*LETTER* Communicated by Erkki Oja

# Input Statistics and Hebbian Cross-Talk Effects

**Anca Rădulescu**
*radulesc@colorado.edu*
*Department of Mathematics, University of Colorado, Boulder, CO 80309-0395, U.S.A.*



**As an extension of prior work, we studied inspecific Hebbian learning using the classical Oja model. We used a combination of analytical tools and numerical simulations to investigate how the effects of synaptic cross talk (which we also refer to as synaptic inspecificity) depend on the input statistics. We investigated a variety of patterns that appear in dimensions higher than two (and classified them based on covariance type and input bias). We found that the effects of cross talk on learning dynamics and outcome is highly dependent on the input statistics and that cross talk may lead in some cases to catastrophic effects on learning or development. Arbitrarily small levels of cross talk are able to trigger bifurcations in learning dynamics, or bring the system in close enough proximity to a critical state, to make the effects indistinguishable from a real bifurcation. We also investigated how cross talk behaves toward unbiased ("competitive") inputs and in which circumstances it can help the system productively resolve the competition. Finally, we discuss the idea that sophisticated neocortical learning requires accurate synaptic updates (similar to polynucleotide copying, which requires highly accurate replication). Since it is unlikely that the brain can completely eliminate cross talk, we support the proposal that is uses a neural mechanism that "proofreads" the accuracy of the updates, much as DNA proofreading lowers copying error rate.**

## 1 Introduction

**1.1 Synaptic Plasticity and Cross Talk.** It is generally believed that synaptic plasticity (i.e., activity-dependent adjustments of synaptic connection strengths) is the basis of most processes in the nervous system, such as development, learning, creation and storage of memories, cognition, and ultimately behavior (Katz & Shatz, 1996). The term plasticity may reflect a variety of phenomena, from actual new synapse creation and deletion, to silencing and unsilencing of existing synapses, to only changes in existing

Color versions of all figures in this letter are presented in the online supplement available at http://www.mitpressjournals.org/doi/suppl/10.1162/NECO_a_00565.

 
doi:10.1162/NECO_a_00565

synapse strengths. In 1949, Hebb proposed that learning occurs in response to local signals, such as the conjoint activity of pre- and postsynaptic neurons: "When an axon of cell A is near enough to excite cell B or repeatedly or consistently takes part in firing it, some growth or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased" (Hebb, 2002).

Those who interpret and use Hebb's rule generally assume that synaptic modifications act in a local, connection-specific manner (i.e., only synapses between the neurons presenting correlated activity are modified, independent of activity at other synaptic sites). In the literature, the most representative models for long-term changes in synaptic efficacy ( Malenka & Bear, 2004; Elliott, 2012) are long-term potentiation (LTP; Bliss & Lømo, 1973) and long-term depression (LTD; Lynch, Dunwiddie, & Gribkoff, 1977). A variety of initial studies of long-term potentiation and depression initially reported synapses updates to be local (i.e., "specific") (Isaac, Nicoll, & Malenka, 1995; Dudek & Bear, 1992). However, ulterior data failed to replicate synaptic specificity (Chevaleyre & Castillo, 2004; Matsuzaki, Honkura, Ellis-Davies, & Kasai, 2004). Rather, they started to suggest that there is "cross talk" that likely occurs during Hebbian plasticity (Kossel, Bonhoeffer, & Bolz, 1990; Bonhoeffer, Staiger, & Aertsen, 1989; Engert & Bonhoeffer, 1997; Schuman & Madison, 1994; Bi, 2002; Bi & Poo, 2001)—that activity-induced synaptic modification may trigger changes in other, unstimulated synapses (possibly the ones that are geometrically close to or adjacent to the target ones). More recent experimental work (Harvey & Svoboda, 2007) has shown quite unequivocally that induction of LTP at one synapse increases the likelihood of LTP to be induced at closely neighboring synapses.

This source of "error," or noise, is believed to be due to the imperfection of chemical synaptic transmission, in which some degree of diffusion of neuromessengers combines with the high synapse density (especially for highly connected neurons), making it difficult, or even impossible, for a triggered synaptic change to remain completely connection specific.

A proposed list of such factors that contribute to cross talk (Elliott, 2012) includes early-phase LTP/LTD presynaptic (Bonhoeffer et al., 1989; Kossel et al., 1990; Schuman & Madison, 1994) or postsynaptic (Engert & Bonhoeffer, 1997; Harvey & Svoboda, 2007) diffusion of intracellular (Harvey & Svoboda, 2007; Harvey, Yasuda, Zhong, & Svoboda, 2008) and extracellular messengers (Lemann, Gottmann, & Heumann, 1994; Korte et al., 1995; Levine, Dreyfus, Black, & Plummer, 1995), as well as late-phase LTP and LTD factors, on longer timescales (Frey & Morris, 1998; Navakkode, Sajikumar, & Frey, 2004; see also section 4). The necessity for close synaptic packing (DeFelipe, Marco, Busturia, & Merchán-Pérez, 1999) creates a geometric conflict. In NMDA-mediated sites, for example, the spine neck must be sufficiently narrow to reduce Ca escape to other sites (Koch & Zador, 1993; Sabatini, Oertner, & Svoboda, 2002), but also sufficiently wide to allow synaptic currents through. In this light,

complete chemical isolation and accuracy seem, and may indeed be, impossible to achieve in the brain.

**1.2 Plasticity Models and the Effects of Cross Talk.** A variety of models have been used to investigate the effects of synaptic cross talk on brain function. Since many different models can produce the same behavior, it is not possible to use behavior to test whether a model is correct; rather, models can be used to determine whether certain types of interactions are capable of replicating certain outcomes, generating testable hypotheses. In our context, modeling is used to predict in principle whether and when cross talk can lead to a complete breakdown in the outcome otherwise obtained in the synapse-specific case.

In most mathematical models of synaptic plasticity, the system develops, or learns, one or more patterns of synaptic configurations, which are typically stable equilibria but could also be cycles or more complex invariant sets in the case of nonlinear models (Wiskott & Sejnowski, 1998; Elliott, 2003). In this framework, synaptic cross talk can be regarded as an internal noise parameter, whose increase may not only alter performance but, past a critical value, may trigger radical crashes (bifurcations) in the system's dynamics, actually destroying its capacity to reach the stable states (the desired developmental or learning outcomes). It has been argued that in order to avoid such crashes, very accurate connection strength adjustments must be required but that such levels of accuracy are biophysically impossible (Cox & Adams, 2009). Furthermore, it has been shown that the critical level of cross talk sufficient to induce bifurcations in these models is very sensitive to the input statistics and postsynaptic connectivity, and in some cases, it can be made arbitrarily small (Elliott, 2012). Either way, many nonlinear models of synaptic plasticity are fatally compromised by even tiny amounts of cross talk (Elliott, 2012), supporting the idea that some parallel circuitry (proofreading) might be necessary to boost robustness to synaptic inspecificity, and thus permit or facilitate useful development and learning, even in the presence of cross talk (see section 4.3 for additional comments on proofreading).

The possibility that synaptic cross talk can have such catastrophic effects makes it very important for us to assess its impact on nonlinear models of synaptic plasticity as a way toward understanding its actual impact in the brain. One cannot expect, however, a generic proof of principle for all learning models, especially given the vastness of the field; rather, one can point out relevant examples of such behavior in models that are biologically plausible.

We study here the effect of cross talk in the Oja rule, a very simple, multiplicative normalization of Hebbian learning. Oja's model is driven only by second-order statistics, hence works as a principal component (PCA) rather than an independent component analyzer (ICA; Cox & Adams, 2009). We are not proposing that the brain actually does PCA, but we consider this

very simple particular case of the general unsupervised learning problem because it is completely tractable by a combination of analytical and numerical tools. While our approach incorporates some aspects of biological realism, many simplifications are made along the way (described in the following sections) with the goal to investigate cross talk in a simple and relevant context rather than to propose a detailed model of biological learning. Although the existence of stable equilibria relates here only to second-order input statistics, this model captures a feature observed in other nonlinear, more elaborate models: synaptic cross talk is able to induce catastrophic breakdowns in learning in a manner that is highly idiosyncratic, depending in a very input-specific and model-specific manner on the learning rule.

The rest of the letter is organized as follows. In section 2, we present the model (the Oja rule in the presence of cross-talk, or "inspecificity") and some properties of the input patterns to be learned, and we provide an overview of the basics of the rule's dynamic behavior. In section 3.1 we investigate numerically the three-dimensional Oja inspecific network; we focus in particular on how it processes different classes of input distributions, preserving some of the dynamical aspects found in the two-dimensional phase plane (Rădulescu & Adams, 2013), but also introducing new features specific to higher dimensions. In section 3.2, we study analytically, in an $n$-dimensional example, the behavior observed numerically in the previous section. In section 4, we put the numerical and analytical results in the biological context of a learning cortical network. Section 4.1 focuses on the meaning and importance of input bias and on its effects in conjunction with cross talk. Section 4.2 discusses the biological plausibility of an Oja-type learning model and reviews a possible biophysical implementation of the rule, as described in the literature. Section 4.3 briefly discusses the analogy between neural cross talk and DNA copying errors, and the necessity of a proofreading mechanism in both cases.

## 2 Methods

**2.1 The Oja Model with Synaptic Cross Talk.** Oja (1982) showed that a simple neuronal model can perform unsupervised learning based on Hebbian synaptic weight updates incorporating an implicit "multiplicative" weight normalization to prevent unlimited weight growth (von der Malsburg, 1973). Oja's rule has been extensively studied and used (Hertz, Krogh & Palmer, 1991; Taylor & Coombes, 1993) in its original or modified forms (Oja & Karhunen, 1985; Diamantaras & Kung, 1996).

Our focus is on studying a single-output network, learning an input distribution according to Oja's rule (Oja, 1982). More precisely, the output neuron receives, through a set of $n$ input neurons, $n$ signals $\mathbf{x} = (x_1, \ldots, x_n)^T$ drawn from an input distribution $\mathcal{P}(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, transmitted via synaptic connections of strengths $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T$. The resulting scalar output $y$ is

generated as the weighted sum of the inputs $y = \mathbf{x}^T\boldsymbol{\omega}$. The synaptic weights $\omega_i$ are modified by implementing first a Hebb-like strengthening proportional to the product of $x_i$ and $y$ ($\Delta w_i = \gamma y x_i$), followed by an approximate "normalization" step, maintaining the Euclidean norm of the weight vector close to one,

$$\omega_i(t+1) = \omega_i(t) + \gamma y(x_i - y\omega_i), \tag{2.1}$$

where $yx_i$ is the effective change in the synaptic strength $w_i$, while $y^2\omega_i$ can be interpreted as a "decay," or "forgetting," term. The input covariance matrix $\mathbf{C} = \langle \mathbf{x}\mathbf{x}^T \rangle$ can be used as an appropriate long-term characterization of the inputs to study the asymptotic convergence of the expected weight vector $\mathbf{w}(t) = \langle \boldsymbol{\omega}(t+1) | \boldsymbol{\omega}(t) \rangle$. Then equation 2.1 becomes

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \gamma[\mathbf{C}\mathbf{w} - (\mathbf{w}^T\mathbf{C}\mathbf{w})\mathbf{w}] \tag{2.2}$$

or, in continuous time, the case studied in this letter,

$$\frac{d\mathbf{w}}{dt} = \gamma[\mathbf{C}\mathbf{w} - (\mathbf{w}^T\mathbf{C}\mathbf{w})\mathbf{w}]. \tag{2.3}$$

Since it depends only on second-order statistics of the incoming input, this model acts as a principal component analyzer for the input distribution (Oja, 1982), one simplified way of modeling data compression and transmission in the brain. Although the normalization is implemented in this equation via an $o(\gamma)$ approximation, one can easily check that $\frac{d\|\mathbf{w}\|^2}{dt} = 0$ when $\|\mathbf{w}\| = 1$, so that the $n$-dimensional sphere $\|\mathbf{w}\| = 1$ is an attracting hypersurface for the system (in particular, the stable equilibria are the two normalized principal eigenvectors of $\mathbf{C}$, which lie on this sphere).

In previous work, we (Rădulescu, Cox, & Adams, 2009; Rădulescu & Adams, 2013) and others (Botelho & Jamison, 2002, 2004) have examined how cross talk affects the Oja model. We formalized the effects of synaptic cross talk via a time-dependent (but not input or weight-dependent) error matrix $\boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{E}}(t)$, whose elements reflect at each time $t$ the fractional contribution that the activity across weight $\omega_i$ makes to the update of $\omega_j$.

When introducing this matrix in the original Hebbian updating rule ($\Delta w_i = \gamma y[\boldsymbol{\mathcal{E}}x]_i$) and performing the same normalization steps as in the error-free rule, one would obtain the order $o(\gamma)$ approximation for each component $\omega_i$:

$$\omega_i(t+1) = \omega_i(t) + \gamma y\left([\boldsymbol{\mathcal{E}}\mathbf{x}]_i - \boldsymbol{\omega}^T[\boldsymbol{\mathcal{E}}\mathbf{x}]\omega_i\right). \tag{2.4}$$

The subtractive term is one way of implementing an approximate nonlocal normalization as part of a local online rule. However, the presence of the same error matrix $\boldsymbol{\mathcal{E}}$ in the forgetting term implies biologically that the

normalizing component "knows" ahead the pattern-to-pattern form of $\boldsymbol{\mathcal{E}}$, which is highly implausible. One should rather consider the Hebbian and normalizing steps to have different error matrices, reflecting their different physical implementation. Here, we assume LTD to be triggered presynaptically by a retrograde messenger, so that diffusion to different synapses located on the same output neuron does not matter, and subsequently the normalizing step is error free, requiring only the calculation of the square of the output $y^2$ and its multiplication by $\boldsymbol{\omega}$ (see section 4.2 for a more extensive discussion of the biophysical implementation of these steps). The average (mean field) form of the rule becomes (since $\boldsymbol{\mathcal{E}}$ is input and weight independent)

$$\frac{d\mathbf{w}}{dt} = \gamma[\mathbf{ECw} - (\mathbf{w}^T\mathbf{Cw})\mathbf{w}], \tag{2.5}$$

where, as before, $\mathbf{w}(t) = \langle \boldsymbol{\omega}(t+1)|\boldsymbol{\omega}(t)\rangle$. The average error matrix $\mathbf{E} = \langle \boldsymbol{\mathcal{E}} \rangle \in \mathcal{M}_n(\mathbb{R})$ has positive entries, and is symmetric and equal to the identity matrix $\mathbf{I} \in \mathcal{M}_n(\mathbb{R})$ for zero cross talk. To fix our ideas, we considered the error matrix $\mathbf{E}$ to be isotropic, that is, of the form

$$\mathbf{E} = \begin{bmatrix} q & \epsilon & \cdots & \epsilon \\ \epsilon & q & \cdots & \epsilon \\ \vdots & & \ddots & \vdots \\ \epsilon & \epsilon & \cdots & q \end{bmatrix}, \tag{2.6}$$

where $0 \leq \epsilon \leq \frac{1}{n}$ represents synaptic cross talk, or "error," and $\frac{1}{n} \leq q \leq 1$ is the synaptic "quality," satisfying $q + (n-1)\epsilon = 1$.

