Projective Clustering Using Neural Networks with Adaptive Delay and Signal Transmission Loss

Jianhong Wu
wujh@mathstat.yorku.ca
MITACS Centre for Disease Modeling, and Department of Mathematics and Statistics, York University, Toronto, Ontario M3J 1P3, Canada

Hossein Zivari-Piran
hzp@mathstat.yorku.ca
Laboratory for Industrial and Applied Mathematics, York University, Toronto, Ontario M3J 1P3, Canada

John D. Hunter
jdh2358@gmail.com
TradeLink Securities, Chicago, Il 60606, U.S.A.

John G. Milton
jmilton@jsd.claremont.edu
W. M. Keck Science Center,
Claremont Colleges, Claremont, CA 91771, U.S.A.

We develop a new neural network architecture for projective clustering of data sets that incorporates adaptive transmission delays and signal transmission information loss. The resultant selective output signaling mechanism does not require the addition of multiple hidden layers but instead is based on the assumption that the signal transmission velocity between input processing neurons and clustering neurons is proportional to the similarity between the input pattern and the feature vector (the top-down weights) of the clustering neuron. The mathematical model governing the evolution of the signal transmission delay, the short-term memory traces, and the long-term memory traces represents a new class of large-scale delay differential equations where the evolution of the delay is described by a nonlinear differential equation involving the similarity measure already noted. We give a complete description of the computational performance of the network for a wide range of parameter values.

1 Introduction

The human nervous system learns about its environment by discovering structures buried in the sensory inputs that it receives. How is the familiar to be distinguished from the unfamiliar in real time while at the same time preserving a sensitivity for novelty and a robustness against variations (Milton & Mackey, 2000)? Yet the nervous system performs this task so successfully that considerations of neurophysiology and neuroanatomy have inspired the development of a large number of neural network–type methods to mine enormously large data sets for interesting structures, a process known as data clustering. There is an interplay between the efforts of computational neuroscientists and neural network theorists: the better such networks are able to mimic the nervous system, the more effective they will likely be for clustering large data sets. Conversely, new developments in data clustering by artificial neural networks shed light on how the living nervous system performs this task.

Carpenter, Grossberg, and their coworkers (Carpenter & Grossberg, 1987a, 1987b, 1987c, 1990; Carpenter, Grossberg, & Reynolds, 1991; Carpenter, Grossberg, & Rosen, 1991a, 1991b; Carpenter, Grossberg, Markuzon, Reynolds, & Rosen, 1992; Williamson, 1996) introduced and developed adaptive resonance theory (ART) to demonstrate how brain networks automatically learn to cluster information presented to it in real time (see section 2.1). The two key elements of an ART network are a selection process and a match-based learning mechanism. The selection process picks up the most likely category (cluster candidate) for an input pattern. If the chosen category's template is sufficiently similar to the input pattern to satisfy a predefined vigilance condition, the category resonates and learns—its template is updated. Otherwise the category is reset, and the next most likely category is created. If no existing category satisfies the match criterion, a new category is created. Thus, ART networks incrementally produce the categories to represent clusters of input patterns. Although ART-type networks have proven very effective for clustering arbitrary sequences of input patterns into recognition codes, they are unable to function efficiently in the high-dimensional spaces that the human visual system typically encounters. The problem is the sparsity of data points, which makes it impossible to find interesting patterns in the full space of dimensions. Pruning off dimensions in advance, as most feature selection procedures do, can lead to significant loss of information, making the obtained classifications unreliable.

Aggarwal and coworkers (Aggarwal, Procopiuc, Wolf, Yu, & Park, 1999; Aggarwal & Yu, 2000) introduced the concept of projective clustering to address the issue of detecting-low dimensional patterns in a high-dimensional data set. A projective cluster consists of two parts: a subset C of data points (the cluster) and a subset of dimension D such that the points in C are closely related in the subspace of dimension D (the projective cluster). To illustrate the concept of a projective subspace, consider the following animals (data input): sheep, dog, cat, sparrow, seagull, viper, lizard, goldfish, red mullet, blue shark, and frog. Projective subspaces can be formed, for example, by classifying the animals into those that bear live progeny and those that do not, the environment that they live in, the existence of lungs, and so on. Cao and Wu (Cao, 2002; Cao & Wu, 2002, 2004) implemented projective clustering into ART network, the result is PART. PART networks outperform ART networks for pattern recognition in high-dimensional spaces. Recent demonstrations include the use of PART networks to classify patterns in neural spike trains (Hunter, Wu, & Milton, 2008) and the application of PART for gene fitting (Takahashi, Kobayashi, & Honda, 2005).

The key feature of a PART network is a hidden layer that incorporates a selective output signal mechanism (SOS) that calculates the similarity between the output (activation) of a given input neuron (which corresponds to a particular component of an input) with the corresponding component of the template (statistical mean) of a candidate cluster neuron and allows the signal (activation of the input neuron) to be transmitted to the cluster neuron only when the similarity measure is sufficiently large. This similarity check is achieved by adding multiple layers of hidden neurons. In addition, in PART, the output signal of an input neuron will be completely prohibited from transmitting to its target cluster neuron if the similarity measure is small, although in practice, this output signal may still play a (relatively minor) role in the final clustering result. However, despite the success of projective clustering for computer-generated data of high dimensionality (Cao & Wu, 2002; Hunter et al., 2008), there is no convincing evidence as yet to support the existence of an SOS mechanism that incorporates hidden layers in the nervous system.

Here we introduce a novel clustering network, termed PART-D, which interprets the SOS mechanism in terms of two recently emphasized properties of the nervous system: the adaptability of transmission time delays (Carr, 1993; Fields, 2005; Stanford, 1987; Stevens, Tanner & Fields, 1998; Zalc & Fields, 2000) and the signal losses that necessarily arise in the presence of transmission delay (Bale & Petersen, 2009; Sincich, Horton, & Sharpee, 2009; see appendix A). Glial-neuron interactions play important roles in determining axonal myelination and hence conduction velocity (Stevens et al., 1998; Fields, 2005; Zalc & Fields, 2000). Thus, self-organization of transmission delays may be an important but underrecognized mechanism for learning (Eurich, Pawelzik, Cowan, & Milton, 1999). We show that a plausible SOS mechanism can arise because the self-organized adaptation of transmission delays is driven by the dissimilarity between the input pattern and the stored pattern (represented by the template of a cluster neuron). Such an adaptation can be regarded as a consequence of the Hebbian learning law (Hebb, 1949), and the dynamic adaptation can be modeled by a nonlinear differential equation. As a result, we obtain a new class of systems of delay differential equations with adaptive delay. The dissimilarity between driven transmission delay and the signal transmission loss due to delay identifies the selective output signaling component of input neuron activations in terms of the self-organization of transmission delays.

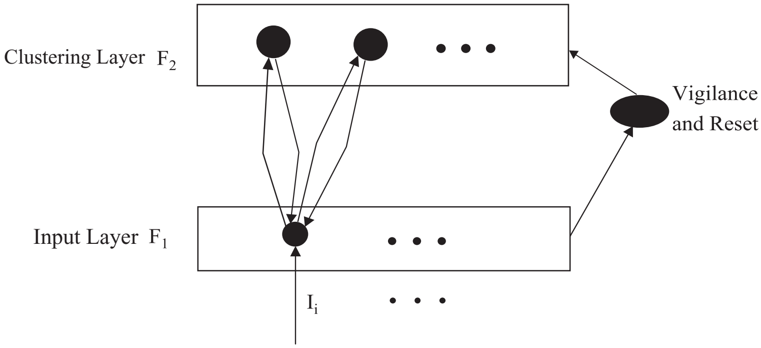


Figure 1: A simplified configuration of ART architecture that consists of an input layer F_1, a clustering layer F_2, and a reset mechanism.

We organize our discussion as follows. In section 2, we briefly review adaptive resonance theory and highlight the key differences among ART, PART, and PART-D networks. In section 3, we derive the equations for a PART-D network, and in section 4, we determine analytically its performance features. Our model takes the form of an unusual system of functional differential equations where the evolution of time delays is governed by a nonlinear differential equation. Detailed qualitative analysis of the system yields an explicit formula for calculating the total amount of selective output signals to a given cluster neuron, and this formula gives precise information about the relative role of the signals with different dissimilarity measures in the final clustering result. Finally in section 5, we discuss some directions for future work.

2 Adaptive Resonant Neural Networks

2.1 ART Neural Networks. The two key components of an ART network are the selection process and the match process: the selection process picks up the most likely category (cluster candidate) for an input pattern. If the chosen category's template is sufficiently similar to the input pattern to satisfy a predefined vigilance condition, the category resonates and learns: its template is updated to respond to the new input pattern. Otherwise the category is reset, and the next most likely category is chosen. If no existing category satisfies the match criterion, a new category is recruited. Thus, ART incrementally produces categories necessary to represent clusters of input patterns.

Figure 1 illustrates the basic ART architecture, which consists of an input processing layer or comparison layer (F_1 layer), a clustering layer (F_2 layer), bottom-up synaptic weights and top-down synaptic weights between the two layers, and a reset mechanism.

The F_1 layer is analogous to cell groups in a sensory area of the cerebral cortex, and the role of this layer is to process the inputs. The number of neurons in the F_1 layer is the same as the number of the components or variables of the input vector, which activates each neuron according to the size of the corresponding component. The neurons in the F_1 layer are not connected to each other, reflecting our standing assumption in this letter that variables in the input vector are independent. This assumption is obviously not met in many applications. A subsequent development of PROCLUS (Aggarwal & Yu, 2000) deals with the case when this assumption is not met. We address the modification of our PART-D network to accommodate this case in a future paper.

The bottom-up weights measure the impact of the output from a neuron in the F_1 layer to the collective effort to activate a candidate cluster neuron. These weights are updated after each input or learning trial according to certain learning laws.

Each neuron in the F_2 layer represents a cluster, and an input vector that eventually activates a given F_2 neuron through the bottom-up connections is assigned to the cluster represented by the F_2 layer neuron. Neurons in the F_2 layer are connected to each other, and the connection topology is determined by the underlying learning rules. Here we adopt the competitive learning rule and the special on-center, off-surround connection topology inspired by visual neurophysiology. Hence, the neuron in the F_2 layer that receives the maximal total amount input signals from the F_1 layer is the winner candidate to represent the input vector. Other connection topologies are possible. For example, a connection topology can be constructed to overcome the problem that ART or PART clustering results can depend on the order in which input vectors are presented to the network (Cao & Wu, unpublished observations).

For each cluster neuron the associated top-down weights represent the statistical mean of the corresponding cluster, and thus these weights give the feature of the cluster. The feature vector will also be updated again after each learning trail according to certain learning rules to record the learning experience.

2.2 PART Neural Networks. The basic architecture of PART (see Figure 2a) is similar to that of the ART neural networks but includes a new feature: selective output signal. This feature is essential for pattern recognition in subspaces (Cao, 2002; Cao & Wu, 2002, 2004). Because of the competitive learning in the F_2 layer, calculating the total inputs from all neurons in the F_1 layer to a given cluster neuron is essential for the choice of a winner candidate cluster neuron to represent the input vector. This is closely related to the clustering criterion to be discovered by the network. This calculation distinguishes PART from its ancestor ART and distinguishes the proposed architecture (PART-D) of this letter from PART. In ART, the focus is to find clusters with respect to all variables of the input vector, and the total number of signals received by a cluster neuron is the weighted (by the bottom-up weights) sum of the outputs of the processing neurons. In PART, a new feature is developed. This SOS mechanism

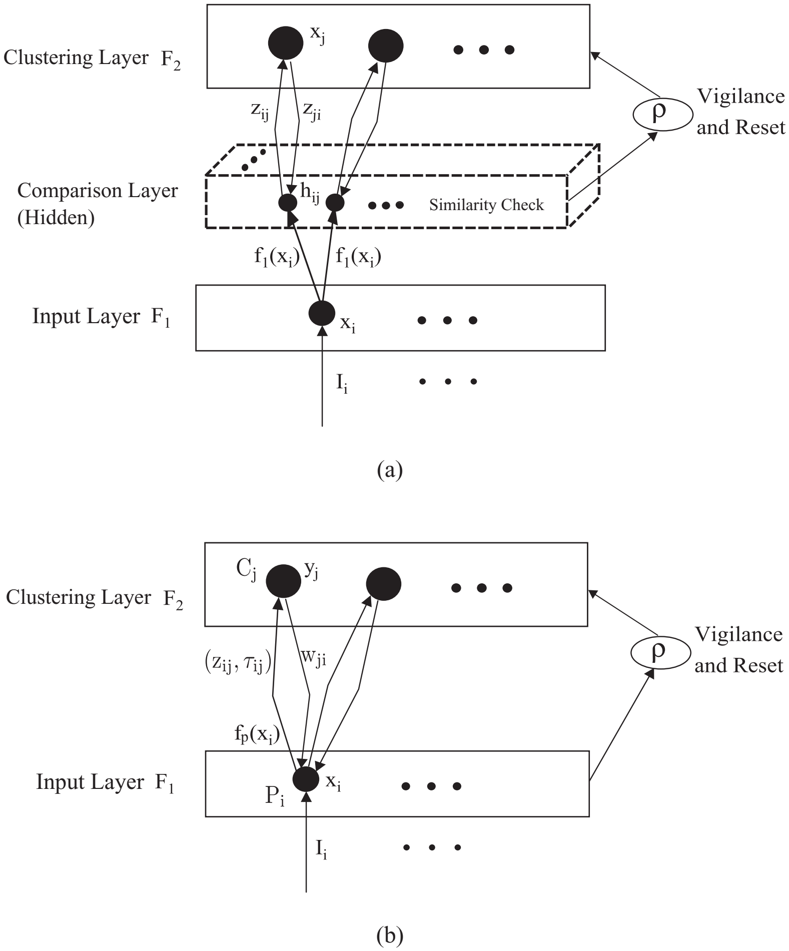


Figure 2: (a) PART architecture. In addition to the usual F_1 layer (input and comparison), F_2 layer (clustering), and a reset mechanism, a hidden layer is associated with each F_1 layer node v_i for similarity check to determine whether the node v_i is active relatively to an F_2 layer node v_j. (Adapted from Cao, 2002.) (b) PART-D architecture. The PART-D neural network replaces the hidden layers by a new concept based on dissimilarity-driven transmission delays and a delay-induced signal loss.

requires the addition of hidden multi layers of neurons to perform the similarity check and ultimately leads to a new reset subsystem.

