Submitted to: *Physcia D*, special issue on "Unconventional Computing"

Reliable Computing with Unreliable Components: Using Separable Environments to Stabilize Long-Term Information Storage

M. A. Nugent[1], R. Porter[2] and G. T. Kenyon[3]

[1] KnowmTech, Santa Fe NM,
[2] International, Space and Response Division, Los Alamos National Laboratory
[3]Physics Division, Los Alamos National Laboratory, Los Alamos, NM

To whom correspondence should be addressed:
Garrett T. Kenyon
Physics Division, P-21,
P.O. Box 1663, MS-D454
LANL
Los Alamos, NM 87545
505-667-1900 (W)
505-665-4507 (Fax)
gkenyon@lanl.gov

Abstract

How, in the face of both intrinsic and extrinsic volatility, can unconventional computing fabrics store information over arbitrarily long periods? Here, we argue that the predictable structure of many realistic environments, both natural and artificial, can be used to maintain useful categorical boundaries even when the computational fabric itself is inherently volatile and the inputs and outputs are partially stochastic. As a concrete example, we consider the storage of binary classifications in connectionist networks, although the underlying principles should be applicable to other unconventional computing paradigms. Specifically, we demonstrate that an unsupervised, activity dependent plasticity rule, AHAH (Anti-Hebbian-And-Hebbian), allows binary classifications to remain stable even when the underlying synaptic weights are subject to random noise. When embedded in environments composed of separable features, the weight vector is restricted by the AHAH rule to local attractors representing stable partitions of the input space, allowing unsupervised recovery of stored binary classifications following random perturbations that leave the system in the same basin of attraction. We conclude that the stability of long-term memories may depend not so much on the reliability of the underlying substrate but rather on the reproducible structure of the environment itself, suggesting a new paradigm for reliable computing with unreliable components.

Introduction

A major motivation for pursuing advances in Unconventional Computing is the prospect of using nano/molecular-scale elements to achieve vast increases in component density (Calude et al., 2006 ; A. Adamatzky, 2007). As individual computational elements become smaller, however, they become increasingly susceptible to random fluctuations that, in the absence of formal mechanisms for local error correction, would eventually erase previously stored information. Moreover, even if it were possible to construct reversible nano/molecular switches that remained stable for many months or years (Miller et al., 2005; Shouval, 2005), it is nonetheless unlikely that such elements could be programmed explicitly in a macroscopic device. Rather—in analogy with biological systems—useful behaviors in nano/molecular scale devices will likely emerge from local plasticity rules designed to reinforce advantageous input/output functions during ongoing interactions with an external environment. Yet in all but the most regular (i.e. unrealistic) circumstances, chance modifications in stored values—arising from "accidental" combinations of stimulus-response pairings representing both false positives and false negatives—would necessarily occur at some finite rate. When operating in realistic, noisy environments, it is simply not possible to learn or modify existing input/output functions without occasionally learning something wrong. In the absence of explicit retraining, the accumulation of such errors could substantially degrade or erase previously stored associations.

As a concrete example, we consider the question of long-term information storage in a connectionist network containing a single layer of modifiable synaptic weights. In most connectionist networks, whether biological or engineered, the collection of synaptic weights represents the primary locus for storing useful—typically learned, classifications (Rummelhart

and McClelland, 1986), although physiological systems exploit other types of plasticity as well (Kim and Linden, 2007). Random degradation of synaptic weights, if such effects are allowed to accumulate over time, must therefore lead to the erasure of stored information and a corresponding collapse in system performance. Yet when operating in noisy, realistic environments, random fluctuations in individual synaptic weights are inevitable. Connection strengths must remain plastic in order to permit new responses to be acquired, or to allow existing responses to be modified, in the face of changing environmental circumstances, yet the same mechanisms underlying useful plastic changes will also be invoked by false coincidences with some finite probability. No connectionist system can perform perfectly in any realistic environment. It is therefore unclear how the random synaptic changes induced by false coincidences could be prevented from accumulating until long-term associations were eventually erased. Moreover, as connectionist networks are reduced in size, individual connection strengths will become increasingly subject to chance fluctuations in nano/molecular-scale components. Biological networks, for example, must overcome random perturbations introduced by the constant recycling of synaptic proteins (Xia et al., 2007). Indeed, any physical implementation of a connectionist network, whether biological or engineered, will be subject to random fluctuations as the individual synaptic elements approach nano/molecular-scales (Goldstein, 2005). How can the information distributed across the ensemble of synaptic weights be preserved for long periods in the face of both extrinsic and intrinsic sources of variability?

In principle, information could be retained for arbitrarily long durations, even in volatile computing fabrics or in noisy environments, via regular retraining of all previously stored stimulus/responses pairings (Abraham and Robins, 2005), although such an approach is unlikely to prove feasible in practice as the original inputs are seldom readily available. Alternatively,

surrogate or random data sets have been employed in the context of *unlearning* (Crick and Mitchison, 1983; Hopfield et al., 1983) and *pseudo-rehearsal* (Robins and McCallum, 1999) protocols that allow novel associations to be acquired with only minimal disruptions to existing memories. However, it is unlikely that surrogate data sets alone, including the simulated periods of "REM sleep" required for *unlearning* protocols, could systematically correct the random fluctuations introduced by both intrinsic and extrinsic sources of noise, given that such procedures are specifically designed to enable both new and old associations to be retained. Likewise, conventional approaches to local error correction based on N-fold redundancy assume access to detailed knowledge of the replicated circuits(Neumann, 1956), information that is unlikely to be available in self-organizing systems.

Here, we describe a mechanism for stabilizing stored memories by adopting plasticity rules that exploit the reproducible structure of the surrounding environment. Our approach is inspired by recent attempts to develop unsupervised learning rules appropriate for physical substrates in which random component failure is likely to be a frequent occurrence (Nugent et al., 2004). Our approach utilizes a bidirectional, activity dependent plasticity rule we refer to AHAH (Anti-Hebbian-And-Hebbian). The AHAH rule is similar to unsupervised plasticity rules that have been shown to maximize the 4th order cumulant, or kurtosis, of the response distribution, leading to the development of receptive fields that are related to independent components in the input data (Bell and Sejnowski, 1997). Here, we show that when embedded in separable environments, such rules can stabilize long-term memories in the absence of specific retraining and despite ongoing random fluctuations in individual synaptic weights.

The manuscript is organized as follows. First, we present a conceptual illustration of the main ideas before providing a formal mathematical derivation of the AHAH rule itself. Next, we present an operational example in the context of a two dimensional XOR classification task in which some percentage of the connections coming from an array of radial basis nodes become unreliable—producing random output—while the remaining weights are allowed to adapt under the auspices of the AHAH rule until the original performance is recovered. To better investigate the effects of redundancy and separability on long-term storage, we then use a one-dimensional neural network classifier to show how the AHAH rule is able to increase the duration of learned memories by repairing random weight fluctuations that would otherwise erase previously stored information. Finally, we look at the action of the AHAH rule in response to inputs that are not linearly separable, confirming that under the proposed framework long-term memory storage depends critically on the presence of strong statistical regularities in the embedding environment. We conclude that the stable long-term storage of binary classifications may be achieved by exploiting the reproducible, albeit noisy, structure of natural inputs.

