Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.

This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/authorsrights

Author's personal copy

Journal of Theoretical Biology 337 (2013) 133–149

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Hebbian crosstalk and input segregation

Anca Rădulescu [a,*], Paul Adams [b,c]

[a] Department of Mathematics, 395 UCB, University of Colorado, Boulder, United States
[b] Department of Neurobiology and Behavior, Stony Brook University, Stony Brook, United States
[c] Kalypso Institute, Stony Brook, NY, United States

HIGHLIGHTS

- Synaptic inspecificity (cross talk) may affect the outcome of a Hebbian learning rule.
- We analyze the equilibria and stability of the 2-dimensional inspecific Oja rule.
- A phase plane bifurcation occurs at critical cross talk value, only for unbiased inputs.
- A different normalization scheme, and a stochastic version are presented.
- We suggest an application to ocular segregation; we compare cortical proofreading with DNA copying.

ARTICLE INFO

Article history:
Received 20 September 2012
Received in revised form
4 August 2013
Accepted 5 August 2013
Available online 14 August 2013

Keywords:
Crosstalk
Hebbian synapses
Pairwise correlations
Sensitivity analysis
Codimension two bifurcation

ABSTRACT

Hebbian synapses respond to input/output correlations, and thus to input statistical structure. However, recent evidence suggests that strength adjustments are not completely connection-specific, and this "crosstalk" could distort, or even prevent, learning processes. Crosstalk would then be a form of adjustment mistake, analogous to mistakes in polynucleotide copying. The mutation rate must be extremely low for successful evolution (which is a type of learning process), and similarly neural learning might require minimal crosstalk. We analyze aspects of the effect of crosstalk in Hebbian learning from pairwise input correlations, using the classical Oja model.

In previous work we showed that crosstalk leads to learning of the principal eigenvector of **EC** (the input covariance matrix pre-multiplied by an error matrix that describes the crosstalk pattern), and found that, with positive input correlations, increasing crosstalk smoothly degrades performance. However, the Oja model requires negative input correlations to account for biological ocular segregation. Although this assumption is biologically somewhat implausible, it captures features that are seen in more complex models. Here, we analyze how crosstalk would affect such segregation.

We show that, for statistically unbiased inputs, crosstalk induces a bifurcation from segregating to non-segregating outcomes at a critical value which depends on correlations. We also investigate the behavior in the vicinity of this critical state and for weakly biased inputs.

Our results show that crosstalk can induce a bifurcation under special conditions even in the simplest Hebbian models, and that even the low levels of crosstalk observed in the brain could prevent normal development. However, during learning pairwise input statistics are more complex, and crosstalk-induced bifurcations may not occur in the Oja model. Such bifurcations would be analogous to "error catastrophes" in genetic models, and we argue that they are usually absent for simple linear Hebbian learning because such learning is only driven by pairwise correlations.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

1.1. Motivation

There has been sustained interest in the possibility that learning and synaptic plasticity might not just be complementary (Baldwin, 1909; Hinton and Nowlan, 1987) to Darwinian evolution and adaptation, but physically analogous (Edelman, 1987; Young, 1979; Jerne, 1994; Changeux et al., 1973.; Calvin, 1996; Fernando et al., 2010). Such analogies would be particularly fruitful if they pointed to new interpretations of hitherto obscure aspects of physiology or anatomy. Perhaps the simplest, most direct analogy, and one with interesting implications for neocortical machinery, is the possibility that mutations (mistakes in copying DNA) are comparable to anatomical errors in the strengthening of synaptic connections thought to underlie learning. By "anatomical errors"

* Corresponding author. Tel.: +1 303 492 7716; fax: +1 303 492 7707.
E-mail addresses: radulesc@colorado.edu, ancu.math@gmail.com (A. Rădulescu).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.08.004

