific illustrations of what the mathematics is telling us as we go through the discussions which follow.

The basis for this analysis is (15.19). We will use Grossberg's enumeration so that X_n is greater than or equal to all other X_i and X_1 is less than or equal to all other X_i. Given initial values $X_i(0)$ for all nodes, let us first examine what happens to $X_n(t)$. Here we have

$$\dot{X}_n = B \cdot X_n \cdot \sum_{k=1}^{n} X_k \cdot \left(g(x_n) - g(x_k)\right).$$

The first thing we can note is that node n does not self-contribute to its own derivative for X_n. In other words, within the sum the $k = n$ term is zero. Because x_n is the largest term among all the x_k, we have $g(x_n) \geq g(x_k)$ for all $k \neq n$. If there is *any* x_k for which $g(x_k) < g(x_n)$, $\dot{X}_n > 0$ and x_n will tend to be pushed to a *relatively* larger value compared to the rest of the distribution. Now, if any x_k meets this condition then x_1 meets the condition because, under Grossberg's enumeration, x_1 is the smallest excitation variable. For this variable we have

$$\dot{X}_1 = B \cdot X_1 \cdot \sum_{k=1}^{n} X_k \cdot \left(g(x_1) - g(x_k)\right)$$

and by the reciprocity effect between node n and node 1, we are guaranteed $\dot{X}_1 < 0$. Thus, x_1 will be pushed to a relatively smaller value. In a sense, we can say x_n and x_1 "repel" one another.

Now, the overall situation is made more complicated by the fact that x, the total excitation, can change (and most of the time *will* change) when any of the x_i change. Recall that (15.19) is expressed in terms of the normalized variables, X_i, rather than in terms of the individual x_i, and that $X_i = x_i/x$. Thus, there can be many subtleties involved in figuring out what the overall limiting response of the system will be. Fortunately, Grossberg was able to prove a number of useful facts

about the system, which he summarized in a series of theorems [GROS14]. Here we summarize these key results. The reader is referred to Grossberg's paper for the proofs.

To get a grasp on the overall dynamic, we begin with a simple special case. Suppose both x_1 and x_n lie in a region 1 of figure 15.7. If this is so, then all other x_k also lie in this region because their values are contained between x_1 and x_n. But if all the x_i lie in a region 1, then all $g(x_i)$ are equal and therefore all $\dot{X}_i = 0$. This is a condition of equilibrium so far as the normalized variables are concerned. Is it also a condition of equilibrium for all the x_i? To see this, consider that there are only two ways to get $\dot{X}_i = 0$: (1) $x = \infty$ or (2) $\dot{x}_i = X_i \dot{x}$. (15.20) forbids the first. The second says the condition of equilibrium for X_i is not necessarily an equilibrium for x_i.

Can this case actually happen? For this we must look at the steady-state value for x. Suppose that for region 1 the activation shape function is $g(x_i) = g_1$. Then

$$G = \sum_{k=1}^{n} X_k \cdot g_1 = g_1 \cdot \sum_{k=1}^{n} X_k = g_1 \cdot 1 = g_1 .$$

Therefore, the steady state value of x is $x = B - A/g_1 > 0$ (if the case is possible). This means we must have $g_1 > A/B$. If this condition is satisfied, the case is possible; otherwise it is not. Thus, for example, in figure 15.7A it is possible to select g_0 and $g_{\max}$ so that this case cannot occur for g_0 but can occur for $g_{\max}$. This result is stated more succinctly and with more precision in the form of a theorem. For this and the following theorems, refer to figure 15.7B.

Theorem 2: If $X_1(0) \cdot \min\{x(0), B - A/g_{\max}\} \geq u^{(1)}$ and $X_n(0) \cdot \max\{x(0), B - A/g_{\max}\} \leq u^{(2)}$ then $\lim_{t \to \infty} X_i(t) = X_i(0)$ for $i = 1, \ldots, n$ and $x(t)$ approaches $B - A/g_{\max}$.

The limiting distribution for this case is called a ***fair distribution*** because $X_i(\infty) = X_i(0)$ for all i.

For the remaining theorems we need a few definitions. Using Grossberg's enumeration and recalling the order-preserving property of this system, the response is said to be ***contour enhancing*** if $\dot{X}_n(t) \geq 0$ and $\dot{X}_1(t) \leq 0$ for all $t \geq 0$ and neither is identically zero. The network's reverberation is said to be ***persistent*** if $x(t) > 0$ for all $t \geq 0$. The reverberation is said to be ***transient*** if $x(t)$ goes to zero. Let $u^{(*)}$ be defined as the value $u < B$ at which $g(u) = g(B)$. Then if $u^{(*)} + u^{(2)} \geq \max\{x(0), B - A/g_{\max}\}$, we say ***condition α holds***. With these definitions in hand, we are ready to summarize Grossberg's main theorems.

Theorem 3: If condition α holds then: (i) $\lim_{t \to \infty} X_i(t)$ exists for all i from 1 to n; (ii) $\dot{X}_n(t) \geq \dot{X}_i(t)$ for all i from 1 to $n - 1$; (iii) $\dot{X}_1(t) \leq 0$.

This theorem tells us that if our activation shape function and initial total signal activity x

conforms to condition α then the dynamics of the system will not produce a final distribution in which all the X_i take on a single uniform value. "Uniformization" is a very undesirable result for a contrast enhancer (because it produces an output with no contrast at all), and theorem 3 tells us this cannot occur if the condition of the theorem is met.

Theorem 4: Define the index $K < n$ by the condition $X_K(0) < X_{K+1}(0) = X_n(0)$. Let $x_n(0) < u^{(2)}$. Then the limits $Q_i = X_i(\infty)$ and $E = x(\infty)$ exist, $\dot{X}_n(t) \geq 0$, $\dot{X}_n(t) \geq \dot{X}_i(t)$, $i < n$, and $\dot{X}_1(t) \leq 0$. If $x_1(t) \geq u^{(1)}$ then all $\dot{X}_i(t) = 0$. Furthermore, if there is some index $L < n$ such that

$$X_L(0) \cdot \min\left\{ x(0), B - \frac{A}{\sum_{i=L}^{n} X_i(0) \cdot g_{\max}} \right\} \geq u^{(1)}$$

and the network's reverberation is persistent, $\dot{X}_i(t) \geq 0$ and $d(X_i/X_j)/dt = 0$ for $i, j \geq L$.

This is a contour enhancement theorem. Contour enhancement brings out changes occurring across the input pattern and suppresses uniform backgrounds. It is one of the primary functions of on-center/off-surround networks, and theorem 4 tells us under what conditions it is guaranteed to occur. We will say the distribution with $d(X_i/X_j)/dt = 0$ for $i, j \geq L$ is ***contrast-enhanced fair***.

Theorem 5: If $X_1(0) \cdot \min\{x(0), B - A/g_{\max}\} \geq u^{(1)}$ then $\lim_{t \to \infty} X_i(t) = X_i(0)$ for $i = 1, \ldots, n$. Otherwise, if $X_i(t) \cdot \min\{x(0), B - A/g_{\max}\} < u^{(1)}$ for some sufficiently large t then $\lim_{t \to \infty} X_i(t) = 0$.

This theorem gives us another condition under which a fair distribution results. But more importantly, it defines the ***quenching threshold*** $\mathrm{QT} = u^{(1)}/(B - A/g_{\max})$. If $X_i < \mathrm{QT}$ for all $t > 0$ then X_i is treated as noise and is quenched.

Theorem 6: If the reverberation is persistent, $B - A/g_{\max} < N \cdot u^{(1)}$ for N such that $1 < N < n$, and $X_{n-N+1}(0) < X_n(0)$ then $X_j(t)$ asymptotically goes to zero for all j from 1 to $n - N + 1$.

Theorem 6 tells us which x_i will be quenched by the network. Note that the condition requiring the reverberation be persistent means we must have $B - A/g_{\max} > 0$. It also means the initial vector $\{x_i\}$ must have sufficient intensity that at least one member of it exceeds QT.

Theorem 7: If condition α holds, the reverberation is persistent, $X_1(0) < X_n(0)$ and we have $X_1(0) \cdot (B - A/g_{\max}) \leq u^{(*)}$ then $X_1(t)$ asymptotically goes to zero.

This theorem tells us how the quenching of the smallest term depends on the shape of the rising portion of the activation shape function in figure 15.7B.

Theorem 8: If condition α holds, the reverberation is persistent, and $X_i(t_i) < X_n(t_i)$ for some large value of t_i then $X_i(t)$ asymptotically goes to zero. Furthermore, if $X_n(0) > X_{n-1}(0)$ then

the only non-zero excitation in the limiting distribution is X_n. Otherwise the final distribution is locally uniform.

This is the extreme case where the network "chooses a winner" in the manner of a MAXNET. Grossberg calls this result a ***0-1 distribution***. The term "locally uniform" means the final distribution has two or more ***non-zero*** "ties" in the final values. Note, too, that unlike the classic MAXNET, the initially largest x_i (which is x_n when we use Grossberg's enumeration) survives (is asymptotically nonzero) regardless of whether or not there is a tie.

Theorem 9: If $X_1(0)\cdot(B - A/g_{\max}) \le u^{(*)}$ then $\dot{X}_1(t) \le 0$ for all $t \ge 0$. Moreover, if the reverberation is persistent and $X_1(0) < X_n(0)$ then $X_1(0)$ asymptotically goes to zero.

This theorem tells us condition α is not necessary for $X_1(0)$ to be quenched.

Theorem 10: If $X_i(t_i)\cdot(B - A/g_{\max}) \le u^{(*)}$, $i < n$, for some sufficiently large t_i then $\dot{X}_i \le 0$ for all $t > t_i$. Furthermore, if the reverberation is persistent and $X_i(t_i) < X_n(t_i)$ then X_i asymptotically goes to zero.

This theorem tells us that even if X_i initially grows in the early stages of the network's dynamical response, this is no guarantee that this X_i will not be eventually quenched. One way to understand this is to consider what happens when X_1 goes to zero. Once this has happened, we can regard the system as being comprised of $n - 1$ nodes with a new "initial condition" at the time t_i when X_1 reaches zero. X_2 then becomes "the new X_1" & etc.

Theorem 11: If $(n-1)\cdot u^{(*)} + u^{(2)} > \max\{x(0), B - A/g_{\max}\}$ then $\dot{X}_1(t) \le 0$ for all $t \ge 0$. Moreover, if the reverberation is persistent and $X_1(t_i) < X_n(t_i)$ then X_1 asymptotically goes to zero.

This theorem gives us yet another way for X_1 to go to zero. Note again that once this has happened we have a "new" system with $n - 1$ nodes. Replacing $n - 1$ by $n - 2$ in the theorem in principle then tells us the fate of X_2 and so on. Of course, if $x(0) > B - A/g_{\max}$ this is more than a bit tricky because we are likely not to know the "new" $x(0)$ (which is why I said "in principle" just now). But if the reverse is true, as it often may well be, the process is straightforward.

These theorems tell us quite a bit about the dynamics of this system, and they provide quantitative means for designing the activation shape function.

§4.3 The Effect of Persistent Inputs on CE[(1)]

Grossberg's 1973 paper treated only the homogeneous form of (15.14). Comparatively little is known about the mathematical properties of the non-homogeneous equation. In this section we will examine some sufficient conditions under which a stable, persistent steady state solution exists for the non-homogeneous equation. First, we recall that

$$F = \sum_{k=1}^{n} f(x_k) = \sum_{k=1}^{n} x_k \cdot g(x_k) = x \cdot \sum_{k=1}^{n} X_k \cdot g(x_k) = x \cdot G .$$

We will also find it convenient to define $I = \Sigma_{i=1:n} I_i$ and $\Gamma = \Sigma_{i=1:n}(X_i \cdot I_i)$.