Methods

Conceptual Model

For purposes of illustration, consider a continuous 2-D input space consisting of features organized as clusters of data points, enumerated 1 through 5 (fig. 1A). The presence of such clusters indicates that only a small fraction of all possible 2-D inputs occur with finite probability. We refer to this set of features as the reproducible "structure" of the environment. In a completely unstructured environment, the input data points would form a uniform cloud. Next consider a neuron with linearly weighted inputs arising from a set of radial basis functions that span the set of input features (fig. 1B). Such a neuron can be represented by a decision boundary that divides the input space into "*in class*" and "*out of class*" partitions. Initially, we assume the synaptic weights are chosen randomly and the decision boundary lies at an arbitrary position and orientation. For simplicity, the decision boundary is drawn as a straight line—corresponding to the case where the input is just the X and Y coordinates of each data point—but in general there may be multiple radial basis functions and, when viewed in the original input space, the partition will be curved, although in the higher dimensional space of the radial basis functions the problem is again assumed to be linearly separable. The essential property of the proposed unsupervised, bidirectional plasticity rule is that, under the influence of ongoing activity, the decision boundary moves until it cleanly divides the input space (remapped, if necessary, by the radial basis functions) into distinct, linearly separable features. Although the weight vector may continue to wander slightly due to random fluctuations in the individual synaptic weights, the decision boundary will be constantly restored, due to unsupervised

bidirectional plasticity, to the nearest stable attractor at which classifications are maximally binary.

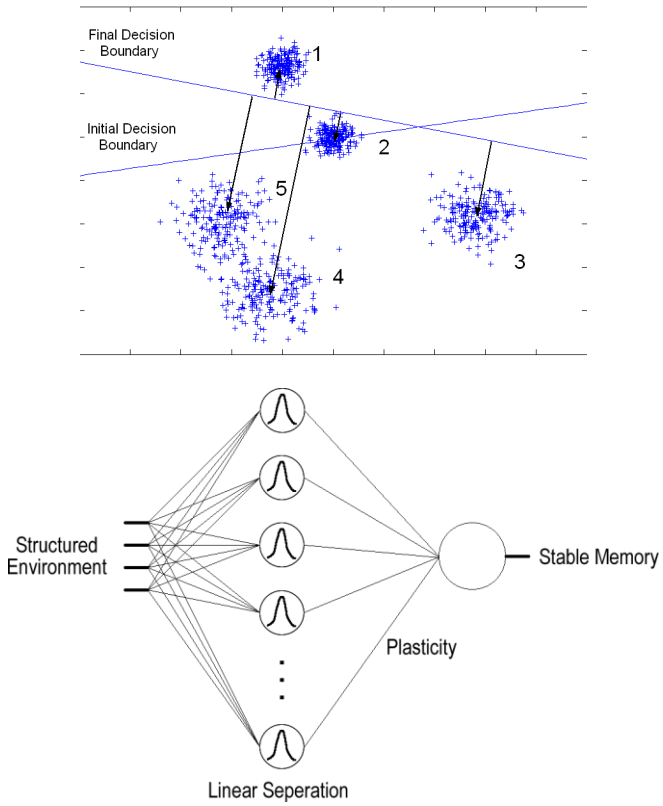

Figure 1. Conceptual model. Top) 2-D input space with separable features represented by clusters of data points, enumerated 1-5. The input space is partitioned by a decision boundary that, as a consequence of unsupervised bidirectional plasticity, moves from its initial location to a stable point corresponding to a local maximum of the margin, indicated by the black arrows. Bottom) Implementation of a general decision surface using a layer of radial basis functions followed by linearly weighted synaptic input to a neuron with binary output.

Other stable decision boundaries are also possible. Choosing among the set of stable decision boundaries represents the stored information. Critically, only certain classifications, namely,

those for which the margin (a measure of the distance between the decision boundary and the collection of input features) is sufficiently large and is at a local maximum, will be stable under random weight fluctuations. Thus, the reproducible structure of the environment determines what classifications can be stored over long periods. Learning, moreover, is restricted to selecting a stable decision boundary, or shattering, of the input space. It is not possible to store arbitrary classifications of random inputs via the proposed mechanism. Nonetheless, in many realistic environments, it is likely that useful classifications will align with stable decision surfaces, as these correspond to discriminations between features that are naturally present.

The AHAH Rule

For a single neuron classifier driven by a set of input lines, the AHAH rule can be derived by considering a general adaptation mechanism of the form:

$$\Delta w_i = \alpha f\left(w_i, x_i, y\right) \tag{1}$$

where Δw_i is the change in the ith weight from one time step to the next, α is a small constant, $f(\bullet)$ is a function that depends on quantities that are local (physically) to the ith synapse, in this case the current weight of the ith synapse, w_i, the activity of the ith input, x_i, and the output of the neuron, y, given by:

$$y = \sum w_i x_i = \vec{w} \cdot \vec{x}, \tag{2}$$

where the sum is over all synaptic inputs to the neuron.

A useful framework for analyzing this type of update rule is to employ a stochastic gradient ascent of an objective function, $J(y)$, in which case (1) can be written as:

$$\Delta w_i = \alpha \frac{d}{dw_i} J(y). \tag{3}$$

Implicit in the following derivation is the assumption that the inputs to the system are chosen randomly on each time step according to a stochastic process. This assumption is necessary because the evolution of the weight vector can only be analyzed with respect to the statistics of a known input distribution.

Denoting the derivative of $J(y)$ with respect to y as $g(y)$, the update rule becomes:

$$\Delta w_i = \alpha x_i g(\vec{w} \cdot \vec{x}) \tag{4}$$

Rules of this type can be useful for finding interesting statistical structure in the input data. As an example, consider the objective function:

$$J(y) = E\{y^2\}, \tag{5}$$

where $\mathrm{E}\{y\}$ is the expectation value of y, formally computed with respect to the input distribution, although in practice we simply allow the weight vector to evolve under the applicable update rule, using the temporal dynamics to approximate the ensemble average. Substituting (5) into (4) and using the time-dependent value of y in place of the expectation value, we derive the Hebbian learning rule:

$$\Delta w_i = \alpha x_i y \tag{6}$$

When the input data has zero mean, this learning rule will cause the weight vector $\vec{w}$ to align with the direction of maximum variance, that is, the direction that maximizes the objective function, $E\left\{y^2\right\}$. Conversely, the anti-Hebbian learning rule $\Delta w_i = -\alpha x_i y$ will cause the vector $\vec{w}$ to align with the direction of minimum variance (i.e. the direction that minimizes the objective function). Assuming that reasonable constraints are placed on the magnitude of the weight vector, $\vec{w}$, the quadratic objective function (5) will have a global maximum (or minimum) and the plasticity rule will converge to a single fixed point, regardless of the initial or desired value of $\vec{w}$. Thus, in the absence of continuous retraining, both Hebbian and anti-Hebbian plasticity mechanisms are incapable of retaining a particular partition of the input space, as the weight vector, under the influence of ongoing random variations in synaptic weights, will always return to a configuration that simply maximizes or minimizes the response variance. Of course, maximizing or minimizing the response variance may, for some applications, correspond to the desired behavior. In general, however, purely Hebbian or anti-Hebbian plasticity rules are poor candidates for mediating the long-term storage of arbitrary classifications of separable inputs.