For each cluster neuron, we add a layer of neurons (the total number is the same as the number of input neurons), the similarity check layer, to check the similarity of the output signal from a given input neuron with the corresponding top-down weight of the target cluster neuron. The output signal is allowed to be broadcast to the cluster neuron only when there is a strong similarity and when the corresponding bottom-up weight (significance factor of the corresponding dimension) is larger than a given constant. Obviously this also determines the projective subspace at the end of the current learning trial for a cluster represented by a cluster neuron.

The role of the reset subsystem is to reset the candidate cluster neuron if the dimension of the projective subspace is smaller than a vigilance parameter. This is natural and necessary, since for a given pair of points in a large data set in high-dimensional spaces, the probability of finding a few dimensions with respect to which the two points are close to each other is large, and thus the vigilance parameter should be relatively large to get rid of this randomness.

2.3 PART-D Neural Networks. The PART-D neural network introduced here (see Figure 2b) replaces the hidden layers by a new concept based on a dissimilarity-driven transmission delay and a delay induced signal loss. These are motivated by the following observations and assumptions:

- Dissimilarity-driven transmission delay: Signal transmission between neurons in two different layers is not instantaneous, and the transmission speed between an input neuron and a cluster neuron is proportional to the similarity between the output from the input neuron to the corresponding component of the top-down weights (feature vector),
- Delay-induced transmission loss: During transmission, the signal decays exponentially, and thus the longer the delay, the larger the loss of the signal.

As a consequence, the signal received by a cluster neuron is only a portion of the output signal from an input neuron. If the transmission is instantaneous, the proportionality constant is 1, as is typically assumed, but if the transmission is delayed, this proportionality constant is $e^{-\alpha\tau}$, where α is the decay rate of the signal and τ is the delay (see appendix A). In PART, this constant is assumed to be zero if the dissimilarity-driven delay is positive. In PART-D the calculation of the proportionality constant utilizes a nonlinear differential equation that describes the adaptation of the time delay. Thus, there is an explicit formula that determines the proportionality constant. In some sense, this formula provides a qualitative formulation, very much similar to the concept of membership in fuzzy clustering, of the dissimilarity and its related delay.

3 Projective Adaptive Resonant Theory with Delay (PART-D)

The PART-D network consists of two layers of neurons, synaptic connections between neurons of these two layers and synaptic connections among cluster neurons, a reset mechanism and a dissimilarity-driven transmission delay, and the resulting loss of the signal (see Figure 2b). Denote the nodes in the F_1 layer (comparison/input processing layer) by $P_i, i \in \Lambda_p := \{1, \ldots, m\}$; nodes in the F_2 layer (clustering layer) by C_j, $j \in \Lambda_c := \{1, \ldots, n\}$ and the activation of the F_1 node P_i by x_i and the activation of the F_2 node C_j by y_j; and the bottom-up weight from P_i to C_j by z_{ij} and the top-down weight (also called a template) from C_j to P_i by w_{ji}.

The short-term memory (STM) equations for neurons in the F_1 layer are given by

$$\epsilon_p \frac{dx_i(t)}{dt} = -x_i(t) + I_i, \quad t \geq -1, \quad i \in \Lambda_p, \tag{3.1}$$

where $0 < \epsilon_p \ll 1$ and I_i is the constant input imposed on P_i. This is based on the assumption that for an isolated neuron, the dynamics is the balance of the internal decay and the external input excitation.

The change of the STM for an F_2 neuron depends on the internal decay, the excitation from self-feedback, the inhibition from other F_2 neurons, and the excitation by the bottom-up filter inputs from F_1 neurons. We have the STM equations for the committed neurons in the F_2 layer:

$$\begin{aligned}\epsilon_c \frac{dy_j(t)}{dt} &= -y_j(t) + [1 - Ay_j(t)][f_c(y_j(t)) + T_j(t)] \\ &\quad -[B + Cy_j(t)] \sum_{k \in \Lambda_c \setminus \{j\}} f_c(y_k(t)), \quad t \geq 0, \ j \in \Lambda_c,\end{aligned} \tag{3.2}$$

where $0 < \epsilon_c \ll 1$, $f_c : R \to R$ is a signal function to be specified later; A, B, and C are nonnegative constants; and the bottom-up filter input T_j is given by

$$T_j(t) = D \sum_{i \in \Lambda_p} z_{ij}(t) f_p(x_i(t - \tau_{ij}(t))) e^{-\alpha \tau_{ij}(t)}, \quad t \geq 0, \tag{3.3}$$

where D is a scaling constant and $f_p : R \to R$ is the signal function of the input layer. It is assumed here that the signal transmissions between two layers are not instantaneous and the signal decays exponentially at a rate $\alpha > 0$ (see appendix A). The exponential decay term $e^{-\alpha \tau_{ij}}$ can be replaced by any parameterized function $g_\alpha = g_\alpha(\tau_{ij})$ as long as $g_\alpha(0) = 1$ and for a fixed $\tau > 0$, $g_\alpha(\tau)$ can be made arbitrarily small if α is sufficiently large.

The term τ_{ij} is the signal transmission delay between the input neuron P_i and the cluster neuron C_j. We assume this delay is driven by the dissimilarity in the sense that the signal processing from the input neuron P_i to the cluster neuron C_j is faster when the output from P_i is similar to the corresponding component of w_{ji} of the feature vector $w_j = (w_{ji})_{i\in\Lambda_p}$ of the cluster neuron C_j. Therefore, we have

$$\beta\frac{d\tau_{ij}(t)}{dt} = -\tau_{ij}(t) + E[1 - h_{ij}(t)], \quad t \geq 0, \quad i \in \Lambda_p, \quad j \in \Lambda_c, \tag{3.4}$$

where $\beta > 0$, $E \in (0, 1)$ are constants and

$$h_{ij}(t) = S(d(f_p(x_i(t)), w_{ji}(t)), z_{ij}(t))$$

is the similarity measure between the output signal $f_p(x_i(t))$ and the corresponding component $w_{ji}(t)$ of the feature vector of the cluster neuron C_j, with respect to the significance factor of the bottom-up synaptic weight $z_{ij}(t)$. Here, d is the usual distance function,

$$d(a, b) = |a - b| \quad \text{for any}\, a, b \in \mathbb{R},$$

and $S : R^+ \times [0, 1] \to [0, 1]$ is a given function, nonincreasing with respect to the first argument and nondecreasing with respect to the second argument. Moreover, $S(0, 1) = 1$ (the similarity measure is 1 with complete similarity and maximal synaptic bottom-up weight) and $S(+\infty, z) = S(x, 0) = 0$ for all $z \in [0, 1]$ and $x \in R^+ := [0, \infty)$(the similarity measure is 0 with complete dissimilarity or minimal bottom-up synaptic weight). Therefore, if $\tau_{ij}(0) = 0$, then from equation 3.4, it follows that $0 \leq \tau_{ij}(t) \leq E$ for all $t \in R^+$. Moreover, if $h_{ij}(t) = 1$ on an interval $[0, b)$ for a given $b > 0$, then $\tau_{ij}(t) = 0$ for all $t \in [0, b)$.

In what follows, we are going to assume

$$h_{ij}(t) = h_\sigma(d(f_p(x_i(t)), w_{ji}(t)))l_\theta(z_{ij}(t)), \quad t \geq 0, \quad i \in \Lambda_p, \quad j \in \Lambda_c. \tag{3.5}$$

In other words, $h_{ij}(t)$ is determined by the distance between the output signal $f_p(x_i(t))$ and the corresponding component $w_{ji}(t)$ of the feature vector of the cluster neuron C_j, multiplied by the significance factor of the bottom-up synaptic weight $z_{ij}(t)$. We also assume that for a given constant σ, h_σ is given by

$$h_\sigma(\xi) = \begin{cases} 1 & \text{if } \xi \leq \sigma \\ 0 & \text{if } \xi > \sigma \end{cases},$$

and for a constant, $\theta > 0$, l_θ is given by

$$l_\theta(\xi) = \begin{cases} 1 & \text{if } \xi \geq \theta \\ 0 & \text{if } \xi < \theta \end{cases}.$$

Therefore, $h_{ij}(t) = 1$ if $|f_p(x_i(t)) - w_{ji}(t)| \leq \sigma$ and $z_{ij}(t) \geq \theta$, and $h_{ij}(t) = 0$ if either $|f_p(x_i(t)) - w_{ji}(t)| > \sigma$ or $z_{ij}(t) < \theta$. As a consequence, the dissimilarity $1 - h_{ij}(t)$ is either 0 or 1. This choice of the dissimilarity measure will significantly simplify the mathematical analysis of the model, as shown in the next section. However, we emphasize that this dissimilarity measure involves the choice of two parameters (θ, σ) and that the binary values 0 or 1 do not reflect the fuzzy nature of the dissimilarity. (More details are provided in section 6.)

The equation governing the change of the weights follows from the synaptic conservation rule of von der Malsburg (1973), and only connections to activated neurons are modified. The top-down weights are modified so that the template will point to the direction of the delayed and exponentially decayed outputs from F_1 layer. Therefore, we have

$$\gamma \frac{dw_{ji}(t)}{dt} = f_c(y_j(t))\big[-w_{ji}(t) + f_p(x_i(t - \tau_{ij}(t)))e^{-\alpha\tau_{ij}(t)}\big],$$
$$t \geq 0, \quad i \in \Lambda_p, \quad j \in \Lambda_c, \tag{3.6}$$

where $\gamma > 0$ is a given constant.

The bottom-up weights are changed according to the competitive learning law and Weber's law rule that says that long-term memory (LTM) size should vary inversely with input pattern scale to present a clustering neuron that has learned a particular pattern from coding every superset pattern (see Carpenter & Grossberg, 1987c). Thus, the LTM equations for committed neurons C_j in F_2 layer are

$$\begin{aligned}\delta \frac{dz_{ij}(t)}{dt} = {} & f_c(y_j(t))[(1 - z_{ij}(t))h_{ij}(t)L - z_{ij}(t)(1 - h_{ij}(t)) \\ & -z_{ij}(t)\textstyle\sum_{k\in\Lambda_p\setminus\{i\}} h_{kj}(t)], \quad t \geq 0, \quad i \in \Lambda_p, \quad j \in \Lambda_c,\end{aligned} \tag{3.7}$$

where $0 < \delta \ll \gamma = O(1)$ and $L > 0$ is a given constant.

The LTM equations for noncommitted candidate node C_j in F_2 layer are

$$\delta \frac{dz_{ij}(t)}{dt} = [1 - z_{ij}(t)]L - z_{ij}(t)(m - 1), \quad t \geq 0, \quad i \in \Lambda_p, \tag{3.8}$$

and

$$\delta \frac{dw_{ji}(t)}{dt} = -w_{ji}(t) + f_p(x_i(t)), \quad t \geq 0, \quad i \in \Lambda_p. \tag{3.9}$$

In the PART-D architecture, this dynamical process is coupled with a reset mechanism. In particular, a candidate (active) F_2 node C_j will be reset if at any given time $t \geq 0$, the degree of match is less than a prescribed vigilance. Reset occurs if and only if

$$\sum_{i \in \Lambda_p} h_{ij}(t) < \rho. \tag{3.10}$$

Here, $\rho \in \{1, 2, \ldots, m\}$ is a vigilance parameter.