Using these substitutions, (15.14) is re-written as

$$\dot{x}_i = -(A + F + I_i) \cdot x_i + B \cdot (f(x_i) + I_i).$$

Summing this expression over i and rearranging terms produces

$$\dot{x} = -(A + \Gamma + F) \cdot x + B \cdot (F + I).$$

For a stable, persistent steady state solution we must simultaneously satisfy $\dot{x} = 0$ and $\dot{x}_i = 0$. We begin with the $\dot{x} = 0$ condition. Define $\mu = (1 - (A + \Gamma)/(BG))^{-2}$. After some minor algebraic manipulation we arrive at

$$x = \frac{1}{2}\left(B - \frac{A + \Gamma}{G}\right) \cdot \left[1 + \sqrt{1 + \mu \cdot \frac{4I}{BG}}\right]. \tag{15.21}$$

Because both G and Γ are functions of the x_i terms, this is a nonlinear equation and must be evaluated numerically. It is worth noting that for $1 - (A + \Gamma)/(BG) = 0$ it reduces to

$$x = \sqrt{\frac{I \cdot B}{G}}$$

and so remains finite. For this to be a valid solution we must also require $x \geq 0$, from which we obtain as a condition $\Gamma \leq BG - A$.

Next we turn to the $\dot{x}_i = 0$ condition. For this we obtain

$$x_i = \frac{B \cdot (f(x_i) + I_i)}{A + F + I_i}. \tag{15.22}$$

It is important to note that if $I_i \neq 0$ then $x_i = 0$ cannot be a steady-state solution. Equally, x_i cannot be negative, and so we have the important result that the non-homogeneous SNI equation must produce a non-zero forced response in the steady state for every node receiving a non-zero input stimulus.

These results prove a fixed point solution exists, but they do not prove it is a *stable* fixed point solution. Empirically, however, we find that the system does indeed settle into a stable final state, i.e. it renormalizes the activities of the x_i variables to meet the required conditions. We can

conceptually understand this from the equation $\dot{x}_i = -(A + F + I_i)\cdot x_i + B\cdot(f(x_i) + I_i)$. The term multiplying x_i is a negative feedback terms and acts as a kind of decay rate for transients in the system. As I_i, and therefore F, increases, this decay rate factor grows larger and tends to put a cap on how high the second term in the expression can drive the excitation variables x_i. Figure 15.8 illustrates a typical charge-up and discharge response for $x(t)$. A fixed excitation pattern was applied at $t = 0$ and the network (initially relaxed) was allowed to charge for 400 iteration steps using $\Delta t = 0.5/(A + n - 1)$, which was $\Delta t \cong 0.02$ for this example. The network was effectively fully charged after around 200 iterations ($t \cong 4$). After 400 iterations the input was set to zero and the network allowed to discharge. The discharge was effectively complete after 400 more iterations. Note the slowing of the discharge rate as $x(t)$ decreases to its steady-state level.

§4.4 An Example of CE$^{(1)}$ Performance

This section presents a performance example of CE$^{(1)}$. The input pattern is a 5 × 5 grid of signals (a "retina") in which a T-shaped pattern is presented in the presence of noise (figure 15.9). The caption of figure 15.9 provides the numerical data on the input signal. The signal was pre-scaled by a 25-node GN$^{(1)}$ stage before being presented for 500 iteration time steps to the CE$^{(1)}$ layer, which consists of 25 SNI$^{(2)}$-type nodes. (The presence of the GN$^{(1)}$ layer actually made no significant difference to the final result; it merely had a minor effect on the final charge-up values

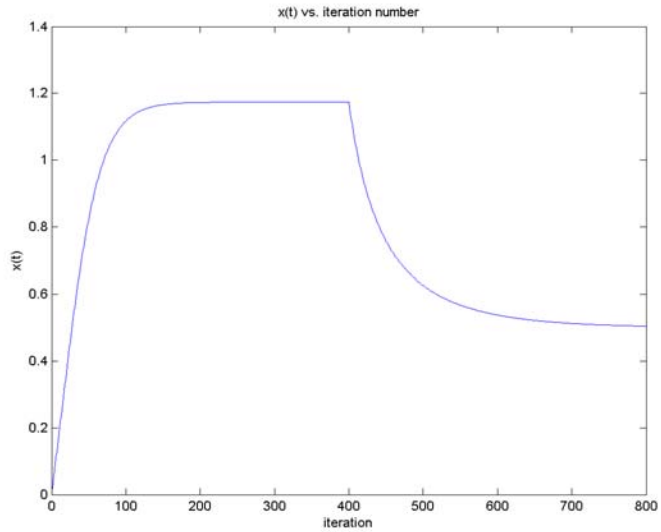


Figure 15.8: Typical charging and discharging $x(t)$ response for CE$^{(1)}$. A steady input pattern is applied at t = 0 and maintained for 400 iteration steps. The input is then set to zero and the network is allowed to discharge for an additional 400 iteration steps. For this example $\Delta t = 0.5/(A + n - 1)$, $B = 1$, $A = 0.5$, $g_{max} = 1$, and QT = 0.1. The input pattern was prescaled by a GN$^{(1)}$ network with the same B and A values. $n = 25$.

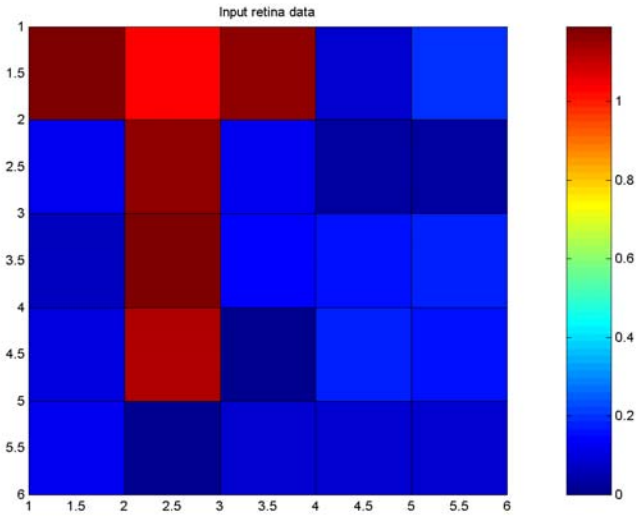


Figure 15.9: Input pattern for $CE^{(1)}$ example. The input was a 5 × 5 grid of "pixels" in which a "T" character was presented in the presence of noise. The pixel noise consisted of random values from a uniform distribution over the range from 0.0 to 0.2. The elements of the "T" had pixel values of 1, and the other pixel elements in the grid were 0 prior to the addition of the noise. Noisy pixel amplitudes are color-coded.

of the x_i after 500 iteration steps). $\Delta t = 0.5/(A + n - 1) \cong 0.02$ was used for the simulation. The $GN^{(1)}$ and $CE^{(1)}$ layers both used $B = 1.0$ and $A = 0.5$ for their parameter values.

The activation shape function for $CE^{(1)}$ was generally of the form of figure 15.7B and was explicitly defined by

$$g(u) = \begin{cases} \dfrac{g_{\max}}{u^{(1)}} \cdot u, & 0 \le u \le u^{(1)} \\ g_{\max}, & u^{(1)} < u \le u^{(2)} \\ \dfrac{g_{\max} \cdot u^{(2)}}{u}, & u > u^{(2)} \end{cases} \tag{15.23}$$

with $g(u) = 0$ for $u < 0$. This results in $f(u) = u \cdot g(u)$ being a sigmoid function with a saturation value $f_{\max} = g_{\max} \cdot u^{(2)}$. The parameter values were $g_{\max} = 1$, $u^{(1)} = 0.05$, and $u^{(2)} = 0.8$. This results in a quenching threshold $QT = u^{(1)}/(B - A/g_{\max}) = 0.10$.

Figure 15.10 shows the results of the simulation at the end of the charge-up period and at the end of the discharge period for both x_i and X_i. The value attained by $x(t)$ at the end of the charge-up period was $x = 1.1993$, which agrees exactly with (15.21).

A close comparison of figure 15.10A with figure 15.9 shows that some amount of signal strength redistribution already takes place during the charge-up period, although the noise is not entirely quenched during this period. (This is qualitatively evident, upon close inspection, from the color shades in these plots; note however that pre-scaling and the "automatic gain control"

character of CE$^{(1)}$ does produce a significant decrease in absolute signal amplitudes. To see this one must refer to the numerical values on the accompanying color bars). One way to quantitatively assess this is to define a pattern-to-background ratio as follows.

We categorize the individual signals into those that below to the T pattern, set p, and those that belong to the non-T background pixels, set b. For each category we sum the squares of the signals belonging to that category. The pattern-to-background ratio, p/b, is merely the ratio of these two sums-of-squares.

As a practical matter, information theorists usually prefer to measure such ratios in units of ***decibels***, i.e. $(p/b)_{\text{dB}} = 10 \cdot \log_{10}(p/b)$. (This permits easier performance comparison with other well known systems such as communication systems). For this example simulation, the initial pattern-to-background ratio was $(p/b)_{\text{dB}} = 15.1$ dB; after charge-up the p/b ratio of the x_i signals was $(p/b)_{\text{dB}} = 20.6$ dB, a 5.5 dB improvement. This would be regarded as excellent improvement.

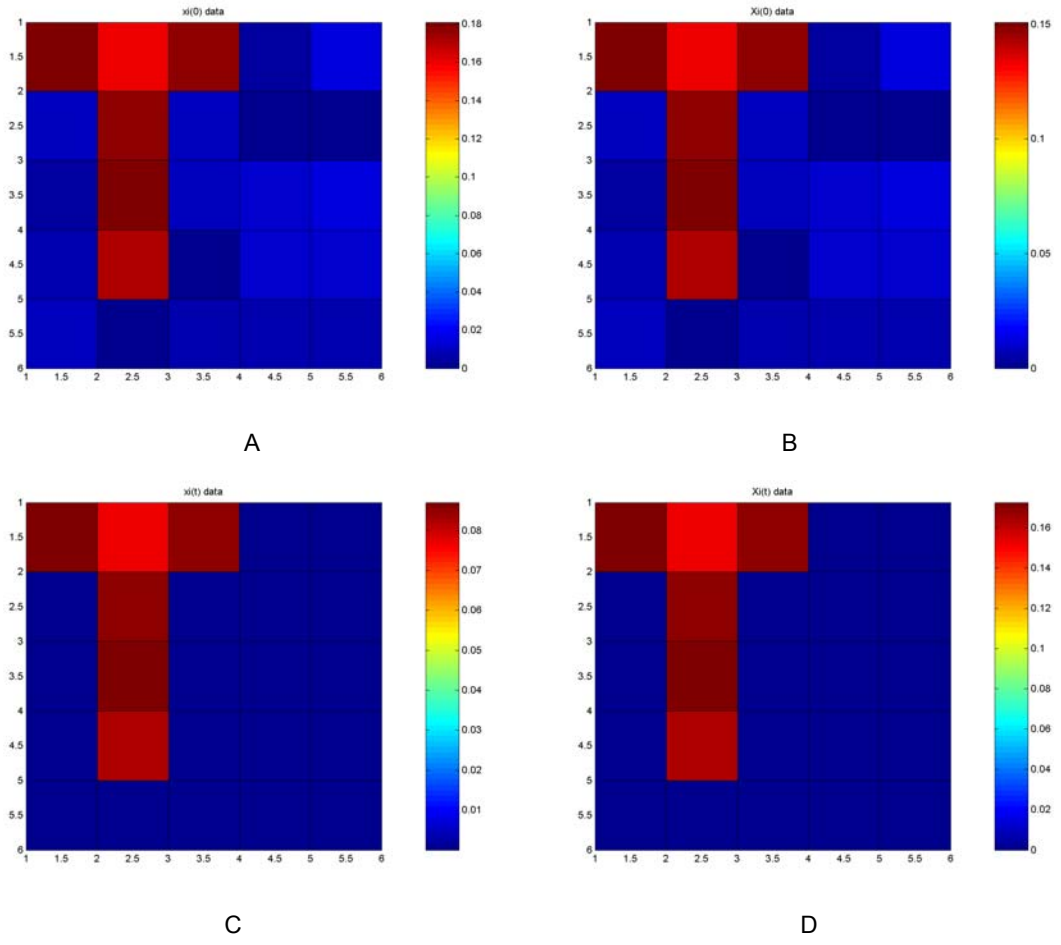
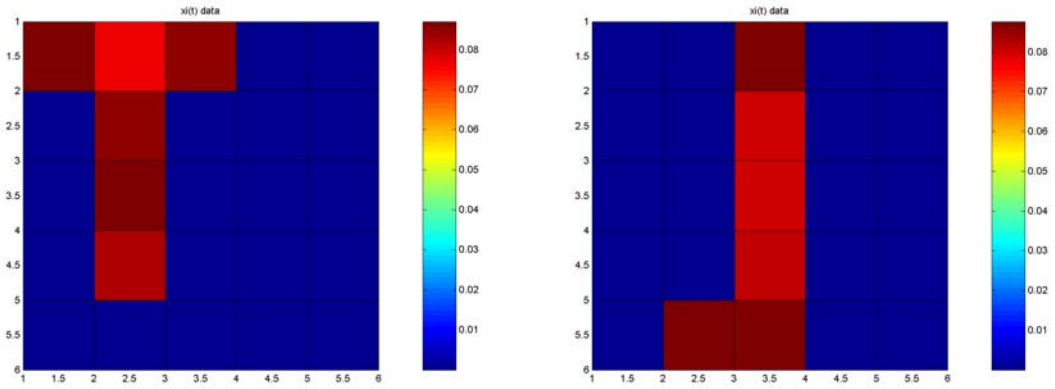


Figure 15.10: Simulation results for the CE$^{(1)}$ network at the end of the charge-up period and at the end of the discharge period. (A) x_i values at end of charge-up. (B) X_i values at end of charge-up. (C) x_i values after discharge. (D) X_i values after discharge.