One can easily show that equation 2.5 preserves the dot product $\langle \cdot, \cdot \rangle_{\mathbf{C}}$ (where $\langle \mathbf{v}, \mathbf{u} \rangle_{\mathbf{C}} = \mathbf{v}^T\mathbf{Cu}$, for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$). Furthermore, an equilibrium for equation 2.5 is an eigenvector of $\mathbf{EC}$, normalized so that $\|\mathbf{w}\|_{\mathbf{C}}^2 = \lambda_{\mathbf{w}}$, where $\lambda_{\mathbf{w}}$ is its corresponding eigenvalue of $\mathbf{EC}$.

Notice that equation 2.5 has equilibria that are tightly related to those of the averaged corresponding form of equation 2.4; our working form is, however, simpler computationally, in the sense that stability of equilibria is more easily tractable. We have shown that the eigenvalues of the Jacobian matrix at an equilibrium $\mathbf{w}$ are given by $-2\gamma\lambda_{\mathbf{w}}$ and $-\gamma[\lambda_{\mathbf{w}} - \lambda_{\mathbf{v}_j}]$, where $\lambda_{\mathbf{w}}$ and $\lambda_{\mathbf{v}_j}, \forall j = \overline{1, n-1}$ are the $n$ eigenvalues of $\mathbf{EC}$ (noting first that $\mathcal{B}_{\mathbf{w}} = \{\mathbf{w}, \mathbf{v_1}, \ldots \mathbf{v_{n-1}}\}$, the completion of $\mathbf{w}$ to a basis of eigenvectors of $\mathbf{EC}$, orthogonal with respect to the dot product $\langle \cdot, \cdot \rangle_{\mathbf{C}}$, also forms an eigenvector basis for the Jacobian). We concluded that if $\mathbf{EC}$ has a unique largest eigenvalue (which is generically true), then a normalized eigenvector $\mathbf{w}$ is a local hyperbolic attracting equilibrium for equation 2.5 iff it corresponds to this maximal eigenvalue. If $\mathbf{EC}$ has a multiple largest eigenvalue, the

system will have a set of nonisolated, neutrally attracting equilibria (all normalized eigenvectors spanning the principal eigenspace in this case of dimension $\geq 2$). Some of the computations are summarized in appendix A (e.g., a description of the attraction basins, supporting the absence of cycles in the phase space) and are expanded in more detail in our previous work (Rădulescu et al., 2009; Rădulescu & Adams, 2013).

Since the nature and position of the equilibria depend on the spectral properties of **EC**, the next task is to study the spectral changes of **EC** when perturbing the system by increasing cross talk. In our previous work on the model, we investigated the effects of cross talk on the system's dynamics and their dependence on the characteristics of the input distribution (correlation sign, degree of bias). However, in our first study, we considered learning only of positively correlated $n$-dimensional input distributions; we found a smooth degradation of the learning outcome with increasing error but no sudden changes in dynamics (Rădulescu et al., 2009). In our second study, we showed that negatively correlated inputs can induce a bifurcation (stability swap of equilibria, through a critical stage) when increasing the error, even in a case as simple as a two-dimensional system. This bifurcation occurred only in the case of unbiased inputs (Rădulescu & Adams, 2013), and we interpreted it in the context of ocular dominance and input segregation.

One would expect that increasing the dimension of the system would bring out interesting new phenomena induced by cross talk. With this goal in mind, we want to extend our existing work and investigate the effects of cross talk in higher-dimensional networks, learning different classes of negatively correlated inputs. We consider the input distribution to be centered (mean zero) and the mutual correlations to be identical. More precisely, we will consider covariance matrices of the form

$$\mathbf{C} = \begin{bmatrix} v+\delta_1 & \pm c & \cdots & \pm c \\ \pm c & v+\delta_2 & \cdots & \pm c \\ \vdots & & \ddots & \vdots \\ \pm c & \pm c & \cdots & v+\delta_n \end{bmatrix}. \tag{2.7}$$

For our general computations, we assume that the inputs have mutual covariances $c$ uniform in absolute value, and small with respect to the diagonal variances. More precisely, we assume $v + \delta_n > (n-1)|c|$, making the matrix diagonally dominant (see section 2.2 for considerations on the input statistics). Throughout the letter, $\delta_i$ will be called the input biases. Without loss of generality, we set $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_n \geq 0$. For any $k \leq n$, we say that the input has bias loss of order $k$ if $\delta_1 = \cdots = \delta_k > 0$. In particular, we say that the input is unbiased if it has bias loss of order $n$, that is, if $\delta_1 = \cdots = \delta_n = 0$. Although the background covariance $\pm c$ is taken for simplicity to be uniform in absolute value, we expect the inspecific learning

rule to lead to interesting dynamics, in particular when the inputs exhibit a certain degree of mutual correlation.

**2.2 Oja's Rule and the Input Statistics.** The goal of this work is to investigate the effects of one particular aspect of biological realism (cross talk) in the context of a model that is otherwise as transparent as possible. We chose the Oja principal component analyzer as a widely known and simple example of a Hebbian model of unsupervised learning, important in cortical processing (Hinton & Sejnowski, 1999) and involving repeated adjustment driven only by statistical properties of the input. While a connectionist model may capture some of the desired basic aspects of learning dynamics, the situation in the brain is far from being this simple.

To begin the Oja model may appear rather unbiological by its very use of a rate-coding scheme and a simple multiplicative Hebbian learning rule, in conjunction with a local (and controversially plausible) normalization procedure (section 4.2 gives more detail on possible empirical bases of the rule and their implementation). While in our approach we incorporated cross talk, we neglected many other biological aspects inherent in synaptic transmission (e.g., timed spikes, external noise, temporal correlations, synaptic homeostasis; Cox & Adams, 2009); a more biologically realistic model would use spike-timing-dependent plasticity and natural inputs (Hyvärinen, Hurri, & Hoyer, 2009). We simply used positive or negative continuous-time activations and weights (one can interpret negative weights as disconnections), and we assumed the input patterns to be zero mean and have identical mutual correlations (Rădulescu et al., 2009). More elaborate models, incorporating detailed spiking patterns, may automatically learn the principal component of the zero-mean inputs, without explicit centering or normalization. Gerstner and Kistler (2002) have developed a model that assumes an Oja-type rate-coding scheme, with Poisson spikes and spike-time-dependent plasticity with LTP and LTD lobes, and postsynaptic spikes triggered by presynaptically generated EPSPs. One could in principle study the effects of cross talk on such a model by applying an error matrix to the LTP or LTD parts; a direct analysis, however, might turn out to be much more difficult than in the case at hand here.

Since our analysis focuses on symmetric matrices $\mathbf{C}$ with positive or negative off-diagonal elements, we have to ask whether and when such a matrix can constitute the covariance matrix of a centered $n$-dimensional distribution. While establishing equivalent conditions may be difficult even for small dimensions (Vasudeva, 1998), one can find sufficient criteria (e.g., any positive semidefinite $\mathbf{C}$ is a covariant matrix).[1]

In our initial computations, we assumed sufficiently weak pairwise correlations to make $\mathbf{C}$ diagonally dominant (in this case, equivalent

[1] If $\mathbf{X}$ is an $n \times 1$ column vector-valued random variable whose covariance matrix is the $n \times n$ identity matrix, then $\mathrm{cov}(\sqrt{\mathbf{C}}\mathbf{X}) = \sqrt{\mathbf{C}}\,\mathrm{cov}(\mathbf{X})\,\sqrt{\mathbf{C}} = \mathbf{C}$.

to $v + \delta_n > (n-1)|c|$). Any symmetric diagonally dominant matrix with nonnegative diagonal entries is automatically positive semidefinite, hence a covariance matrix. Such segregated inputs can be found in a variety of contexts in the brain. For example, studies of cortico-striate projections (Yim, Aertsen, & Kumar, 2011) have observed weak pairwise correlations within the pool of inputs to individual striatal neurons, which are believed to enhance the saliency of signal representation in the striatum. On the other hand, **C** will not remain diagonally dominant for strong pairwise correlations, which are also likely to occur biologically. A known example of cells with strongly correlated activity is that of retinal ganglion cells, placed in topographic proximity of each other and innervating the same cell in the LGN (Mastronarde, 1989; Trong & Rieke, 2008). Our work in sections 3.1 and 3.2 assumes diagonal dominance (as a mathematical convenient assumption that allows us to establish a useful classification and illustrate typical behaviors that can occur in the system). In appendix C, we complete our analysis with a numerical approach to a larger collection of matrices, with extended parameter ranges.

**2.3 The Error Matrix.** Together with the uniform magnitude of input cross-correlations (i.e, uniform absolute value $|c|$ of the off-diagonal elements of **C**), we also assumed, for simplicity, uniform error (the Hebbian adjustment of any weight was equally affected by error and did not depend on either the strength of that weight or on geometry). Such "isotropicity" seems like a reasonable basic assumption and has been discussed in our previous work (Rădulescu et al., 2009; Rădulescu & Adams, 2013). Furthermore, it allowed us to identify other features of the input distribution, crucially consequential on the learning dynamics and outcome: the sign of the mutual correlations and the input bias. However, cross talk has been documented experimentally, for technical reasons, mostly between synapses that are anatomically neighboring each other (Harvey & Svoboda, 2007; Bi, 2002; Bonhoeffer et al., 1989). In previous work (Rădulescu et al., 2009), we have justified isotropicity based on the fact that individual cortical connections are composed of multiple synapses scattered over the dendritic tree (Varga, Jia, Sakmann, & Konnerth, 2011; Chen, Leischner, Rochefort, Nelken, & Konnerth, 2011; Jia, Rochefort, Chen, & Konnerth, 2010), but we have also considered other (more metric-dependent although all symmetric) forms of **E**. Cross-talk effects could probably be captured when using more general, nonisotropic forms for **E** without affecting the main conclusions. In this letter, the distinction between local and global cross talk is not that relevant, since our main results concern a low- (three-)dimensional network.

## 3 Results

**3.1 Classes of Inputs and Bias Effects on Three-Dimensional Dynamics.** In this section, we study how input patterns can influence the effects

of cross talk in driving the dynamics of a three-dimensional network—the lowest dimension for which the question applies but which seems to capture the essence of this problem even in higher-dimensional systems. In this section, we inspect all combinatorial possibilities of input bias and correlation sign (as defined below) and determine the effect of increasing cross talk on dynamics in each case. In section 3.2 and the appendixes, we support with rigorous proofs some of the results obtained through numerical simulations (we used Matlab software package, version 7.2.1).

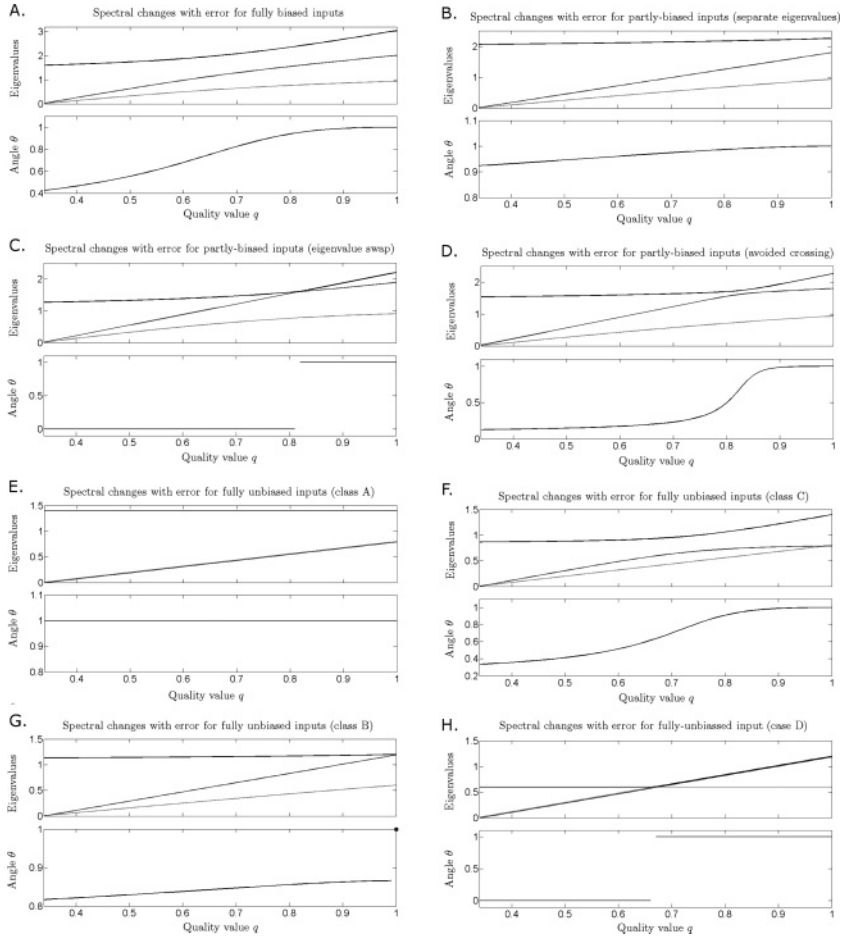
We considered all combinatorial possibilities of covariance matrices of the form

$$\mathbf{C} = \begin{bmatrix} v + \delta_1 & \pm c & \pm c \\ \pm c & v + \delta_2 & \pm c \\ \pm c & \pm c & v \end{bmatrix}, \tag{3.1}$$

where $\delta_1 \geq \delta_2 \geq 0$ (i.e., allowing bias of any order). We first analyzed the case $v > 2|c|$, corresponding to the diagonal dominance assumption discussed in section 2.2. A subsequent numerical analysis extended the results to encompass a wider variety of input distributions. In order to simplify the notation, we assumed without loss of generality that $\delta_3 = 0$, which can be easily justified by a change of variables.

We found that in special highly unbiased cases, cross talk has no effect on the presence and position of the asymptotic attractors (see Figure 1E). In other cases, the depreciation of the asymptotic outcome with error is so slow that small levels of cross talk have virtually no effect on learning (see Figure 1A and 1B; also see Figure 2 for a phase-space illustration). Other significant classes of inputs, however, showed a sudden change of the attractor states, from a reliable principal component estimator to an almost orthogonal direction. This occurred either in the form of an eigenvalue swapping bifurcation in dynamics (producing the instantaneous loss of learning accuracy at a critical error value; see Figures 1C and 3 for an illustration of phase-space transitions) or in the milder form of an eigenvalue "avoided crossing," (inducing a smooth yet very steep depreciation of the learned direction at a specific error; see Figures 1D and 1G). As discussed in our previous work, bifurcations and avoided crossings can be practically indistinguishable: learning works reasonably well for small enough errors. For errors past the crash value, the outcome becomes irrelevant to the input statistics, and the system is essentially encoding information on the cross-talk pattern itself.