4 Performance of PART-D Neural Networks

Equations 3.1 to 3.7 describe a system of functional differential equations where the dynamics of the delay $\tau_{ij}(t)$ are adaptive and are described by the nonlinear equation 3.4. To determine a solution, we need to specify the initial condition. We assume that all neurons in both layers are set to their normalized equilibrium states and that the transmission delays are initially set to zero:

$$\begin{aligned}
&x_i(t) = 0, \quad i \in \Lambda_p, t \leq -1,\\
&y_j(t) = 0, \quad j \in \Lambda_c, t \in [-1, 0],\\
&\tau_{ij}(t) = 0, \quad i \in \Lambda_p, j \in \Lambda_c, t \in [-1, 0].
\end{aligned}$$

Unlike most of the delay differential systems investigated in the literature, the delay is state dependent, and, in fact, the evolution of the delay is governed by a differential equation involving the dissimilarity measure mentioned above, which is related to the status of input neurons and the top-down weights (see Hartung, Krisztin, Walther, & Wu, 2006, for a recent survey of using state-dependent delays). We show that the model equation has regular dynamics that are essential for the algorithm development:

- The winner-take-all paradigm. The F_2 node with the largest bottom-up filter input becomes the winner, and only this node is activated after some finite time.
- An F_2 node will never be reset during the whole trial if it is not initially reset.
- The synaptic weights are updated following specific formulas.
- The dimension of a specific projective cluster is nonincreasing in time.

In what follows, we make the following assumptions:

H1: Constants A, B, and C satisfy $A = 1$, $B = 0$, $C > 0$.

H2: The signal function $f_p : R \to R$ is nondecreasing and satisfies the Lipschitz condition (with a given constant $K > 0$):

$$|f_p(x) - f_p(\hat{x})| \leq K|x - \hat{x}|, \quad x, \hat{x} \in \mathbb{R}.$$

H3: The signal function $f_c : R \to R$ satisfies, for a constant $\eta_c \in (0, 1)$, that

$$f_c(y) = \begin{cases} 1 & \text{if } y \geq \eta_c \\ 0 & \text{if } y < \eta_c \end{cases}.$$

The dynamics of the PART-D neural network are described by the following two theorems:

Theorem 1. *Let*

$$\tau_{ij}^* = E[1 - h_{ij}(0)],$$
$$T_j^* = D \sum_{i \in \Lambda_p} z_{ij}(0) f_p(I_i) e^{-\alpha \tau_{ij}^*}.$$

Assume that

$$\frac{L}{L + m - 1} > \theta$$

and that there exists $J \in \Lambda_c$ such that $T_j^* < T_J^*$ for all $j \in \Lambda_c \setminus \{J\}$. Let $M = \sum_{i \in \Lambda_p} z_{iJ}(0) f_p(I_i)$. Assume further that there exists $T_{min} > 0$ so that

$$\frac{DL}{L + m - 1} f_p(I_i) > T_{min}, \quad i \in \Lambda_p,$$
$$T_J^* > T_{min} + DMe^{-\alpha E},$$
$$\frac{1 + T_j^*}{2 + T_j^* + C} < \eta_c < \min\left\{ \frac{T_J^*}{1 + T_J^*}, \frac{1 + T_{min}}{2 + T_{min}} \right\}, \quad j \in \Lambda_c \setminus \{J\}.$$

Then we can choose ϵ_p, ϵ_c, and δ sufficiently small so that the following results hold:

i. Inhibition of noncandidate neurons: For $j \neq J$ and $t \geq 0$, $y_j(t) < \eta_c$, and $f_c(y_j(t)) = 0$.

ii. Sustained excitation of the candidate neuron: There exists $\Gamma > 0$ such that $y_J(t) < \eta_c$ and $f_c(y_J(t)) = 0$ when $t < \Gamma$, and $y_J(t) \geq \eta_c$ and $f_c(y_J(t)) = 1$ when $t \geq \Gamma$.

iii. Invariance of similarity: For any $i \in \Lambda_p$, $j \in \Lambda_c$ and $t \geq 0$, $h_{ij}(t) = h_{ij}(0)$.

iv. (Learning at Infinity): For any $i \in \Lambda_p$ and $j \in \Lambda_c$ with $j \neq J$, $z_{ij}(t)$ and $w_{ji}(t)$ remain unchanged for all $t \geq 0$. But

$$\lim_{t \to \infty} z_{iJ}(t) = \begin{cases} 0 & \text{if } h_{iJ}(0) = 0 \\ \dfrac{L}{L + l_i} & \text{if } h_{iJ}(0) = 1 \end{cases},$$

and

$$\lim_{t\to\infty} w_{Ji}(t) = f_p(I_i)e^{-\alpha\tau_{iJ}^*},$$

where $\quad l_i = \#\{k \in \Lambda_p \setminus \{i\};\, h_{kJ}(0) = 1\}.$

Theorem 2. *If $\epsilon_p, \epsilon_c, \delta$ are sufficiently small, we have*

v. *Fast excitation:* $\Gamma \in (0, 1)$.

vi. *Fast learning: Write $z_{ij}^{\epsilon_p,\epsilon_c,\delta}$ and $w_{ji}^{\epsilon_p,\epsilon_c,\delta}$ to indicate explicitly the dependence on $(\epsilon_p, \epsilon_c, \delta)$. Then we have*

$$\lim_{\delta\to 0} z_{iJ}^{\epsilon_p,\epsilon_c,\delta}(1) = \begin{cases} 0 & \text{if } h_{iJ}(0) = 0 \\ \dfrac{L}{L + l_i} & \text{if } h_{iJ}(0) = 1 \end{cases}$$

and

$$\lim_{\epsilon_p\to 0, \beta\to 0} w_{Ji}^{\epsilon_p,\epsilon_c,\delta}(1) = (1-q)w_{Ji}(0) + qf_p(I_i)e^{-\alpha\tau_{iJ}^*},$$

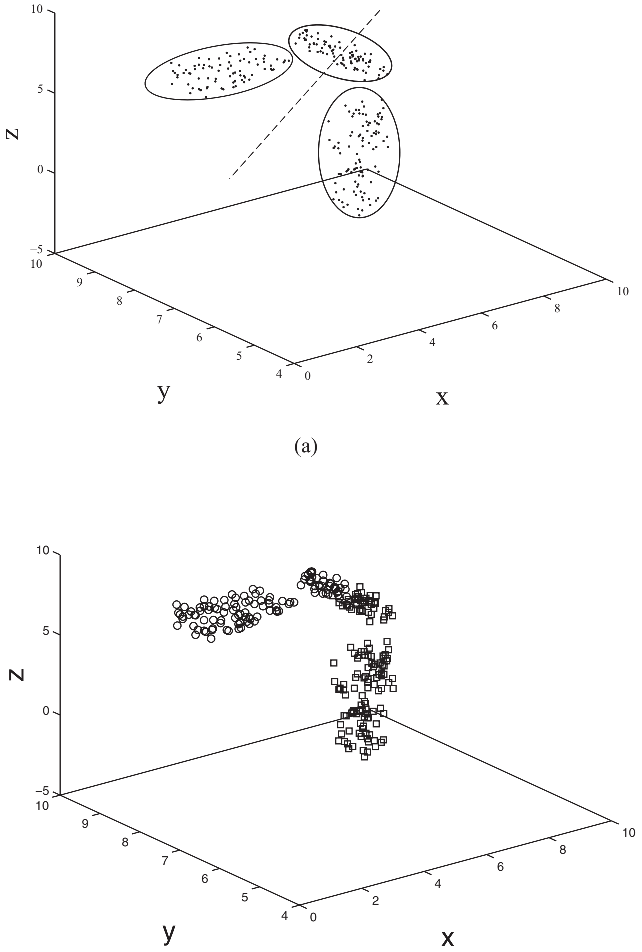
where $q = 1 - e^{-1/\gamma}$.

vii. *Convergence of projective subspace: For any $j \in \Lambda_c$, define $D_j(t) = \{i \in \Lambda_p;\, l_\theta(z_{ij}(t)) = 1\}$. Then, as $\epsilon_p, \epsilon_c, \delta \to 0$, we have*

$$\begin{aligned} D_j(t) &= D_j(0) && \text{for any } j \neq J, \\ D_J(t_2) &\subseteq D_J(t_1) && \text{if } t_2 \geq t_1 \geq 0, \\ D_J(t) &= D_J(1) && \text{for all } t \geq 1. \end{aligned}$$

Theorem 1 shows that only the node C_J in the cluster layer that receives the maximal sum (T_J^*) of selective and delayed output signals from all input neurons $P_i, i \in \Lambda_p$, will be eventually activated and the corresponding top-down and bottom-up weights will be updated. In real applications, the time interval between successive inputs of signals to the network is finite, and thus for the PART-D network to function efficiently, it is necessary that it stabilizes very fast for each given constant input. Theorem 2 describes the transient behaviors of the PART-D network. As Cao and Wu (2002) show, these transient behaviors are the basis for the development of effective algorithms for projective clustering. In particular, formulas in part vi of theorem 2 give us the explicit expression for the updating of synaptic weights corresponding to the activated clustering neuron C_J, and part vii confirms that the dimension of the associated projective subspace is decreasing and eventually stabilizes.

Following a similar procedure, we have developed a Matlab code that implements our PART-D clustering model. Figure 3b shows the output of PART-D applied to a data set in R^3, where it successfully recovers the projected subspaces and their respective clusters.



(a)

(b)

Figure 3: (a) A data set with two projected clusters. The horizontal plane together with points on the right-hand side of the dashed line gives one projective cluster, and the vertical *yz*-plane together with points on the left-hand side gives another projective cluster. The application of ART to clusters in full three-dimensional space will result in a false conclusion of three clusters (depicted by points enclosed by each ellipse). (b) PART-D successfully recovers the two projected subspaces and their respective clusters, shown as small squares and small circles.

4.1 Proof of Theorem 1. Theorem 1 is a summary of the following five lemmas:

Lemma 1. *Limited variance of activation level for F_1 neurons: For a given input $I = (I_1, \ldots, I_m)^T$ and a given $\zeta > 0$, there exists $\epsilon_p^0 > 0$ so that $|f_p(x_i(t)) - f_p(I_i)| < \zeta$ for all $t \geq 0$, provided $\epsilon_p \in (0, \epsilon_p^0)$.*

Proof. We first note that from equation 3.1 and $x_i(-1) = 0$, it follows that

$$x_i(t) = I_i\left[1 - e^{-(t+1)/\epsilon_p}\right], \quad t \geq -1, \quad i \in \Lambda_p. \tag{4.1}$$

Using equation 4.1 and Lipschitz continuity of f_p (see H2), we have

$$|f_p(x_i(t)) - f_p(I_i)| \leq K|x_i(t) - I_i| = KI_i e^{-(t+1)/\epsilon_p} \leq KI_i e^{-1/\epsilon_p}, \quad t \geq 0.$$

If $I_i \neq 0$, $\frac{\zeta}{KI_i} < 1$, and

$$0 < \epsilon_p < \epsilon_p^0 := -\frac{1}{\ln\left(\frac{\zeta}{KI_i}\right)},$$

then we have

$$|f_p(x_i(t)) - f_p(I_i)| \leq KI_i e^{-1/\epsilon_p} < KI_i e^{-1/\epsilon_p^0} = \zeta, \quad t \geq 0.$$

On the other hand, if $I_i = 0$, then

$$|f_p(x_i(t)) - f_p(I_i)| = 0 < \zeta, \quad t \geq 0$$

for any $\epsilon_p > 0$, and if $\frac{\zeta}{KI_i} \geq 1$, then

$$|f_p(x_i(t)) - f_p(I_i)| \leq KI_i e^{-1/\epsilon_p} < \zeta, \quad t \geq 0,$$

since $e^{-1/\epsilon_p} < 1$ for any $\epsilon_p > 0$.

Lemma 2. *Uniform variance of activation level for F_1 neurons: For a given input $I = (I_1, \cdots, I_m)^T$: and $t_2 \geq t_1 \geq 0$ and a given $\epsilon_p > 0$, there exists $0 < \hat{\epsilon}_p \leq \epsilon_p$ so that $x_i(t_1; \hat{\epsilon}_p) = x_i(t_2; \epsilon_p)$.*

Proof. Choose

$$\hat{\epsilon}_p = \frac{t_1 + 1}{t_2 + 1}\epsilon_p.$$

Note that $y_j(0) = 0 < \eta_c$ for all $j \in \Lambda_c$. Therefore,

$$\Gamma := \sup\{t \geq 0; y_j(t) < \eta_c \text{ for every } j \in \Lambda_c\} > 0. \tag{4.2}$$

In other words, Γ is the first instant when at least one cluster neuron is activated. We show later that $\Gamma < \infty$.