A B

Figure 15.11: CE$^{(1)}$ response to a "J" pattern input overwriting the previous "T" pattern in STM. (A) Initial x_i variables at start of simulation. (B) Final x_i variables after application and removal of a noisy "J" pattern.

Figures 15.10C and 15.10D show the x_i and X_i results after the discharge period. The background noise is completely quenched, i.e. the signals belonging to set b are zero. Nothing remains except the T-pattern itself, although as can be clearly seen in the figures there is some residual noise remaining in the T-pattern signals.

Inspection of figure 15.10C reveals a possible cause of concern. There is some very noticeable attenuation in the absolute levels of the x_i signals. The CE$^{(1)}$ network would typically be a sub-system within a much larger network system model, and so attenuation of this sort could potentially be a problem for the overall system. For example, in an extreme case the $\{x_i\}$ pattern might altogether fall below the quenching threshold of some downstream network. This potential problem can, fortunately, be easily dealt with. What one does is use parameter B to control the absolute signal levels, and then ***scale*** the other system parameters relative to B. Proper scaling will ensure the performance of the CE$^{(1)}$ network will be altered in no way other than the absolute levels of the x_i signals. For the case of the CE$^{(1)}$ system described here, the scaling rules are as follows. Let $B = 1$ be the reference (unscaled) system with parameters A, etc. The scaling transformations

$$\begin{aligned} 1 &\Rightarrow B \\ A &\Rightarrow B \cdot A \\ g_{\max} &\Rightarrow B \cdot g_{\max} \\ u^{(2)} &\Rightarrow B \cdot u^{(2)} \\ u^{(1)} &\Rightarrow u^{(1)} \cdot (B - A/g_{\max})/(1 - A/g_{\max}) \end{aligned} \qquad (15.24)$$

will preserve the performance of the CE$^{(1)}$ network in its entirety except for the signal amplitudes. Note that the scaling rule for $u^{(1)}$ leaves the quenching threshold QT unchanged.

Finally, persistent reverberation in CE$^{(1)}$ after the input pattern is removed means the network

will not spontaneously decay back to a relaxed state. However, the arrival of a new input pattern of sufficient intensity (strong enough to overcome the QT) will overwrite the old x_i pattern in short term 'memory' (STM). Figure 15.11 illustrates this. Figure 15.11A is the STM pattern resulting from an initial "T" input retina (applied for 400 iteration steps followed by a 400 step discharge after the "T" was removed). Figure 15.11B shows the final x_i resulting from the application of a noisy "J" input retina (applied for 400 iteration steps, followed by a 400 step discharge after removal of the input). The figures clearly show that the stored "T" is erased and replaced by the new STM "J" pattern. $CE^{(1)}$ does not require a reset, as the MAXNET and Mexican Hat networks do, to operate on new input data provided that data is large enough to overcome the quenching threshold. If it is not, $CE^{(1)}$ will retain its original STM after the second pattern is removed, i.e. it ignores subsequent weak subthreshold input patterns. Thus it preserves the strongest recent input pattern, provided that pattern exceeded the QT.

§ 5. The Grossberg Classifiers $CL^{(1)}$ and $CL^{(2)}$

The Grossberg classifier is the ultimate level of on-center/off-surround SNI network prior to moving up to full-blown ART networks. Indeed, it was the analysis of "learning" stability – or, more accurately, "learning" *in*stability – for this network in [GROS5] that set the stage for the theory of ART networks in [GROS6]. Figure 15.12 illustrates the simplest of these classifiers, which we will call $CL^{(1)}$. Almost everything we need to understand in order to appreciate how an ART network must function we can learn from the characteristics of the Grossberg classifiers.

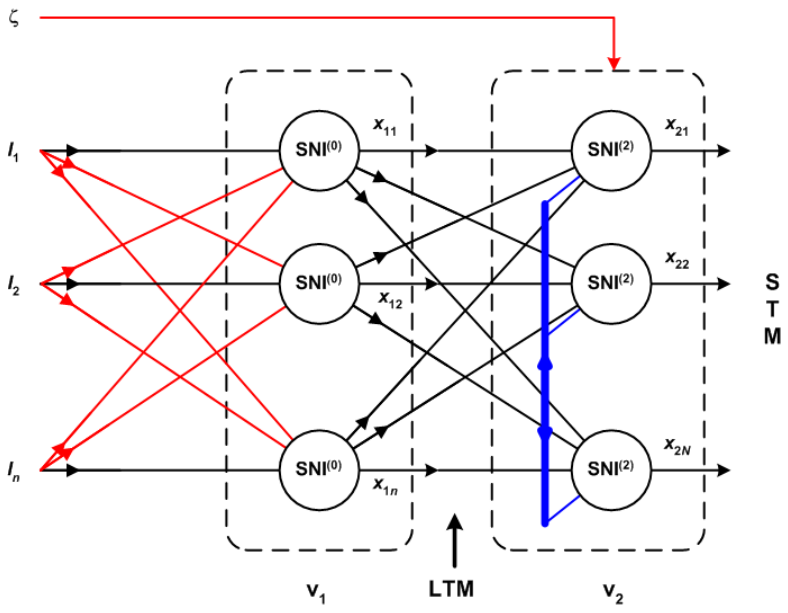


Figure 15.12: Basic Grossberg classifier $CL^{(1)}$. Red-colored connections are inhibitory, black are excitatory. STM is 'short term memory.' LTM is 'long term memory.' ζ is a non-specific inhibition applied to v_2.

The basic classifier is composed of two layers, v_1 and v_2. Layer v_1 is a Grossberg normalizer, either $GN^{(1)}$ or $GN^{(2)}$. Figure 15.12 uses $GN^{(1)}$. This layer has n $SNI^{(0)}$ maps, one for each input tract I_i. Its output pattern is a vector $\Theta = [\theta_1 \;\; \theta_2 \cdots \theta_n]^T$ where

$$\theta_i = \frac{B_1 \cdot I^{(1)}}{A_1 + I^{(1)}} \cdot \frac{I_i}{I^{(1)}}, \quad I^{(1)} = \sum_{i=1}^{n} I_i \tag{15.25}$$

for the case where v_1 is the $GN^{(1)}$ network. An expression of the form (15.12) results for $GN^{(2)}$.

The v_2 layer is a contrast enhancer with N $SNI^{(2)}$ nodes, $CE^{(1)}$ in the simplest case. However, its main job in $CL^{(1)}$ is to ***categorize*** input signal vectors $\{I_i\}$. For this purpose, typically $N < n$, and the total number of non-overlapping categories[3] that can be represented by v_2 is N. Note from figure 15.12 how each $SNI^{(2)}$ receives projections from each $SNI^{(0)}$. These projections are ***weighted*** so that the contribution I_{ij} from the *i-th* node in v_1 to the *j-th* node in v_2 is given by $I_{ij} = w_{ij} \cdot \theta_i$. Letting $W_j = [w_{1j} \;\; w_{2j} \cdots w_{nj}]^T$, the total input to the *j-th* node in v_2, corresponding to one of the I_j inputs in figure 15.6, is given by $I_{(2)j} = [\theta_1 \;\; \theta_2 \cdots \theta_n]^T \cdot W_j$ or, in vector form, $I_{(2)j} = \Theta^T \cdot W_j$. The overall N-element vector input $I_{(2)}$ presented to v_2 is written

$$I_{(2)} = \begin{bmatrix} W_1^T \\ \cdots \\ W_2^T \\ \cdots \\ \vdots \\ \cdots \\ W_N^T \end{bmatrix} \cdot \Theta = \mathbf{W} \cdot \Theta \,. \tag{15.26}$$

The matrix of weights **W** is called the ***long term memory*** or LTM of the network.

Layer v_2 is also modified to include a non-specific inhibitory input signal ζ. Its dynamical equation is thereby modified and is given by

$$\dot{x}_{2j} = -A \cdot x_{2j} + \left(B - x_{2j}\right) \cdot \left(f\left(x_{2j}\right) + I_{(2)j}\right) - x_{2j} \cdot \sum_{k \neq j} f\left(x_{2k}\right) - \zeta \,. \tag{15.27}$$

This inhibitory input is applied to all the nodes in v_2 and is used to clear the STM storage of the node.

When $\zeta = 0$, v_2 responds to the $I_{(2)}$ input vector in precisely the same way as $CE^{(1)}$ responded to its input vector in section 4. The *functional* difference between v_2 and $CE^{(1)}$ lies with how we interpret this new $I_{(2)}$ vector. $CL^{(1)}$ is often called a ***feature detector***, and to understand this it is important we first understand the idea of a ***feature***. In psychology a feature is an attribute of something (usually an object or event) that is critical to distinguishing that thing from other

[3] In ART terminology, categories are said to be non-overlapping when exactly one v_2 SNI represents the category, with the other SNI nodes having their $x_j = 0$. It is a form of "grandmother cell" encoding.

things. (For this reason it is also called a ***distinctive feature***). For example, the "distinctive feature" of a triangle is that it has three sides.

Now, this is a nice, intuitively-agreeable, and *vague* description. As it stands, this dictionary "definition" of "feature" cannot be reduced to mathematics – which is another way of saying it is vague since mathematics can be regarded as a language for saying things in a very, very precise way. What we need to be able to grasp is, so to speak, "what is the feature that distinguishes what a 'feature' is?" Phrased this way, one can easily be alert to the potential we have here for ending up with a circular definition. To avoid this and to attain to a specific *meaning* for this word, the approach one needs to take is to define 'feature' in terms of something we hold to be a *consequence* to which a 'feature' must lead. This is a *practical definition*, i.e. a definition that can be put into practice. Making the translation from equivocal language to mathematical language in this instance was handsomely accomplished in Grossberg's early work on ***embedding fields***.

> When a standard English speaking adult hears a word spoken or speaks a word himself, the word seems to occur at a single instant of time. That is, we can say either that the word has, or has not, been said at a given time in a perfectly definite way. Moreover, no more than a finite number of words are spoken in a lifetime. Thus, both "spatially" (the number of verbal units) and "temporally" (the number of time instants at which verbal units occur), language seems to have many properties of a finite, or discrete, phenomenon.
>
> One of the most vital uses of language is to report our sensory experiences, such as variations in tactile pressure, light intensity, loudness, taste, etc. Many of these sensory impressions seem to vary in a continuous way both in space and in time. A basic characteristic of much sensory experience is that it seems to be *spatio-temporally continuous*. . . The representation by language of sensations requires that the two kind of phenomena interact, and so, mathematically speaking, we must envisage the interaction of spatio-temporally discrete and continuous processes of such a kind that the relatively discrete process provides an adequate representation of the relatively continuous process. Moreover, although each sensory modality seems to provide us with essentially different varieties of experience, the very same language tools are adequate for describing at least the rudiments of all of these various modalities. Thus, the discrete representation of continuous processes must be a *universal representation* of some kind [GROS2].

This is the first crucial consideration for the practical meaning of a feature, namely that it is produced as a discrete representation of a continuous process. But this, by itself, is not sufficient.