None of these possibilities is a priori excluded in the brain, but previous work has suggested that nature may favor bias. Segregated outcomes (disconnected completely, forming wiring patterns that are then subject to more subtle synaptic learning) are considered to be an important part of normal development. In our previous work, we argued that cross talk seems to act




against this desymmetrizing tendency and prevent segregation, especially for inputs close to unbiased. We viewed this as a limitation of symmetry-breaking mechanisms that generate specific wiring, and we further argued that other factors, such as strong mutual inhibition (large negative correlations) or special specificity-enhancing circuitry ("proofreading"), might act to overcome the equalizing effect of cross talk. The current study completes this idea with new aspects.

One can say, then, that efficient cross-talk-induced segregation happens in our model for a balance of positive and negative correlations in the input distribution. Since the presence, number, and strength of the negative correlations appeared to be crucial in determining the behavior of the system, we defined a formal classification of all possible correlation matrices based on

Figure 1: Spectral changes induced by increasing inspecificity, for various inputs schemes. In all panels, we show, with respect to the quality $q = 1 - 2\epsilon$, the evolution of the eigenvalues, with black for the largest eigenvalue, red for the second largest, and green for the lowest (top subplot); the cosine of the angle between the inspecific stable vector and the correct attracting direction(s) (bottom subplot). In all panels, $v = 1$, $|c| = 0.2$. The classification is as follows. (A) For fully biased inputs ($\delta_1 = 2$, $\delta_2 = 1$), the three eigenvalues remain separated. For partly biased inputs ($\delta_1 = \delta_2 = 1$), there are three cases, depending on the number of negative cross-correlations and on their placement: the leading eigenvalues can remain separated (B). They can cross at a critical value of $q = q^*$ (C) or approach significantly for some value of $q$ but "avoid" crossing (D). For fully unbiased inputs, we found four cases, classified simply by the number of negative off-diagonal cross-correlations: all positive cross-correlations, and leading eigenvalues remain separated (E); one negative cross-correlation, where leading eigenvalues coincide only at $q = 1$ and immediately separate (F); two negative cross-correlations, where leading eigenvalues may approach each other in an avoided crossing of magnitude depending on parameters, but remain separated (G); all negative cross-correlations, where leading eigenvalues coincide on a whole interval, as quality depreciates from $q = 1$ to a critical value (H). In panel H, the system has a curve of half-neutral attractors, which persists until $q$ reaches the critical value, when a different, orthogonal eigenvector takes over as the stable direction. (Please refer to online supplement for color version of this figure.)

the number of negative upper-diagonal entries of **C** and then used the three classes to understand the corresponding behavior with respect to cross talk.

We distinguished four combinatorial classes: **Class** $(+, +, +)$, comprising the unique matrix configuration with all positive entries; **Class** $(+, +, -)$, made of the three matrix configurations with one negative upper-diagonal entry; **Class** $(+, -, -)$, for the three configurations with two negative upper-diagonal entries; **Class** $(-, -, -)$, for the one configuration with all negative off-diagonal entries. We studied the matrix **EC**, and the differences that occur in its spectrum when considering different classes of input, in conjunction with different degrees of bias: from fully biased ($\delta_1 > \delta_2 > 0$) to partly biased ($\delta_1 = \delta_2 > 0$) to fully unbiased ($\delta_1 = \delta_2 = 0$). In this section, $q$ will be restricted to the interval $(1/3, 1]$ (representing quality higher than error). Based on these combinatorial classes of input, we distinguished three main qualitative behaviors: separated leading eigenvalues, crossing leading eigenvalues and "avoided crossing."[2]

[2]Since the spectra depend qualitatively on all parameter values, we present here the results of a numerical investigation rather than a rigorous analytical study, which would be extremely cumbersome. The only case in which the computations are more tractable and for which we preferred an analytical approach is the fully unbiased case, presented in appendix A.

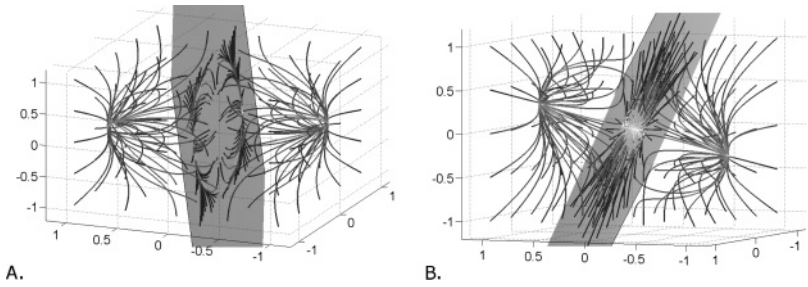



Figure 2: Phase-Space trajectories for fully biased inputs. (A) In the absence of error, the system converges generically to the two normalized vectors in the principal direction $\mathbf{w_C}$ of the covariance matrix $\mathbf{C}$. The attraction basins are separated by the subspace $\langle \mathbf{w}, \mathbf{w_C} \rangle = 0$ (the shaded plane). (B) For error $\epsilon = 0.2$, the system converges generically to the two normalized vectors in the principal direction $\mathbf{w_{EC}}$ of the modified covariance matrix $\mathbf{EC}$. The attraction basins are separated by the subspace $\langle \mathbf{w}, \mathbf{w_{EC}} \rangle = 0$ (the shaded plane). Parameters: $v = 1$, $c = 0.2$, $\delta_1 = 2$, $\delta_2 = 1$. Color coding: trajectories evolve in time from darker to lighter shades. (Please refer to online supplement for color version of this figure.)

*3.1.1 Separated Leading Eigenvalues.* The largest eigenvalue remains separated from the second largest eigenvalue for the whole range of $q$ (as illustrated in Figure 1A, top panel), determining the corresponding leading eigenvector to gradually drift from the direction of the principal component of $\mathbf{C}$, as $q$ decreases (blue curve in Figure 1A, bottom panel). For any value of $q$, the system has two hyperbolically attracting equilibria: the normalized principal eigenvectors of $\mathbf{EC}$, whose basins are separated by an invariant plane. In Figure 2, we show the evolution of a set of trajectories to illustrate convergence to the two attractors in the phase space, as well the dynamics within the separating plane.

In the presence of cross talk, the network will process the input in a very similar qualitative fashion as in absence of cross talk, observing the main statistical trends, even though the quantitative outcome might be slightly or more substantially altered, depending on the input pattern and the degree of cross talk. Depending on parameters, the eigenvalue curves with respect to $q$ may exhibit a significant point of minimal separation, where the learning outcome (leading eigenvector of $\mathbf{EC}$) deteriorates very fast (see section 3.1.3).

This case is generally associated with biased inputs (the only possible behavior when $\delta_1 > \delta_2 > 0$). That is, no negative correlations are required to maintain segregated inputs in their segregated state when cross talk is introduced. However, this behavior can be found in conjunction with loss of bias, provided the mutual negative correlations are limited: it also appears in partial loss of bias ($\delta_1 = \delta_2 > 0$) for class $(+, +, +)$ (see Figures 1B and 2),

as well as in full loss of bias ($\delta_1 = \delta_2 = 0$) for classes $(+, +, +)$ and $(+, +, -)$ (see Figures 1G, 1E, and 2).

An interesting, quite extreme case of separated eigenvalues occurs for symmetric inputs that are fully unbiased and all positively correlated: the leading eigenvalue is separated from the second eigenvalue (which has multiplicity two), but neither the leading eigenvalue nor the corresponding eigenvector of **EC** changes when the cross talk is increased. Hence, in this case, the learning is fully accurate for any degree of cross talk (see Figure 1E); one may argue that this particular class of input statistics is completely error proof.

*3.1.2 Crossing of Leading Eigenvalues.* This behavior sits, in a sense, at the opposite pole of the "separated eigenvalues" case, and in its most standard form, it is typical to partial loss of bias ($\delta_1 = \delta_2 = \delta > 0$) in combination with all negative correlations, that is, class $(-, -, -)$; see Figure 1C and section 3.2. The term describes an instantaneous swap of the attractors from one eigendirection to another direction that could be as much as orthogonal to the original principal component swap, which produces a crash in the learning outcome. This behavior occurs when the two leading eigenvalue branches cross and switch at a critical value of the quality $q^* = \frac{v+\delta+c}{v+\delta-c}$. (We have described this phenomenon in a two-dimensional model in Rădulescu & Adams, 2013.) Very small levels of cross talk ($q > q^*$) in fact have very little effect on learning in this case. Although the leading eigenvalue changes, the direction of the leading and attracting eigenvector is preserved, so that the system will converge to the same outcome as in the absence of error.

This may seem like a very desirable input distribution to learn in the presence of low cross talk; however, one has to keep in mind that if the cross-correlations are small in absolute value $|c|$ with respect to the variance $v$, then the critical $q^*$ gets arbitrarily close to 1. Such perfect learning will therefore happen only when inspecificity is infinitesimally small, which makes this scenario lose its appeal, especially when we recall that at the end of the "good" interval lies the bifurcation, crashing the equilibrium to a direction completely irrelevant to the input statistics. In this light, one might expect the network to have an additional, quite precise estimator of the degree of cross talk involved, so that when learning an irrelevant outcome, it would at least be aware of it. Any slight error of the system toward miscalculating the limits for the permissible error could have dire consequences.

In Figure 3, we represent three phase-space plots: before, at, and after the bifurcation point $q = q^*$. While Figures 3A and 3C illustrate the typical phase space with two hyperbolically stable equilibria (one representing accurate, error-free learning and the other inaccurate learning for a postcritical error), the phase space at the bifurcation point is qualitatively different: the system has no hyperbolic attractors but rather a closed curve (ellipse) of half-stable equilibria (neutral along the direction of the curve). Clearly, the outcome

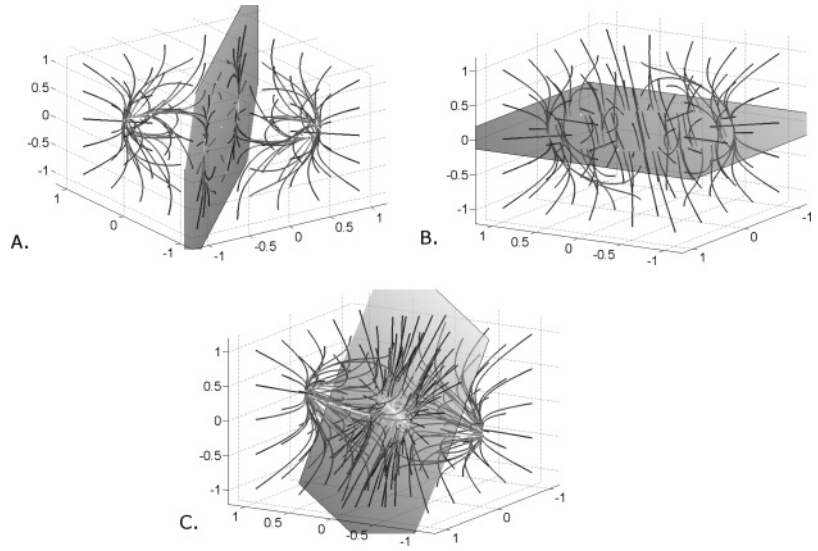



Figure 3: Bifurcation in attractor dynamics for partly biased inputs, all negative cross-correlations. (A) For small error, the attractors (the two normalized principal eigenvectors of **EC**) do not differ much from the correct attractors (the two normalized principal eigenvectors of panel **C**). The attraction basins are separated by the subspace $\langle \mathbf{w}, \mathbf{w}_{\mathbf{C}} \rangle = 0$ (the shaded plane). (B) For critical error $\epsilon = \frac{-c}{v+\delta-c}$, the system exhibits an ellipse of neutrally stable equilibria (yellow curve contained in the shaded plane). (C) For error past the critical value, the attractors have moved significantly far from the correct positions. Parameters: $v = 1$, $c = 0.2$, $\delta = \delta_1 = \delta_2 = 1$. Color coding: trajectories evolve in time from darker to lighter shades. (Please refer to online supplement for color version of this figure.)

of learning is in this case extremely dependent on the initial conditions (although, as we commented in Rădulescu & Adams, 2013, the stochastic version of the system will have noise-driven stationary solutions that drift around this neutrally attracting ellipse).

The neutrally attracting ellipse phase-plane dynamics is not specific to this critical bifurcation state (and thus it cannot be ignored as improbable in the context of generic behavior). For some classes of inputs, such an attracting-ellipse slice represents the natural state of the cross-talk free system and persists for an entire inspecificity range (see Figure 4). This is the case for bias of order two ($\delta_1 = \delta_2 = 0$) when occurring in conjunction with substantial negative correlations, that is, classes $(+, -, -)$ and $(-, -, -)$. The computations are quite simplified in the absence of any bias, so for the case of fully unbiased inputs we carried out analytically a complete classification in theorem 1 in appendix A. We describe these two fully unbiased cases in more detail below.

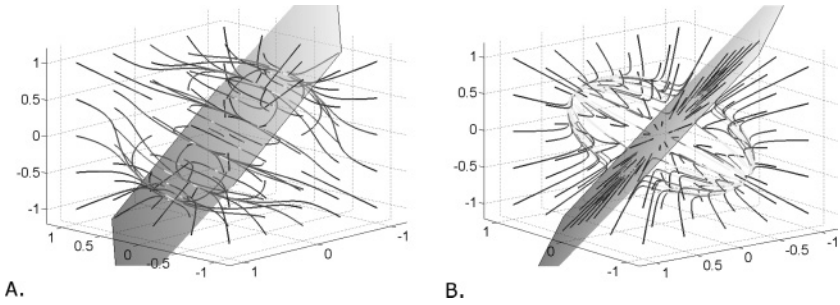



Figure 4: Bifurcation in attractor dynamics for partly biased inputs, all negative cross-correlations. (A) For small error, the system has an ellipse of neutrally stable equilibria (yellow curve). This ellipse is stable in the sense that it persists for a whole interval of errors, from $\epsilon = 0$ to $\epsilon = \frac{-c}{v-c}$. (B) For error past the critical value, the ellipse is destroyed, but the new attractors are significantly far from the plane of the ellipse. Parameters: $v = 1, c = 0.2, \delta_1 = 2 = \delta_2 = 0$. Color coding: trajectories evolve in time from darker to lighter shades. (Please refer to online supplement for color version of this figure.)

We found that in instances of highly unbiased inputs, learning may lead to an ambiguous outcome even in the absence of cross talk (see Figures 1F, 1H, and 4). Indeed, in the cross-talk-free class $(+, -, -)$, the matrix $\mathbf{C}$ has a double leading eigenvalue to begin, and the system has a whole closed curve of neutrally attracting equilibria (in the eigenplane spanned by the corresponding eigenvectors). When cross talk is introduced, the two leading eigenvalues segregate, and one of the eigenvectors takes over, which determines an immediate complete switch in the learning outcome. In this case, even the smallest degree of inspecificity leads to favoring one specific direction, slightly detaching off the plane that contains the curve of accurate equilibria (notice that the cosine of the accuracy angle, represented by the blue curve in Figure 1F, does not fall too far off the perfect value $\cos(\theta) = 1$).