On $[0, \Gamma)$, we have

$$f_c(y_j(t)) = 0 \text{ for } j \in \Lambda_c. \tag{4.3}$$

Therefore, using equations 3.6 and 3.7, we get

$$z_{ij}(t) = z_{ij}(0), \quad w_{ji}(t) = w_{ji}(0) \quad \text{for } t \in [0, \Gamma), \quad i \in \Lambda_p, j \in \Lambda_c. \tag{4.4}$$

Using lemma 1, we can find $\epsilon_p^0 > 0$, so that if $0 < \epsilon_p < \epsilon_p^0$, then $f_p(x_i(t)) \to f_p(I_i)$ and $f_p(x_i(0)) \to f_p(I_i)$, both from the same side because of the monotonicity of f_p. Hence, if $|f_p(I_i) - w_{ji}(0)| > \sigma$, then $|f_p(x_i(t)) - w_{ji}(t)| > \sigma$ and $|f_p(x_i(0)) - w_{ji}(0)| > \sigma$, and we have

$$h_\sigma(|f_p(x_i(t)) - w_{ji}(t)|) = h_\sigma(|f_p(x_i(0)) - w_{ji}(0)|) = 0, \qquad i \in \Lambda_p, j \in \Lambda_c, t \in [0, \Gamma).$$

Similarly, if $|f_p(I_i) - w_{ji}(0)| < \sigma$, then $|f_p(x_i(t)) - w_{ji}(t)| < \sigma$, and $|f_p(x_i(0)) - w_{ji}(0)| < \sigma$, and we have

$$h_\sigma(|f_p(x_i(t)) - w_{ji}(t)|) = h_\sigma(|f_p(x_i(0)) - w_{ji}(0)|) = 1, \qquad i \in \Lambda_p, j \in \Lambda_c, t \in [0, \Gamma).$$

Finally if $|f_p(I_i) - w_{ji}(0)| = \sigma$, then we consider two possible cases:

Case 1: $f_p(x_i(t)) \neq f_p(I_i)$ for all $t \geq 0$ and all $\epsilon_p > 0$. In this case, either $f_p(x_i(t)), f_p(x_i(0)) \in (w_{ji}(0) - \sigma, w_{ji}(0) + \sigma)$, resulting in

$$h_\sigma(|f_p(x_i(t)) - w_{ji}(t)|) = h_\sigma(|f_p(x_i(0)) - w_{ji}(0)|) = 1,$$

or $f_p(x_i(t)), f_p(x_i(0)) \notin [w_{ji}(0) - \sigma, w_{ji}(0) + \sigma]$, which means

$$h_\sigma(|f_p(x_i(t)) - w_{ji}(t)|) = h_\sigma(|f_p(x_i(0)) - w_{ji}(0)|) = 0.$$

Case 2: There exist some $\tilde{t} \geq 0$ and $\tilde{\epsilon}_p > 0$ for which $f_p(x_i(\tilde{t}); \tilde{\epsilon}_p) = f_p(I_i; \tilde{\epsilon}_p)$. In this case, for $0 \leq t \leq \tilde{t}$, using lemma 2, we can choose $\epsilon_p^0 = \min_{t \geq 0}\{\frac{t+1}{\tilde{t}+1}\tilde{\epsilon}_p\} = \frac{\tilde{\epsilon}_p}{\tilde{t}+1}$ to get

$$f_p(x_i(t); \epsilon_p) = f_p(x_i(\tilde{t}); \tilde{\epsilon}_p) \quad \text{for all } 0 < \epsilon_p < \epsilon_p^0,$$

which is also true for $t > \tilde{t}$, since $f_p(x_i(t))$ is nondecreasing and $x_i(t) \leq I_i$. Therefore,

$$\begin{aligned} h_\sigma(|f_p(x_i(t)) - w_{ji}(t)|) &= h_\sigma(|f_p(x_i(0)) - w_{ji}(0)|) \\ &= h_\sigma(|f_p(I_i) - w_{ji}(0)|) = 1. \end{aligned}$$

In all cases, we can find $\epsilon_p^0 > 0$, so if $0 < \epsilon_p < \epsilon_p^0$, then

$$\begin{aligned} h_\sigma(|f_p(x_i(t)) - w_{ji}(t)|) &= h_\sigma(|f_p(x_i(0)) \\ &\quad - w_{ji}(0)|), \quad i \in \Lambda_p, \quad j \in \Lambda_c, t \in [0, \Gamma). \end{aligned} \tag{4.5}$$

Therefore,

$$h_{ij}(t) = h_{ij}(0), \quad t \in [0, \Gamma), \quad i \in \Lambda_p, \quad j \in \Lambda_c. \tag{4.6}$$

Note also that $\tau_{ij}(0) = 0$. Therefore, using equation 4.6, we get

$$\tau_{ij}(t) = E[1 - h_{ij}(0)][1 - e^{-t/\beta}], \quad t \in [0, \Gamma], \quad i \in \lambda_p, \quad j \in \Lambda_c. \tag{4.7}$$

Here and in what follows, $[0, \Gamma] = [0, \infty)$ if $\Gamma = \infty$. Furthermore, from equation 4.2, we get

$$\epsilon_c \frac{dy_j(t)}{dt} = -y_j(t) + [1 - y_j(t)]T_j(t), \quad t \in [0, \Gamma), \quad i \in \Lambda_c, \quad j \in \Lambda_p. \tag{4.8}$$

It is natural to introduce

$$\tau_{ij}^* = E[1 - h_{ij}(0)] \tag{4.9}$$

and

$$T_j^* = D \sum_{i \in \Lambda_p} z_{ij}(0) f_p(I_i) e^{-\alpha \tau_{ij}^*}. \tag{4.10}$$

Obviously,

$$\begin{aligned} 0 &\leq \tau_{ij}^* - \tau_{ij}(t) \\ &= \tau_{ij}^* e^{-t/\beta} \leq \tau_{ij}^* e^{-t_s/\beta} \to 0 \text{ uniformly on } [t_s, \Gamma] \text{ as } \beta \to 0, \end{aligned} \tag{4.11}$$

for any $0 < t_s < \Gamma$.

Using equation 4.7, we have

$$0 \leq \tau_{ij}(t) \leq E < 1, \quad 0 \leq t \leq \Gamma, \quad i \in \Lambda_p, \quad j \in \Lambda_c \tag{4.12}$$

and

$$t - \tau_{ij}(t) + 1 \geq 1 - \tau_{ij}(t) > 0, \quad t \in [0, \quad \Gamma], \quad i \in \Lambda_p, \quad j \in \Lambda_c.$$

Therefore, using equations 4.1, 4.4, and 4.11, we get

$$\begin{aligned}|T_j(t) - T_j^*| &\leq D \sum_{i \in \Lambda_p} z_{ij}(0) |f_p(x_i(t - \tau_{ij}(t))) e^{-\alpha\tau_{ij}(t)} - f_p(I_i) e^{-\alpha\tau_{ij}^*}| \\ &\leq D \sum_{i \in \Lambda_p} z_{ij}(0) (f_p(I_i) |e^{-\alpha\tau_{ij}(t)} - e^{-\alpha\tau_{ij}^*}| + |f_p(x_i(t - \tau_{ij}(t))) \\ &\quad - f_p(I_i)| e^{-\alpha\tau_{ij}(t)}), \quad \text{for } t \in [0, \Gamma],\end{aligned} \tag{4.13}$$

where we used the inequality

$$|ac - bd| \leq b|c - d| + |a - b|c \quad \text{for } a, b, c, d \geq 0 \in \mathbb{R}.$$

On the other hand, using H2 and equation 4.1,

$$\begin{aligned}|f_p(x_i(t - \tau_{ij}(t))) - f_p(I_i)| &\leq K |x_i(t - \tau_{ij}(t)) - I_i| \\ &= K I_i e^{-(t - \tau_{ij}(t) + 1)/\epsilon_p} \\ &= K I_i e^{-t/\epsilon_p} e^{(\tau_{ij}(t) - 1)/\epsilon_p} \\ &\leq K I_i e^{-t/\epsilon_p} \\ &\leq K I_i e^{-t_s/\epsilon_p} \quad \text{for } t \in [t_s, \Gamma], \quad \forall\, 0 < t_s < \Gamma,\end{aligned} \tag{4.14}$$

and using equation 4.7,

$$\begin{aligned}e^{-\alpha\tau_{ij}(t)} - e^{-\alpha\tau_{ij}^*} &= e^{-\alpha\tau_{ij}^*} (e^{\alpha\tau_{ij}^* e^{-t/\beta}} - 1) \\ &\leq e^{\alpha E e^{-t/\beta}} - 1 \\ &\leq \frac{7}{4} \alpha E e^{-t/\beta} \\ &\leq \frac{7}{4} \alpha E e^{-t_s/\beta} \quad \text{for } t \geq t_s,\end{aligned} \tag{4.15}$$

where we used the inequality

$$e^a - 1 \le \frac{7}{4}a \quad \text{for } 0 \le a < 1$$

and assumed $0 < t_s < \Gamma$ is chosen such that

$$\alpha E e^{-t_s/\beta} < 1. \tag{4.16}$$

Now we can combine equations 4.13, 4.14, and 4.15 to get

$$|T_j(t) - T_j^*| \le P e^{-t_s/\epsilon_p} + Q e^{-t_s/\beta} \ \to 0 \text{ uniformly on } [t_s, \Gamma] \text{ as } \beta \to 0 \text{ and } \epsilon_p \to 0, \tag{4.17}$$

where $0 < t_s < \Gamma$ must satisfy equation 4.16, and

$$P = DK \sum_{i \in \Lambda_p} z_{ij}(0) I_i$$

$$Q = \frac{7}{4}\alpha DE \sum_{i \in \Lambda_p} z_{ij}(0) f_p(I_i)$$

are nonnegative constants.

Note that t_s in equation 4.11 or 4.17 can be made arbitrarily small as long as $t_s \to 0$ is slower than both $\beta \to 0$ and $\epsilon_p \to 0$. This condition can be satisfied, for example, by choosing $t_s = \max\{\sqrt{\beta}, \sqrt{\epsilon_p}\}$. However, if t_s changes as either β or ϵ_p changes, then the uniform convergence cannot be stated. In the rest of this letter, whenever we consider t_s as a dependent of β or ϵ_p, we will only use the upper-bound expressions introduced in equations 4.11 and 4.17.

Lemma 3. *Excitation of a cluster neuron: Assume that there exists $J \in \Lambda_c$ so that*

$$0 \le T_j^* < T_J^*, \quad j \in \Lambda_c \setminus \{J\} \tag{4.18}$$

and

$$\frac{T_J^*}{1 + T_J^*} > \eta_c. \tag{4.19}$$

Then there exist $\hat{t}_s \in [0, \Gamma)$, $\hat{\epsilon}_p > 0$, and $\hat{\beta} > 0$ so that if $0 < \epsilon_p < \hat{\epsilon}_p$ and $0 < \beta < \hat{\beta}$, then Γ given in equation 4.2 is finite and

$$y_J(\Gamma) = \eta_c > y_j(\Gamma), \quad \dot{y}_J(\Gamma) > 0, \quad j \in \Lambda_c \setminus \{J\} \tag{4.20}$$

and

$$y_j(t) < y_J(t) < \eta_c, \quad j \in \Lambda_c \setminus \{J\}, \quad t \in [\hat{t}_s, \Gamma). \tag{4.21}$$

Furthermore, if $\hat{\epsilon}_p$ and $\hat{\beta}$ are small, then Γ can be made arbitrarily small.

Proof. Fix $\mu > 0$ so that

$$\frac{T_J^*}{1+T_J^*} - (1+\eta_c)\mu > \eta_c. \tag{4.22}$$

Using equations 4.18 and 4.17, we can find $\hat{\beta} > 0$ and $\hat{\epsilon}_p > 0$, so that if $0 < \epsilon_p < \hat{\epsilon}_p$ and $0 < \beta < \hat{\beta}$, then

$$|T_J(t) - T_J^*| < \mu, \ \ t \in [t_s, \Gamma] \tag{4.23}$$

and

$$0 \le T_j(t) < T_J(t), \quad t \in [t_s, \Gamma], \quad j \in \Lambda_c \setminus \{J\} \tag{4.24}$$

for any given $0 < t_s < \Gamma$.

For every $j \in \Lambda_c \setminus \{J\}$, we consider three possible cases,

Case 1: $\forall\, t \in (t_s, \Gamma)$, $y_j(t) < y_J(t)$. Simply take $t_s^{\{j\}} := t_s$. Figure 4 shows an example of this typical situation.