> Since different behavioral sequences in different stages of learning can often coexist, all intermediaries between continuity and discreteness can in principle coexist at any time. . . Properties of discreteness and continuity coexist at every stage of learning. The continuous background is never wholly eliminated. We must study how certain processes superimposed on this background become increasingly discrete relative to an initially prescribed standard of continuity, and will have at our disposal at least two different levels of dynamical graining such that the degree of continuity of one level takes on meaning only relative to the degree of continuity of the other [GROS2].

The eventual logical endpoint of this requirement of "increasingly discrete processes super-imposed on a continuous background" is the idea of ***context***. That is to say, a "feature" *is* a feature

only with respect to some sort of more continuous context. A year earlier (1968) Grossberg already had in hand an approach for saying this mathematically:

> The psychological postulates that lead to the equations which describe our learning machines *M* are quite simple. The following discussion heuristically describes these postulates in the case of learning a list of "simple" letters or events, such as the alphabet *ABC...Z*.
>
> (a) The letter *A* is never decomposed into two or more parts in daily speech and listening. It is a "simple" behavioral unit. Thus we assign to every simple behavioral unit r_i a single abstract point v_i in *M* . . .
>
> (b) *M* must react to presentation of behavioral units at specified times. Hence a real-valued function of time $x_i(t)$ is assigned to each point v_i. The value of $x_i(t)$ at any time describes how recently r_i has been presented to *M*.
>
> (c) Consider *M's* response to presentation of *A*, then *B*, and then *C* at a speed ω. If ω is small (say $\omega \cong 2$ sec), then the influence of *A* and *B* to *M's* response to *C* is substantial. As ω increases the influence of *A* and *B* on *M's* response gradually changes and ultimately becomes negligible. Since the effects of prior presentations of events wear off gradually, each $x_i(t)$ is continuous. . .
>
> (f) *Before M* has learned the list *AB*, other responses than *B* to *A* must exist, or else *B* would already be the only response to *A*. Thus a function $w_{AB}(t)$ exists which can distinguish the presentation or non-presentation of *AB* and lets only *B* occur in response to *A* after *AB* has been learned. Since $w_{AB}(t)$ grows only if *A* and then *B* are presented to *M*, $w_{AB}(t)$ *correlates* (prescribed) past values of x_A with $x_B(t)$. $w_{AB}(t)$ therefore occurs at the only position at which past x_A and present x_B values exist, namely, at the end of the pathway leading from v_A to v_B [GROS18].

It would seem to be no coincidence that the notations used by Grossberg in 1968 and those he used to introduce ART in 1976 were in all essentials the same. Cutting to the bottom line here, we will say *the signal representations from layer v_1 contain representations of features and the relative intensity θ_i measures the relative importance of the feature coded by v_i in any given input pattern* [GROS17]. The signal inputs $I_{(2)j}$ to each node in v_2 are therefore *measures of how much each of these features are correlated with the discrete "unit"* represented by the *j-th* node of v_2.

In the language of feature-detector theory, each W_j is called a ***classifying vector***. Layer v_2 is said to ***recode*** the *n* feature representations presented by v_1, which in the language of the early papers quoted above, amounts to "making a signaling process increasingly discrete." To grasp the *system function* of a Grossberg classifier amounts to understanding three aspects of it: (1) How does the recoding work? (2) how is it controlled and (3) how stable is this coding? We will take on these questions one by one in the following sections.

§5.1 Feature Detector Coding

The *length* of any vector *V* of any number of dimensions is $|V| = (V^T V)^{1/2}$. The *direction* of a vector is defined to be the unit vector $V/|V|$. It is a fundamental property of vectors that for any two vectors *V* and *U* of equal dimension, $V^T U \le |V| \cdot |U|$ with equality if and only if either (1) one of the vectors has zero length or (2) both vectors have the same direction, i.e. $V/|V| = U/|U|$. It follows from this that one can always define an angle $\varphi_{V,U}$ such that $\cos(\varphi_{V,U}) = V^T U /(|V| \cdot |U|)$.

Applying these properties to the $I_{(2)j}$ terms defined above, $\Theta^T \cdot W_j = |\Theta| \cdot |W_j| \cdot \cos(\varphi_{\Theta,W})_j$. Thus if all weight vectors W_j have the same length but different directions, for all input patterns Θ_k of the same length, the maximum $I_{(2)j}$ will be the one for which $|\varphi_{\Theta,W}|$ is the least. Put another way, an input pattern belong to a set $\{\Theta_k$ such that $|\Theta_k| = |\Theta|\}$ will represent a total set of features "most like" that classifying vector W_j for which the magnitude of $\varphi_{\Theta,W}$ is least (under the condition that all W_j have the same length).

For the time being let us assume that the condition $|W_j|$ = constant for all $j = 1, \ldots, N$ is ensured by some mechanism of the network system. (We will later come back to look at this assumption and determine if and how well it can be realized). The issue then becomes: *How much "like"* the classifying vector must a Θ vector be in order for v_2 to classify Θ as belonging to the input space partition defined by the classifying vector?

Here it is helpful to contrast $CL^{(1)}$ against the Instar-MAXNET classifier of chapter 14. The conditions just stated above are the same as those which we said were necessary for the Instar-MAXNET to function as a useful classifier. In the case of that network, we recall that the competition of the MAXNET layer is winner-take-all. Regardless of how "close" two of the MAXNET inputs may be, the competition will produce a single winner (unless there is a tie, in which case it produces *no* winner). Thus, the classification choice is entirely based on the *relative* intensities of the MAXNET inputs.

The $CE^{(1)}$ competition is of a very different sort. As we saw in section 4, every v_2 node that receives a sufficiently strong input to overcome the quenching threshold will set up a persistent reverberation in v_2 and be stored in STM after Θ is removed. A sufficient condition for this is

$$X_1(0) \cdot \min\{x(0),\, B - A/g_{\max}\} \geq u^{(1)}.$$

Let us again use Grossberg's enumeration and suppose that only X_{2N} satisfies the conditions of theorem 4 and no other X_{2j} violates the condition of theorem 8. In this case, only x_{2N} is stored in STM, all other x_{2j} decay to zero, and the resulting final distribution after removal of $I_{(2)}$ is called a **0-1 distribution**. This is the Grossberg classifier's version of a winner-take-all result. Furthermore, we have $x_{2N} = x = B - A/g_{\max}$.

There is also another condition, expressed in another Grossberg theorem in the 1973 paper, for the occurrence of a 0-1 distribution:

> **Theorem 12:** Let $g(u)$ be continuous, non-negative and strictly monotone increasing, and let the X_i be enumerated according to Grossberg's enumeration. If $X_n(0) = X_1(0) = 1/n$ then $X_n(t) = X_1(t)$ for all $t > 0$. Otherwise $X_n(t)$ is monotone increasing faster than any other function $X_i < X_n$ and X_1 is monotone decreasing. Furthermore, if the reverberation is persistent then the limiting distribution is either 0-1 or locally uniform.

Neither distribution in figure 15.7 meets the condition of $g(u)$ being strictly monotone increasing. However, if the input pattern is not uniform, $x_N > x_{i<N}$, and no $x_i(t)$ ever grows large enough to reach $u^{(2)}$, the situation is effectively the same as that of theorem 12 and a 0-1 distribution will result, provided the input is strong enough to evoke STM, even if the maximum $X_i(0)$ is not large enough to reach the quenching threshold. Figure 15.13 illustrates $x(t)$ for an example of this.

One of the interesting features of this $x(t)$ is the minima that occurs at approximately the 450-th time step in the iteration. Here $x(t)$ has momentarily dipped below the steady-state value for persistent reverberation and is now beginning to climb again. This is a consequence of the largest $X_j(t)$ beginning to grow as the lower-valued nodes in the network continue to decrease. In other words, the dip in the curve illustrates the energy-redistributing property of the $CE^{(1)}$ network. If the input signal been sufficiently less intense, the reverberation would have been transient.

Because the outputs of v_1 are normalized and we are assuming all the W_j have different directions, the occurrence of a 0-1 distribution in STM for sufficiently large inputs can be guaranteed by setting up a large QT through the selection of parameter $u^{(1)}$. Because the input signals can be written as $I_{2j} = |\Theta| \cdot |W_j| \cdot \cos(\varphi_{\Theta,W})_j$, the setting of QT determines the maximum angle $\varphi_{\Theta,W}$ by which Θ can differ from W_j in order for the feature set represented in Θ to be classified by W_j. However, because of the weak form of theorem 15.12 it is not easy to state what the minimum level of $|\Theta|$ must be to sustain persistent reverberation, and therefore not a simple matter to state the quenching level or $\varphi_{\Theta,W}$ in terms of $I^{(1)}$. Further complicating the situation is the

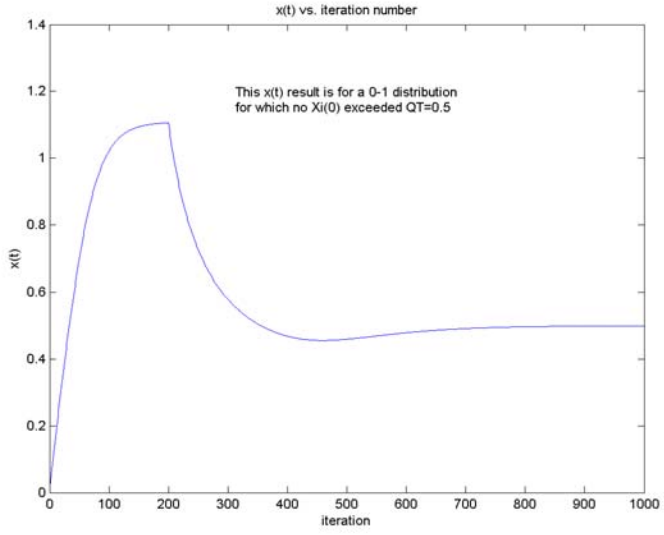


Figure 15.13: $x(t)$ vs. t for a 0-1 distribution case under the weak form of theorem 15.12.

fact that $x(0)$ is a function of how long the stimuli I_{2j} are applied in charging up v_2, and therefore whether or not STM is evoked has the duration of the stimulus as another of its factors.

The re-coding of the $CL^{(1)}$ input by means of a 0-1 distribution is called a ***compressed code***. In contrast to the classifiers of chapter 14, $CL^{(1)}$ does not attempt to partition the entire input signal space into contiguous decision regions. Rather, each W_j classifies a ***convex cone*** [GROS5]

$$P_j = \left\{ \Theta \text{ such that } \Theta^T W_j > \max\left(\varepsilon, \Theta^T W_k, k \neq j\right) \right\}$$

where ε is the minimum value of I_{2j} required to exceed the quenching threshold. Inputs for which no I_{2j} exceeds the quenching threshold do not produce STM (the reverberation is transient) and are left *unclassified* (uncoded) by $CL^{(1)}$.

It is also possible to arrange the set of W_j vectors in such a way that more than one v_2 node might exceed the QT in response to one or more Θ vectors. This results in ***partial contrast*** in the STM output (more than one x_{2j} being non-zero). In this case, v_2 is said to have a ***tuning curve***, i.e. a maximal response to certain input patterns and sub-maximal responses to others [GROS5]. Although this is the more general case, in the sense that a 0-1 distribution can be regarded as just a special case of it, in practice tuning curves were not used by the earliest practical versions of ART networks, ART1 and ART2 [CARP4].

§5.2 Weight Adaptation and the Issues with $CL^{(1)}$

Grossberg classifiers use the Instar adaptation rule (IAR) to adapt the weights. Indeed, the IAR was developed by Grossberg in his work that led up to adaptive resonance theory. In the notation for $CL^{(1)}$,

$$\dot{W}_j = \eta \cdot \left(\Theta - W_j\right) \cdot x_{2j} \tag{15.28}$$

where η is the learning rate parameter, $0 < \eta < 1$. In difference equation form this becomes

$$W_j\left(t + \Delta t\right) = W_j\left(t\right) + \Delta t \cdot \eta \cdot \left(\Theta\left(t\right) - W_j\left(t\right)\right) \cdot x_{2j}\left(t\right). \tag{15.29}$$

Generally speaking, η should be small enough so that weight changes are dominated by stable STM patterns. (This requirement for small η is consistent with the biology of postsynaptic LTP/LTD phenomena, which likewise take place slowly through the metabotropic signaling mechanisms we discussed in chapter 12). Otherwise the changes in *all* the W_j that would take place while v_2 is charging and discharging would interact with these dynamics, leading to unpredictable outcomes and almost surely guaranteeing that no *stable* encoding of the features in Θ would result. This is because changing W_j changes $I_{(2)j}$, which in turn affects the x_{2j}. In the ART

literature it is often presumed that (15.28) is applied after the STM pattern at v_2 stabilizes.