Alternatively, several higher-order objective functions have been suggested, motivated from a number of fields including statistics (Hyvarinen et al., 2001), information-theory (Bell and Sejnowski, 1997), and biology (Nathan and Leon, 1992). These higher-order objective functions are characterized by multiple local maxima, such that the corresponding learning rules may converge to one of many fixed points. The existence of multiple fixed points, or attractors, in

higher-order learning rules provides a novel mechanism for stabilizing learned memories in the face of intrinsic and extrinsic variability. In particular, we seek an objective function that will stabilize any classification of the input space into linearly separable categories. (For the sake of the following argument, we will assume that the environment contains linearly separable features given an appropriate set of radial basis functions). If one imagines that a neuron has already been taught a desired classification corresponding to a linearly separable division of the input space, then we seek a plasticity rule for maintaining that classification. Several objective functions have been suggested that admit multiple stable fixed points (Hyvarinen et al., 2001). For example, the following objective function uses the fourth-order cumulant, or kurtosis (Bell and Sejnowski, 1997):

$$J(y) = E\left\{y^4\right\} - 3E\left\{y^2\right\}^2 \tag{7}$$

The weight vector $\vec{w}$ must be bounded to produce a stable learning rule. One way to do this to place a constraint on the variance:

$$E\left\{y^2\right\} = 1 \tag{8}$$

This constraint also reduces the sensitivity to second order statistics, thereby eliminating the need for preprocessing steps such as whitening (i.e. de-correlating the data). A constraint on the variance can be included in the stochastic ascent learning rule by introducing a penalty term with scaling β:

$$J(y) = \left(E\left\{y^4\right\} - 3\right) - \beta E\left\{y^2\right\} \tag{9}$$

Taking the derivative of (9) with respect to the output of the neuron, y, substitute into (3) and multiplying by -1 (since we want the learning rule to minimize kurtosis) we arrive at

$$\Delta w_i = \alpha x_i y \left(b - a y^2 \right) \tag{10}$$

where a and b are constants. The constant a can be factored out of the RHS of (10) and absorbed into the learning rate, α. The fixed points of the above learning rule, determined by the condition $\Delta w_i = 0$, are then given by the expression, $y = \pm\sqrt{b/a}$. In the following, we choose our example categorization tasks such that the ratio $(b/a) = 1$, so that the fixed points of the learning rule become $y = \pm 1$, although in general the determination of the associated Lagrange multiplier depends on the statistics of the input distribution (Nugent et al., 2004). Of critical importance is the fact that the fixed points at $y = \pm 1$ are stable. For example, if $y > 1$ and $x_i > 0$, then $\Delta w_i < 0$ (anti-Hebbian plasticity) and the value of y will decline. Likewise, if $y < 1$ then $\Delta w_i > 0$ (Hebbian plasticity) and the value of y will increase. A similar analysis confirms that the fixed point at -1 is stable as well. Thus, the plasticity rule given by (10) mixes both Hebbian and anti-Hebbian terms, which together attempt to force the output of the neuron to a state that partitions the input space into two classes, here denoted by the labels $\{\pm 1\}$. Moreover, these categorizations, once established, are inherently stable under random perturbations of the weight vector as long as the input space is linearly separable and the size of the perturbation is not too large relative to the individual margins.

The behavior of the above rule has been examined mostly in the context of how receptive fields might arise naturally (i.e. in an unsupervised manner), as a consequence of the structure of the

input data (Bell and Sejnowski, 1997). Here, we mainly consider a modified version of the above rule that is better suited for maintaining linearly separable partitions of the input space:

$$\Delta w_i = \alpha x_i z(y)\left(b - az^2(y)\right) \tag{11}$$

where the function z(y) is of a sigmoid form, the details of which are not critical (we use a *tanh*(·) function). The sigmoidal term makes the induction of plasticity maximally sensitive to inputs near the margin, or decision boundary, by reducing the weight updates produced by inputs that are already strongly classified one way of the other.

The above learning rules (10, 11) implememnt a generalized Hebbian mechanism where the activity dependent quantity $b - ay^2$ controls the transition from Hebbian to anti-Hebbian learning and thus falls into the same general category as the BCM rule, which has been used to explain the development of occular dominance columns in the visual cortex (Bienenstock et al., 1982; Nathan and Leon, 1992). We refer to the above class of local, unsupervised, bidirectional plasticity mechanisms as examples of the Anti-Hebbian and Hebbian rule (AHAH rule!). For our purposes, the key property of such AHAH rules is that the dynamical evolution of the weight vector is governed by multiple local minima, or attractors, each corresponding to a distinct, linearly separable partition of the input space. Some of these local minima will be useful for a particular classification task and others will not. A learning algorithm will typically be required to select the optimal partition or, if there are multiple output neurons, the set of partitions that best facilitate the desired classification. Here, we treat the design of an appropriate learning algorithm for chosing an optimial partition as a separate issue from the question of how to stabilize the weight vector once it has been positioned at the desired location.

Results

XOR Example

To illustrate the ability of the AHAH rule specified by equation (10) to maintain stored classifications in the face of synaptic volatility, we constructed a synthetic 2-D XOR problem, with the structure of the input space consisting of two distinct features represented by Gaussian distributions, both with a variance of 0.5, centered along the diagonals, at (1,1) and (-1,-1), respectively. A Support Vector Machine (SVM) was used to initialize the network (Hastie et al., 2001), setting both the number and shape of the radial basis functions as well as the synaptic weights to, and bias of, the target neuron. Additional details regarding the nature of the SVM and how specific parameters of the AHAH rule were determined have been published elsewhere

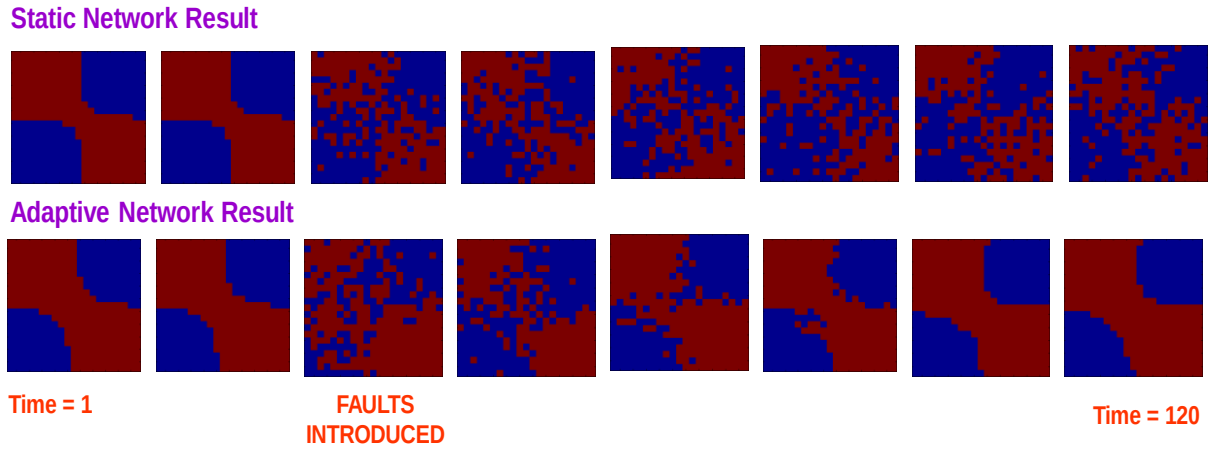

Figure 2. XOR example. The decision surface through time for a static network (top), which lacked plasticity, and for a modifiable network (bottom), with plasticity governed by the AHAH rule (10). At the time step indicated, faults were injected into first layer nodes (radial basis functions), causing the corresponding synapses to become unreliable. Unsupervised bidirectional plasticity was able to approximately restore the original decision surface.

(Nugent et al., 2004). Here, our objective is merely to illustrate the autonomous error-correcting behavior of a specific AHAH rule. Classification performance was monitored by sampling the input space and plotting the class-label predicted by the output neuron using a binary shading scheme (fig. 2). "Good" performance is indicated by the presence of two clearly defined decision surfaces well matched to the XOR dichotomy.

At the time step indicated, a substantial fraction (25%) of the synaptic weights became unreliable, producing random output that varied from time step to time step over the allowed range. In the static case, there was no adaptation (i.e. the network made no attempt to account for errant synapses) and the degradation in performance persisted throughout the experiment. In the adaptive case, all synaptic weights, including those producing unreliable output, were continuously updated according to the AHAH rule (10). Remarkably, unsupervised bidirectional plasticity was able to restore performance on the XOR classification task until the decision surface approximately recovered its initial configuration. More precisely, under the action of the AHAH rule, the decision surface evolved to the nearest stable attractor, which in this case matched closely the requirements of the XOR classification task. This example shows how an unsupervised, activity-dependent bidirectional plasticity rule can stabilize stored classifications even in the face of extreme volatility in individual synapses.