Case 2: $\forall\, t \in (t_s, \Gamma)$, $y_j(t) > y_J(t)$. Then

$$\begin{aligned}\epsilon_c \frac{d}{dt}[y_J(t) - y_j(t)] &= -y_J(t) + y_j(t) + [1-y_J(t)]T_J(t) - [1-y_j(t)]T_j(t)\\ &\ge [1-y_j(t)]T_J(t) - [1-y_j(t)]T_j(t)\\ &= [1-y_j(t)][T_J(t) - T_j(t)]\\ &\ge (1-\eta_c)[T_J^* - T_j^* - (P_J + P_j)e^{-t_s/\epsilon_p} - (Q_J + Q_j)e^{-t_s/\beta}]\\ &\ge (1-\eta_c)[T_J^* - T_j^* - \phi(t_s, \epsilon_p, \beta)] \quad \text{for } t \in (t_s, \Gamma),\end{aligned} \tag{4.25}$$

where

$$\phi(t_s, \epsilon_p, \beta) = (P_J + P_j)e^{-t_s/\epsilon_p} + (Q_J + Q_j)e^{-t_s/\beta}.$$

On the other hand,

$$\begin{aligned}\epsilon_c \left| \frac{d}{dt}[y_J(t) - y_j(t)] \right| &= |-y_J(t) + y_j(t) + [1-y_J(t)]T_J(t)\\ &\quad - [1-y_j(t)]T_j(t)|\\ &\le |-y_J(t) + y_j(t)\| + |T_J(t)| + |T_j(t)|\\ &\le 1 + T_J^{**} + T_j^{**} \quad \text{for } t \in (0, t_s),\end{aligned} \tag{4.26}$$

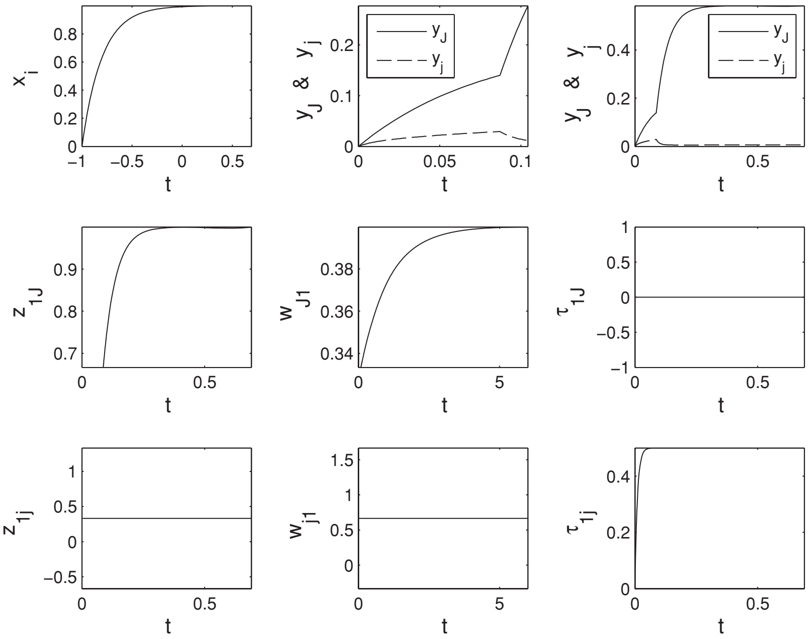


Figure 4: Plots of simulation results for the case where $y_J(t) > y_j(t)$ for $0 < t \le \Gamma$ (case 1 in the proof of lemma 3). Here $J = 1$, $j = 2$, and $\Gamma \approx 0.0871$. The top middle plot provides a closer look at the top right plot for $t \le \Gamma$. Note the wider range chosen to show the dynamics of w_{J1} and w_{j1} (LTM variables) compared to others (STM variables). The real winner is the expected winner, C_J. The expected winner in this case has a larger initial bottom-up weight ($z_{1J} > z_{1j}$) and a better similarity measure ($h_{1J} = 1$, $h_{1j} = 0$), which results in shorter signal delays ($\tau_{1J} = 0$, $\tau_{1j} \to E = 0.5$). See appendix B for the details of parameter values used.

where

$$T_J^{**} = D \sum_{i \in \Lambda_p} z_{iJ}(0) f_p(I_i),$$

$$T_j^{**} = D \sum_{i \in \Lambda_p} z_{ij}(0) f_p(I_i),$$

are nonnegative constants.

Using equation 4.25, while integrating over $[0, t_s]$ and equation 4.26 over $[t_s, \Gamma]$, we get

$$0 \ge \epsilon_c[y_J(\Gamma) - y_j(\Gamma)] \ge -(1 + T_J^{**} + T_j^{**})t_s + (1 - \eta_c) [T_J^{*} - T_j^{*} - \phi(t_s, \epsilon_p, \beta)](\Gamma - t_s) \quad (4.27)$$

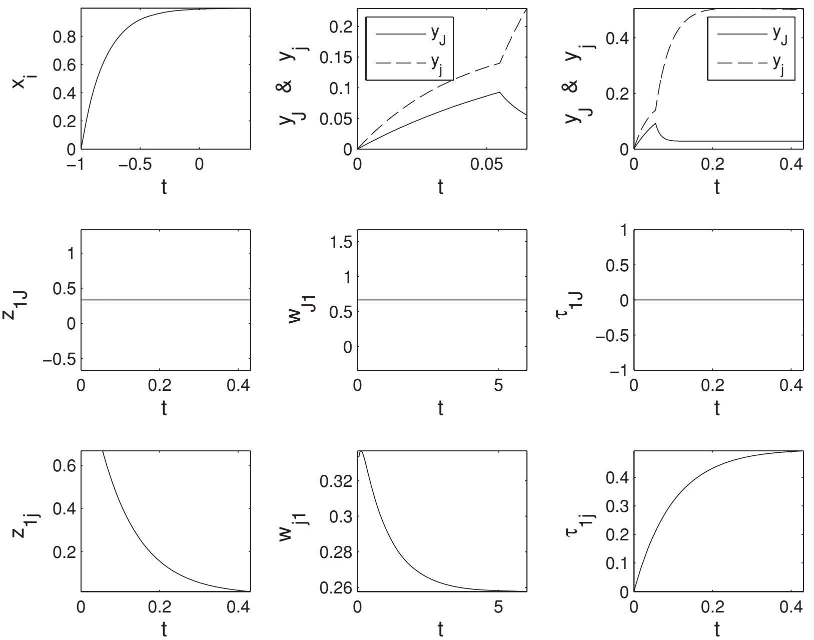


Figure 5: Plots of simulation results for the case where $y_j(t) > y_J(t)$ for $0 < t \leq \Gamma$ (case 2 in the proof of lemma 3). Here $J = 2$, $j = 1$, $\beta = 0.1$, and $\Gamma \approx 0.0550$. The top middle plot is a closer look at the top right plot for $t \leq \Gamma$. This situation can be prevented by choosing a sufficiently small value for the parameter β (e.g., $\beta = 0.01$, as shown in Figure 6). It can also be seen that w_{j1} exhibits a nonuniform behavior compared to the expected uniform behavior of the other standard cases. Note the wider range chosen to show the dynamics of w_{J1} and w_{j1} (LTM variables) compared to others (STM variables). The real winner is not the expected winner, C_J. The expected winner in this case has a smaller initial bottom-up weight ($z_{1J} < z_{1j}$) but a better similarity measure ($h_{1J} = 1, h_{1j} = 0$), which results in shorter signal delays ($\tau_{1J} = 0$, $\tau_{1j} \rightarrow E = 0.5$). However, the chosen grow rate for τ_{1j} (proportional to $1/\beta$) is not fast enough to affect the selection made by initial bottom-up weights. See appendix B for details on the parameter values used.

or

$$(1 + T_J^{**} + T_j^{**})t_s \geq (1 - \eta_c)[T_J^* - T_j^* - \phi(t_s, \epsilon_p, \beta)](\Gamma - t_s). \tag{4.28}$$

If t_s, ϵ_p and β are small enough, then equation 4.28 cannot hold, meaning that this case cannot occur. Figure 5 shows an example of this situation, where a (relatively) large chosen β causes the predictions to fail.

Case 3: $\exists\, t^* \in (t_s, \Gamma)$ such that $y_j(t^*) = y_J(t^*)$. Then

$$\epsilon_c \frac{d}{dt}[y_J(t) - y_j(t)]|_{t=t^*} = T_J(t^*) - T_j(t^*) > 0,$$

from which it follows that $y_J(t) > y_j(t)$ for $t > t^*$ close to t^*. We claim that $y_J(t) > y_j(t)$ for all $t \in (t^*, \Gamma)$. If, by way of contradiction, there exist $t^{**} \in (t^*, \Gamma)$ so that $y_J(t^{**}) = y_j(t^{**})$ and $y_J(t) > y_j(t)$ for all $t \in (t^*, t^{**})$, then

$$\frac{d}{dt}[y_J(t) - y_j(t)]|_{t=t^{**}} \leq 0.$$

But by equation 4.8, we have

$$\epsilon_c \frac{d}{dt}[y_J(t) - y_j(t)]|_{t=t^{**}} = [1 - y_J(t^{**})][T_J(t^{**}) - T_j(t^{**})] > 0$$

since $y_J(t^{**}) \leq \eta_c < 1$ and $T_J(t^{**}) > T_j(t^{**})$, a contradiction.

In this case, take $t_s^{\{j\}} := t^*$ (to be used later). Figure 6 shows an example of this situation.

Finally, by taking $\hat{t}_s = \max_{j \in \Lambda_c}\{t_s^{\{j\}}\}$, we have, $y_J(t) > y_j(t)$ on $[\hat{t}_s, \Gamma)$ and for every $j \in \Lambda_c \setminus \{J\}$.

We now show that $\Gamma < \infty$. If, by way of contradiction, $\Gamma = \infty$, then $y_j(t) < \eta_c$ for $j \in \Lambda_c$ and $t \geq 0$. Thus, by equation 4.8, we get

$$\begin{aligned}\epsilon_c \frac{d}{dt} y_J(t) &= -y_J(t) + [1 - y_J(t)]T_J(t)\\ &= -[1 + T_J(t)]y_J(t) + T_J(t)\\ &= -[1 + T_J^*]y_J(t) + T_J^* + q(t)\end{aligned}$$

with

$$|q(t)| \leq |T_J(t) - T_J^*|(|y_J(t)| + 1) \leq (\eta_c + 1)\mu, \ \text{for}\, t \in [\hat{t}_s, \Gamma).$$

Therefore, using the variation-of-constants formula, we obtain

$$\begin{aligned}&\left| y_J(t) - y_J(\hat{t}_s)e^{-(1+T_J^*)(t-\hat{t}_s)/\epsilon_c} - \frac{T_J^*}{1+T_J^*}(1 - e^{-(1+T_J^*)(t-\hat{t}_s)/\epsilon_c}) \right|\\ &\qquad \leq \frac{(\eta_c + 1)\mu}{1 + T_J^*} \leq (\eta_c + 1)\mu, \ \text{for}\, t \in [\hat{t}_s, \Gamma),\end{aligned}$$

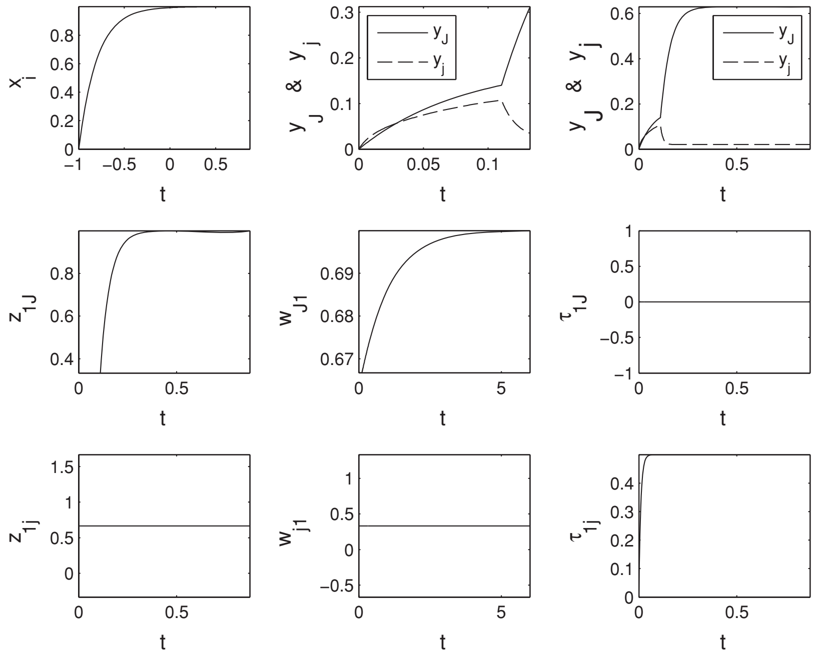


Figure 6: Plots of simulation results for for the case where $\exists\, t^* \in (t_s, \Gamma)$, such that $y_j(t^*) = y_J(t^*)$ (case 3 in the proof of lemma 3). Here $J = 2$, $j = 1$, $\beta = 0.01$, $t^* \approx 0.0292$, and $\Gamma \approx 0.1103$. The top middle plot is a closer look at the top right plot for $t \leq \Gamma$. Note the wider range chosen to show the dynamics of w_{J1} and w_{j1} (LTM variables) compared to others (STM variables). The real winner is the expected winner C_J. The expected winner in this case has a smaller initial bottom-up weight ($z_{1J} < z_{1j}$) but a better similarity measure ($h_{1J} = 1, h_{1j} = 0$), which results in shorter signal delays ($\tau_{1J} = 0$, $\tau_{1j} \to E = 0.5$). The increase of τ_{1j} finally overcomes the selection made by initial bottom-up weights. See appendix B for the details of the parameter values used.

from which it follows that

$$\liminf_{t\to\infty} y_J(t) \geq \frac{T_J^*}{1+T_J^*} - (\eta_c + 1)\mu > \eta_c,$$

a contradiction to $y_J(t) \leq \eta_c$ for all $t \geq 0$. Therefore, $\Gamma < \infty$ and $y_J(\Gamma) = \eta_c$. Note that the above argument also shows that if ϵ_c is small, then Γ can be made arbitrarily small.

Using a similar argument as above, we can also establish $y_j(\Gamma) < y_J(\Gamma)$ if $j \neq J$.