Signal x_{2j} acts as a "gating mechanism" for adaptation in (15.28-15.29). Suppose the final STM pattern distribution in v_2 is the 0-1 distribution. Then all the x_{2j} except one are zero. The non-zero x_{2j} then "gates on" the adaptation for the W_j in its Instar, while change in the W_k of the other Instars is prevented. This is effectively the same type of partitioning control we saw in chapter 14 for other types of competitive networks. Furthermore, we note that due to persistence in the STM, the adaptation of W_j continues all the while the STM persists.

As we saw in our earlier discussions of the IAR, W_j will adapt in such a way as to converge asymptotically to the input vector Θ. If the STM distribution is 0-1, only the "chosen" SNI will encode Θ in its weights. If the distribution is not 0-1 (that is, if there is partial contrasting rather than "choice" in the v_2 output pattern), then all the non-zero output SNI nodes will adapt in the "direction" of setting their weights equal to Θ. This is not necessarily a bad thing, but it is full of potential for producing bad results, e.g. the convergence of two or more SNI weight vectors to the same value. The more advanced case of partial contrast ("partially compressed") codes is discussed by Carpenter and Grossberg in [CARP4].

Assuming the STM is a 0-1 distribution, let $\{\Theta_1, \Theta_2, \cdots \Theta_J\}_j$ be the set of input pattern vectors that result in $x_{2j} \neq 0$, $x_{2k} = 0$ ($k \neq j$). Let every other SNI node have its own characteristic set $\{\Theta_1, \Theta_2, \cdots \Theta_K\}_k$ of input vectors to which it responds. We will say a ***stable encoding*** of the input patterns exists if the intersect $\{\Theta_1, \Theta_2, \cdots \Theta_J\}_j \cap \{\Theta_1, \Theta_2, \cdots \Theta_K\}_k$ is the empty set for every pair j, k with $k \neq j$.[4] Throughout this and the next section we will assume a stable encoding exists and we will examine the validity of this assumption in section 5.4.

Under these conditions, if the underlying statistics of the input signals are stationary and the successive presentations of inputs $\Theta(t)$ are statistically independent, then each W_j will converge to the expected value $E[\{\Theta_1, \Theta_2, \cdots \Theta_J\}_j]$ of its partition of the Θ input space. We saw the proof of this earlier in the text.

This adaptation dynamic is similar in many ways to that of the RBF-MAXNET network of chapter 14. For the RBF-MAXNET, if the input vector is changed without the MAXNET layer being reset for a new tournament, the selected RBF-Instar will start to learn the *new* input vector, to the detriment of its learning of the input vector that "won" the competition. Thus, *the resetting of the MAXNET layer must be coordinated with changes in the input signal and adaptation must be disabled after every change in input until a new tournament is run*. Of course, if the new input

[4] In our notation, the Θ_1 in the $\{...\}_j$ set is not equal to the Θ_1 in the $\{...\}_k$ set, etc. The notation is an abbreviation that saves us from having to write $\Theta_{1(j)}$ vs. $\Theta_{1(k)}$ etc.

to the RBF-MAXNET is significantly different, such that the radial basis function activation of the RBF-Instar falls to a low level, undesired learning will be hampered by the low value of activation that follows the input change. But if, as in the examples presented in chapter 14, the coverage region of the radial basis function significantly overlaps nearby partitions of the input space, then the activation function will often enough *not* be small and the selected RBF-Instar will infringe upon the encoding that should belong to one of its neighbors, leading to *unstable encoding*. To prevent this requires the intervention of the attentional subsystem in figure 13.3.

Now let us consider the application of the IAR to $CL^{(1)}$. Here we find that some special considerations must be applied. First, recall from the earlier discussion of $CE^{(1)}$ that the non-homogeneous equation produces nonzero values of x_{2j} whenever the node's input $I_{(2)j}$ is non-zero. Thus the $CE^{(1)}$ layer ***cannot produce a 0-1 distribution until after Θ is removed***. While the stimulus is being applied, *all* the x_{2j} values are nonzero for which the $I_{(2)j}$ are nonzero. We will look at two situations: (1) quenching threshold set low enough that a 0-1 distribution develops in STM after Θ is removed; (2) quenching threshold set high enough that no STM is maintained after Θ is removed.

Figures 15.14 illustrate the first case. The parameters of the system are $B = 1$, $A = 0.5$, $g_{\max} = 1$ and $u^{(1)} = 0.5$ (giving a quenching threshold of QT = 1). The input pattern is applied for 400 time steps and then set to zero thereafter. As figure 15.14B illustrates, the $CE^{(1)}$ layer produces a post-stimulus STM in response to the signal. As we shall soon see, this is an undesirable behavior for adaptation of a Grossberg classifier and will necessitate the use of the reset signal ζ if the classifier weights are adapted.

The values of the x_{2j} nodes at the 400-th iteration step are shown in figure 15.14C. As we can see, all the non-zero inputs produce non-zero responses in their respective x_{2j} values. By comparing figures 15.14A and C, we can also see that significant contrast enhancement has taken place, and that the x_{2j} node that received the largest input stimulus is significantly larger than the other x_{2k} nodes. It is this node that survives in the 0-1 distribution that follows the removal of the input stimulus, figure 15.14D.

This contrast enhancement in figure 15.14C obviously suggests a simple modification to the IAR if we wish to have the adaptation performed with a 0-1 distribution. The idea is to introduce an adaptation threshold into the IAR, i.e.

$$w_{ij}(t+\Delta t) = w_{ij}(t) + \Delta t \cdot \eta \cdot \left(\theta_j(t) - w_{ij}(t)\right) \cdot h\left(x_{2j}(t) - \Omega\right) \tag{15.30}$$

where Ω is the adaptation threshold and $h(u)$ is the Heaviside extractor function (15.10).

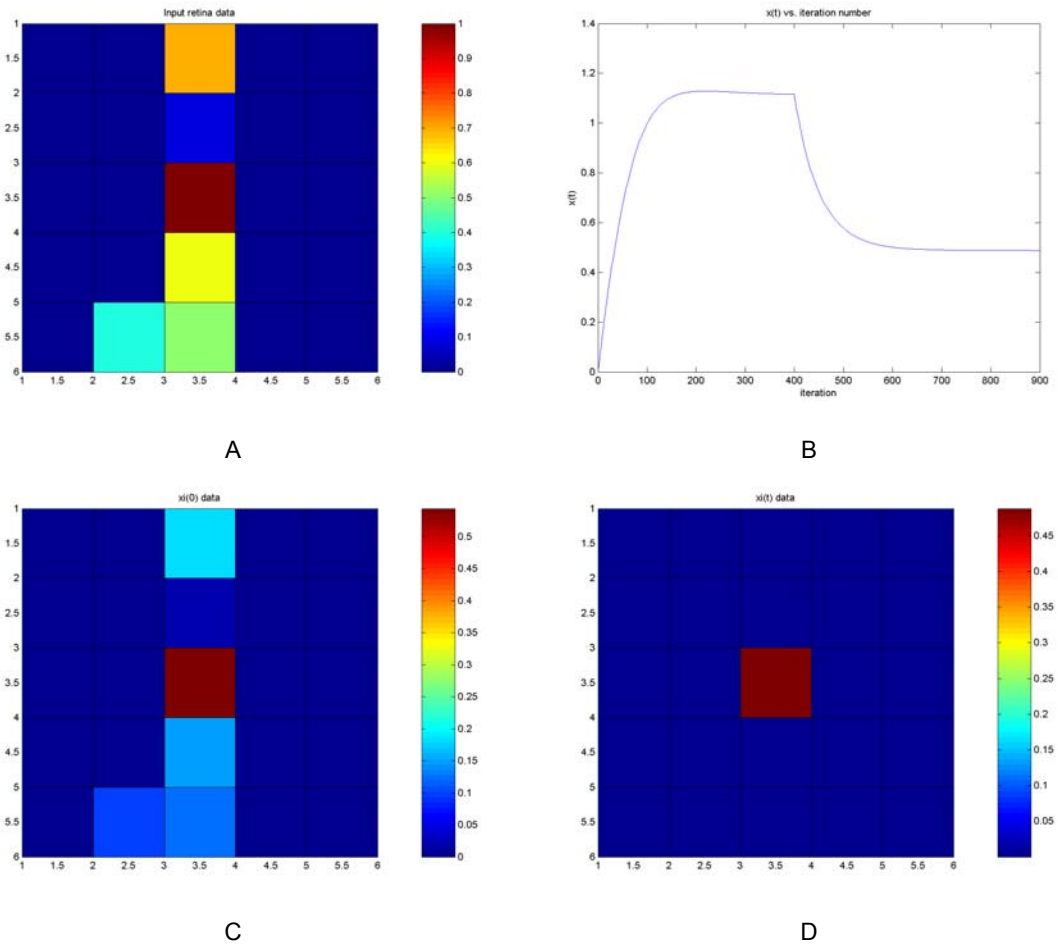
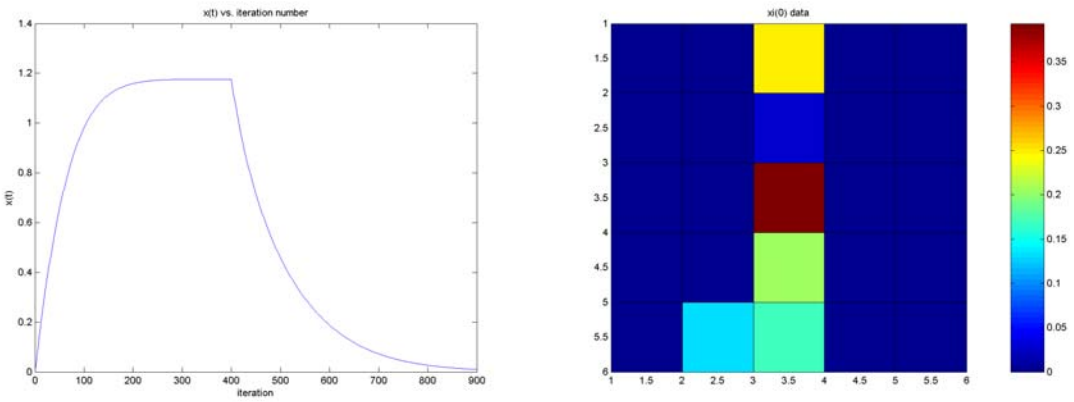


Figure 15.14: CL$^{(1)}$ layer v_2 input pattern, $x(t)$ curve, and x_{2j} values while $\Theta \neq 0$ (C) and $\Theta = 0$ (D).

After the stimulus is removed the Θ in the IAR expression (15.30) is zero. However, because of the persistence of the STM, if the adaptation of (15.30) is allowed to continue then W_j will begin to "learn" the all-zeros input pattern. This is clearly undesirable, and so the reset signal ζ must be asserted to abolish the STM before the encoding of W_j becomes compromised. In network systems like those of chapter 14, this would be the function of the attentional subsystem we saw was necessary for those networks. In the ART literature, Grossberg and his colleagues generally refer to this function as an ***arousal mechanism*** (he reserves the term "attentional subsystem" to mean a specific anatomy in a full-blown ART system, which CL$^{(1)}$ is not). [GROS5] discusses arousal mechanisms in terms of inputs insufficient to evoke STM. The function of such a mechanism as discussed there is aimed at triggering a "search" for a suitable classification of the input. But, as we can see here, there is a dual side to this, which we might term an "anti-arousal" mechanism. Alternatively, arousal might be used to *enable* adaptation.

Because this issue arises because of the post-stimulus STM persistence, an obvious question is



A B

Figure 15.15: $CL^{(1)}$ layer v_2 response with QT raised high enough to suppress post-stimulus STM.