Single-Layer Classifier

To explore the behavior of the AHAH rule in a relatively tractable context, we constructed a single layer neural network classifier (fig. 3) whose weights were subject to both random fluctuations and to various types of damage. Inputs to the classifier consisted of a set of M input

vectors, presented one per time step in random order, where M was also the number of synaptic inputs to the neuron (not including redundant inputs, defined below). Each input vector consisted of a normally distributed pattern of activity, with unit magnitude and standard deviation, σ, centered on the respective input line. Conceptually, the input space might be thought of as the keys on a piano, with the restriction that allowed chords consisted only of Gaussian shaped clusters centered on each note (a rather dissonant sound!). If σ is kept very small, the inputs will be non-overlapping and each note is effectively played individually, leading to maximum separability. With high σ, on the other hand, neighboring chords overlap extensively and are very difficult to separate. In this way, we are able to simulate a highly structured environment in which the separability of neighboring inputs can be controlled by a single variable, σ.

To apply the above neural network to a particular problem, we designed a simplified classification task by labeling each of the M input patterns as either "+" or "-". Class labels were assigned in an alternating fashion according to a spatial wavelength, λ. For example, for $\lambda = 1$, the class labels were {1,-1,1,-1,1,-1,…}. Similarly, for $\lambda = 2$, the class labels were {1,1,-1,-1,1,1, -1,-1…}. In practice, the desired classification was implemented by setting the sign of each synaptic weight (always of unit magnitude initially) to that of the class label of the associated input line. For small values of λ, the class labels alternated rapidly, thus making the inputs more difficult to separate, whereas for large values of λ, neighboring inputs tended to fall in the same class. The two free parameters, σ and λ, thus provide a high level of control over both the separability of the input space and the degree of frustration of the target classification problem.

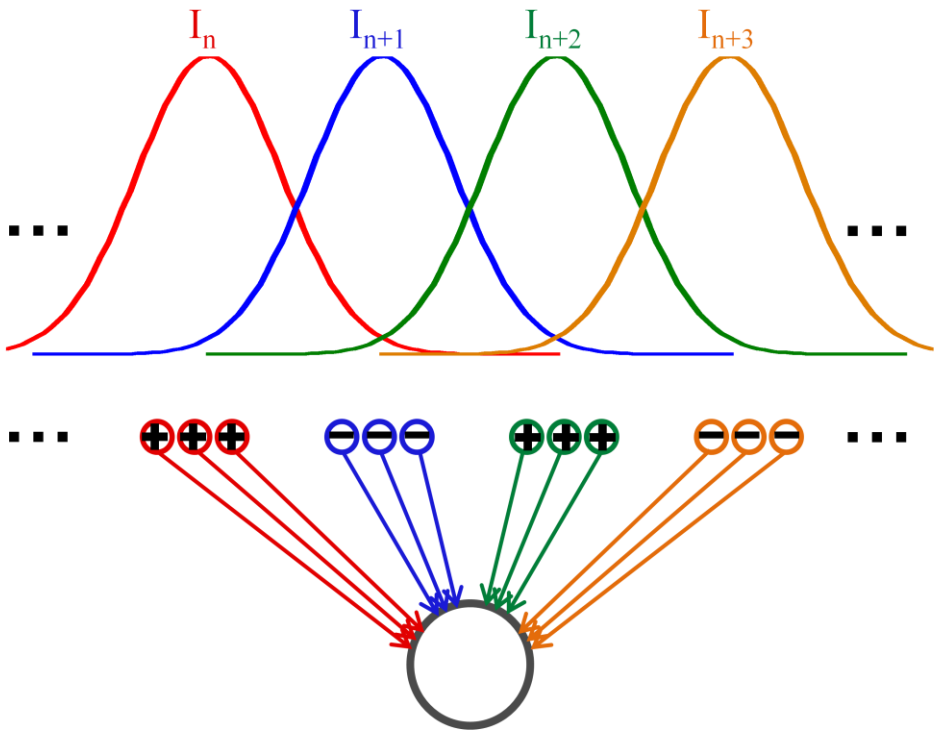

Figure 3. Single layer 1-D classifier. On each time step, one input is selected from among *M* Gaussian distributions (4 shown with $\sigma = 1$), which control how strongly the 1-D array of input lines are activated. Each input line is repeated *R* times (here *R*=3) at the same location (spread out horizontally for purposes of illustration). Class labels, either + of -, alternated with wavelength λ (here λ=1). Arrow heads mark adjustable weights subject to the AHAH rule specified by equation (11).

Finally, to study the effects of redundancy, each input line was replicated *R* times, so that the total number of connections was given by $M \times R$. Although substantial overlap between synaptic weights could also be achieved by appropriate choices of σ and λ, it was nonetheless advantageous to be able to directly modulate the degree of redundancy without affecting the other parameters, changes in which affected the separability and degree of frustration as well. Each redundant input line was activated identically by the stimulus, although the corresponding

synaptic weights were subject to independent random variation and stochastic faults, as described below.

Synaptic Fluctuations

We began by using the simple single-layer neural network classifier with *M* inputs to label orthogonal, non-overlapping input patterns (i.e. σ was effectively set equal to zero, so that each set of *R* input lines were activated separately). Normally distributed random values were added to each synaptic weight on every time step, representing ongoing fluctuations due to various sources of intrinsic variability. The weight vector was allowed to evolve under the auspices of the AHAH plasticity rule, which tended to stabilize the initial classification despite the random fluctuations in individual synaptic strengths. Each synapses was replicated *R* times, so that the information encoded by each synapse could be distributed across an identically activated population, thereby allowing redundant connections to repair themselves by enforcing conformity. Because there was no frustration, the total number of input lines, *M*, also the number of input vectors, as well as the linear distance between oppositely labeled classes, λ, could be chosen arbitrarily (M=16, λ=1). The noise amplitude was set to $A_f = 1/M$ and the learning rate to $\alpha = 3A_f$.

To assess the ability of the AHAH rule to stabilize stored classifications, we measured the mean time (number of time steps) to the first incorrectly labeled input as a function of the redundancy, *R* (fig. 4, solid line). As a control, we also measured the mean time to first error as a function of *R* in a network without plasticity but subject to the same distribution of random fluctuations.