Finally, we note that at $t = \Gamma$, we have $y_J(t) = \eta_c$ and hence

$$\begin{aligned}\epsilon_c \frac{d}{dt} y_J(t) &= -y_J(t) + [1 - y_J(t)][1 + T_J(t)] \\ &= -[2 + T_J(t)]y_J(t) + 1 + T_J(t) \\ &= -(2 + T_J^*)y_J(t) + 1 + T_J^* + p(t)\end{aligned}$$

with $|p(t)| \leq (\eta_c + 1)\mu$. Note that

$$\frac{1 + T_J^*}{2 + T_J^*} \geq \frac{T_J^*}{1 + T_J^*}.$$

Thus, equation 4.19 implies $\dot{y}_J(\Gamma) > 0$.

Lemma 4. *Sustained excitation: Assume equations 4.18 and 4.19 hold. In addition, assume*

$$\frac{L}{L + m - 1} > \theta \tag{4.29}$$

and that there exists $T_{min} > 0$ so that

$$\frac{DL}{L + m - 1} f_p(I_i) > T_{min}, \quad i \in \Lambda_p, \tag{4.30}$$

$$\frac{1 + T_j^*}{2 + T_j^* + C} < \eta_c < \frac{1 + T_{min}}{2 + T_{min}}, \quad j \in \Lambda_c \setminus \{J\}, \tag{4.31}$$

and

$$T_J^* > DMe^{-\alpha E} + T_{min}, \tag{4.32}$$

where

$$M = \sum_{i \in \Lambda_p} z_{iJ}(0) f_p(I_i).$$

Then if ϵ_p, β and δ are small, $y_J(t) > \eta_c > y_j(t)$ for all $t > \Gamma$ and all $j \in \Lambda_c \setminus \{J\}$. Moreover, $h_{ij}(t) = h_{ij}(0)$ for all $t \geq 0$.

Proof. Using lemma 3 and equations 4.17 and 4.32, if β and ϵ_p are small, then there exists $\Delta > 0$ such that $y_J(t) > \eta_c > y_j(t)$ and $T_J(t) > T_{\min}$ for all $t \in (\Gamma, \Gamma + \Delta)$. We now claim that the supremum of such Δ is infinity.

By way of contradiction, if $\sup\{\Delta\} < \infty$, then for $t^* = \Gamma + \sup\{\Delta\}$, we have

$$y_J(t) > \eta_c > y_j(t), \quad T_J(t) > T_{\min}, \quad t \in [\Gamma, t^*),$$

and at least one of the following three equalities must hold:

$$y_J(t^*) = \eta_c, \;\; y_j(t^*) = \eta_c \text{ for some } j \in \Lambda_c \setminus \{J\}, \;\; T_J(t^*) = T_{\min}.$$

Step 1. We now show that $y_J(t^*) = \eta_c$ is impossible. By way of contradiction, if $y_J(t^*) = \eta_c$, then $\dot{y}_J(t^*) \leq 0$. On the other hand, we have

$$\begin{aligned}\epsilon_c \dot{y}_J(t^*) &= -y_J(t^*) + [1 - y_J(t^*)][1 + T_J(t^*)] \\ &= [2 + T_J(t^*)]\left[\frac{1 + T_J(t^*)}{2 + T_J(t^*)} - y_J(t^*)\right].\end{aligned}$$

Note that $T_J(t^*) \geq T_{\min}$; we have

$$\frac{1 + T_J(t^*)}{2 + T_J(t^*)} \geq \frac{1 + T_{\min}}{2 + T_{\min}} > \eta_c,$$

and hence, we also have

$$\epsilon_c \dot{y}_J(t^*) = [2 + T_J(t^*)]\left[\frac{1 + T_J(t^*)}{2 + T_J(t^*)} - \eta_c\right] > 0,$$

a contradiction.

Step 2: We now show that for any given $j \in \Lambda_c \setminus \{J\}$, $y_j(t^*) = \eta_c$ is impossible. Again, by way of contradiction, if $y_j(t^*) = \eta_c$, then $\dot{y}_j(t^*) \geq 0$. On the other hand, we note that equations 3.6 and 3.7 imply $\delta \dot{z}_{ij}(t) = \gamma \dot{w}_{ji}(t) = 0$, and, hence, $z_{ij}(t) = z_{ij}(\Gamma) = z_{ij}(0)$ and $w_{ji}(t) = w_{ji}(\Gamma) = w_{ji}(0)$ for all $t \in [0, t^*)$. Therefore, $h_{ij}(t) = h_{ij}(\Gamma) = h_{ij}(0)$ for all $t \in [0, t^*)$. By equation 3.4, we get $\beta \dot{\tau}_{ij}(t) = -\tau_{ij}(t) + E[1 - h_{ij}(0)]$. Hence, we can use a similar argument that led us to equation 4.17 and show that $T_j(t) \to T_j^*$ as $\beta \to 0$ and $\epsilon_p \to 0$ uniformly on $[t_s, t^*)$ for any $t_s \in (0, t^*)$. Therefore, if ϵ_p and β are small,

$$\frac{1 + T_j(t^*)}{2 + T_j(t^*) + C} \to \frac{1 + T_j^*}{2 + T_j^* + C} < \eta_c.$$

This shows that

$$\begin{aligned}\epsilon_c \dot{y}_j(t^*) &= -y_j(t^*) + [1 - y_j(t^*)][1 + T_j(t^*)] - C y_j(t^*) \\ &= -[2 + T_j(t^*) + C] y_j(t^*) + 1 + T_j(t^*) \\ &= [2 + T_j(t^*) + C)] \left[\frac{1 + T_j(t^*)}{2 + T_j(t^*) + C} - y_j(t^*) \right] < 0,\end{aligned}$$

a contradiction.

In what follows, we show that $T_J(t^*) = T_{\min}$ is impossible.

Step 3: We claim that $h_{ij}(0) = 1$ implies $h_{ij}(t) = 1$ for all $t \in [0, t^*)$. The case where $j \neq J$ is trivial. We now deal with the case where $j = J$. We must have that $h_{iJ}(\Gamma) = 1$ and, hence, $z_{iJ}(\Gamma) \geq \theta$ and $|f_p(x_i(\Gamma)) - w_{Ji}(\Gamma)| \leq \sigma$.

In the case where $z_{iJ}(\Gamma) = \theta$, we have at $t = \Gamma$ that

$$\begin{aligned}\delta \dot{z}_{iJ}(t) &= [1 - z_{iJ}(t)]L - z_{iJ}(t) \sum_{k \in \Lambda_p \setminus \{i\}} h_{kJ}(t) \\ &= L - [L + \sum_{k \in \Lambda_p \setminus \{i\}} h_{kJ}(t)] z_{iJ}(t) \\ &= L - [L + \sum_{k \in \Lambda_p \setminus \{i\}} h_{kJ}(t)] \theta \\ &\geq L - [L + m - 1]\theta > 0\end{aligned}$$

since $\theta < L/(L + m - 1)$. Hence, $z_{iJ}(t) > \theta$ for all $t > \Gamma$ but close to Γ.

In the case where $f_p(x_i(\Gamma)) - w_{Ji}(\Gamma) = \sigma$, we have

$$\begin{aligned}\gamma \dot{w}_{Ji}(\Gamma) &= -w_{Ji}(\Gamma) + f_p(x_i(\Gamma - \tau_{iJ}(\Gamma))) e^{-\alpha \tau_{iJ}(\Gamma)} \\ &= -w_{Ji}(\Gamma) + f_p(x_i(\Gamma)) \\ &= \sigma > 0,\end{aligned}$$

and hence $f_p(x_i(t)) - w_{Ji}(t) < \sigma$ for all $t > \Gamma$ and t close to Γ. Similarly, we show that $f_p(x_i(\Gamma)) - w_{Ji}(\Gamma) = -\sigma$ leads to $f_p(x_i(t)) - w_{Ji}(t) > -\sigma$ for all $t > \Gamma$ and t close to Γ.

Therefore, in any case, we have $z_{iJ}(t) > \theta$ and $|f_p(x_i(t)) - w_{Ji}(t)| < \sigma$ for all $t > \Gamma$ and t close to Γ.

If the above claim is not true, then there exists the first $\hat{t} \in (\Gamma, t^*)$ so that either $z_{iJ}(\hat{t}) = \theta$ or $|f_p(x_i(\hat{t})) - w_{Ji}(\hat{t})| = \sigma$. If $z_{iJ}(\hat{t}) = \theta$ and $|f_p(x_i(\hat{t})) - w_{Ji}(\hat{t})| < \sigma$, then $\dot{z}_{iJ}(\hat{t}) \leq 0$. Using the same argument as above, we get $\delta \dot{z}_{iJ}(\hat{t}) > 0$, a contradiction. Therefore, if the claim is not true, then

$|f_p(x_i(\hat{t})) - w_{Ji}(\hat{t})| = \sigma$. In this case, if $f_p(x_i(\hat{t})) - w_{Ji}(\hat{t}) = \sigma$, then using $\tau_{iJ}(t) = 0,\ t \in (\Gamma, \hat{t})$, we have

$$\gamma \dot{w}_{Ji}(t) = -w_{Ji}(t) + f_p(x_i(t)),\ \ t \in (\Gamma, \hat{t}),$$

which can be solved analytically to get

$$\begin{aligned}
w_{Ji}(\hat{t}) &= w_{Ji}(\Gamma)e^{-(\hat{t}-\Gamma)/\gamma} + e^{-(\hat{t}-\Gamma)/\gamma}\int_{\Gamma}^{\hat{t}} \frac{e^{(t-\Gamma)/\gamma}}{\gamma} f_p(x_i(t))\,dt \\
&\geq w_{Ji}(\Gamma)e^{-(\hat{t}-\Gamma)/\gamma} \\
&\quad + e^{-(\hat{t}-\Gamma)/\gamma}\int_{\Gamma}^{\hat{t}} \frac{e^{(t-\Gamma)/\gamma}}{\gamma}(f_p(x_i(\hat{t})) - K[x_i(\hat{t}) - x_i(t)])\,dt \\
&= w_{Ji}(\Gamma)e^{-(\hat{t}-\Gamma)/\gamma} + (1 - e^{-(\hat{t}-\Gamma)/\gamma})f_p(x_i(\hat{t})) \\
&\quad - Ke^{-(\hat{t}-\Gamma)/\gamma}\int_{\Gamma}^{\hat{t}} \frac{e^{(t-\Gamma)/\gamma}}{\gamma}[x_i(\hat{t}) - x_i(t)]\,dt \\
&\geq w_{Ji}(\Gamma)e^{-(\hat{t}-\Gamma)/\gamma} + (1 - e^{-(\hat{t}-\Gamma)/\gamma})f_p(x_i(\hat{t})) - K(1 - e^{-(\hat{t}-\Gamma)/\gamma})\xi,
\end{aligned}$$

where $|x_i(t_1) - x_i(t_2)| \leq \xi,\ t_1, t_2 \in (\Gamma, \hat{t})$, and hence, using $w_{Ji}(\hat{t}) = f_p(x_i(\hat{t})) - \sigma$, we get

$$\begin{aligned}
-\sigma &\geq w_{Ji}(\Gamma)e^{-(\hat{t}-\Gamma)/\gamma} - e^{-(\hat{t}-\Gamma)/\gamma} f_p(x_i(\hat{t})) - K(1 - e^{-(\hat{t}-\Gamma)/\gamma})\xi \\
&\geq w_{Ji}(\Gamma)e^{-(\hat{t}-\Gamma)/\gamma} - e^{-(\hat{t}-\Gamma)/\gamma} f_p(x_i(\Gamma)) - Ke^{-(\hat{t}-\Gamma)/\gamma}(x_i(\hat{t}) \\
&\quad - x_i(\Gamma)) - K(1 - e^{-(\hat{t}-\Gamma)/\gamma})\xi \\
&\geq -\sigma e^{-(\hat{t}-\Gamma)/\gamma} - K\xi,
\end{aligned}$$

or, equivalently,

$$\sigma \leq \sigma e^{-(\hat{t}-\Gamma)/\gamma} + K\xi,$$

which cannot hold for a small ϵ_p, as $\xi \to 0$ when $\epsilon_p \to 0$, a contradiction. Similarly, we can show that $f_p(x_i(\hat{t})) - w_{Ji}(\hat{t}) = -\sigma$ leads to a contradiction, since

$$\begin{aligned}
w_{Ji}(\hat{t}) &= w_{Ji}(\Gamma)e^{-(\hat{t}-\Gamma)/\gamma} + e^{-(\hat{t}-\Gamma)/\gamma}\int_{\Gamma}^{\hat{t}} \frac{e^{(t-\Gamma)/\gamma}}{\gamma} f_p(x_i(t))\,dt \\
&\leq w_{Ji}(\Gamma)e^{-(\hat{t}-\Gamma)/\gamma} + e^{-(\hat{t}-\Gamma)/\gamma}\int_{\Gamma}^{\hat{t}} \frac{e^{(t-\Gamma)/\gamma}}{\gamma}(f_p(x_i(\Gamma)) \\
&\quad + K[x_i(t) - x_i(\Gamma)])\,dt
\end{aligned}$$

$$
\begin{aligned}
&= w_{Ji}(\Gamma)e^{-(\hat{t}-\Gamma)/\gamma} + (1-e^{-(\hat{t}-\Gamma)/\gamma}) f_p(x_i(\Gamma))\\
&\quad + Ke^{-(\hat{t}-\Gamma)/\gamma}\int_{\Gamma}^{\hat{t}}\frac{e^{(t-\Gamma)/\gamma}}{\gamma}[x_i(t)-x_i(\Gamma)]\,dt\\
&\le \sigma e^{-(\hat{t}-\Gamma)/\gamma} + f_p(x_i(\Gamma)) + K(1-e^{-(\hat{t}-\Gamma)/\gamma})\xi,
\end{aligned}
$$

which, using $w_{Ji}(\hat{t}) = f_p(x_i(\hat{t})) + \sigma$, leads to

$$
f_p(x_i(\hat{t})) + \sigma \;\le\; \sigma e^{-(\hat{t}-\Gamma)/\gamma} + f_p(x_i(\Gamma)) + K(1-e^{-(\hat{t}-\Gamma)/\gamma})\xi
$$

or

$$
\begin{aligned}
\sigma &\le \sigma e^{-(\hat{t}-\Gamma)/\gamma} + (f_p(x_i(\Gamma)) - f_p(x_i(\hat{t}))) + K(1-e^{-(\hat{t}-\Gamma)/\gamma})\xi\\
&\le \sigma e^{-(\hat{t}-\Gamma)/\gamma} + K(1-e^{-(\hat{t}-\Gamma)/\gamma})\xi.
\end{aligned}
$$

But this cannot hold for a small ϵ_p, as $\xi \to 0$ when $\epsilon_p \to 0$, a contradiction. This verifies the claim.