"Why allow post-stimulus STM to occur at all?" The reason it occurs in this example is because the QT is set low enough to allow persistent reverberation in v_2. This persistence can be abolished by raising the QT, which is easily accomplished by increasing $u^{(1)}$. What happens when we do this?

Figure 15.15 illustrates this second case. Here all system parameters, including the input pattern to v_2, are the same except for $u^{(1)}$, which is raised to $u^{(1)} = 0.90$. The $x(t)$ plot, figure 15.15A, illustrates that post-stimulus STM is indeed suppressed. The price that is paid for this is seen by comparing the x_{2j} values at time step 400, figure 15.15B, with their counterparts in figure 15.14C. It is clear that much less contrast enhancement is present in 15.15B, and therefore it becomes significantly more difficult to find a reliable adaptation threshold Ω to achieve the effect of adaptation under a 0-1 distribution. *The same dynamics that abolished post-stimulus STM also place the stimulated response in the region of operation where a* ***fair distribution*** *is generated*.

In general this is a much more severe and difficult problem than is the problem of setting up an attentional ("arousal/anti-arousal") subsystem for the first case. This is not to say the attentional subsystem problem for the persistently reverberating version of $CL^{(1)}$ is trivial. To say so is to underestimate the complexity issues that can arise when $\Theta(t)$ merely changes to a *different* but non-zero input value, especially when that value does not have sufficient strength to abolish the STM and replace it with its own. ($\Theta = 0$ is merely the extreme case of this). In the next section we will look at a far easier solution to these problems, namely the system $CL^{(2)}$.

To illustrate this, let us look at another interesting aspect in the dynamics of $CL^{(1)}$. In the discussion above we assumed $\Theta(t)$ was applied long enough to establish a non-homogeneous steady-state response by v_2. As it turns out, this is not a necessary condition for establishing an STM. Let us suppose a $\Theta(t)$ is applied briefly and for a time not long enough for v_2 to come to the

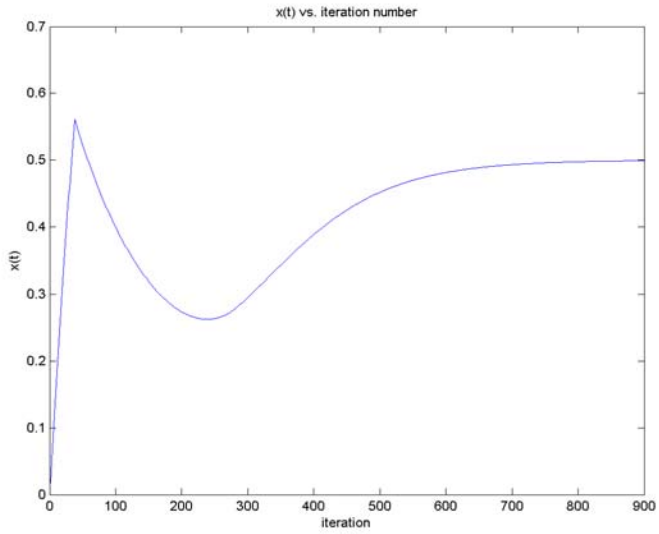


Figure 15.16: Stimulation of persistent reverberation (STM) by a brief pulse of input stimulus.

plateau pictured in figure 15.14B. If the magnitude of $\Theta(t)$ is sufficiently large, it is possible for the redistribution of energies in v_2 to produce an STM anyway. An example of this is shown in figure 15.16. The input vector was applied to an initially relaxed v_2 for 38 iteration steps and then returned to an all-zeros level. The reverberations set up in v_2 by this brief stimulus nonetheless succeed in establishing a 0-1 final distribution even in the absence of continued stimulus-driven excitation.

These considerations – the lag between Θ application and removal, the variety of different charge-up and discharge times that pass before a 0-1 distribution is established, and general control of the adaptation process – must lead us to consider *control structures* for an adaptive $CL^{(1)}$ feature detector. The situation here is not unlike the one we encountered with the RBF-MAXNET classifier earlier. The principal difference is that for $CL^{(1)}$ the situation is much harder to analyze. These issues that face $CL^{(1)}$ are brought about by the nature of the dynamics of contrast enhancer $CE^{(1)}$. While $CE^{(1)}$ is a very good network in its own proper sphere of application, namely non-adaptive contrast enhancement, it is not so well suited to serve the feature *learning* function of a Grossberg classifier. And this brings us to our next level of refinement – one that will lead us directly into ART networks in the next chapter.

§5.3 $SNI^{(3)}$ and Grossberg Classifier $CL^{(2)}$ with Large Surround Inhibition

The solution to the problems encountered in the previous section involves so tiny a change it is understandable if one feels stunned that so substantial a change in network behavior can come from such a seemingly small difference. We will make a small modification to the SNI.

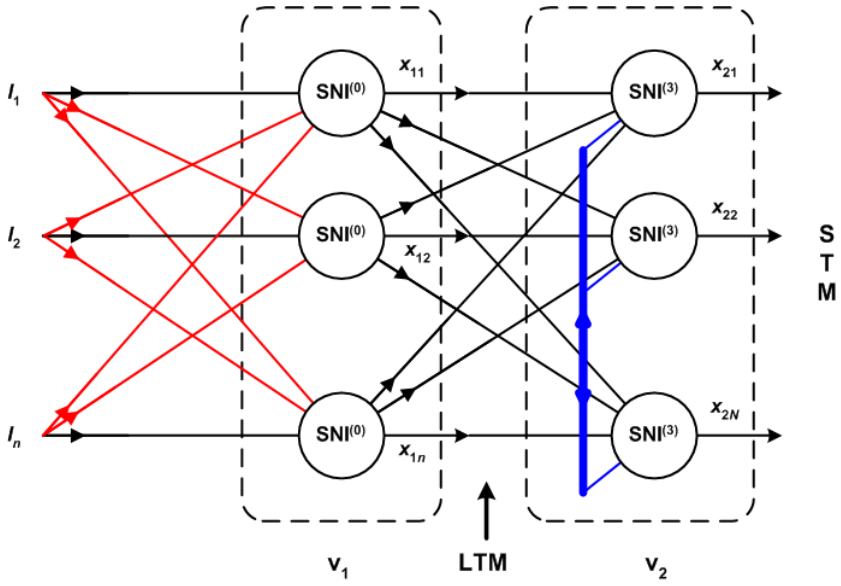


Figure 15.17: Grossberg Classifier $CL^{(2)}$.

The fundamental cause of the problems we have just seen all stem from one source: the inability of $CE^{(1)}$ to produce a 0-1 distribution when the stimulus is non-zero. In turn, this inability stems from the nature of the dynamics of shunting node Instar $SNI^{(2)}$. Let us consider an alternate form of SNI, which we will call $SNI^{(3)}$. The dynamical equation for $SNI^{(3)}$ is

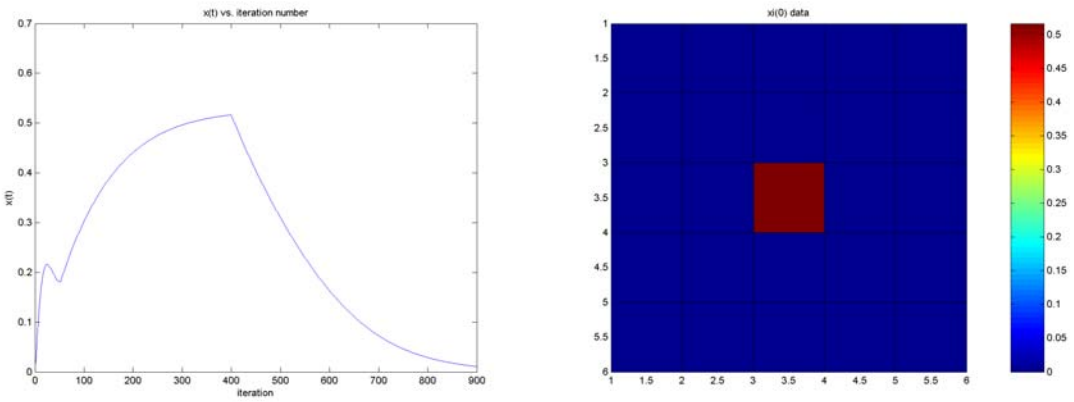
$$\dot{x}_{2j} = -A \cdot x_{2j} + \left(B - x_{2j}\right) \cdot \left(f\left(x_{2j}\right) + I_{(2)j}\right) - \left(x_{2j} + D\right) \cdot \sum_{k \neq j} f\left(x_{2k}\right). \tag{15.31}$$

This equation is identical to that of $SNI^{(2)}$ except for the addition of the constant D in the right-most term, $D \geq 0$. Grossberg is fond of likening D to the Nernst potential for potassium in neurons, but this is mere romance. D increases inhibition from the off-surround nodes in v_2. When we replace the nodes in v_2 with $SNI^{(3)}$ we obtain classifier $CL^{(2)}$ as depicted in figure 15.17. First we will examine a large-D case ($D = 49 \cdot B$); then we will look at small D effects.

The effect of the D term is most clearly seen in the steady state response. Setting the derivative in (15.31) to zero, we obtain

$$x_{2j} = \frac{B \cdot \left(f\left(x_{2j}\right) + I_{(2)j}\right) - D \cdot \left(F - f\left(x_{2j}\right)\right)}{A + F + I_{(2)j}} \tag{15.32}$$

where F is defined as before. The steady state solution x_{2j} is now capable of equaling zero in the presence of stimulus input $I_{(2)j}$ and therefore $CL^{(2)}$ is capable of producing a 0-1 distribution in response to non-zero stimulation. Moreover, the steady state solution for x_{2j} is now formally capable of being negative.



A B

Figure 15.18: Response of $CL^{(2)}$ to the previous input signal example. The parameter settings for this simulation were $D = 49$ and $u^{(1)} = 0.95$ with all other parameters the same as in the previous examples.

The possibility of negative values for x_{2j} presents the same interpretational difficulties as we previously encountered for $GN^{(2)}$, including the strange consequence that $x(t)$ can now be negative. Grossberg et al. commonly allow negative node values to occur in their networks. Negative-valued x_{2j} terms do not affect the dynamics of positive x_{2j} nodes because the activation function is zero for $x_{2j} < 0$. But there is no performance advantage in allowing negative node values and some analysis and interpretation advantages in preventing them. Therefore we will introduce clipping to ensure $x_{2j} \geq 0$ by writing the difference equation as

$$x_{2\,j}(t+\Delta t) = \max\left\{0,\quad x_{2\,j}(t) + \Delta t \cdot \left[-\left(A + F + I_{(2)j}\right)\cdot x_{2\,j} + B\cdot\left(f\left(x_{2\,j}\right) + I_{(2)j}\right) - D\cdot\left(F - f\left(x_{2\,j}\right)\right)\right]\right\} \tag{15.33}$$

This is equivalent to replacing the v_2 node variables by an activation variable $h(x_{2j})$ where h is the Heaviside extractor.

Figure 15.18 illustrates the response of this system to the example input pattern used for the examples in the previous section. Figure 15.18B gives the values of the x_{2j} terms, using (15.33), at the 400-th time step in the simulation. We can easily see that a 0-1 distribution has been produced. Figure 15.18A shows the time course of $x(t)$, and we see from this that post-stimulation STM has been abolished. Thus, with one small change to the SNI the principal problems we saw in the previous section have been removed.

We must still consider what the effect on the adaptation will be during the relatively long time course of the decay of $x(t)$ after the 400-th time step in figure 15.18A. Let us assume we use an unmodified IAR (that is, we do not employ the artifice of an adaptation threshold). With $\Theta = 0$, the adaptation difference equation becomes

$$W_j(t+\Delta t) = \left(1 - \Delta t \cdot \eta \cdot x_{2j}(t)\right) \cdot W_j(t)$$

and the length of the weight vector becomes

$$\left|W_j(t+\Delta t)\right| = \left(1 - \Delta t \cdot \eta \cdot x_{2j}(t)\right) \cdot \left|W_j(t)\right| .$$

The direction of the updated weight vector is therefore

$$\frac{W_j(t+\Delta t)}{\left|W_j(t+\Delta t)\right|} = \frac{W_j(t)}{\left|W_j(t)\right|} .$$

In other words, although the length of the weight vector relaxes during the decay of $x(t)$, but does not reach zero owing to the decay of $x_{2j}(t)$, the *direction* of the weight vector is *unaltered*. Here it is important for us to remember that the Grossberg classifier is based on the encoding of features in the *direction* of the input signal Θ. Thus, ***the relaxation of the weight vector in response to zero stimulus does not alter the encoding of the patterns by CL$^{(2)}$***.