The AHAH rule produced supralinear growth in mean memory lifetime as a function of redundancy whereas memory duration grew only linearly with increasing redundancy in the absence of plasticity. As expected, the differential improvements resulting from unsupervised bidirectional plasticity were negligible for small values of redundancy. In the present context, the AHAH rule allows groups of redundant synapses to "pull back" any deviant weights whose values have diverged, ultimately bringing them into conformity with the prevailing response. As

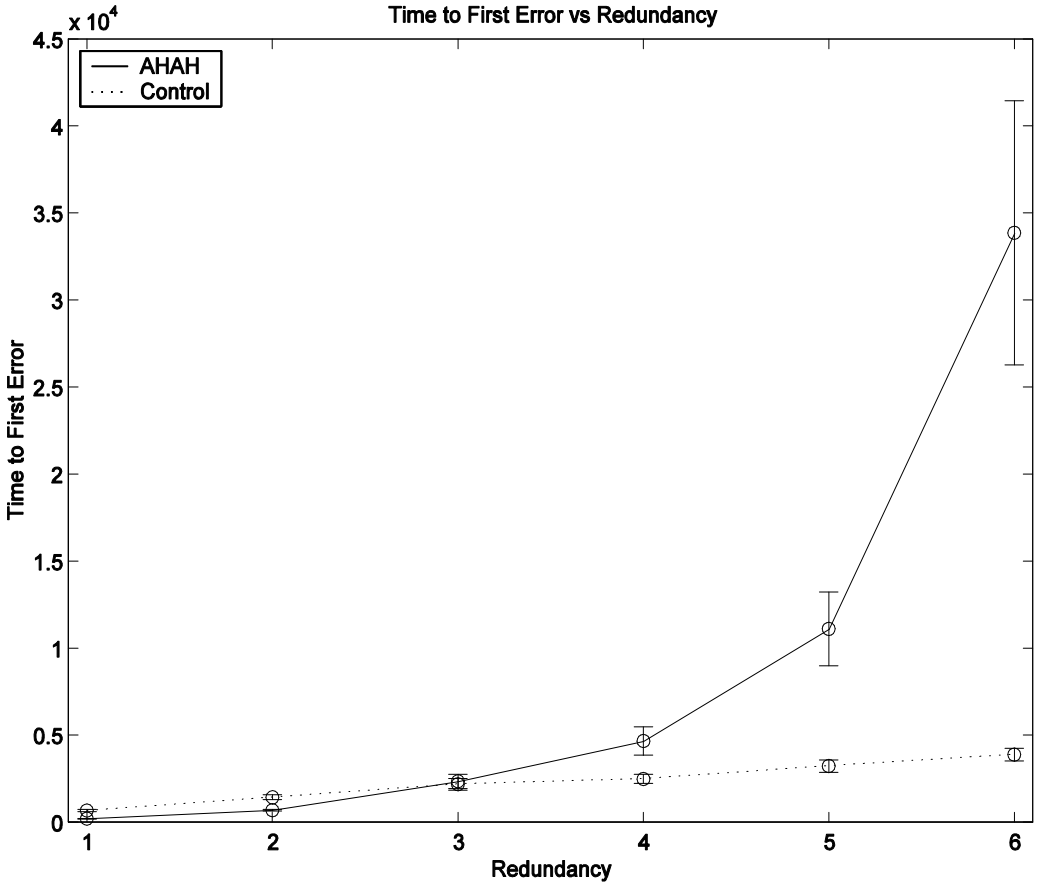

Figure 4. Mean life time vs. weight redundancy. Synaptic weights were randomly varied by a small amount on each time step. Unsupervised adaptation via the AHAH rule produced a supralinear increase in mean memory lifetime compared to the time to first classification error in the absence of plasticity.

long the random fluctuations in individual weight values remains sufficiently small, so that it is unlikely that a majority of the redundant connections all switch sign on the same time step, the AHAH rule will act to restore the weight vector to the vicinity of its original location.

Our results show that the AHAH rule is able to repair random fluctuations in individual connection strengths as long as the necessary information is distributed redundantly across a sufficient number of synapses. However, the assumption of no frustration between the individual synaptic connections (i.e. each weight is activated by only a single input vector), is unrealistic. We therefore investigated whether the AHAH could preserve stored information even when input vectors were overlapping and the target classification imposed a finite amount of frustration, which occurs when the same synapses are activated by oppositely classified patterns (fig. 5). Holding the classification wavelength fixed at λ=2, , the standard deviation of the input vectors, σ, was systematically increased, thereby increasing the overlap between neighboring patterns and thus the degree of frustration in the network as a whole. Performance was monitored as a fraction of all M input vectors that were classified incorrectly after subjecting individual synaptic weights to random fluctuations as described above. The network was run for 1000 time steps under random synaptic fluctuations, where the last 500 time steps was used to find the classification error, defined as the average disagreement between the faulted network and an equivalent network with no synaptic weight fluctuations and no plasticity. This process was then repeated for N_t=50 trials, to obtain a mean classification error and corresponding error bar. For these studies, the redundancy, R, was set to 1 to isolate the effects of pattern width.

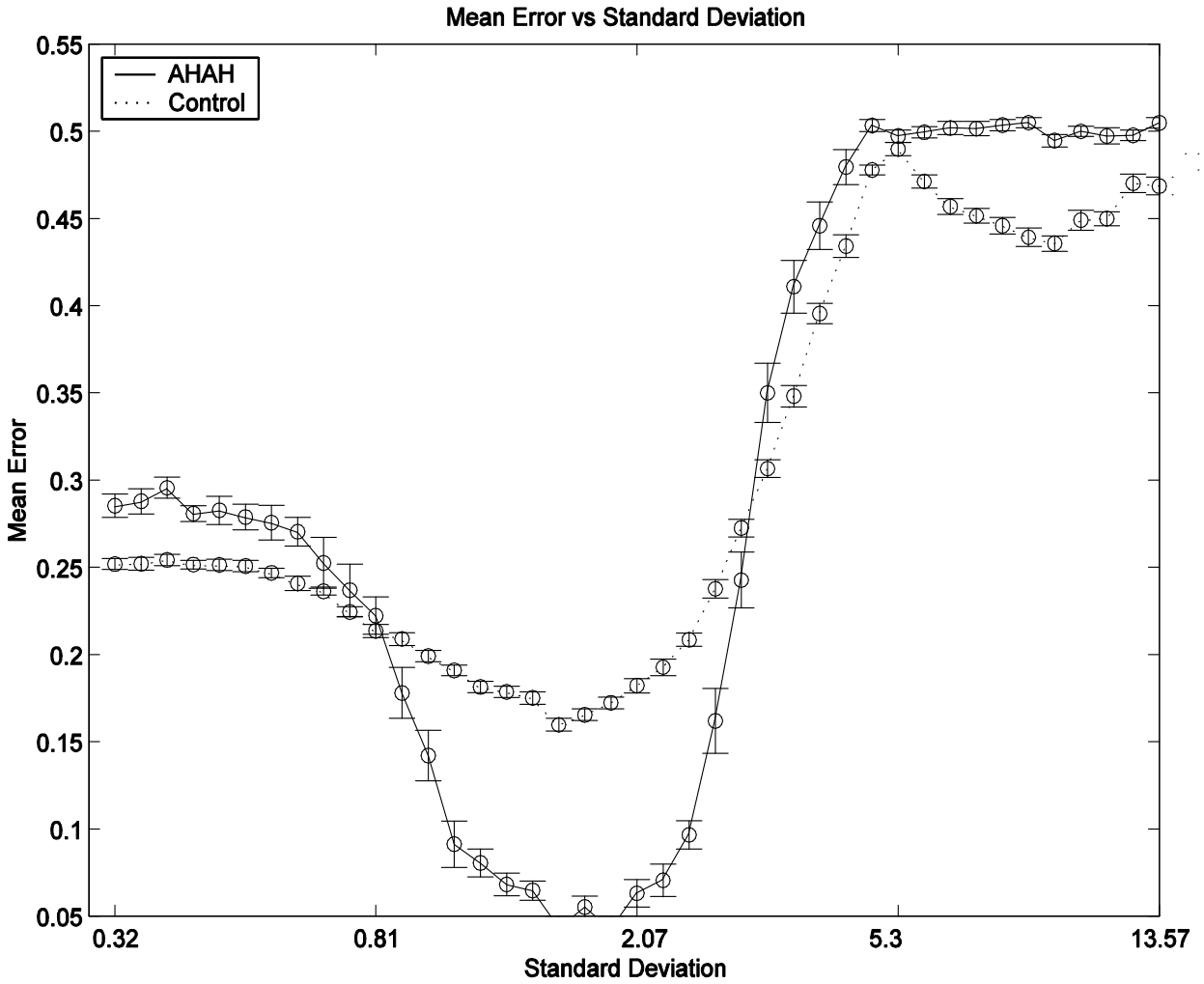

Figure 5. Mean error vs. separability. Overlap between neighboring input features, determined by σ, was systematically increased, producing a region where the additional redundancy counteracted the effects of reduced separability and unsupervised plasticity could correct random fluctuations in individual synaptic weights.