Step 4: We now prove that if $h_{iJ}(0) = 1$ and $z_{iJ}(0) > \frac{L}{L+m-1}$ for some $i \in \Lambda_p$, then $T_J(t^*) > T_{\min}$. By the result in step 3, we know that $h_{iJ}(t) = 1$ for all $t \in [0, t^*)$, and hence,

$$
\begin{aligned}
\delta\dot{z}_{iJ}(t) &= [1 - z_{iJ}(t)]L - z_{iJ}(t)\textstyle\sum_{k\in\Lambda_p\setminus\{i\}} h_{kJ}(t)\\
&= [1 - z_{iJ}(t)]L - z_{iJ}(t)l_i = (L + l_i)\left[\frac{L}{L+l_i} - z_{iJ}(t)\right],
\end{aligned}
$$

for all $t \in [0, t^*)$, where $l_i = \sum_{k\in\Lambda_p\setminus\{i\}} h_{kJ}(t) \le m-1$ is a constant integer. Therefore, if $z_{iJ}(0) \ge \frac{L}{L+l_i}$, then $z_{iJ}(t)$ is decreasing, but always $z_{iJ}(t) \ge \frac{L}{L+l_i}$ for all $t \in [0, t^*)$, and if $\frac{L}{L+m-1} \le z_{iJ}(0) < \frac{L}{L+l_i}$ then $z_{iJ}(t)$ is increasing for all $t \in [0, t^*)$, and hence, $z_{iJ}(t) > z_{iJ}(0) > \frac{L}{L+m-1}$. Therefore,

$$
\begin{aligned}
T_J(t) &\ge Dz_{iJ}(t)f_p(x_i(t))\\
&\ge \frac{DL}{L+m-1}f_p(x_i(t))\\
&\to \frac{DL}{L+m-1}f_p(I_i) \;\text{ as } \epsilon_p \to 0\\
&> T_{\min}
\end{aligned}
$$

for all $t \in [0, t^*]$.

Step 5: We claim that if δ is sufficiently small and if $h_{ij}(0) = 0$, then $h_{ij}(t) = 0$ for all $t \in [0, t^*)$. This is clearly true if $j \ne J$. If $h_{iJ}(0) = 0$, then at Γ, we have either $z_{iJ}(\Gamma) < \theta$ or $|f_p(x_i(\Gamma)) - w_{Ji}(\Gamma)| > \sigma$. In the latter case, we can find $Q > 0$ so that on $[\Gamma, \Gamma + Q] \subset [\Gamma, t^*)$, we have $|f_p(x_i(t)) - w_{Ji}(t)| > \sigma$ and, hence, $h_{iJ}(t) = 0$. This implies that

$$\delta \dot{z}_{iJ}(t) = -z_{iJ}(t) - z_{iJ}(t) \sum_{k \in \Lambda_p \setminus \{i\}} h_{kJ}(t) \leq -z_{iJ}(t),$$

from which it follows that

$$z_{iJ}(t) \leq z_{iJ}(0) e^{-(t-\Gamma)/\delta}.$$

Consequently, $z_{iJ}(\Gamma + Q) < \theta$ provided $z_{iJ}(0)e^{-Q/\delta} < \theta$. In any case, if $h_{iJ}(0) = 0$ and if δ is small, then we can find small $\theta > 0$ so small that $h_{iJ}(t) = 0$ for $t \in [\Gamma, \Gamma + Q]$ and $z_{iJ}(\Gamma + Q) < \theta$. If $z_{iJ}(t) < \theta$ does not hold for all $t \in [\Gamma + Q, t^*)$, then there is the first $s \in (\Gamma + Q, t^*)$ so that $z_{iJ}(s) = \theta$. However, on $[\Gamma + Q, s)$, we have $h_{iJ}(t) = 0$ and thus $\delta \dot{z}_{iJ}(t) \leq 0$, from which it follows that $z_{iJ}(s) = \theta$ is impossible. Thus, $z_{iJ}(t) < \theta$ for all $t \in [\Gamma + Q, t^*)$, and hence, $h_{iJ}(t) = 0$ for all $t \in [0, t^*)$.

Step 6: We now show that $T_J(t^*) = T_{\min}$ is impossible. This has been shown already in step 4 if $h_{iJ}(0) = 1$, $f_p(I_i) \geq 1$ and $z_{iJ}(0) > \frac{L}{L+m-1}$ for some $i \in \Lambda_p$.

For any given $i \in \Lambda_p$ with $h_{iJ}(0) = 1$ and $z_{iJ}(0) \leq \frac{L}{L+m-1}$, we have

$$\delta \dot{z}_{iJ}(t) = [1 - z_{iJ}(t)]L - z_{iJ}(t) l_i = (L + l_i)\left[\frac{L}{L + l_i} - z_{iJ}(t)\right],$$

where $l_i = \sum_{k \in \Lambda_p \setminus \{i\}} h_{kJ}(t) \leq m - 1$ is a constant integer. Therefore, since $z_{iJ}(0) \leq \frac{L}{L+m-1} \leq \frac{L}{L+l_i}$, we must have that $z_{iJ}(t)$ is increasing on $[0, t^*)$.

In conclusion, if there exists no $i \in \Lambda_p$ so that $h_{iJ}(0) = 1$ and $z_{iJ}(0) > \frac{L}{L+m-1}$, then we have

$$\begin{aligned}
T_{\min} &< T_J^* - DMe^{-\alpha E} \\
&= D \sum_{i \in \Lambda_p} z_{iJ}(0) f_p(I_i) e^{-\alpha \tau_{iJ}^*} - D \sum_{i \in \Lambda_p} z_{iJ}(0) f_p(I_i) e^{-\alpha E} \\
&= D \sum_{i \in \Lambda_p, h_{iJ}(0)=1} z_{iJ}(0) f_p(I_i)(e^{-\alpha \tau_{iJ}^*} - e^{-\alpha E}) \\
&\quad + D \sum_{i \in \Lambda_p, h_{iJ}(0)=0} z_{iJ}(t) f_p(I_i)(e^{-\alpha \tau_{iJ}^*} - e^{-\alpha E}) \\
&= D \sum_{i \in \Lambda_p, h_{iJ}(0)=1} z_{iJ}(0) f_p(I_i)(1 - e^{-\alpha E})
\end{aligned}$$

$$
\begin{aligned}
&\le D \sum_{i\in\Lambda_p, h_{iJ}(0)=1} z_{iJ}(0) f_p(I_i)\\
&\le D \sum_{i\in\Lambda_p, h_{iJ}(0)=1} z_{iJ}(0)(f_p(x_i(t)) + K(I_i - x_i(t)))\\
&= D \sum_{i\in\Lambda_p, h_{iJ}(0)=1} z_{iJ}(0) f_p(x_i(t)) + DK \sum_{i\in\Lambda_p, h_{iJ}(0)=1} z_{iJ}(0)(I_i - x_i(t))\\
&\le D \sum_{i\in\Lambda_p, h_{iJ}(0)=1} z_{iJ}(t) f_p(x_i(t)) + DKe^{-(t+1)/\epsilon_p} \sum_{i\in\Lambda_p, h_{iJ}(0)=1} z_{iJ}(0) I_i\\
&\le D \sum_{i\in\Lambda_p} z_{iJ}(t) f_p(x_i(t-\tau_{ij}(t)))e^{-\alpha\tau_{ij}(t)}\\
&\quad + DKe^{-(t+1)/\epsilon_p} \sum_{i\in\Lambda_p, h_{iJ}(0)=1} z_{iJ}(0) I_i\\
&= T_J(t) + DKe^{-(t+1)/\epsilon_p} \sum_{i\in\Lambda_p, h_{iJ}(0)=1} z_{iJ}(0) I_i, \quad \text{for } t \in [0, t^*),
\end{aligned}
$$

where we used $\tau_{ij}(t) = 0$ if $h_{ij}(0) = 1$, for $t \in [0, t^*)$. Taking limits when $\epsilon_p \to 0$, we get

$$
T_{\min} \ < \ T_J^* - DMe^{-\alpha E} \ \le \ T_J(t), \quad \text{for } t \in [0, t^*).
$$

Therefore, $T_{\min} = T_J(t^*)$ is impossible. This completes the proof.

Lemma 5. *Learning at infinity. Assume all conditions of lemma 4 are satisfied, and assume that $\beta, \epsilon_p, \delta$ are sufficiently small. Then for any $i \in \Lambda_p$ and $j \in \Lambda_c$ with $j \ne J$, $z_{ij}(t)$ and $w_{ji}(t)$ remain unchanged for all $t \ge 0$. But*

$$
\lim_{t\to\infty} z_{iJ}(t) = \begin{cases} 0 & \textit{if } h_{iJ}(0) = 0 \\ \dfrac{L}{L + l_i} & \textit{if } h_{iJ}(0) = 1 \end{cases}
$$

and

$$
\lim_{t\to\infty} w_{Ji}(t) = f_p(I_i) e^{-\alpha\tau_{iJ}^*},
$$

where $l_i = \#\{k \in \Lambda_p \setminus \{i\}; h_{kJ}(0) = 1\}$.

Proof. This is now obvious after lemma 4, since for $j \in \Lambda_c \setminus \{J\}$, we have $\gamma \dot{w}_{ji}(t) = \delta \dot{z}_{ij}(t) = 0$, and on $[\Gamma, \infty)$, we have

$$\begin{aligned}
\gamma \dot{w}_{Ji}(t) &= -w_{Ji}(t) + f_p(x_i(t - \tau_{iJ}(t)))e^{-\alpha\tau_{iJ}(t)},\\
\delta \dot{z}_{iJ}(t) &= [1 - z_{iJ}(t)]L - z_{iJ}(t)l_i, \quad \text{if } h_{iJ}(0) = 1,\\
\delta \dot{z}_{iJ}(t) &= -z_{iJ}(t) - z_{iJ}(t)l_i, \quad \text{if } h_{iJ}(0) = 0.
\end{aligned}$$

5 Proof of Theorem 2

Note that on $[\Gamma, \infty)$, we have

$$\begin{aligned}
\gamma \dot{w}_{Ji}(t) &= -w_{Ji}(t) + f_p(x_i(t - \tau_{iJ}(t)))e^{-\alpha\tau_{iJ}(t)},\\
\delta \dot{z}_{iJ}(t) &= [1 - z_{iJ}(t)]L - z_{iJ}(t)l_i, \quad \text{if } h_{iJ}(0) = 1,\\
\delta \dot{z}_{iJ}(t) &= -z_{iJ}(t) - z_{iJ}(t)l_i, \quad \text{if } h_{iJ}(0) = 0.
\end{aligned}$$

Note also that $\Gamma \to 0$ as $\epsilon_p, \beta, \epsilon_c \to 0$. Therefore,

$$\begin{aligned}
&w_{Ji}(1) \to e^{-1/\gamma} w_{Ji}(0) + f_p(I_i)e^{-\alpha\tau^*_{iJ}}(1 - e^{-1/\gamma}), \quad \text{as } \epsilon_p \to 0, \quad \beta \to 0,\\
&z_{iJ}(1) = e^{-(L+l_i)/\delta} z_{iJ}(0) + \frac{L}{L + l_i}(1 - e^{-(L+l_i)/\delta}) \to \frac{L}{L + l_i} \quad \text{as } \delta \to 0,\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } h_{iJ}(0) = 1,\\
&z_{iJ}(1) = e^{-(1+l_i)/\delta} z_{iJ}(0) \to 0 \quad \text{as } \delta \to 0, \quad \text{if } h_{iJ}(0) = 0.
\end{aligned}$$

This proves the first part of theorem 2.