We can obtain a ballpark estimate of how bad the decay in W_j may be from examining the difference equation

$$u(k+1) = \left(1 - a \cdot r^k\right) \cdot u(k) .$$

Here a corresponds to $\Delta t \cdot \eta$ and the geometric ratio factor r is chosen so that this difference equation is any reasonable approximation of the $x(t)$ decay curve in figure 15.18A. What is important to note is that for slow adaptation $a \cdot r^k << 1$. Solving the difference equation by recursion and manipulating this solution gives us

$$\lim_{k\to\infty} \ln\left(\frac{u(k)}{u(0)}\right) = \sum_{n=0}^{\infty} \ln\left(1 - ar^n\right) \cong \sum_{n=0}^{\infty} -ar^n = \frac{-a}{1-r}$$

from which we obtain $u_{\text{final}} \cong \exp(-a/(1-r)) \cdot u(0)$. For slow adaptation the argument of the exponential will have a magnitude of order unity or less, and so the decay of the weights is controllable even if CL$^{(2)}$ is exposed to a prolonged period of no stimulation. Even this decay can be reduced by employing an attentional subsystem as we did in chapter 14 for RBF-MAXNET.

In a sense, this process corresponds to a kind of gain control for the W_j vectors. During the adaptation process the W_j will move to, on the average, increasing values of $|W_j|$ and then "sag" when the stimulus is removed. It is also instructional to look at the dynamics when Θ(t) is not removed but rather changes to a different non-zero value. In this case, CL$^{(2)}$ will always respond to the second pattern because it lacks the ability for persistent STM. Figure 15.19 is an illustration

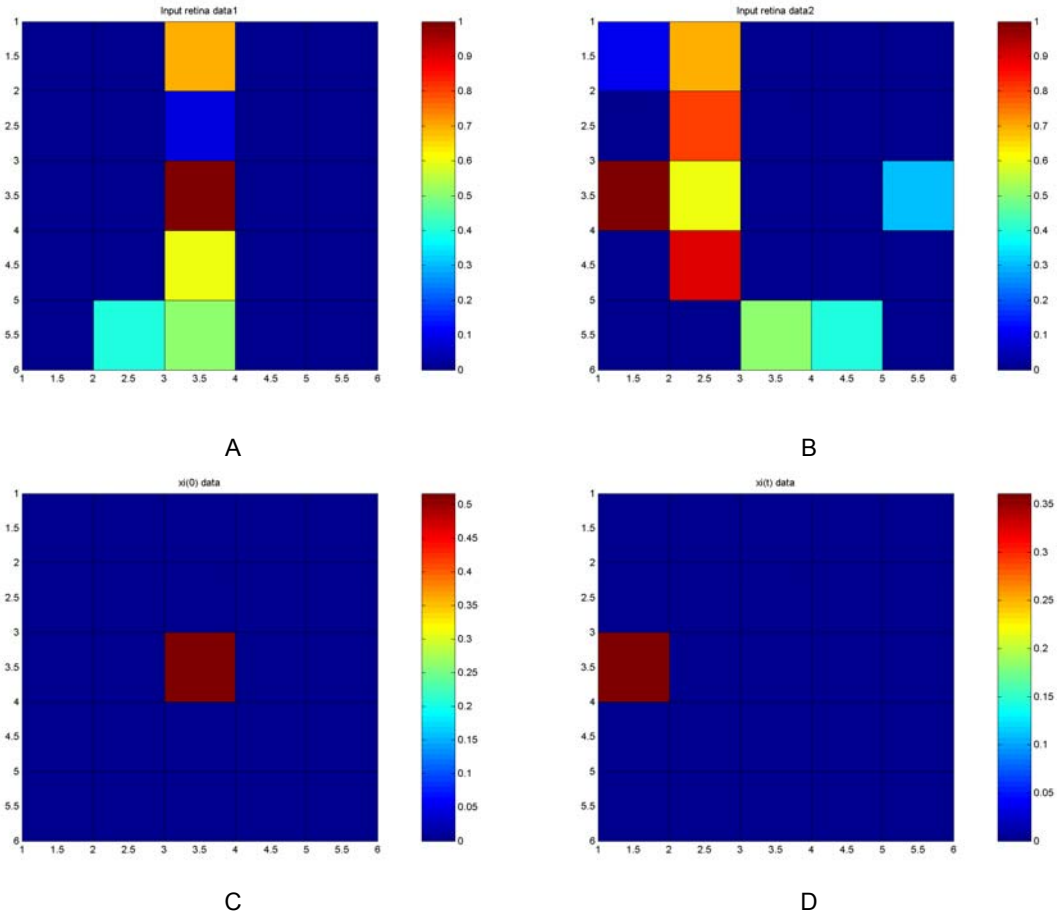


Figure 15.19: Dynamical response of $CL^{(2)}$ when the I_{2j} inputs switch from one non-zero value to another. (A) first input pattern; (B) second input pattern; (C) x_{2j} values at step 400; (D) x_{2j} values at step 1000.

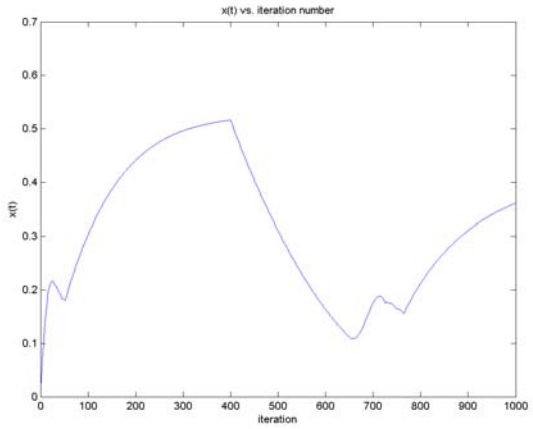


Figure 15.20: $x(t)$ for the example of figure 15.19. The pattern switches from input 1 to input 2 at 400 time steps into the simulation. Both patterns achieve their steady state values within 0.5% within the simulation.

of two successive $\{I_{2j}\}$ input patterns and the 0-1 responses to each. Figure 15.20 illustrates $x(t)$ for the simulation. The first pattern (15.19A) is applied until time step 400. The second pattern

(15.19B) is applied at time step 401 and maintained through the rest of the simulation. The reverberation responses to both patterns stabilize to within 0.5% of steady state within the time spans of the pattern applications. Both result in 0-1 distributions.

The simulation illustrates that $CL^{(2)}$ requires a transition time to switch between pattern sets. Pattern 2, applied at time step 401, does not begin to take over the $CL^{(2)}$ activation patterns until approximately time step 650. This is indicated by the local minimum in $x(t)$ at that time step. The end of the reverberation and establishment of the second 0-1 distribution occurs at the second minimum around time step 770. The simulation reveals the importance of ***dwell time*** for the input patterns for successful weight adaptation. For the application times shown in the example, it is clear that input patterns must be applied for a time period significantly longer than the transient interval revealed in figure 15.20 if slow adaptation is to be dominated by the 0-1 distributions. This property of the network system is consistent with LTP/LTD phenomena observed at the synaptic level.

Finally, let us look at the noise characteristics of $CL^{(2)}$. The mere addition of the term D to the SNI dynamic, and the fact that this term permits some x_{2j} to be zero in the face of non-zero inputs to $SNI^{(3)}$, does not mean the quenching threshold is a *squelching* threshold for noise. In general some of the x_{2j} nodes will be non-zero in the presence of even low-level input signals.

Figures 15.21 illustrate the effect for four different test cases. In all cases the system parameters are the same as those of the previous examples. The first input pattern is identical to that of figure 15.19A. The second input pattern consists of 25 random input signals. For figures 15.21A and 15.21C, the input noise is uniformly distributed in the range from 0 to 0.2. For figures 15.21B and 15.21D the input noise is uniformly distributed in the range from 0 to 0.02.

The other difference introduced in this example is the normalizer. In all the previous example cases shown above, the input normalizer was $GN^{(1)}$ with parameters $B_1 = 1$, $A_1 = 0.5$. This normalizer is also used for the test cases in figures 15.21A and 15.21B. For figures 15.21C and 15.21D the normalizer is $GN^{(2)}$ with the same B_1 and A_1 parameter values and parameter C set to $B_1/24$.

Both test cases with the larger noise values (15.21A and 15.21C) underwent reverberation dynamics that resulted in a 0-1 distribution for the second (noise) pattern. In most cases the largest noise term survives the competition and is "chosen" by v_2. However, if it should happen that the noise signal corresponding to the 0-1 choice from the first pattern is among the larger noise values (not necessarily the largest), then occasionally *it* will be selected over the globally largest noise signal (and thus it interferes with the learning of pattern 1). In any event, the noise response will alter the weight vector associated with the chosen v_2 node.

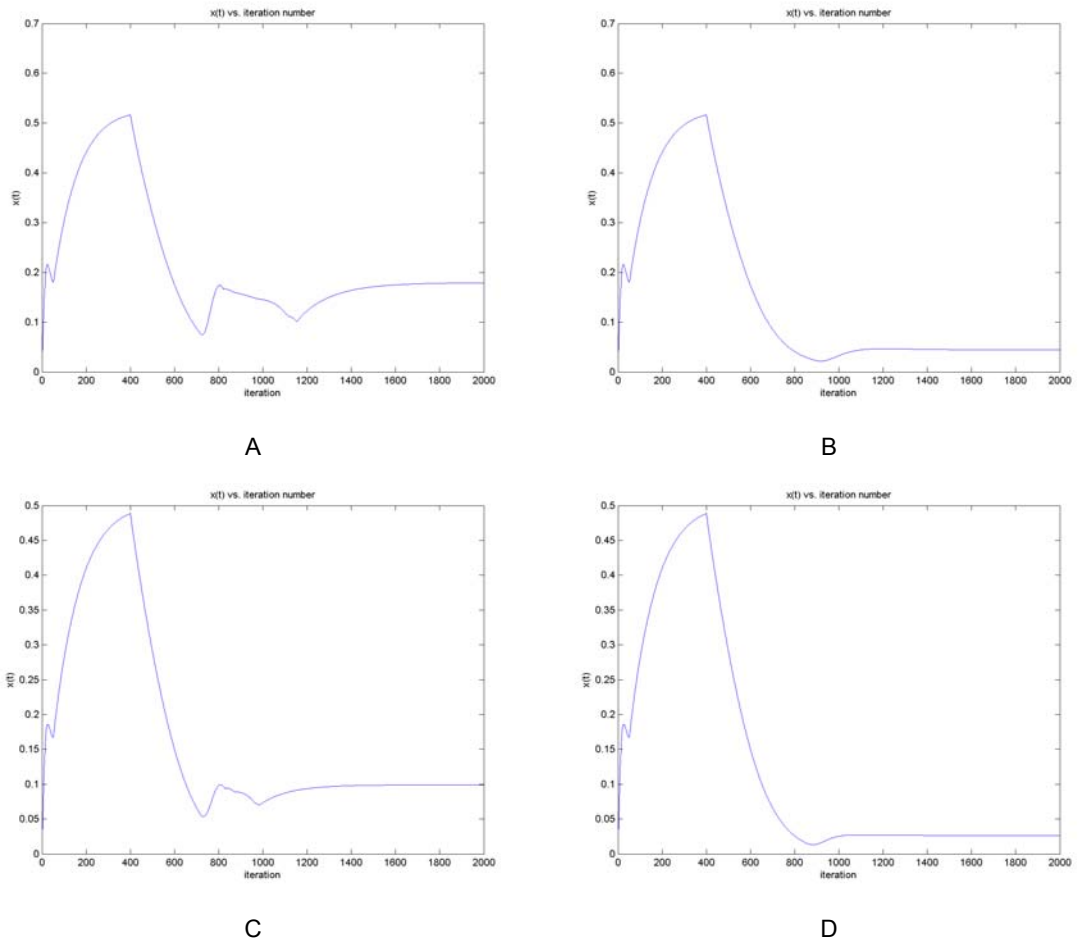


Figure 15.21: Four test cases where the first input pattern is followed by a pattern of small random values. (A) noise in second pattern uniformly distributed over (0, 0.2) with $GN^{(1)}$ normalizer; (B) noise in second pattern uniformly distributed over (0, 0.02) with $GN^{(1)}$ normalizer; (C) noise in second pattern uniformly distributed over (0, 0.2) with $GN^{(2)}$ normalizer; (D) noise in second pattern uniformly distributed over (0, 0.02) with $GN^{(2)}$ normalizer.