Our results show clearly than even in the presence of finite frustration, the AHAH rule was able to stabilize stored classifications, relative to an identically initialized control network without bidirectional plasticity, over a broad range of input patterns. When the width of the Gaussian

envelope fell below approximately 0.8, there was insufficient overlap between neighboring input vectors to allow mutual reinforcement, so that the AHAH rule actually produced a small increase in the classification error relative to the control. For very large overlap, on the other hand, corresponding to values of σ greater than approximately 3.0, the high degree of frustration similarly prevented mutual reinforcement between neighboring input patterns, so that the fractional classification error tended to the level predicted by pure chance, 0.5. Between these limits, however, the information encoded by each synapses was effectively distributed across its neighbors, providing a source of redundancy that allowed the AHAH rule to recover stored memories that would otherwise have been lost due to random fluctuations in the individual synaptic weights.

Lesions vs. Malfunctions, Gradual vs. Sudden Faults

We have so far illustrated the ability of the AHAH rule to restore the decision surface to its original configuration under the influence of small random fluctuations in individual synaptic weights. To assess the ability of the AHAH rule to recover from discontinuous faults, we used the same single layer classifier to investigate the effects of *lesions* and *malfunctions* (fig. 6). The input space was designed to be easily separable into distinct features ($\lambda = 2, \sigma = 1$) and there were no random synaptic fluctuations other than the two types of faults under consideration. If plasticity was present, weights were updated by the AHAH rule for the first 1000 time steps, in order to damp out any start up artifacts. From time steps 1001 to 9000 a set percentage of the input lines were faulted, either suddenly or gradually. From time steps 9001 to 10000 the average classification error was measured. If turned on gradually, each faulty input line was introduced one at a time, distributed over the specified period. If turned on suddenly, the faults

were introduced all at once at time step 1001. *Lesions* were modeled as an input going to zero whereas *malfunctions* replaced the correct input on each time step with a random input picked from a uniform distribution between -1 and 1. In all cases, redundancy, *R*, was set to 5, which

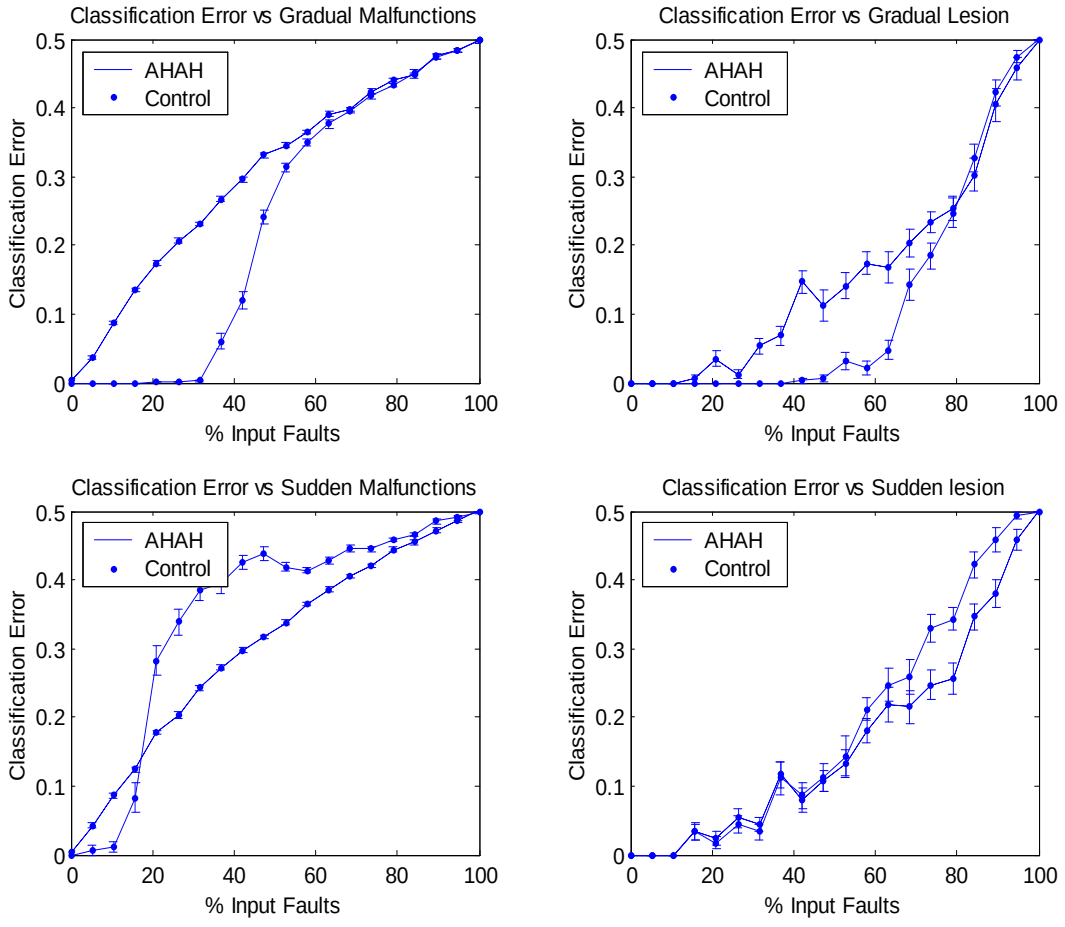

Figure 6. Recovery from different types of faults. *Malfunctions* refer to synapses that have become unreliable (random output) whereas *lesions* refer to synapses that have simply been eliminated (output set identically to zero). The AHAH rule was exceptionally good at stabilizing stored classification when faults were introduced gradually, whereas sudden malfunctions were highly disruptive and more likely to knock the decision surface out of the local attractor.

was found to be optimal for the chosen set of input parameters.

Our findings indicate that whereas unsupervised adaptation governed by the AHAH rule can stabilize stored classifications in response to discontinuous faults that are introduced gradually, sudden random faults are much more disruptive. When more than 20% of the input lines are suddenly damaged due to malfunctions, but not lesions, the decision boundary is removed from its local attractor and catastrophic error results. Unsupervised adaptation via the AHAH rule is able to fully compensate for the same percentage of faults if introduced gradually. Indeed, with bidirectional plasticity the network retains essentially perfect classification for gradually introduced lesions up to approximately 50% gradually introduced faults.

Unstructured Environments

If long-term memories can only be maintained in environments possessing reproducible structure, it follows that the absence of stable, separable features should, over time, lead to the loss of stored information. Specifically, if embedded in a completely unstructured environment consisting of only random, uncorrelated inputs, unsupervised plasticity should actually erase existing memories. To illustrate this point, we subjected the single-layer classifier to varying periods of random input in the absence of any intrinsic weight fluctuations or discontinuous faults (fig 7). The only changes in individual synaptic weights were due to unsupervised bidirectional plasticity. Although the persistence time of stored classifications could be prolonged by increasing redundancy, there always exited a time period over which information was lost.