We may interpret this as the error helping the system "make up its mind" in the presence of too much ambiguity in the input statistics. This is an occurrence we have not encountered in our previous, more restrictive versions of the model, since it requires inputs with concomitant negative cross-correlations and loss of bias of order >2. This ambiguity can be interpreted as the basis of a competitive process in which any input channel has equal chances to win. Competitive dynamics has been studied at large in developmental and learning models in the context of imposed (by means of multiplicative or subtractive normalization) or emergent competition. It has become clear that a linear Hebb rule, even when coupled with a multiplicative normalization or winner-takes-all type nonlinearities, is not able to produce segregation of positively correlated inputs (von der Malsburg, 1973; Goodhill & Barrow, 1994; Miller & MacKay, 1994). When used in

conjunction with unbiased inputs, it will lead to an equal-weight outcome (Dayan & Abbott, 2002). A variety of known nonlinear mechanisms can break the inherent symmetry, even when the input per se does not favor segregated outcomes (Elliott, 2003), including subtractive normalization (Miller & MacKay, 1994; Goodhill & Barrow, 1994), the BCM rule (Bienenstock, Cooper, & Munro, 1982) and spike-time-dependent-plasticity (Elliott, 2008). As interpreted in one of our previous discussions on ocular dominance wiring (Rădulescu & Adams, 2013), such mechanisms may lead, for example, to ocular segregation under unbiased statistics (the two eyes are likely receiving similar, positively correlated inputs from the visual field). One context that permits segregation under multiplicative normalization is having negatively correlated inputs.

Our current analysis illustrates this issue and shows that when sufficient negative correlations are present, the fashion in which the cross talk handles inherent input ambiguity or competition depends quite significantly on the number (and, to a lesser extent, the positions) of the negative mutual correlations within the input. In our model, at least two negative mutual correlations are necessary for cross talk to produce segregation of symmetric inputs. For two out of three negative correlations, even the smallest degree of cross talk helps the system make an asymptotic selection for one particular direction in the eigenspace spanned by the multiple eigenvalue. For all negative correlations, no small degree of cross talk can resolve this competitive state. The level of critical cross talk that can finally destroy the curve of neutrally stable equilibria also pushes the system to learn an orthogonal direction, hence becomes irrelevant to the main features of the original input statistics. Indeed, in the cross-talk-free class $(-, -, -)$, the matrix $\mathbf{C}$ has a double leading eigenvalue, and the system again has a whole ellipse of neutral equilibria, contained in the corresponding eigenplane. When subject to errors up to a critical value $q^* = \frac{v+c}{v-c}$, the two larger eigenvalues change but remain equal; furthermore, the subspace spanned by the two corresponding eigenvectors remains unchanged, hence the learning process preserves the original ambiguity. Past the critical error value, the eigenvalues swap, and the eigendirection for the new leading eigenvalue (of multiplicity one) is orthogonal to the previous plane (see Figure 1H). In other words, past the critical error value, the system will finally choose a particular direction to learn, but this direction will be highly inaccurate, and thus the task of learning the input statistics will be performed very poorly.

*3.1.3 "Avoided Crossing" of Leading Eigenvalues.* This can be seen as a hybrid case in which the principal eigenvalues never actually swap but get very close (arbitrarily close, depending on the values of $v$ and $|c|$), so that learning has a significantly rapid depreciation around the critical value $q^*$ (which also depends on all other parameter values; see the blue curve in Figure 1D). This situation can be observed when the input has partial bias loss in mixed cases from classes $(+, +, -)$ and $(+, -, -)$.

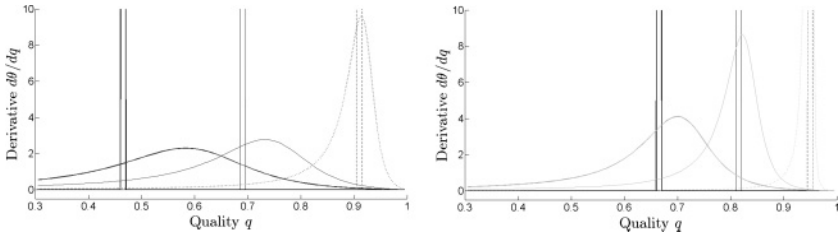



Figure 5: Comparison between real and fapp bifurcations and their dependence on input patterns. We illustrate the network performance, or "sensitivity," with respect to $q$ (measured as the derivative $\frac{d\theta}{dq}$ of the angle $\theta$ between the network attracting direction with and without cross talk) for two types of input statistics with order one bias loss ($\delta_1 = \delta_2 = \delta > 0$): one of class $(-,-,-)$ (giving rise to eigenvalue swap bifurcations, plotted in blue) and one of class $(+,-,-)$ (producing fapp bifurcations, plotted in red in panel A and green in panel B, respectively). The input base variance was fixed to $v = 1$ in both panels, but the bias was $\delta = 1$ (right panel) and $\delta = 0.1$ (left panel). We also inspected several values of the mutual cross-correlations: $c = 0.4$ (thick solid curve), $c = 0.2$ (thin solid curve), and $c = 0.05$ (dotted curve). As $c$ decreases, the fapp bifurcations for the $(+,-,-)$ input are getting arbitrarily close to $q = 1$, and approximate better and better the discontinuous blow-up of the corresponding real bifurcation obtained for the $(-,-,-)$ input. This effect is more evident when the partial bias is increased (from $\delta = 0.1$ to $\delta = 1$). (Please refer to online supplement for color version of this figure.)

Biologically, such a "pseudobifurcation," if occurring over a narrow enough range of $q$, is indistinguishable from a real bifurcation, induced by crossing eigenvalues; for this reason, we refer to it as a for-all-practical-purposes (fapp) bifurcation. Since it represents a sudden (although smooth) depreciation of the principal direction, one may consider calculating the "susceptibility" or "sensitivity" $\frac{d\theta}{dq}$ of the angle $\theta$ with respect to the quality $q$.

In Figure 5, we illustrate the difference between the discontinuous breakdown of the derivative $\frac{d\theta}{dq}$ in the case of a real bifurcation (discontinuity of $\theta$) and the continuous blow-up of $\frac{d\theta}{dq}$ in the case of a fapp bifurcation ($\theta$ has a significant although finite variation over a narrow interval of $q$). One may regard this dichotomy to be in principle analogous to the difference between discontinuous and continuous phase transitions. Formally, an avoided crossing can be defined to produce a fapp bifurcation if the size of the blow-up exceeds a certain threshold (which may depend on the particular network and the accuracy level desired for learning).

With this definition, there are circumstances in which fapp bifurcations can occur even at arbitrarily small cross talk ($q$ arbitrarily close to 1). For example, Figure 5 shows the difference between the effect of cross talk in

the case of two input distributions, both with loss of bias of degree one. For the first type of distribution, class $(-,-,-)$, the all-negative mutual cross-correlations determine eigenvalue crossing (the blue curves, which exhibit discontinuous blow-ups). The second type, class $(+,-,-)$, can lead to avoided crossing. We compared the behavior of the network in these two situations, inspecting a few values of the bias $\delta = \delta_1 = \delta_2 > 0$ (left panel versus right panel), and mutual cross-correlation values $|c|$ (different curves in the same panel, as explained in the caption). We found that increasing the bias $\delta$ and decreasing the cross-correlations $|c|$ transports the point of maximum sensitivity (the location of the blow-ups) closer to $q = 1$. Moreover, the size of the continuous blow-up (the height of the finite peak in the case of avoided crossing) gets larger as $q$ migrates toward 1, so that the smaller the values of $|c|$, the lower the level of cross talk sufficient to produce a blow-up, and the more indistinguishable the fapp bifurcation looks from the bifurcation-induced discontinuity. This reiterates the idea that a fapp bifurcation can be as detrimental to learning as a real bifurcation, especially since it can arise at arbitrarily small levels of cross-talk, just like an actual bifurcation.

In appendix C, consider inputs with stronger pairwise correlations (so that **C** is no longer diagonally dominant). When we consider high negative mutual correlations, the fapp bifurcation, associated with arbitrarily small levels of cross talk, appears in conjunction with an actual bifurcation, at very high cross-talk levels. This suggests that for such inputs, after undergoing the fapp degradation in outcome, the system may suddenly reverse to accurate computation of the learning attractor at very high cross-talk levels.

**3.2 An Analytical Application in Higher Dimensions.** In this technical section, we work out an analytical $n$-dimensional computation, with the aim of showing that the phenomena described in section 3.1 may also apply to describe behavior in a higher-dimensional Oja learning model with cross talk. In previous work (Rădulescu, Cox, & Adams, 2009), we have investigated the $n$-dimensional case for all positively cross-correlated inputs and showed that it does not induce stability swapping bifurcations, even in higher dimensions. Since negative correlations are the key ingredient for the presence of bifurcations, we consider for our application in this section the case of all negative cross-correlations:

$$\mathbf{C} = \begin{bmatrix} v+\delta_1 & -c & \cdots & -c \\ -c & v+\delta_2 & \cdots & -c \\ \vdots & & \ddots & \vdots \\ -c & -c & \cdots & v+\delta_n \end{bmatrix}, \tag{3.2}$$

where here $c > 0$. Our three-dimensional numerical results suggest that combined covariance matrices that encompass other patterns of positive

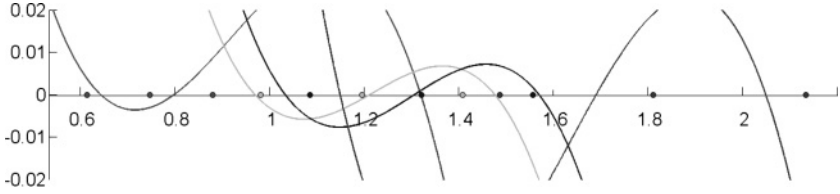



Figure 6: A simple example of how the characteristic polynomial $\Delta$ of **EC** and its roots change as the quality $q$ decreases, for dimension $n = 3$ and fixed parameters $v = 1$, $c = -0.2$, $\delta_j = j/3$, for $j \in \overline{1,3}$, so that $\epsilon_1^* \sim 0.091$, $\epsilon_2^* \sim 0.107$, $\epsilon_3^* \sim 0.130$. Each different color represents a different value of $q$: $q = 0.98$ (red), $q = 0.805$ (blue), $q = 0.76$ (green), and $q = 0.6$ (pink). The continuous curves correspond to the graph of the polynomial for different $q$'s, and the bullets represent (along the $x$-axis) the points $\lambda_j = (q - \epsilon)(v + \delta_j - c)$, for $j = \overline{1,3}$. The figure shows how the order of the roots of $\Delta$ changes with respect to the points of the partition $\lambda_3 < \lambda_2 < \lambda_1$ (which in turn travel down the axis as $q$ decreases). For $q = 0.98$ (i.e., $\epsilon = 0.01 < \epsilon_1^*$), $\lambda_1 > \xi_1 > \lambda_2 > \xi_2 > \lambda_3 > \xi_3$. For $q = 0.805$ (i.e., $\epsilon = 0.0975 \in [\epsilon_1^*, \epsilon_2^*]$), $\xi_1 > \lambda_1 > \lambda_2 > \xi_2 > \lambda_3 > \xi_3$. For $q = 0.76$ (i.e., $\epsilon = 0.12 \in [\epsilon_2^*, \epsilon_3^*]$), $\xi_1 > \lambda_1 > \xi_2 > \lambda_2 > \lambda_3 > \xi_3$. For $q = 0.6$ (where $\epsilon = 0.2 > \epsilon_3^*$), $\xi_1 > \lambda_1 > \xi_2 > \lambda_2 > \xi_3 > \lambda_3$. (Please refer to online supplement for color version of this figure.)

and negative mutual cross-correlations are expected to produce hybrid dynamics between these two extreme ends. In a higher-dimensional network, the dynamics may depend strongly not only on the number of negative correlations, but also on their distribution and geometry within the covariance matrix. A random matrix approach may help classify the behavior for all input patterns, but this is not within the scope of this study.

In this section, we present only the main analytical results we obtained for our application; proofs of the statements and additional comments can be found in appendix D. Propositions 1 and 2 differentiate between behaviors in response to biased versus unbiased $n$-dimensional negatively correlated inputs, and illustrate a situation that extends the behavior found in the three-dimensional model. As before, in the case of biased inputs, the eigenvalues remain separated, and the attracting direction degrades smoothly as the cross talk increases. Moreover, also similar to the three-dimensional case, order one loss of bias is not enough to trigger an eigenvalue-crossing bifurcation (for which bias loss of order $\geq 2$ is required), but may be enough to produce fapp bifurcations. Depending on the parameter values, both actual and fapp bifurcations can occur for arbitrarily small levels of cross talk (see Figure 6).

*3.2.1 Fully Biased Case.* Consider the covariance biases $\delta_j$'s to be distinct: $\delta_1 > \delta_2 > \cdots > \delta_n = 0$. The characteristic polynomial of **EC** can be

expressed as

$$\Delta(\lambda) = \det(\mathbf{EC} - \lambda\mathbf{I}) = \begin{vmatrix} X_1(\lambda) & f_2 & \cdots & f_n \\ f_1 & X_2(\lambda) & \cdots & f_n \\ \vdots & & \ddots & \vdots \\ f_1 & f_2 & \cdots & X_n(\lambda) \end{vmatrix},$$

where for all $j = \overline{1,n}$, we called $f_j = \epsilon(v + \delta_j - c) + c$ and $X_j(\lambda) = q(v + \delta_j - c) + c - \lambda$.

We consider $\lambda_j = (q - \epsilon)(v + \delta_j - c)$; clearly: $\lambda_1 > \lambda_2 > \cdots > \lambda_n$. In appendix B, we show how these values can be used to partition the real line and separate the roots of $\Delta$. This leads to:

**Proposition 1.** *In the biased case $\delta_1 > \delta_2 > \cdots > \delta_n$, the matrix **EC** has n real distinct eigenvalues $\xi_1 > \xi_2 > \cdots > \xi_n$, for any error $\epsilon \in (0, 1/n)$.*

We can define, as in the two- and three-dimensional applications, the critical error values, for which $f_j(\epsilon_j^*) = 0,\ \forall j \in \overline{1,n}$,

$$\epsilon_j^* = \frac{-c}{v + \delta_j - c}, \tag{3.3}$$

so that $0 < \epsilon_1^* < \epsilon_2^* < \cdots < \epsilon_n^*$ (since $\delta_1 > \delta_2 > \cdots > \delta_n$). Clearly, for all $j \in \overline{1,n}$, we have $f_j > 0$ iff $\epsilon > \epsilon_j^*$. As $\epsilon$ increases from 0 to $1/n$, it traverses the values $\epsilon = \epsilon_j^*$. When $\epsilon$ is in the intervals between two consecutive critical values $\epsilon_j^*$, each two consecutive roots of $\Delta$ are separated by at least one $\lambda_j$. When $\epsilon$ reaches each critical value $\epsilon_j$, the root $\xi_j$ crosses from one interval to another through the stage $\xi_j = \lambda_j$.