The main difference between figures 15.21A and C is the significantly lower level of noise-driven STM. In all cases, $GN^{(2)}$ provides superior performance over $GN^{(1)}$. This same advantage is also found for the cases of figures 15.21B and D. In these two cases, the level of noisy pattern 2 was low enough that a 0-1 distribution did *not* result. The distribution was a contrast-enhanced fair distribution in which the lowest noise terms were squelched (driven to zero), but the higher noise signals were pattern-enhanced.

The principal conclusions to be drawn from this example are these: Even with $CL^{(2)}$ there is an advantage to employing an attentional subsystem similar to that used with the RBF-MAXNET in chapter 14. The advantage is that this helps to suppress noise-driven corruption of the classifying vectors W_j. Unlike the case for $CL^{(1)}$, an attention threshold is relatively easy to pick and robust in the face of differences in the input pattern sequences when D is large. Thus large-D $CL^{(2)}$ is

superior in performance to $CL^{(1)}$. The second principal conclusion is $GN^{(2)}$ provides overall better performance through its ability to eliminate the spatial average background value of the incoming patterns.

§5.4 The Effect of Small *D* on $CL^{(2)}$ Dynamics

The lateral inhibition parameter D has the effect of promoting the production of 0-1 final distributions in layer v_2. The value used in the previous section, $D = 49B$, was chosen so that the level of lateral inhibition was on the order of 100× larger than the shunting contribution of x_{2j} in the most active nodes of v_2. At the other extreme end, $D = 0$, $SNI^{(3)}$ reduces to $SNI^{(2)}$ and $CL^{(2)}$ reduces to $CL^{(1)}$. Clearly, then, there is a continuum of performance changes that takes place as D is increased from 0 to large values. Figure 15.22 illustrates typical responses for D values of B, $2B$, $5B$, and $10B$, respectively. Two trends in particular are noteworthy in these examples. The first is that D affects when a 0-1 pattern first begins to establish in pattern 1. Larger D values tend

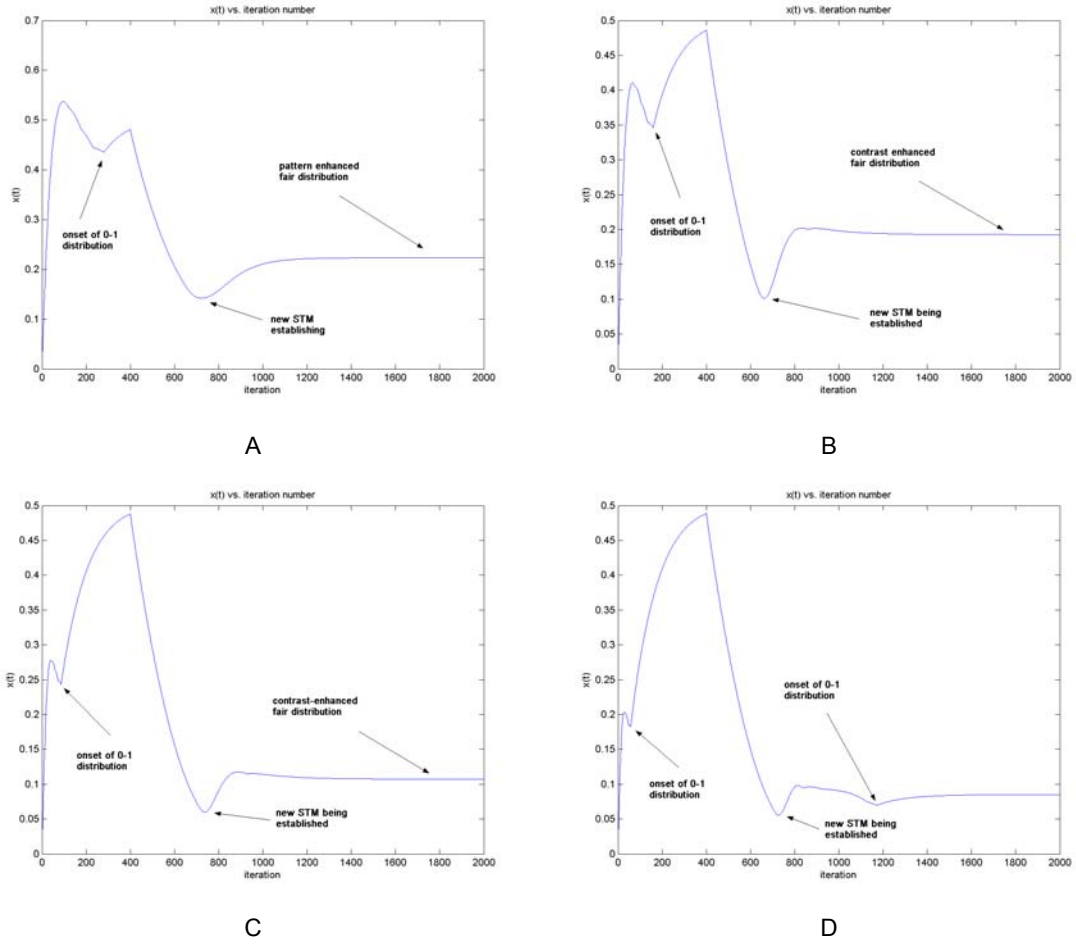


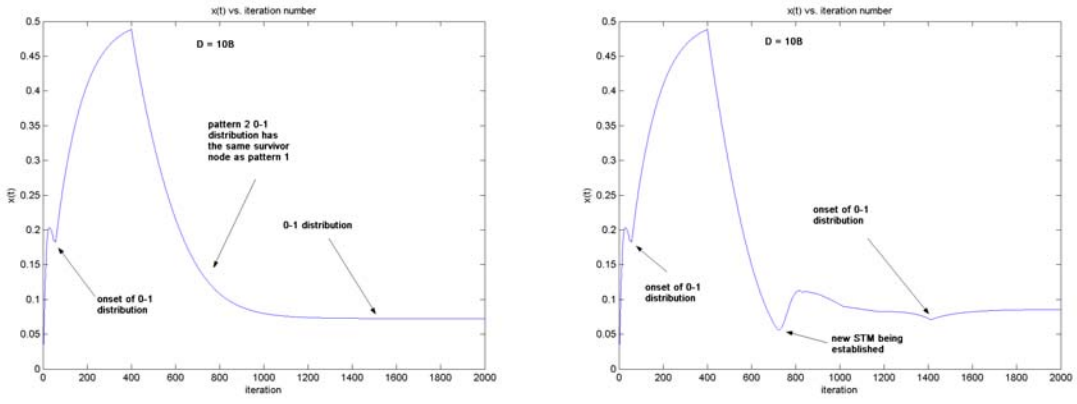
Figure 15.22: Effect of increasing D on the performance of $CL^{(2)}$. Four example cases are illustrated where the second pattern consists of random pixel values in the range from (0, 0.2) as in case 15.21C in the previous section. (A) $D = B$; (B) $D = 2B$; (C) $D = 5B$; (D) $D = 10B$.

to produce the first occurrence of a 0-1 distribution earlier in the charge-up phase of pattern 1. There is, as the figures illustrate, a diminishing returns effect with this and after $D = 10B$ there is not much additional speed-up in establishing the first 0-1 distribution pattern.

The second noteworthy feature is the size of D required to produce a 0-1 distribution in the second pattern. The examples of figure 15.22 correspond to the noisy second pattern in the range from (0, 0.2) using $GN^{(2)}$ of figure 15.21C of the previous section. For $D < \approx 10B$ the noisy second pattern results in a contrast-enhanced fair distribution of final values. For $D \approx 10B$ we see a 0-1 distribution forming for the second pattern. We also see that larger values of D tend to delay the onset of the STM for the second pattern, although this is not a particularly strong effect. For a random second pattern there is a noticeable degree of variability in time required to form a 0-1 distribution for the second pattern. This is illustrated in figure 15.23. Figure 15.23A is one extreme case where the 0-1 distribution captured for the second pattern results in the same surviving x_{2j} node as in pattern 1. Figure 15.23B shows a case where a different x_{2j} survivor remains in the 0-1 distribution. There is considerable pattern-dependent variance in the settling dynamics for the second pattern.

§5.5 The Stability-Plasticity Dilemma in $CL^{(2)}$

In a number of ways, the behavior of $CL^{(2)}$ is similar to the Instar-MAXNET competitive network of chapter 14, and in some ways it is inferior. In order to selectively target classification vectors W_j it is desirable to operate v_2 in the 0-1 distribution mode, which effectively abolishes persistent STM when the inputs $I_{(2)j} = 0$. This also happens at the Instar outputs of the Instar-MAXNET network of chapter 14 but happens much more swiftly because the first layer Instars in



A B

Figure 15.23: Two more example cases for $D = 10B$ with a (0, 0.2) random second pattern. Compare especially figure 15.23B against figure 15.22D and note the times for the onset of the 0-1 distributions.

that network do not interact. We saw that if the W_j vectors were unequal in length in this network then it was possible for those with larger values of $|W_j|$ to dominate the Instars with smaller weight vector lengths even if these were closer in direction to the input vector. Only if the weights were constrained by some type of normalization is it possible for the Instar-MAXNET to respond only to $\cos(\varphi_{\Theta,W})_j$. It was for this reason the RBF-MAXNET was preferred in chapter 14.

$CL^{(2)}$ faces the same issue. Even if all the classifying vectors W_j start off initially with the same lengths, over time the adaptation process will produce unequal lengths. If all the input vectors Θ were of equal length, maintenance of equality of classifying vector lengths cannot be guaranteed because the IAR takes ΔW_j along a straight line path toward Θ. Putting this another way, were we to trace "the tip of the W_j vector" as it moves in an N-dimensional vector space, we would find its trajectory to lie on a chord rather than an arc. If the direction cosines among all the classifying vectors are small (i.e., the classifying vectors point in maximally different directions) and if an attentional subsystem control for the IAR adaptation is employed, the undesirable effects of changes in vector lengths could in principle be avoided if all the Θ are of equal length. But, of course, they are not, and this hints that some form of ***gain control*** might be needed. In the next chapter we will see that a gain control mechanism *is* made part of an ART network.

However, even all input vectors – or at least all *selected* input vectors – are of equal length, it is possible that an unfortunate *sequence* of successive Θ vectors could under certain conditions lead to the non-existence of a stable coding. Grossberg has shown [GROS5] that only if the number of patterns Θ is small and properly partitioned relative to the number of classifying vectors is a stable encoding possible with $CL^{(2)}$. It is naive to think that biological systems *in vivo* will meet up with the mathematical conditions required for stable coding. In other words, $CL^{(2)}$ does not escape the stability-plasticity dilemma, and this is the observation that led Grossberg to go on to develop ART networks.

Lest the reader feel that somehow the efforts he has had to exert to follow the theory in this chapter has been wasted, let me assure you: It has not been in vain. The material that has been presented here is propaedeutic to being able to understand adaptive resonance theory. The systems we have examined here are not ART networks, but they are the foundations *for* ART networks. It is possible that many practitioners of neural network theory and technique have decided ART is extremely complicated – some think too complicated – merely because not a sufficient amount of education has been provided (or obtained through self-study) of the non-ART systems discussed here. (After all, this theory in its entirety was developed during the 'dark age' when neural network funding in the U.S. was almost nil and U.S. based research had almost died out). It is true that the theory itself is extremely elegant; it is not true that ART networks are

unduly complex. In chapter 16 we will take the next step and examine ART networks proper.