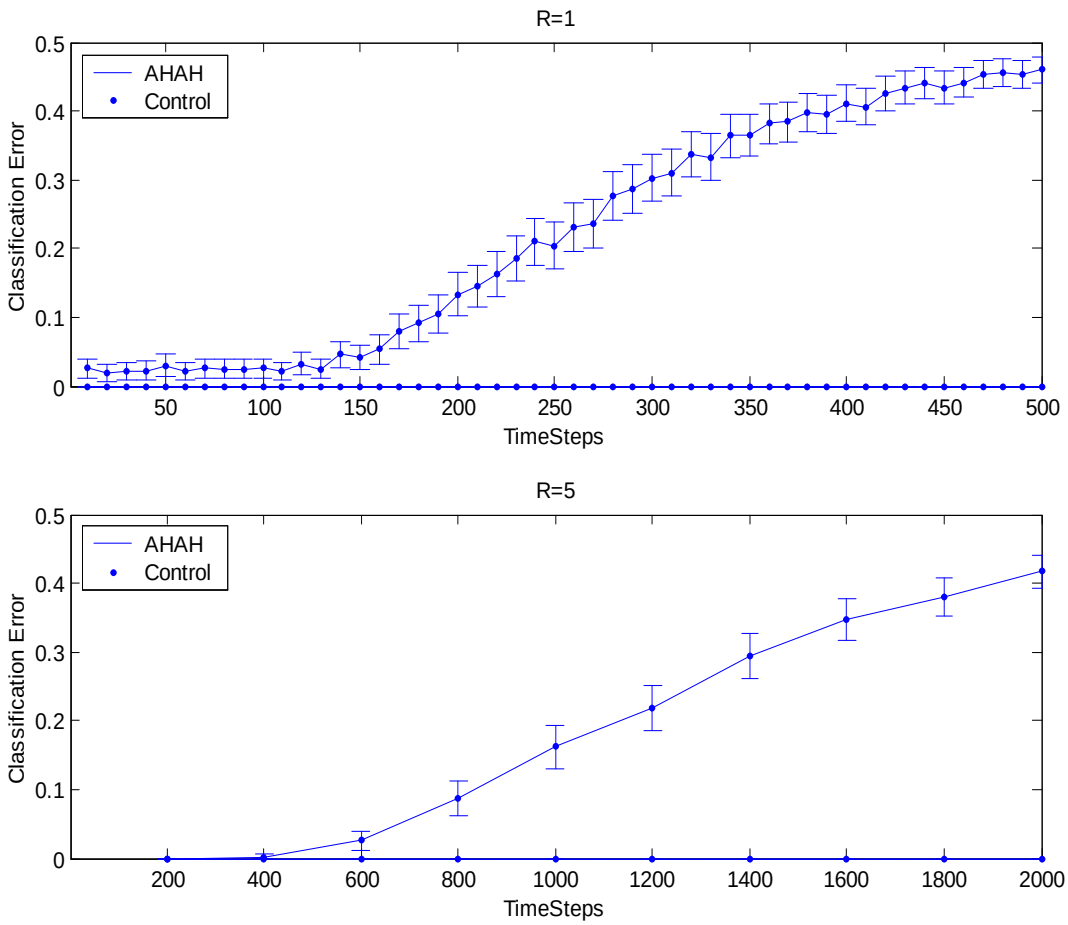

Figure 7. Unstructured environments. The single-layer classifier was exposed to varying amounts of randomized input data and the classification error was measured for two cases of redundancy. After sufficient exposure to unstructured inputs the stored information was eventually lost even though there were no random weight fluctuations or discontinuous faults, as indicated by the absence of classification errors in the control network that lacked plasticity.

Indeed, unsupervised bidirectional plasticity can lead to the loss of stored information even in non-random environments if input features are not linearly separable at the final classification stage. Again using the single-layer classifier in the absence of either random weight fluctuations or discontinuous faults, we varied the Gaussian width, σ, in order to systematically decrease the

separability between neighboring input features (fig. 8). For several different choices of λ, there was always a corresponding value of σ at which the environment became sufficiently unstructured so that stored classifications were lost under the action of unsupervised plasticity. Although ongoing bidirectional plasticity is necessary for properly adapting to changing environments and for restoring the decision boundary in the face of random weight fluctuations and discontinuous faults, in unstructured environments such plasticity can actually cause stored information to be irretrievably lost.

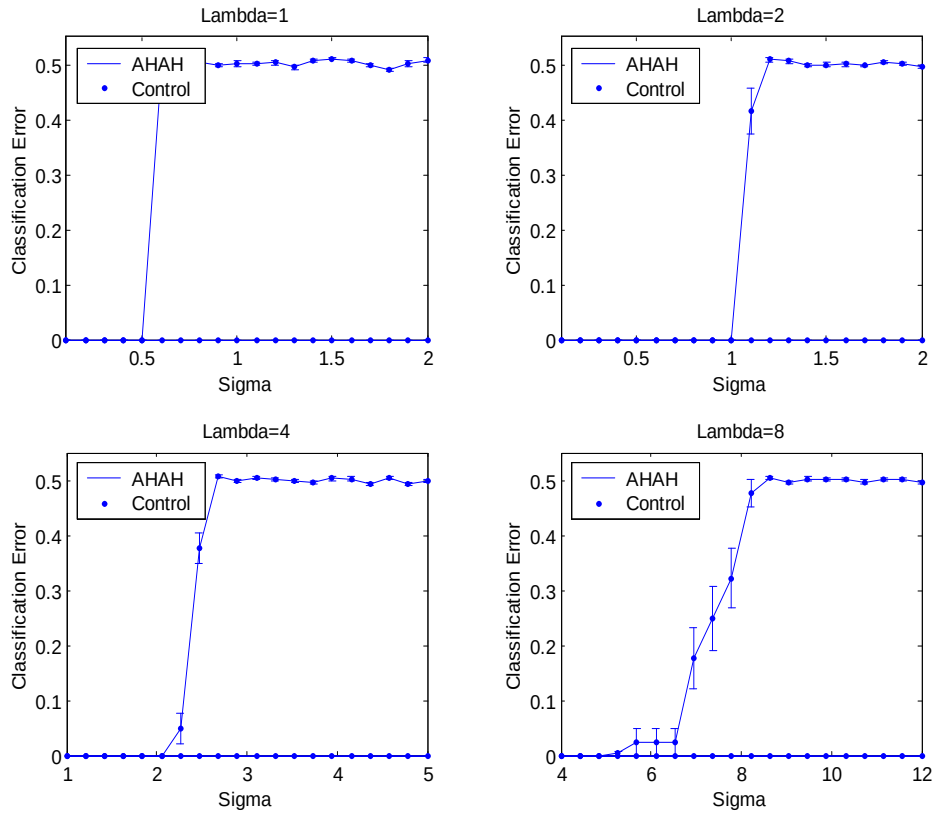

Figure 8. Long-term storage as a function of separability. Using the single layer classifier, overlap between input features was systematically varied by increasing σ, which determined pattern width, for several different values of λ, which controlled the spatial period over which pattern labels alternated between ±1. Classification error determined after 1000 time steps. Each data point represents 20 trials. Redundancy, R, set to 1 to isolate the effects of pattern width. Unsupervised adaptation eventually erased stored information as the separability of input features was reduced. There were no classification errors in the control condition (no plasticity) and thus the corresponding curves always coincided with the x-axis.

Discussion and Conclusions

Intrinsic and extrinsic volatility is a problem likely to be encountered by many Unconventional Computing fabrics, especially at the nano/molecular scale (Goldstein, 2005). Our results suggest

a possible solution to this problem using unsupervised, bidirectional plasticity rules which automatically repair previously established decision boundaries. The present approach is not applicable to the self-repair of computing devices that must function in arbitrary environments, as the unsupervised learning rules we describe can only act to stabilize previously stored input classifications that can be usefully partitioned into linearly separable components. However, it may be that attempts to build completely general devices that can operate in any environment might, in some instances, be usefully redirected to constructing devices that instead are designed to operate in naturalistic environments with sufficient structure to support stable information storage. Indeed, if a memory could *not* be referenced to a separable feature of its environment, one could argue that it might be of no practical use to store.