*3.2.2 Losing the Bias.* Suppose now that for $j \in \overline{1,n-1}$, $\delta_j = \delta_{j+1} + \zeta_j$, and allow some of the $\zeta_j \to 0$; in the limit, this results in a loss of bias in the covariance matrix $\mathbf{C}$ ($v + \delta_j = v + \delta_{j+1}$ for some index $j$). In consequence, $\lambda_j - \lambda_{j+1} \to 0$. It follows that in the limit of $\zeta = 0$ and $\xi = \lambda_1 = \lambda_2$, so that the maximal eigenvalue of **EC** preserves its multiplicity =1. This situation changes if we introduce an order two bias loss $\delta_1 = \delta_2 = \delta_3$ (i.e., if we make both $\zeta_1$ and $\zeta_2$ approach zero simultaneously). Then $\lambda_1 - \lambda_2 \to 0$ and $\lambda_2 - \lambda_3 \to 0$, so that the two leading roots collide into a double root $\lambda_3 = \xi_2 = \lambda_2 = \xi_1 = \lambda_1$. This justifies the following proposition:

**Proposition 2.** *Suppose $\epsilon < \epsilon_1^*$. An order k bias loss of the covariance matrix **C** of the type $\delta_1 = \cdots = \delta_k$ results in a leading eigenvalue of multiplicity $k - 1$ for the modified covariance matrix **EC**.*

## 4 Discussion

**4.1 Specific Comments on Our Model.** In this study, we considered a learning network based on the classical unsupervised learning model of Oja, extended to incorporate synaptic cross talk; we aimed to show how different input patterns can exacerbate or, on the contrary, efface the effects of cross talk on the asymptotic outcome of learning. We gave central attention to differences in second-order input statistics, studied how cross talk affected the outcome in each case, and observed that the effects can vary widely depending on these second-order statistics.

Efficient cross-talk-induced segregation happens in our model for a balance of positive and negative correlations. It could be argued that the model itself may artificially impose such a condition by being linear Hebbian, with multiplicative normalization. To address this critique, one may chose to study an equivalent model with subtractive normalization; that would, however, produce a different collection of issues, since subtractive normalization may be less biologically plausible. A better solution would be performing a similar cross-talk analysis on a extended nonlinear model with multiplicative normalization. The fact that certain nonlinear Hebbian models are reducible to linear Hebbian models (Miller, 1990; Elliott & Shadbolt, 2002) has led to a general belief that no Hebbian model, linear or nonlinear, can segregate positively correlated afferents under multiplicative normalization. Recently, Elliott and Shadbolt (2002) offered an explicit counterexample.

In this letter, we focus on a rule that is based only on second-order statistics, but the concept of unbiased distribution can be generalized for nonlinear Hebbian rules, sensitive to a lack of bias of higher order. The work of Elliott and others has shown that segregated outcomes are quite typical of nonlinear Hebbian rules with unbiased statistics (Elliott, 2003), and that cross talk can induce bifurcations in these cases (Elliott, 2012). We have suggested before the example of radially symmetric distributions considered by Lyu and Simoncelli (2009), with joint PDF equal density contour lines being nested hyperspheres with nongaussian spacings. We expect that in this setup, completely unbiased (spherical) input statistics would favor no particular direction in the weight space, so that the outcomes would be signed combinations of equal magnitude weights, nontrivially determined by the higher-order correlations. The presence of enough cross talk in the processing of such inputs may amount to suddenly switching the outcome between two such states.

**4.2 Some Biophysical Aspects of Oja's Rule.** Since our focus is on a biological realistic phenomenon (cross talk), it may seem odd to study a linear Hebbian model with multiplicative normalization, which may appear to be very formal and unbiological. But as argued in Rădulescu

and Adams (2013), Oja's rule is not as biophysically implausible as first appears.[3]

In our analysis of the Oja rule, we allowed both inputs and weights to be negative. However, if only positive patterns are allowed, the Hebbian part of the rule would always be positive (and correspond to LTP only), and the normalizing part of the rule would always be negative (and represent LTD only). It seems that in the brain, the negative and positive parts of signals are represented using different neurons, such that the two halves of the Oja rule would operate biologically with fixed and opposite polarities (LTP and LTD). However, the overall effect of the biological implementation would be the same as in our version of Oja's rule, which allows either polarity in both parts of the rule.

Experimental studies at single synapses suggest that reliable LTP may be implemented through repeated pairing of correctly timed pre- and postsynaptic spikes, which occur in an all-or-none manner (Petersen, Malenka, Nicoll, & Hopfield, 1998; Markram, Lübke, Frotscher, Roth, & Sakmann, 1997). Averaged over the many synapses comprising a connection, the overall outcome would be the multiplicative Hebbian rule. A simple mechanism for such batching would be if the coincidence-induced calcium increase at a synapse activated (by binding of Ca-Calmodulin) some fraction of its CaMKinase molecules, as follows: after each calcium pulse, Ca-Calmodulin would dissociate but leave some of the CaMKinase molecules phosphorylated; with successive pulses, enough would eventually be activated that the entire set of CaMKinases would fully autophosphorylate, triggering strengthening (Lisman, 1989, 1994; De Koninck & Schulman, 1998).

The normalizing (LTD) part of the Oja rule is, on the other hand, an elegant implementation of an approximate nonlocal normalization step that leads to a purely local online rule. Two obvious requirements of its biophysical implementation are the calculation of $y^2$ and the multiplication by $\boldsymbol{\omega}$. Recent work in neocortex (Sjöström, Turrigiano, & Nelson, 2003, 2004) suggests that LTD occurs in the following way: backpropagating spikes lead to a synapse-related calcium signal that triggers endocannabinoid release from the local dendrite, which then diffuses back to the presynaptic specialization, where it activates a G-protein-coupled endocannabinoid receptor. If there is near-simultaneous activation of presynaptic NMDARs by spike-release glutamate, transmitter release is depressed. This dismisses a previously favored theory (Nevian & Sakmann, 2006) that the level of the spine calcium achieved by LTP or LTD is a sign determinant of the strength change (Lisman, 1994; Shouval, Bear, & Cooper, 2002). This explanation of LTD seems well suited to meet the two biophysical requirements of the normalizing part of the Oja rule (and in this sense, the rule would be more

[3]Thanks to Paul Adams for the useful conversations and generous contributions to this section.

than a formal description). The calcium-dependent endocannabinoid enzyme triggered by calcium entering through voltage-dependent channels activated by backpropagating spikes would implement $y^2$, and the multiplication would be achieved by the requirement for simultaneous activation of the NMDAR. The dependence on $\boldsymbol{\omega}$ could be achieved in two ways: the endocannabinoid signal might be proportional to the postsynaptic strength of the synapse, or the extent of activation of the presynaptic NMDAR could depend on the amount of glutamate released, which would depend on the extent of the active zone, which is known in the long term to adjust to match the PSD area (and hence presumably the synaptic strength). Thus, the synaptic strength would slowly adjust, by a combination of matched but distinct post- and presynaptic adjustments, to reflect the arriving spikes, in the way required by the Oja rule (Rădulescu & Adams, 2013).

This background is necessary to discuss the important issue of the accuracy of the normalizing part of the Oja rule. Clearly if LTD is triggered presynaptically by a retrograde messenger, one must consider the possibility of extracellular LTD cross talk. If the LTD part of the rule is implemented as described above, errors in the diffusion of retrograde messenger to different synapses on the same neuron would not matter, although diffusion to synapses located on other neurons would matter. This problem is avoided because the readout of the weight by the requirement for presynaptic NMDAR activation by simultaneously released glutamate is itself dependent on the occurrence of appropriately timed presynaptic spikes. If instead (but presumably less biologically) the weight is read out postsynaptically and the combined signal $y^2\boldsymbol{\omega}$ is then retrogradely back propagated to the "correct" presynaptic structure, diffusion of the retrograde signal would cause normalization errors. In a nutshell, this could be modeled by adding a new error matrix $\mathbf{F}$ so the averaged rule would become

$$\frac{d\mathbf{w}}{dt} = \mathbf{F}(\mathbf{F}^{-1}\mathbf{E}\mathbf{C}\mathbf{w} - (\mathbf{w}^{\mathrm{T}}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{w})\mathbf{w}).$$

At first glance, it appears that the normalization errors could cancel out the Hebbian errors if $\mathbf{F}$ is appropriately matched to $\mathbf{E}$ (i.e., both "error-onto-all" with adjustment of quality). Such cancelation would correspond to a weight erroneously "forgetting" exactly what it erroneously learns for each pattern. The problem is that while the averaged values of $\mathbf{E}$ and $\mathbf{F}$ are simple and closely related, the instantaneous values $\mathcal{E}$ and $\mathcal{F}$ can be, at least locally, quite different, because one involves intracellular diffusion and the other extracellular diffusion. Furthermore, the stability of the algorithm will also be affected. The observed biological implementation appears to avoid these problems in an elegant way.

**4.3 General Comments.** In previous work (Rădulescu et al., 2009; Rădulescu & Adams, 2013), we have suggested an analogy and between the Oja

rule (even without cross talk) and Eigen's equation of DNA replication and mutation. Indeed, biologically, Darwinian evolution and neural learning are both adaptive processes, encoding inputs based on repeated interactions with the environment (Baum, 2004; Volkenshteĭ, 1991; Adami, 1998), and mathematically, both models describe normalized growth. However, we have argued that unlike Eigen's model, Oja's equation shows a bifurcation at a critical cross-talk value in only very narrow conditions. We have further suggested that while there may not be an actual "isomorphism" (Fernando & Szathmáry, 2009; Fernando, Goldstein, & Szathmáry, 2010) (or other formal mathematical equivalence) between the two models in all parameter ranges, their analogy resides in their common need for accuracy in the adaptation process. While biology is well known for instances in which it affords to be inaccurate, polynucleotide copying requires superaccuracy, and neural learning also seems to require superaccurate synaptic updates (Elliott, 2012; Adams & Cox, 2012).

Indeed, successful and effective reproduction requires copying the entire genome, with an appropriately small error per base rate. The known "proofreading" operation of this replication process is essential in lowering the copying error rate to acceptable levels. The proofreading mechanism copies bases twice, and replication is allowed only when coincidence of the two results is detected. Since proofreading seems to be in general an effective strategy for overcoming physical limitations, it has been proposed that the same operation is being performed in the neocortex in order to ensure the synaptic specificity necessary for effective learning. The mechanism underlying "neural proofreading," as proposed by Adams and Cox (2012), assigns to each thalamocortical connection (responsible for the tuned responses of cortical neurons) a corticothalamic "proofreading neuron," which receives and detects "coincidence" between the input and output spikes arriving at that connection and then sends a double signal to both sides of the connection, confirming the validity of the synaptically detected coincidence. Other aspects, consequences, challenges, and limitations of this elaborate neocortical proofreading circuitry are further investigated in Adams and Cox (2012).

## 5 Conclusion

A lot of work has been aimed recently toward finding key biological factors that may explain the network architectures and computational algorithms that the brain develops to perform learning. The fact that the activity-dependent processes that lead to synaptic strength adjustments cannot be completely synapse specific constitutes a central problem for biological learning. While this model considers only a very simple setup, it helps us better illustrate an important idea, which we have formulated previously (Rădulescu et al., 2009; Rădulescu & Adams, 2013): a performant synaptic updating algorithm may not suffice for accurate learning, and the process

may fail (partly or completely, depending on the input pattern to be learned) even when faced with only infinitesimal amounts of synaptic cross talk. It appears therefore increasingly possible that high-level (e.g., neocortical) learning may require not only performant learning algorithms but also special apparatus for enhancing specificity (Adams & Cox, 2006). The brain may thus have to dedicate comparable effort to developing proofreading for its plasticity machinery (all the more necessary in the face of inaccuracy that seems to not merely degrade learning but rather is able to prevent it altogether). Our model does not exclude either possibility but suggests that learning problems (and perhaps, more generally, all problems of survival or reproduction) are so diverse that no single algorithm can solve them all, so that no universal or canonical cortical circuit should be expected.



## Appendix A: Stability of Equilibria in the Oja Model

The symmetric, positive definite matrix $\mathbf{C} \in \mathcal{M}_n(\mathbb{R})$ defines a dot product in $\mathbb{R}^n$ as

$$\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbf{C}} = \mathbf{v}^T \mathbf{C} \mathbf{w}.$$

Although both $\mathbf{C}$ and $\mathbf{E}$ are symmetric, the product $\mathbf{EC}$ is not symmetric in the Euclidean metric. However, in a new metric defined by the dot product $\langle \cdot, \cdot \rangle_{\mathbf{C}}$, $\mathbf{EC}$ is symmetric. Indeed, for any pair of vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, we have

$$\langle \mathbf{ECu}, \mathbf{v} \rangle_{\mathbf{C}} = (\mathbf{ECu})^t \mathbf{Cv} = \mathbf{u}^t \mathbf{C}^t \mathbf{E}^t \mathbf{Cv} = \mathbf{u}^t \mathbf{CECv} = \langle \mathbf{u}, \mathbf{ECv} \rangle_{\mathbf{C}}.$$

In consequence, $\mathbf{EC}$ has a basis of eigenvectors, orthogonal with respect to the dot product $\langle \cdot, \cdot \rangle_{\mathbf{C}}$.

The following theorem, describing the equilibria of system 2.5, is immediate.