To prove part vi of theorem 2, we first note that if $j \neq J$, then $z_{ij}(t) = z_{ij}(0)$ for all $t \geq 0$, and hence $D_j(t) = D_j(0)$ for all $t \geq 0$. On the other hand, $z_{iJ}(t) = z_{iJ}(0)$ on $[0, \Gamma)$. The case where $h_{iJ}(0) = 0$ is simple since on $[0, \infty)$, we have

$$\delta \dot{z}_{iJ}(t) = -z_{iJ}(t) - z_{iJ}(t) \sum_{k \neq i} h_{kJ}(t) \leq -z_{iJ}(t).$$

For the case where $h_{iJ}(0) = 1$, we have on the interval $[\Gamma, \infty)$, $\delta \dot{z}_{iJ}(t) = (L + l_i)[\frac{L}{L+l_i} - z_{iJ}(t)]$. Therefore, if $z_{iJ}(s) \geq \frac{L}{L+l_i}$ for some $s \in [\Gamma, \infty)$, then $z_{iJ}(t) \geq \frac{L}{L+l_i}$ for all $t \geq s$, and if $\theta \leq z_{iJ}(\Gamma) < \frac{L}{L+l_i}$, then $z_{iJ}(t)$ is increasing for all $t \geq \Gamma$, from which the rest of part vii follows naturally.

6 Conclusion

We have shown that the selective output signaling mechanism (SOS) in a PART network can be identified with two physiologically relevant properties of living neural networks: the adaptability of transmission delays and transmission information loss. Thus, it is not necessary to assign the SOS to a hypothetical hidden layer of neurons. The key concept is the similarity between the level of activation of an input neuron and the corresponding component of the template of the considered cluster neuron. This similarity is measured by h_{ij}, which assumes values of either 1 or 0. This simple measure of the similarity leads to an explicit formula, 4.7, for the dissimilarity-driven signal transmission delay, derived from the equation governing the evolution of the time lag. Such an explicit formula of the time lag determines the value of the delays for all future time by looking at the initial values of the network. Our starting point is that time delays in the signal transmission of the proposed neural network are adaptive, following rule 3.4, which describes the delay shift guided by the dissimilarity. Interestingly, the choice of the dissimilarity measure 3.5 enables us to conclude that this transmission delay is either zero or quickly stabilizes to a positive constant determined by the dissimilarity. As a consequence, the network works as it follows the delay selection mechanism, suggested in Eurich et al. (1999). This seems to suggest that in some cases, the delay selection mechanism can be regarded as a limit of the delay shift mechanism. The assumption that signal strength decays if the transmission is delayed can be replaced by the mechanism of delay selection as follows: replacing equation 3.3 by $T_j(t) = D_f \sum_{i\in\Lambda_p,\tau_{ij}(t)=0} z_{ij}(t) f_p(x_i(t)) + D_d \sum_{i\in\Lambda_p,\tau_{ij}(t)>0} z_{ij}(t) f_p(x_i(t - \tau_{ij}(t)))$ with $D_f \gg D_d$. This definition makes sense, as we have shown that either $\tau_{ij}(t) = 0$ for all $t \geq 0$ or $\tau_{ij}(t)$ is close to a positive constant for all $t \geq 0$.

We emphasize, however, that our choice of the similarity-dissimilarity measure represents a gross approximation of an otherwise fuzzy concept. More precisely, the similarity between the output of an input neuron (in the F_1 layer) and the corresponding component of the feature vector of its target clustering neuron (in the F_2 layer) does not necessarily take the value of either 1 (similar) or 0 (dissimilar); there should be a gray scale for this similarity, like Zadel's metric for fuzzy mathematics. A natural solution is to consider a more general similarity measure function $h_{ij}(t)$ such as $h_{ij}(t) = M(|f_p(x_i(t)) - w_{ji}(t)|)N(z_{ij}(t))$ with $M : R^+ \to [0, 1]$ being monotonically nonincreasing and $N : [0, 1] \to [0, 1]$ being monotonically nondecreasing, and such that $M(0) = 1 > 0 = M(\infty)$ and $N(0) = 0 < 1 = N(1)$. This would also have the advantage for the software implementation of our neural network architecture for clustering, as the two parameters, σ and θ, will no longer be needed. Unfortunately, such a general similarity measure leads to a very complicated system of delay differential equations with adaptive delay. A general qualitative theory and numerical package for such a system is not available at this time and thus should be developed in the future.

Appendix A: Signal Loss with Delayed Transmission

The cable equation for a linear RC membrane cylinder, extending infinitely in both directions, is given by the partial differential equation

$$\frac{\partial}{\partial t}v(t,x) - \frac{\partial^2}{\partial x^2}v(t,x) + v(t,x) = \delta(t)\delta(x).$$

This can be solved using the Fourier transform to yield the Green's function, which gives the voltage response as a function of t and x after an impulse stimulation at time $t = 0$ at point $x = 0$ for

$$v(t,x) = \frac{\mathcal{H}(t)}{\sqrt{4\pi t}} \exp\left(-t - \frac{x^2}{4t}\right), \tag{A.1}$$

where $\mathcal{H}$ is the Heaviside step function, the distance x is normalized by the space constant, and time t is normalized by the time constant.

To compute the delay from stimulus onset to maximal response, we differentiate this function with respect to t, ignoring the scalar term $\sqrt{4\pi}$, which does not affect the maximum, and obtain

$$\begin{aligned}\frac{\partial}{\partial t}v(t,x) &= -1/2t^{-3/2}\exp\left(-t - \frac{x^2}{4t}\right)\\ &\quad + t^{-1/2}\exp\left(-t - \frac{x^2}{4t}\right)\left(-1 + \frac{x^2}{4t^2}\right).\end{aligned} \tag{A.2}$$

Setting the left-hand side equal to zero and solving for t gives the quadratic equation $4t_m^2 + 2t_m - x^2 = 0$, where t_m is the time of the maximum, with positive time solution

$$\text{delay} = t_m = \frac{-1 + \sqrt{1 + 4x^2}}{4}.$$

Solving this equation for x and substituting the result into equation A.1 gives the amplitude (the voltage at the maximum) as a function of the delay t_m, yielding

$$a(t_m) = \frac{e^{-1/2}}{\sqrt{4\pi t_m}} e^{-2t_m}.$$

Appendix B: Details of Simulations

Here we give the details of parameters and prestored patterns we used for the simulations shown in Figures 3b and 4 to 6.

We chose $f_p(x) = x$ for all the simulations. Some of the parameters are set the same for all the experiments:

$$\alpha = 2.0, \ \ \epsilon_p = 0.2, \ \ \epsilon_c = 0.1, \ \ E = 0.5, \ \ \gamma = 1.0, \ \ \delta = 0.1, \ \ L = 2.0.$$

In what follows, the index J refers to the index for which $T_J^* > T_j^*$, for $j \in \Lambda_c \setminus \{J\}$.

Example 1: Details of parameters for the clustering shown in Figure 3b. $m = 3$, 270 data points. $\eta_c = 0.78$, $C = 1.74$, $\beta = 0.01$, $\sigma = 1.1$, $\theta = 0.00001$, $\rho = 2$. Since our results are based on the nonnegativity of the input, the points are shifted along the z-axis before being fed to the network (we used the transformation $z \to z + 6.0$).

Example 2: Details of parameters for the simulations shown in Figure 4. $m = 1, n = 2, J = 1$. $z_{11}(0) = 2/3, z_{12}(0) = 1/3, w_{11}(0) = 1/3, w_{21}(0) = 2/3$, $I_1 = 0.4$, $\eta_c = 0.14$, $C = 7.0$, $\beta = 0.01$, $\sigma = 0.17$, $\theta = 0.05$. Output is: winner = 1, $\Gamma \approx 0.0871$.

Example 3: Details of parameters for the simulations shown in Figure 5. $m = 1, n = 2, J = 2$. All initial values and parameters as in example 2, except $I_1 = 0.7$, $\beta = 0.1$. Output is: winner = 1, $\Gamma \approx 0.0550$. The winner is not C_J since β is not chosen small enough.

Example 4: Details of parameters for the simulations shown in Figure 6. $m = 1, n = 2, J = 2$. All initial values and parameters as in example 3, except $\beta = 0.01$. Output is: winner = 2, $\Gamma \approx 0.1103$, $t^* \approx 0.0292$.

Acknowledgments

This research was partially supported by the William R. Kenan, Jr. Foundation (J.M.), the Canada Research Chairs Program (J.W.), Mathematics for Information Technology and Complex Systems (J.W.), the Natural Sciences and Engineering Research Council of Canada (H.Z., J.W.), the National Institutes of Mental Health (J.H., J.M.; NIMH-47542), and the National Science Foundation (J.M.; NSF-0617072 and NSF-1028970).

References

Aggarwal, C. C., Procopiuc, C., Wolf, J. L., Yu, P. S., & Park, J. S. (1999). Fast algorithms for projective clustering. In *Proceedings of the 1999 ACM SIGMOID International Conference on Management of Data* (pp. 61–72). New York: ACM Press.

Aggarwal, C. C., & Yu, P. S. (2000). Finding generalized projected clusters in high dimensional spaces. In *Proceedings of the 2000 ACM SIGMOID International Conference on Management of Data* (pp. 70–81). New York: ACM Press.

Bale, M. R., & Petersen, R. S. (2009). Transformation in the neural code for whisker deflection direction along the lemniscal pathway. *Journal of Neurophysiology, 102*, 2771–2780.

Cao, Y. (2002). *Neural networks for clustering: Theory, algorithm and applications*. Unpublished doctoral dissertation, York University.

Cao, Y., & Wu, J. (2002). Projective ART for clustering data sets in high dimensional spaces. *Neural Networks, 15*, 105–120.

Cao, Y., & Wu, J. (2004). Dynamics of projective adaptive resonance theory model: The foundation of PART algorithm. *IEEE Transactions on Neural Networks, 15*, 245–260.

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

Carpenter, G. A., & Grossberg, S. (1987b). ART2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics, 26*, 4919–4930.

Carpenter, G. A., & Grossberg, S. (1987c). Neural dynamics of category learning and recognition: Attention, memory and amnesia. In S. Grossberg (Ed.), *The adaptive brain I* (pp. 256–258). New York: Elsevier.

Carpenter, G. A., & Grossberg, S. (1990). ART3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural Networks, 3*, 129–152.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks, 3*, 698–713.

Carpenter, G. A., Grossberg, S., & Reynolds, J. H. (1991). ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks, 4*, 565–588.

Carpenter, G. A., Grossberg, S., & Rosen, D. B. (1991a). ART2—A: An adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks, 4*, 493–504.

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

Carr, C. E. (1993). Processing of temporal information in the brain. *Annual Reviews of Neuroscience, 16*, 223–243.

Eurich, C. W., Pawelzik, K., Cowan, J. D., & Milton, J. G. (1999). Dynamics of self-organized delay adaptation. *Phys. Rev. Lett., 82*, 1594–1597.

Fields, R. D. (2005). Myelination: An overlooked mechanism of synaptic plasticity? *Neuroscientist, 11*(6), 528–531.

Hartung, F., Krisztin, T., Walther, H. O., & Wu, J. (2006). Functional differential equations with state-dependent delays: Theory and applications. In A. Canada, P. Drabek, & A. Fonda (Eds.), *Handbook of differential equations* (Vol. 3, pp. 435–545). New York: Elsevier.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Hunter, J. D., Wu, J., & Milton, J. G. (2008). Clustering neural spike trains with transient responses. In *Proceedings of the 47th IEEE Conferfence on Decision and Control* (pp. 2000–2005). Piscataway, NJ: IEEE Press.

Milton, J. G., & Mackey, M. C. (2000). Neural ensemble coding and statistical periodicity: Speculations on the operation of the mind's eye. *Journal of Physiology (Paris), 94*, 489–503.

Sincich, L. C., Horton, J. C., & Sharpee, T. O. (2009). Preserving information in neural transmission. *J. Neurosci., 29*(19), 6207–6216.

Stanford, L. R. (1987). Conduction velocity variations minimize conduction time differences among retinal ganglion cell axons. *Science, 238*, 358–360.

Stevens, B., Tanner, S., & Fields, R. D. (1998). Control of myelination by specific patterns of neural impulses. *Journal of Neuroscience, 18*(22), 9303–9311.

Takahashi, H., Kobayashi, T., & Honda, H. (2005). Construction of robust prognostic predictors by using projective adaptive resonance theory as a gene filtering method. *Bioinformatics, 21*(2), 179–186.

von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik, 14*, 85–100.

Williamson, J. R. (1996). Gaussian ARTMAP: A neural network for fast incremental learning of noisy multidimensional maps. *Neural Networks, 9*, 881–897.

Zalc, B., & Fields, R. D. (2000). Do action potentials regulate myelination? *Neuroscientist, 6*(1), 5–12.

Received July 29, 2010; accepted October 22, 2010.