Although our formal analysis was limited to the consideration of connectionist networks, it seems likely that the underlying principles could be extended to a variety of unconventional computing paradigms. Future analog computers, for example, will necessarily rely on a large number of volatile parameters. To the extent that such parameters cannot be programmed explicitly, but rather will have to be acquired through experience via activity-dependent mechanisms, AHAH-like plasticity may be useful for maintaining system performance.

Learning

We did not directly consider the problem of learning optimal decision surfaces, only the issue of whether such surfaces, once established, could remain stable under random fluctuations and other perturbations. However, the unsupervised, bidirectional plasticity rules we analyzed in the context of self-repair could likely be extended to account for supervised learning under the action

of a teacher signal (an external input that forced the output of a given neuron to a desired state in the presence of a particular input pattern). Whether an extended version of the AHAH plasticity rule could both learn quickly and remain stable is a question for additional study but in principle the conjecture seems reasonable.

Biological Relevance

Might AHAH-like plasticity be operating at biological synapses? Colloquially, we know that biological networks have evolved mechanisms for storing long-term memories in the absence of explicit retraining. For example, many people have experienced riding a bicycle, or playing a musical instrument, with reasonable competence after years without practice. Under more controlled conditions, it has been demonstrated that learned or adapted behaviors can persist for many months in the absence of explicit retraining (Schreurs, 1993; Nicholson et al., 2003). How biological networks are able to store long-term memories at volatile synapses is unknown, but multiple physiological mechanisms may be involved. Most simply, a certain percentage of synapses could become fixed over time and thus no longer subject to the same degree of random variation. Indeed, such stabilizing mechanisms are implied by the existence of "critical periods" in early development during which neural circuits are thought to be exceptionally modifiable (Crair and Malenka, 1995). However, connectionist networks operating in real-world environments must retain the ability to store new memories while erasing information that has become obsolete. A long-term storage strategy based solely on holding an increasing percentage of synapses at a fixed strength would inevitably compromise the ability to adapt to changing internal and/or external conditions. Alternatively, long-term storage might be accomplished by transferring memories between multiple sites of plasticity, each operating on a substantially

different time scale (Kenyon, 1997; Fusi et al., 2005). While such strategies might improve the stability of long-term storage, they do not address the basic problem of how to retain memories at synapses that are themselves intrinsically unstable.

If AHAH-like mechanisms are truly at work in biological systems, then only structured information could be remembered for any significant length of time, leading to the following hypothetical experiment. If an animal were trained on a task specific to a particular sensory data stream, exposure to an alternate data stream void of structure for a sufficient time should erase the memory for that task. Thus, relevance of the AHAH plasticity rule for biological systems may be amenable to behavioral investigation.

Implications

In silico implementations of neural circuits provide an example of a computational fabric in which the AHAH-like unsupervised learning rules could potentially be implemented in an existing physical device (Fusi et al., 2000). As future implementations of neuromorphic circuits are targeted toward non-silicon-based substrates, the issues of self-repair and local plasticity will become much more critical, as with virtually all unconventional computing devices. The development of autonomous robots, for example, will ultimately require relatively small (brain sized) portable devices able to perform accurate pattern classifications in natural environments. Such devices will likely employ astronomical numbers of elements (approaching, for example, the 10^{15} synapses in the human brain) yet will need to operate on low power (approaching, for example, the 20 Watts consumed by the human brain). Because AHAH-like plasticity rules are unsupervised and depend only on local activity, the proposed self-repair mechanisms should

scale to systems of arbitrary size. Extending the present results to large-scale neuromorphic systems will require incorporating local learning rules capable of acquiring useful behaviors and the extension of our stability analysis to hierarchical neural networks.

References

A. Adamatzky BDLC, L. Bull, S. Stepney, C. Teuscher, ed (2007) Unconventional Computing 2007: Luniver Press.

Abraham WC, Robins A (2005) Memory retention - the synaptic stability versus plasticity dilemma. Trends in Neurosciences 28:73-78.

Bell AJ, Sejnowski TJ (1997) The "independent components" of natural scenes are edge filters. Vision Res 37:3327-3338.

Bienenstock EL, Cooper LN, Munro PW (1982) Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. J Neurosci 2:32-48.

Calude CS, Dinneen MJ, Pˇaun G, Rozenberg G, Stepney S, eds (2006) Unconventional Computation: 5th International Conference, UC 2006, York, UK, September 4-8, 2006, Proceedings: Springer.

Crair MC, Malenka RC (1995) A critical period for long-term potentiation at thalamocortical synapses. Nature 375:325-328.

Crick F, Mitchison G (1983) The function of dream sleep. Nature 304:111-114.

Fusi S, Drew PJ, Abbott LF (2005) Cascade models of synaptically stored memories. Neuron 45:599-611.

Fusi S, Annunziato M, Badoni D, Salamon A, Amit DJ (2000) Spike-Driven Synaptic Plasticity: Theory, Simulation, VLSI Implementation. Neural Comp 12:2227-2258.

Goldstein SC (2005) The impact of the nanoscale on computing systems. In: International Conference on Computer-Aided Design, pp 655-661. San Jose, CA: IEEE/ACM

Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. New York: Springer.

Hopfield JJ, Feinstein DI, Palmer RG (1983) 'Unlearning' has a stabilizing effect in collective memories. Nature 304:158-159.

Hyvarinen A, Karhunen J, Oja E (2001) Independent component analysis. New York: J. Wiley.

Kenyon GT (1997) A model of long-term memory storage in the cerebellar cortex: a possible role for plasticity at parallel fiber synapses onto stellate/basket interneurons. Proc Natl Acad Sci U S A 94:14200-14205.

Kim SJ, Linden DJ (2007) Ubiquitous Plasticity and Memory Storage. Neuron 56:582-592.

Miller P, Zhabotinsky AM, Lisman JE, Wang XJ (2005) The stability of a stochastic CaMKII switch: dependence on the number of enzyme molecules and protein turnover. PLoS Biol 3:e107. Epub 2005 Mar 2029.

Nathan I, Leon NC (1992) Objective function formulation of the BCM theory of visual cortical plasticity: statistical connections, stability conditions. Neural Netw 5:3-17.

Neumann Jv (1956) Probabilistic logics and synthesis of reliable organisms from unreliable components. In: Automata Studies (McCarthy CSaJ, ed), pp 43-98: Princeton University Press.

Nicholson DA, Sweet JA, Freeman JH, Jr. (2003) Long-term retention of the classically conditioned eyeblink response in rats. Behav Neurosci 117:871-875.

Nugent A, Kenyon G, Porter R (2004) Unsupervised adaptation to improve fault tolerance of neural network classifiers. In: 2004 NASA/DoD Conference on Evolvable Hardware (R. S. Zebulum DG, G. Hornby, D. Keymeulen, J. Lohn, A. Stoica, ed), pp 246-249. Seattle, Washington: IEEE Computer Society.

Robins A, McCallum S (1999) The consolidation of learning during sleep: comparing the pseudorehearsal and unlearning accounts. Neural Netw 12:1191-1206.

Rummelhart DE, McClelland JL (1986) Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations.: MIT Press.

Schreurs BG (1993) Long-term memory and extinction of the classically conditioned rabbit nictitating membrane response. Learning and Motivation 24:293-302.

Shouval HZ (2005) Clusters of interacting receptors can stabilize synaptic efficacies. Proc Natl Acad Sci U S A 102:14440-14445. Epub 12005 Sep 14427.

Xia Y, Nawy S, Carroll RC (2007) Activity-Dependent Synaptic Plasticity in Retinal Ganglion Cells. J Neurosci 27:12221-12229.