**Theorem 1.** *An equilibrium for the system is any vector* $\boldsymbol{w} = (w_1 \dots w_n)^T$ *such that* $\boldsymbol{ECw} = (\boldsymbol{w}^T \boldsymbol{Cw})\boldsymbol{w}$*, that is, an eigenvector of* $\boldsymbol{EC}$ *(with corresponding eigenvalue* $\lambda_{\boldsymbol{w}}$*), normalized with regard to the norm* $\| \cdot \|_{\boldsymbol{C}}^2 = \langle \cdot, \cdot \rangle_{\boldsymbol{C}}$*, so that* $\|\boldsymbol{w}\|_{\boldsymbol{C}}^2 = \lambda_{\boldsymbol{w}}$*:*

$$\boldsymbol{ECw} = \lambda_{\boldsymbol{w}} \boldsymbol{w}, \quad \|\boldsymbol{w}\|_{\boldsymbol{C}}^2 = \lambda_{\boldsymbol{w}}.$$

*If we additionally assume (generically) that* $\boldsymbol{EC}$ *has a strictly positive maximal eigenvalue of multiplicity one, then the corresponding eigendirection is orthogonal in* $\langle \cdot, \cdot \rangle_{\boldsymbol{C}}$ *to all other eigenvectors of* $\boldsymbol{EC}$*.*

Take $\mathbf{w}$ to be an equilibrium of system (2.5)—an eigenvector of $\mathbf{EC}$, with eigenvalue $\lambda_{\mathbf{w}} = (\mathbf{w}^T\mathbf{C}\mathbf{w}) > 0$. To establish stability, we calculate the Jacobian matrix at $\mathbf{w}$ to be

$$Df_{\mathbf{w}}^{\mathbf{E}} = \gamma\left[\mathbf{EC} - 2\mathbf{w}(\mathbf{C}\mathbf{w})^T - (\mathbf{w}^T\mathbf{C}\mathbf{w})\mathbf{I}\right].$$

Then we get the following:

**Theorem 2.** *Suppose* $\boldsymbol{EC}$ *has multiplicity one largest eigenvalue. An equilibrium* $\boldsymbol{w}$ *(i.e., by theorem 1, an eigenvector of* $\boldsymbol{EC}$ *with eigenvalue* $\lambda_{\boldsymbol{w}}$*, normalized so that* $\|\boldsymbol{w}\|_C^2 = \lambda_{\boldsymbol{w}}$*) is a local hyperbolic attractor for equation 2.5 iff it is an eigenvector corresponding to the maximal eigenvalue of* $\boldsymbol{EC}$*.*

**Proof.** Fix an eigenvector $\mathbf{w}$ of $\mathbf{EC}$, with $\mathbf{ECw} = \lambda_{\mathbf{w}}\mathbf{w}$. Then:

$$Df_{\mathbf{w}}^{\mathbf{E}}\mathbf{w} = -2\gamma\lambda_{\mathbf{w}}\mathbf{w}.$$

Recall that the vector $\mathbf{w}$ can be completed to a basis $\mathcal{B}$ of eigenvectors, orthogonal with respect to the dot product $\langle\cdot,\cdot\rangle_{\mathbf{C}}$. Let $\mathbf{v} \in \mathcal{B}$, $\mathbf{v} \neq \mathbf{w}$, be any other arbitrary vector in this basis, so that $\mathbf{ECv} = \lambda_{\mathbf{v}}\mathbf{v}$, and $\langle\mathbf{w},\mathbf{v}\rangle_{\mathbf{C}} = \mathbf{w}^t\mathbf{C}\mathbf{v} = 0$. We calculate

$$Df_{\mathbf{w}}^{\mathbf{E}}\mathbf{v} = -\gamma[\lambda_{\mathbf{w}} - \lambda_{\mathbf{v}}]\mathbf{v}.$$

So $\mathcal{B}$ is also a basis of eigenvectors for $Df_{\mathbf{w}}^{\mathbf{E}}$. The corresponding eigenvalues are $-2\gamma\lambda_{\mathbf{w}}$ (for eigenvector $\mathbf{w}$) and $-\gamma[\lambda_{\mathbf{w}} - \lambda_{\mathbf{v}}]$ (for any other eigenvector $\mathbf{v} \in \mathcal{B}$, $\mathbf{v} \neq \mathbf{w}$). An equivalent condition for $\mathbf{w}$ to be a hyperbolic attractor for system 2.5 is that all the eigenvalues of $Df_{\mathbf{w}}^{\mathbf{E}}$ are $< 0$. Since $\gamma, \lambda_{\mathbf{w}} > 0$, this condition is equivalent to having $-\gamma(\lambda_{\mathbf{w}} - \lambda_{\mathbf{v}}) < 0$ for all $\mathbf{v} \in \mathcal{B}$, $\mathbf{v} \neq \mathbf{w}$. In conclusion, an equilibrium $\mathbf{w}$ is a hyperbolic attractor if and only if $\lambda_{\mathbf{w}} > \lambda_{\mathbf{v}}$ for all $\mathbf{v} \neq \mathbf{w}$ (i.e., $\lambda_{\mathbf{w}}$ is the maximal eigenvalue—in other words, if $\mathbf{w}$ is in the direction of the principal eigenvector of $\mathbf{EC}$).

Such attractors always exist provided that the condition of theorem 2 is met (i.e., $\mathbf{EC}$ has a maximal eigenvalue of multiplicity one). Then the network learns, depending on its initial state, one of the two stable equilibria, which are the two (opposite) maximal eigenvectors of the modified input distribution, normalized so that $\|\mathbf{w}\|_{\mathbf{C}}^2 = \lambda_{\mathbf{w}}$. Next, we aim to show that these two attractors are the system's only hyperbolic attractors.

**Theorem 3.** *Suppose the the modified covariance matrix* $\boldsymbol{EC}$ *has a unique maximal eigenvalue* $\lambda_1$*. Then the two eigenvectors* $\pm\boldsymbol{w}_{\boldsymbol{EC}}$ *corresponding to* $\lambda_1$*, normalized such that* $\|\boldsymbol{w}\|_{\mathbf{C}}^2 = \lambda_1$*, are the only two attractors of the system. More precisely, the phase space is divided into two basins of attraction, of* $\boldsymbol{w}_{\boldsymbol{EC}}$ *and* $-\boldsymbol{w}_{\boldsymbol{EC}}$*, respectively, separated by the subspace* $\langle\boldsymbol{w}, \boldsymbol{w}_{\boldsymbol{EC}}\rangle = 0$*.*

**Proof.** We make the change of variable $\mathbf{u} = \sqrt{\mathbf{C}}\mathbf{w}$. The system then becomes

$$\dot{\mathbf{u}} = \mathbf{A}\mathbf{u} - (\mathbf{u}^t\mathbf{u})\mathbf{u}, \tag{A.1}$$

where $\mathbf{A} = \sqrt{\mathbf{C}}\mathbf{E}\sqrt{\mathbf{C}}$ is a symmetric matrix, having the same eigenvalues as **EC**. More precisely, **w** is an eigenvector of **EC** with eigenvalue $\mu$ iff $\sqrt{\mathbf{C}}\mathbf{v}$ is an eigenvector of **A** with eigenvalue $\mu$; hence, any two distinct eigenvectors of **A** are orthogonal in the regular Euclidean dot product.

Consider **v** to be the leading eigenvector of **A**, and let $\mathbf{u} = \mathbf{u}(t)$ be a trajectory of the system A.1. We want to observe the evolution in time of the angle between the variable vector **u** and the fixed vector **v**, measured as

$$\cos\theta = \frac{\langle \mathbf{v}, \mathbf{u}\rangle}{\|\mathbf{v}\| \cdot \|\mathbf{u}\|}.$$

We differentiate and obtain

$$-\|\mathbf{v}\|\sin(\theta)\dot{\theta} = \frac{(\mathbf{v}^t\dot{\mathbf{u}})\|\mathbf{u}\|^2 - (\mathbf{v}^t\mathbf{u})(\mathbf{u}^t\mathbf{u})}{\|\mathbf{u}\|^3}.$$

The numerator of this expression is

$$h(\mathbf{u}) = (\mathbf{u}^t\mathbf{u})(\mathbf{v}^t\mathbf{A}\mathbf{u}) - (\mathbf{v}^t\mathbf{u})(\mathbf{u}^t\mathbf{A}\mathbf{u}). \tag{A.2}$$

We are interested in the sign of $h(\mathbf{u})$. To make our computations simpler, we can diagonalize **A** in a basis of orthogonal eigenvectors $\mathbf{A} = \mathbf{P}^t\mathbf{D}\mathbf{P}$, where **D** is the diagonal matrix of eigenvalues and **P** is an orthogonal matrix whose columns are the eigenvectors. Then

$$h(\mathbf{u}) = (\mathbf{z}^t\mathbf{z})(\mathbf{y}^t\mathbf{D}\mathbf{z}) - (\mathbf{y}^t\mathbf{z})(\mathbf{z}^t\mathbf{D}\mathbf{z}),$$

where $\mathbf{y} = \mathbf{P}\mathbf{v}$ and $\mathbf{z} = \mathbf{P}\mathbf{u}$, so that $\mathbf{D}\mathbf{y} = \mathbf{D}\mathbf{P}\mathbf{v} = \lambda_1\mathbf{y}$ (where $\lambda_1$ is the largest eigenvalue of **EC**, assumed to have multiplicity one). Then

$$h(\mathbf{u}) = (\mathbf{y}^t\mathbf{z})\sum_{j=2}^{n}(\lambda_1 - \lambda_j)z_j^2.$$

Hence, if $\mathbf{y}^t\mathbf{z} > 0$, then $h(\mathbf{u}) > 0$. In other words, if $\mathbf{v}^t\mathbf{u} > 0$, $-\|\mathbf{v}\|\sin(\theta)\dot{\theta} > 0$—hence, $\dot{\theta} < 0$. For our original system, this means that any trajectory starting at a **w** with $\langle \mathbf{w}, \mathbf{w}_{\mathbf{EC}}\rangle > 0$ converges in time toward the principal eigenvector $\mathbf{w}_{\mathbf{EC}}$ of the matrix **EC**.

## Appendix B: A Direct Computation for Unbiased Inputs

**Theorem 4.** *For order two input bias $\delta_1 = \delta_2 = 0$, the dynamic behavior of the system is classified by the classification of the input covariance sign:* $(+,+,+)$, $(+,+,-)$, $(+,-,-)$ *and* $(-,-,-)$.

**Proof.** For unbiased inputs (i.e., order two input bias), all classes can be generated from three symbolic structures:

$$\mathbf{C}_1 = \begin{bmatrix} v & c & c \\ c & v & c \\ c & c & v \end{bmatrix}, \quad \mathbf{C}_2 = \begin{bmatrix} v & -c & c \\ -c & v & c \\ c & c & v \end{bmatrix} \quad \text{and} \quad \mathbf{C}_3 = \begin{bmatrix} v & c & -c \\ c & v & c \\ -c & c & v \end{bmatrix}.$$

Class $(+,+,+)$ represents Structure $\mathbf{C}_1$ with $c > 0$, and class $(-,-,-)$ represents Structure $\mathbf{C}_1$ with $c < 0$. Class $(+,+,-)$ can be obtained from Structures $\mathbf{C}_2$ and $\mathbf{C}_3$ with $c > 0$, while class $(+,-,-)$ can be obtained from Structures $\mathbf{C}_2$ and $\mathbf{C}_3$ with $c < 0$.

Computing directly the spectrum for $\mathbf{C}_1$, we get one simple error-independent eigenvalue $\xi_1 = v + 2c$ (whose eigenvector is also error independent) and one double eigenvalue $\xi_2 = (1 - 3e)(v - c)$. If $c > 0$ (class $(+,+,+)$), $\xi_1$ always dominates (see Figure 1E). If $c < 0$ (class $(-,-,-)$), the double eigenvalue $\xi_2 = (1 - 3e)(v - c)$ takes over for error smaller than the critical value $\epsilon < \frac{-c}{v-c}$ (see Figure 1H).

Also by direct computation, one notices that $\mathbf{C}_1$ and $\mathbf{C}_2$ have the same spectral decomposition. One eigenvalue is given by $\xi_1 = (1 - 3e)(v + c)$, while the other two, $\xi_2 \geq \xi_3$, are the roots of the quadratic polynomial $P(X) = X^2 + (c - 2v - 5ec + 3ev)X + (6ec^2 - cv - 3ev^2 - 2c^2 + v^2 + 3ecv)$. It is easy to see that $P(\xi_1) = -8ec(1 - 3e)(v + c)$. If $c > 0$ (class $(+,+,-)$), then $P(\xi_1) < 0$; hence, $\xi_2 \geq \xi_1$, with equality at $\epsilon = 0$, and $\xi_1 \geq \xi_3$, with equality when $\epsilon = 1/3$ (see Figure 1G). If $c < 0$ (class $(+,-,-)$), then $P(\xi_1) > 0$ and $\xi_1 < (\xi_2 + \xi_3)/2$, hence $\xi_1 \leq \xi_2 < \xi_3$, with equality when $\epsilon = 0$ and $\epsilon = 1/3$ (see Figure 1F).

## Appendix C: A Numerical Extension to Weakly Correlated Inputs

In this section, we loosen the assumption of weakly mutually correlated three-dimensional inputs (i.e., of a diagonally dominant input covariance matrix $\mathbf{C}$) and investigate numerically the behavior of the system under a wider class of input schemes, corresponding to larger ranges for the parameters $c$, $\delta_1$, $\delta_2$, and $q$. We will be studying sensitivity to these parameters in all four combinatorial input classes: $(+,+,+)$, $(+,+,-)$, $(+,-,-)$, and $(-,-,-)$.

Without losing generality, we will be normalizing our matrix $\mathbf{C}$ so that $v = 1$, which will be considered fixed throughout this analysis. The range

for the mutual covariance $c$ will be extended in each case to the largest interval for which **C** remains positive definite. While the parameter $q$ was restricted before to live in the interval $[1/3, 1]$ (representing the constraint for the quality to be larger than the error), in the following illustrations, we will allow $q$ to change within $[0, 1]$. This allows us to better understand how bifurcations and fapp bifurcations appear in the more plausible biological interval $[1/3, 1]$ and also reveals interesting behavior that occurs in the poor-quality range for strongly negatively correlated inputs.

As before, in order to quantify and illustrate the effects of cross talk (error) on the outcome of learning, we use the cosine of the angle $\theta$ between the system's attractors with and without cross talk (i.e., between the directions of the leading eigenvectors of the matrices **EC** and **C**, respectively). Generally the behavior of the system with respect to error, as observed in section 3.1, extends naturally to the range of high mutual correlations within the input distribution. The learning outcome depreciates when gradually increasing the error (decreasing $q$). As discussed in section 3.1, this decay is smooth for some types of input distributions, but for others, it exhibits jump discontinuities (corresponding to bifurcations in the dynamics) or just smooth but very sharp drops (fapp bifurcations) with very steep but bounded slope at the inflection point. We have discussed, in the context of small mutual correlations $c$ (**C** had been assumed to be diagonally dominant, i.e., with $|c| < \frac{v}{n-1}$), that both fapp and actual bifurcations can appear at arbitrarily small cross-talk values ($q$ arbitrarily close to 1). While these effects still occur for higher values of $|c|$, the presence of highly negatively correlated inputs introduces an interesting new effect that is not accounted for by the analysis in the main text.

Figure 7 shows a few instances of bifurcations and fapp bifurcations for one negative pairwise correlation and the slight differences between its two possible off-diagonal positions (next to the diagonal or in the corner of the matrix **C**). When increasing $|c|$ past the value $\frac{v}{n-1}$, while keeping it within the range that preserves positive definiteness of **C**, the behavior of $\cos(\theta)$ with respect to $q$ remains qualitatively the same, whether it is a smooth depreciation of the output when decreasing $q$ (for biased inputs) or a sharp drop (some unbiased inputs trigger bifurcations; see the pink curve in Figure 7A), with only the position and shape of the transitions being altered in the process.

When increasing the number of negative pairwise correlations, the results change qualitatively, in particular for very high levels of cross talk, as shown in Figures 8 and 9. Typically for $(+, -, -)$, there is a fapp bifurcation at low values of cross talk, which in fact can shift to arbitrarily small levels of cross talk depending on the bias parameters. When increasing $|c|$ past $\frac{v}{n-1}$ in class $(+, -, -)$, a bifurcation appears in the low $q$ range, so that after having passed the inflection point (fapp) in its degradation from the correct attractor, the system suddenly reverses, for very large levels of cross talk, to computing the principal direction of **C** more accurately (the cosine is

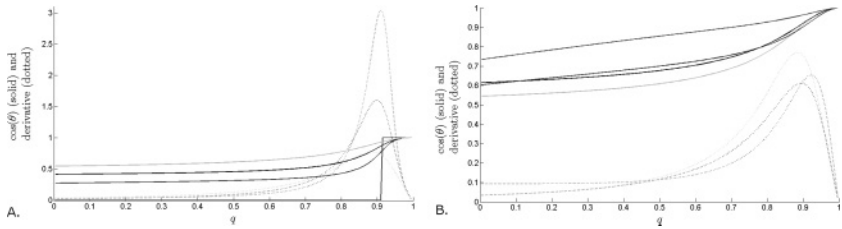



Figure 7: Processing three-dimensional inputs with one negative pairwise correlation. The two panels represent the two distinct possibilities for off-diagonal positions for the negative entry within **C**. In both panels, the solid lines represent $\cos\theta$ represented as a function of $q$. We considered biased inputs: $\delta_1 = 0.4$, $\delta_2 = 0.2$, for $c = 0$ (green), $c = 0.1$ (blue) and $c = 0.55$ (red), as well as an instance of input bias loss of order one: $\delta_1 = \delta_2 = 0.4$, for $c = 0.55$ (pink). The dotted lines measure the sensitivity of the cosine to changes in $q$ by illustrating its derivative with respect to $q$ and are convenient for locating fapp bifurcations. (Please refer to online supplement for color version of this figure.)

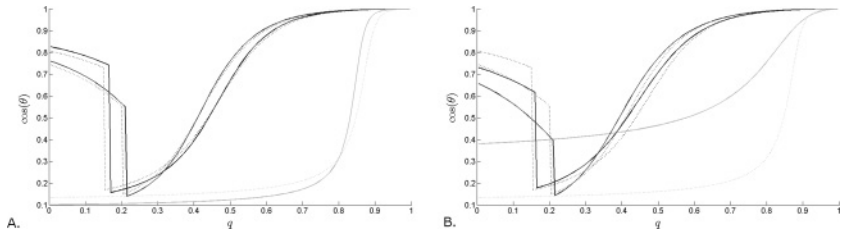



Figure 8: Processing three-dimensional inputs with two negative pairwise correlations. The two panels represent the two distinct possibilities for off-diagonal configurations of the negative entries within **C**. In both panels, the solid lines represent $\cos\theta$ for biased inputs corresponding to $\delta_1 = 0.4$ and $\delta_2 = 0.2$; the dotted lines represent $\cos\theta$ for bias loss of order one, corresponding to $\delta_1 = \delta_2 = 0.4$. The color coding is $c = 0.1$ (green), $c = 0.8$ (blue), and $c = 1$ (red). (Please refer to online supplement for color version of this figure.)

close to 1 for small values of $q$). While this jump discontinuity also exists in class $(+, +, -)$, it does not appear in Figure 7 because it occurs for $q < 0$. For class $(+, -, -)$, this high cross-talk bifurcation is brought within the interval $q \in [0, 1]$ by the increase in the number of negative correlations, together with increasing the pairwise-correlation strength.

The effect is exacerbated when increasing the number of negative pairwise correlations further and observing class $(-, -, -)$. The high cross-talk bifurcations shown in Figure 9 are more pronounced and occur for higher values of $q$ (i.e., more biologically plausible levels of cross talk).

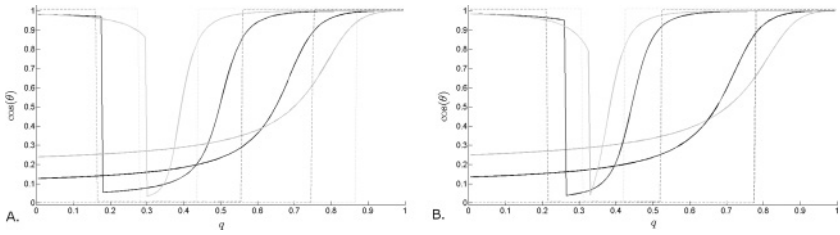



Figure 9: Processing three-dimensional inputs with all negative pairwise correlations. (A) The solid lines represent $\cos\theta$ for biased inputs corresponding to $\delta_1 = 0.4$ and $\delta_2 = 0.2$; the dotted lines represent $\cos\theta$ for bias loss of order one, corresponding to $\delta_1 = \delta_2 = 0.4$. The color coding is $c = 0.1$ (green), $c = 0.2$ (blue), $c = 0.4$ (red), and $c = 0.55$ (cyan). (B) The solid lines represent $\cos\theta$ for biased inputs corresponding to $\delta_1 = 0.6$ and $\delta_2 = 0.4$; the dotted lines represent $\cos\theta$ for bias loss of order one, corresponding to $\delta_1 = \delta_2 = 0.6$. The color coding is $c = 0.1$ (green), $c = 0.2$ (blue), $c = 0.5$ (red), and $c = 0.65$ (cyan). (Please refer to online supplement for color version of this figure.)

## Appendix D: An Extension to Higher Dimensions

We want a concise description of the modified input matrix **EC**. To begin, we can express the matrices **E** and **C** individually as $\mathbf{E} = \epsilon\mathbf{M} + (q - \epsilon)\mathbf{I}$ and $\mathbf{C} = c\mathbf{M} + (v - c)\mathbf{I} + \Sigma\delta_j\mathbf{A}_j$, where **I** is the $n \times n$ identity matrix, **M** is the $n \times n$ matrix with uniform unit entries, and, for any $j = \overline{1, n}$, $\mathbf{A}_j$ is the matrix with zero entries except $A_j(j, j) = 1$. Note, for future computations, that $\mathbf{M}^2 = n\mathbf{M}$ and that $\mathbf{MA}_j$ is the matrix with the only nonzero entries being ones along the $j$th column. Unless otherwise specified, the summations are for $j = \overline{1, n}$. The product **EC** will then be

$$\mathbf{EC} = [\epsilon(v - c) + c(q - \epsilon) + \epsilon cn]\mathbf{M} + (q - \epsilon)(v - c)\mathbf{I} + \epsilon \sum \delta_j \mathbf{MA}_j + (q - \epsilon) \sum \delta_j \mathbf{A}_j.$$

In matrix form, this translates as

$$\mathbf{EC} = \begin{bmatrix} X_1(\lambda) & f_2 & \cdots & f_n \\ f_1 & X_2(\lambda) & \cdots & f_n \\ \vdots & & \ddots & \vdots \\ f_1 & f_2 & \cdots & X_n(\lambda) \end{bmatrix},$$

where $\forall j = \overline{1, n}$, we called $f_j = \epsilon(v + \delta_j - c) + c$ and $X_j(\lambda) = q(v + \delta_j - c) + c - \lambda$.

**D.1** Fully Biased Case. We first consider the covariance biases $\delta_j$'s to be distinct: $\delta_1 > \delta_2 > \cdots > \delta_{n-1} > \delta_n = 0$. We will prove that the polynomial $\Delta$ has $n$ real roots $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_n$, and we will find approximating bounds for their positions on the real line.

We remark first that the end behavior of $\Delta(\lambda)$ is given by $\lim_{\lambda \to -\infty} \Delta(\lambda) = \infty$ and $\lim_{\lambda \to +\infty} \Delta(\lambda) = (-1)^n \infty$. Consider $\lambda_j = (q - \epsilon)(v + \delta_j - c)$; clearly: $\lambda_1 > \lambda_2 > \cdots > \lambda_n$. We will use these values to partition the real line and separate the roots of $\Delta$. To begin, we calculate, for all $i$, $j = \overline{1, n}$,

$$X_i(\lambda_j) = f_i + (q - \epsilon)(\delta_i - \delta_j). \tag{D.1}$$

In particular, $X_j(\lambda_j) = f_j$, $\forall j = \overline{1, n}$. By raw and column manipulations, it can be shown that, $\forall j = \overline{1, n}$,

$$\Delta(\lambda_j) = f_j (q - \epsilon)^{n-1} \prod_{i \neq j} (\delta_i - \delta_j). \tag{D.2}$$

In consequence, $\operatorname{sign}(\Delta(\lambda_j)) = \operatorname{sign}(f_j)(-1)^{n-j}$.

Recall that $f_j = \epsilon(v + \delta_j - c) + c$; hence, $f_1 > f_2 > \cdots > f_n$. To continue our discussion and establish the signs of $\Delta$ at all partition points $\lambda_j$, we need to establish the index $j$ for which the values $f_j$ switch sign.

For each $j \in \overline{1, n}$, consider the "critical" error values, for which $f_j(\epsilon_j^*) = 0$, $\forall j \in \overline{1, n}$:

$$\epsilon_j^* = \frac{-c}{v + \delta_j - c}, \tag{D.3}$$

so that $0 < \epsilon_1^* < \epsilon_2^* < \cdots < \epsilon_n^*$ (since $\delta_1 > \delta_2 > \cdots > \delta_n$). Clearly, for all $j \in \overline{1, n}$, we have $f_j > 0$ iff $\epsilon > \epsilon_j^*$.

**Remark.** The diagonal dominance assumption $v > (n-1)|c|$ allows us to study all cases that may appear, since it guarantees $\epsilon_j^* < 1/n, \forall j \in \overline{1, n}$. This ensures a complete discussion, since then $\epsilon \in [0, 1/n]$ is allowed to reach and cross over all the critical values $\epsilon_j^*$, creating a possible swap in the order of the eigenvalues of **EC**, as we will show later. The proof for the other cases will be omitted, since it is just a simplification of the argument. In fact, the only crossover of true interest to us is $\epsilon = \epsilon_1^*$, where the eigenvalue swap involves the two largest eigenvalues and thus affects the position of the system's attracting equilibria, corresponding to the normalized eigenvectors of the maximal eigenvalue. The other critical values $\epsilon = \epsilon_j^*$, for $j \geq 2$, affect only the stable and unstable spaces of the saddle-equilibria. In this

light, the condition on the entries of the covariance matrix can be loosened to $v > (n-1)|c| - \delta_1$.

We distinguish the following cases:

1. For $0 \le \epsilon < \epsilon_1^*$. This implies $f_j < 0,\ \forall j \in \overline{1,n}$. Then

$$\operatorname{sign}(\Delta(\lambda_j)) = \operatorname{sign}(f_j)(-1)^{n-j} = (-1)(-1)^{n-j} = (-1)^{n-j+1} \tag{D.4}$$

From equations D.1, D.2, and D.4, we obtain the following sign table:

| $\lambda$ | $-\infty$ | $\lambda_n$ | $\lambda_{n-1}$ | $\dots$ | $\lambda_2$ | $\lambda_1$ | $+\infty$ |
|---|---|---|---|---|---|---|---|
| $\operatorname{sign}(\Delta(\lambda))$ | $(+)$ | $(-)$ | $(+)$ | $\dots$ | $(-1)^{n-1}$ | $(-1)^n$ | $(-1)^n$ |

From the intermediate value theorem and the fundamental theorem of algebra, it follows that the polynomial $\Delta(\lambda)$ has $n$ real roots $\xi_1 > \xi_2 > \cdots > \xi_n$, such that

$$-\infty < \xi_n < \lambda_n < \xi_{n-1} < \lambda_{n-1} < \cdots < \lambda_2 < \xi_1 < \lambda_1 < \infty. \tag{D.5}$$

2. For $\epsilon_p^* < \epsilon < \epsilon_{p+1}^*$. Then $f_1, \dots, f_p > 0$ and $f_{p+1}, \dots, f_n < 0$. Similarly as in Case 1, we have

| $\lambda$ | $-\infty$ | $\lambda_n$ | $\lambda_{n-1}$ | $\dots$ | $\lambda_{p+1}$ | $\lambda_p$ | $\dots$ | $\lambda_1$ | $+\infty$ |
|---|---|---|---|---|---|---|---|---|---|
| $\operatorname{sign}(\Delta(\lambda))$ | $(+)$ | $(-)$ | $(+)$ | $\dots$ | $(-1)^{n-p}$ | $(-1)^{n-p}$ | $\dots$ | $(-1)^{n-1}$ | $(-1)^n$ |

hence, the polynomial $\Delta(\lambda))$ has roots $\xi_1 > \xi_2 > \cdots > \xi_n$, such that

$$\begin{aligned} -\infty < \xi_n < \lambda_n < \xi_{n-1} < \lambda_{n-1} < \cdots < \xi_{p+1} \\ < \lambda_{p+1} < \lambda_p < \xi_p < \cdots < \lambda_1 < \xi_1 < \infty \end{aligned} \tag{D.6}$$

3. For $\epsilon_n^* < \epsilon < 1/n$. Then $f_1, \dots, f_n > 0$, and we have

| $\lambda$ | $-\infty$ | $\lambda_n$ | $\lambda_{n-1}$ | $\dots$ | $\lambda_2$ | $\lambda_1$ | $+\infty$ |
|---|---|---|---|---|---|---|---|
| $\operatorname{sign}(\Delta(\lambda))$ | $(+)$ | $(+)$ | $(-)$ | $\dots$ | $(-1)^n$ | $(-1)^{n-1}$ | $(-1)^n$ |

and the polynomial $\Delta(\lambda)$ has roots $\xi_1 > \xi_2 > \cdots > \xi_n$, such that

$$-\infty < \lambda_n < \xi_n < \lambda_{n-1} < \xi_{n-1} < \cdots < \lambda_1 < \xi_1 < \infty. \tag{D.7}$$

In particular, we have proved the following proposition in the main text:

**Proposition 1.** *In the biased case $\delta_1 > \delta_2 > \cdots > \delta_n$, the matrix **EC** has n real distinct eigenvalues $\xi_1 > \xi_2 > \cdots > \xi_n$, for any error $\epsilon \in (0, 1/n)$.*

**D.2** Losing the Bias. Suppose now that for $j \in \overline{1,n}$, $\delta_j = \delta_{j+1} + \zeta_j$, and allow some of the $\zeta_j \to 0$; in the limit, this results in a loss of bias in the covariance matrix $\mathbf{C}$ ($v + \delta_j = v + \delta_{j+1}$ for some index $j$). In consequence, $\lambda_j - \lambda_{j+1} \to 0$.

Let us study the changes of the maximal root $\xi_1$ as $\zeta_1 \to 0$ (i.e., we eliminate the bias between the two most correlated components of the matrix $\mathbf{C}$). Suppose $\epsilon \in [0, \epsilon_1^*]$. This calculation can be extended to the other intervals for $\epsilon$; however, we will discuss here only the case $\epsilon \in [0, \epsilon_1^*]$, since it is the only one that relates directly to the position and multiplicity of the leading root of $\Delta$. It also agrees with our goal to study the behavior of the system for small enough errors. According to equation D.5, we have

$$-\infty < \xi_n < \lambda_n < \xi_{n-1} < \lambda_{n-1} < \cdots < \lambda_2 < \xi_1 < \lambda_1 < \infty.$$

Since $\lambda_1 \to \lambda_2$, it follows that in the limit of $\zeta = 0$ and $\xi = \lambda_1 = \lambda_2$, so that the maximal eigenvalue of $\mathbf{EC}$ preserves its multiplicity $= 1$. This situation changes if we introduce an order two bias loss $\delta_1 = \delta_2 = \delta_3$ (i.e., if we make both $\zeta_1$ and $\zeta_2$ approach zero simultaneously). Then $\lambda_1 - \lambda_2 \to 0$ and $\lambda_2 - \lambda_3 \to 0$, so that the two leading roots collide into a double root $\lambda_3 = \xi_2 = \lambda_2 = \xi_1 = \lambda_1$. This justifies the following proposition:

**Proposition 2.** *Suppose $\epsilon < \epsilon_1^*$. An order $k$ bias loss of the covariance matrix $\boldsymbol{C}$ of the type $\delta_1 = \cdots = \delta_k$ results in a leading eigenvalue of multiplicity $k - 1$ for the modified covariance matrix $\boldsymbol{EC}$.*

This proposition can be generalized to encompass bias loss anywhere in the inputs and any interval for the error $\epsilon$. Below, we give a more general statement, which follows by repeating the argument for the case we already analyzed but could also be proved more directly.

**Theorem 5.** *Suppose that the matrix $\boldsymbol{C}$ is allowed to exhibit bias loss in all possible ways, so that it can be written in block form as equation 3.2, where there exist $k_1, k_2, \ldots, k_N \in \overline{1,n}$, with $\sum_{j=1}^{N} k_j = n$ and such that*

$$\begin{aligned}
\delta_1 &= \cdots = \delta_{k_1} = \nu_1 \\
\delta_{k_1+1} &= \cdots = \delta_{k_2} = \nu_2 \\
&\vdots \\
\delta_{k_{N-1}+1} &= \cdots = \delta_{k_N} = \nu_N
\end{aligned}$$

*with*

$$\nu_1 > \nu_2 > \cdots > \nu_N.$$

*Then the characteristic polynomial $\Delta$ of **EC** has all real eigenvalues. More precisely, these eigenvalues are $\lambda_j = (q - \epsilon)(v + \delta_j - c)$ with multiplicity $k_j - 1$, for all $j \in \overline{1, N}$, and N additional eigenvalues $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_N$.*

**Remark.** The order of these eigenvalues, depending on the the error value $\epsilon$ with respect to the critical error values $\nu_j^* = \frac{-c}{v+\nu_j-c}$, is the same as described in cases 1 to 3.

## References


Adami, C. (1998). *Introduction to artificial life*. New York: Springer-Verlag.

Adams, P., & Cox, K. (2006). A neurobiological perspective on building intelligent devices. *Neuromorphic Eng.*, *3*, 2–8.

Adams, P., & Cox, K. (2012). From life to mind: Two prozaic miracles. *Integral Biomathics*, *67*, 2.

Baum, E. B. (2004). *What is thought?* Cambridge, MA: MIT Press.

Bi, G. (2002). Spatiotemporal specificity of synaptic plasticity: Cellular rules and mechanisms. *Biological Cybernetics*, *87*(5), 319–332.

Bi, G.-q., & Poo, M.-m. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience*, *24*(1), 139–166.

Bienenstock, E., Cooper, L., & Munro, P. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, *2*(1), 32–48.

Bliss, T. V., & Lømo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology*, *232*(2), 331–356.

Bonhoeffer, T., Staiger, V., & Aertsen, A. (1989). Synaptic plasticity in rat hippocampal slice cultures: Local" Hebbian" conjunction of pre- and postsynaptic stimulation leads to distributed synaptic enhancement. *Proceedings of the National Academy of Sciences*, *86*(20), 8113–8117.

Botelho, F., & Jamison, J. E. (2002). A learning rule with generalized Hebbian synapses. *Journal of Mathematical Analysis and Applications*, *273*(2), 529–547.

Botelho, F., & Jamison, J. E. (2004). Qualitative behavior of differential equations associated with artificial neural networks. *Journal of Dynamics and Differential Equations*, *16*(1), 179–204.

Chen, X., Leischner, U., Rochefort, N. L., Nelken, I., & Konnerth, A. (2011). Functional mapping of single spines in cortical neurons in vivo. *Nature*, *475*(7357), 501–505.

Chevaleyre, V., & Castillo, P. E. (2004). Endocannabinoid-mediated metaplasticity in the hippocampus. *Neuron*, *43*(6), 871–881.

Cox, K., & Adams, P. (2009). Hebbian crosstalk prevents nonlinear unsupervised learning. *Frontiers in Computational Neuroscience*, *3*.

Dayan, P., & Abbott, L. (2002). Theoretical neuroscience: Computational and mathematical modeling of neural systems. *Philosophical Psychology*, *15*(4), 563–577.

DeFelipe, J., Marco, P., Busturia, I., & Merchán-Pérez, A. (1999). Estimation of the number of synapses in the cerebral cortex. Methodological considerations. *Cerebral Cortex*, *9*(7), 722–732.

De Koninck, P., & Schulman, H. (1998). Sensitivity of CAM kinase II to the frequency of CA2+ oscillations. *Science*, *279*(5348), 227–230.

Diamantaras, K. I., & Kung, S. Y. (1996). *Principal component neural networks: Theory and applications*. New York: Wiley.

Dudek, S. M., & Bear, M. F. (1992). Homosynaptic long-term depression in area CA1 of hippocampus and effects of N-methyl-D-aspartate receptor blockade. *Proceedings of the National Academy of Sciences*, *89*(10), 4363–4367.

Elliott, T. (2003). An analysis of synaptic normalization in a general class of Hebbian models. *Neural Computation*, *15*(4), 937–963.

Elliott, T. (2008). Temporal dynamics of rate-based synaptic plasticity rules in a stochastic model of spike-timing-dependent plasticity. *Neural Computation*, *20*(9), 2253–2307.

Elliott, T. (2012). Cross-talk induces bifurcations in nonlinear models of synaptic plasticity. *Neural Computation*, *24*(2), 455–522.

Elliott, T., & Shadbolt, N. R. (2002). Multiplicative synaptic normalization and a nonlinear Hebb rule underlie a neurotrophic model of competitive synaptic plasticity. *Neural Computation*, *14*(6), 1311–1322.

Engert, F., & Bonhoeffer, T. (1997). Synapse specificity of long-term potentiation breaks down at short distances. *Nature*, *388*(6639), 279–284.

Fernando, C., Goldstein, R., & Szathmáry, E. (2010). The neuronal replicator hypothesis. *Neural Computation*, *22*(11), 2809–2857.

Fernando, C., & Szathmáry, E. (2009). Chemical, neuronal and linguistic replicators. In M. Pigliucci & G. Müller (Eds.), *Towards an extended evolutionary synthesis* (pp. 209–249). Cambridge, MA: MIT Press.

Frey, U., & Morris, R. (1998). Weak before strong: Dissociating synaptic tagging and plasticity-factor accounts of late-LTP. *Neuropharmacology*, *37*(4), 545–552.

Gerstner, W., & Kistler, W. M. (2002). Mathematical formulations of Hebbian learning. *Biological Cybernetics*, *87*(5–6), 404–415.

Goodhill, G. J., & Barrow, H. G. (1994). The role of weight normalization in competitive learning. *Neural Computation*, *6*(2), 255–269.

Harvey, C., & Svoboda, K. (2007). Locally dynamic synaptic learning rules in pyramidal neuron dendrites. *Nature*, *450*(7173), 1195–1200.

Harvey, C. D., Yasuda, R., Zhong, H., & Svoboda, K. (2008). The spread of RAS activity triggered by activation of a single dendritic spine. *Science Signaling*, *321*(5885), 136–140.

Hebb, D. O. (2002). *The organization of behavior: A neuropsychological theory*. Mahwah, NJ: Erlbaum.

Hertz, J. A., Krogh, A. S., & Palmer, R. G. (1991). *Introduction to the theory of neural computation* (vol. 1), Boulder, CO: Westview Press.

Hinton, G. E., & Sejnowski, T. J. (1999). *Unsupervised learning: Foundations of neural computation*. Cambridge, MA: MIT Press.

Hyvärinen, A., Hurri, J., & Hoyer, P. O. (2009). *Natural image statistics*. New York: Springer.

Isaac, J. T., Nicoll, R. A., & Malenka, R. C. (1995). Evidence for silent synapses: Implications for the expression of LTP, *Neuron*, *15*(2), 427–434.

Jia, H., Rochefort, N. L., Chen, X., & Konnerth, A. (2010). Dendritic organization of sensory input to cortical neurons in vivo. *Nature*, *464*(7293), 1307–1312.

Katz, L. C., & Shatz, C. J. (1996). Synaptic activity and the construction of cortical circuits. *Science*, *274*(5290), 1133–1138.

Koch, C., & Zador, A. (1993). The function of dendritic spines: Devices subserving biochemical rather than electrical compartmentalization. *J. Neurosci.*, *13*(2), 413–422.

Korte, M., Carroll, P., Wolf, E., Brem, G., Thoenen, H., & Bonhoeffer, T. (1995). Hippocampal long-term potentiation is impaired in mice lacking brain-derived neurotrophic factor. *Proceedings of the National Academy of Sciences*, *92*(19), 8856–8860.

Kossel, A., Bonhoeffer, T., & Bolz, J. (1990). Non-Hebbian synapses in rat visual cortex. *NeuroReport*, *1*(2), 115–118.

Lemann, V., Gottmann, K., & Heumann, R. (1994). BDNF, and NT-4/5 enhance glutamatergic synaptic transmission in cultured hippocampal neurones. *Neuroreport*, *6*(1), 21–25.

Levine, E. S., Dreyfus, C. F., Black, I. B., & Plummer, M. R. (1995). Brain-derived neurotrophic factor rapidly enhances synaptic transmission in hippocampal neurons via postsynaptic tyrosine kinase receptors. *Proceedings of the National Academy of Sciences*, *92*(17), 8074–8077.

Lisman, J. (1989). A mechanism for the Hebb and the anti-Hebb processes underlying learning and memory. *Proceedings of the National Academy of Sciences*, *86*(23), 9574–9578.

Lisman, J. (1994). The CAM kinase II hypothesis for the storage of synaptic memory. *Trends in Neurosciences*, *17*(10), 406–412.

Lynch, G., Dunwiddie, T., & Gribkoff, V. (1977). Heterosynaptic depression: A postsynaptic correlate of long-term potentiation. *Nature*, *266*(5604), 737–739.

Lyu, S., & Simoncelli, E. (2009). Nonlinear extraction of independent components of natural images using radial gaussianization. *Neural Computation*, *21*(6), 1485–1519.

Malenka, R. C., & Bear, M. F. (2004). LTP and LTD: An embarrassment of riches. *Neuron*, *44*(1), 5–21.

Markram, H., Lübke, J., Frotscher, M., Roth, A., & Sakmann, B. (1997). Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex. *Journal of Physiology*, *500*(Pt. 2), 409–440.

Mastronarde, D. N. (1989). Correlated firing of retinal ganglion cells. *Trends in Neurosciences*, *12*(2), 75–80.

Matsuzaki, M., Honkura, N., Ellis-Davies, G. C., & Kasai, H. (2004). Structural basis of long-term potentiation in single dendritic spines. *Nature*, *429*(6993), 761–766.

Miller, K. D. (1990). Derivation of linear Hebbian equations from a nonlinear Hebbian model of synaptic plasticity. *Neural Computation*, *2*(3), 321–333.

Miller, K. D., & MacKay, D. J. (1994). The role of constraints in Hebbian learning. *Neural Computation*, *6*(1), 100–126.

Navakkode, S., Sajikumar, S., & Frey, J. U. (2004). The type IV-specific phosphodiesterase inhibitor rolipram and its effect on hippocampal long-term potentiation and synaptic tagging. *Journal of Neuroscience*, *24*(35), 7740–7744.

Nevian, T., & Sakmann, B. (2006). Spine CA2+ signaling in spike-timing-dependent plasticity. *Journal of Neuroscience*, *26*(43), 11001–11013.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*(3), 267–273.

Oja, E., & Karhunen, J. (1985). On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, *106*(1), 69–84.

Petersen, C. C., Malenka, R. C., Nicoll, R. A., & Hopfield, J. J. (1998). All-or-none potentiation at CA3-CA1 synapses. *Proceedings of the National Academy of Sciences*, *95*(8), 4732–4737.

Rădulescu, A., & Adams, P. (2013). *Hebbian crosstalk and input segregation.* arXiv:1207.7257.

Rădulescu, A., & Adams, P. (2013). Hebbian crosstalk and input segregation. *Journal of Theortical Biology*, *337*, 133–149.

Rădulescu, A., Cox, K., & Adams, P. (2009). Hebbian errors in learning: An analysis using the Oja model. *Journal of Theoretical Biology*, *258*(4), 489–501.

Sabatini, B. L., Oertner, T. G., & Svoboda, K. (2002). The life cycle of $CA^{2+}$ ions in dendritic spines. *Neuron*, *33*(3), 439–452.

Schuman, E. M., & Madison, D. V. (1994). Locally distributed synaptic potentiation in the hippocampus. *Science*, *263*(5146), 532–536.

Shouval, H. Z., Bear, M. F., & Cooper, L. N. (2002). A unified model of NMDA receptor-dependent bidirectional synaptic plasticity. *Proceedings of the National Academy of Sciences*, *99*(16), 10831–10836.

Sjöström, P. J., Turrigiano, G. G., & Nelson, S. B. (2003). Neocortical LTD via coincident activation of presynaptic NMDA and cannabinoid receptors. *Neuron*, *39*(4), 641–654.

Sjöström, P. J., Turrigiano, G. G., & Nelson, S. B. (2004). Endocannabinoid-dependent neocortical layer-5 LTD in the absence of postsynaptic spiking. *Journal of Neurophysiology*, *92*(6), 3338–3343.

Taylor, J., & Coombes, S. (1993). Learning higher order correlations. *Neural Networks*, *6*(3), 423–427.

Trong, P. K., & Rieke, F. (2008). Origin of correlated activity between parasol retinal ganglion cells. *Nature Neuroscience*, *11*(11), 1343–1351.

Varga, Z., Jia, H., Sakmann, B., & Konnerth, A. (2011). Dendritic coding of multiple sensory inputs in single cortical neurons in vivo. *Proceedings of the National Academy of Sciences*, *108*(37), 15420–15425.

Vasudeva, R. (1998). How negative can the product-moment correlation coefficient be? *Resonance*, *3*(5), 73–75.

Volkenshte, M. V. (1991). *Physical approaches to biological evolution*. New York: Springer.

von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Biological Cybernetics*, *14*(2), 85–100.

Wiskott, L., & Sejnowski, T. (1998). Constrained optimization for neural map formation: A unifying framework for weight growth and normalization. *Neural Computation*, *10*(3), 671–716.

Yim, M. Y., Aertsen, A., & Kumar, A. (2011). Significance of input correlations in striatal function. *PLoS Computational Biology*, *7*(11), e1002254.

Received April 6, 2013; accepted October 15, 2013.