Author's personal copy

we mean that the spiking activity of particular connections might induce changes (strengthening or weakening) not only of that connection, as envisaged in most theoretical models of synaptic plasticity, but also of other connections (for example, of those that are composed of synapses that are physically very close to the synapses directly involved in the active connection). Such Hebbian "inspecificity" or "crosstalk" might involve changes in existing but inactive connections, or even creation of new connections (Adams, 1998; Le Bé and Markram, 2006). Molecular evolution models show that asexual adaptation can only occur if the *per base* mutation rate is, roughly, less than the reciprocal genome length (Eigen, 1971). In particular, the equilibrium behavior of such models shows a dynamical bifurcation – an "error catastrophe" – at a critical mutation rate, at least in the large genome limit (Saakian and Hu, 2004; Schuster and Swetina, 1988). Indeed, successive improvements in copying fidelity have marked many of the major biological transitions (Smith and Szathmary, 1997), such as that from the RNA world to the DNA/protein world, and the appearance of proofreading replicases. Indeed, there is some evidence that sex itself is prevalent because it relaxes the critical error limit (Ridley, 2001; Otto, 2009).

Synaptic plasticity is often Hebbian (dependent on both input and output firing) and perhaps the simplest formal model of Hebbian learning is that of Oja (1982), who showed that a Hebbian neuron could act as a principal component analyzer, since linear Hebbian learning is driven by the input covariance matrix. In this model, the input second order correlations essentially constitute the learning environment. It has been recently suggested that the Oja equation (even without errors) might have mathematical analogies to the Eigen replication/mutation equation (Fernando and Szathmáry, 2009; Fernando et al., 2010). However, we have shown (Radulescu et al., 2009) that connectional plasticity inspecificity (also known as "crosstalk") does not usually generate a dynamical bifurcation in the Oja model: the leading eigenvector of the controlling matrix (the product of the error matrix and the covariance matrix) remains stable at all error rates, and changes direction only in a smooth manner. Here we analyze one important exception to this conclusion: when the input correlations are both negative and uniform. This case is of interest because it has been previously discussed in relation to the well-known development of binocular input segregation (Dayan and Abbott, 2002; Miller and MacKay, 1994; Elliott, 2003, 2008; Bienenstock et al., 1982). Nevertheless, we conclude that the crosstalk Oja model does not generally show the type of "error catastrophe" seen in molecular evolution. However, rather than showing that this invalidates possible analogies between Darwinian evolution and Hebbian learning, it might merely highlight the limitations of linear Hebbian learning, which is driven only by second order correlations. We propose that the full set of input correlations, both second and higher-order, constitutes the learning environment. Other recent work shows that Hebbian learning driven by higher order correlations does indeed show a crosstalk-induced dynamical bifurcation (Cox and Adams, 2009; Elliott, 2012), suggesting that mechanisms analogous to proofreading and sex (Adams and Cox, 2012, 2006) might play an important role in neural adaptation.

1.2. Background

Learning is thought to occur as a result of changes in synaptic strength triggered by pre- and postsynaptic neural activity, in a "Hebbian" manner. Such changes are not completely specific to the synapses at which the activity occurs (Harvey and Svoboda, 2007; Bi, 2002; Engert and Bonhoeffer, 1997), because of inevitable albeit minimal second-messenger diffusion.

Oja (1982) showed that a simple model neuron could perform unsupervised Hebbian learning of the first principal component of an input distribution. In this model, unlimited weight growth is prevented using an additional term in the learning rule, producing an implicit, "multiplicative" weight normalization (Malsburg, 1973). Biological synapses do show Hebbian properties, using well-understood, spike-coincidence detection machinery, raising the possibility that real neurons can exhibit similar unsupervised learning. Finding principal components, or related second order statistics, could be very useful in the brain for data compression and transmission (Atick and Redlich, 1992; Srinivasan et al., 1982). Furthermore, representational learning often requires that inputs will be pairwise decorrelated. Hebbian learning can also explain developmental changes, such as the segregation of visual input to central neurons (Wimbauer et al., 1997a,b, Wimbauer et al., 1998).

Recent data suggest (Harvey and Svoboda, 2007; Bi, 2002; Engert and Bonhoeffer, 1997) that weight updates may be affected by each other, for example due to unavoidable residual second messenger diffusion between closely spaced synapses. We have suggested that such crosstalk is analogous to mutation in genetics, and that cortical circuitry may be specialized to reduce it (Adams and Cox, 2002a, 2006). However, it is not clear that learning would be subject to an "error catastrophe" such as that occurring in genetic systems (Eigen, 1971). If complete learning failure does not occur at a critical, low, crosstalk level, such circuitry might not be necessary.

In a recent paper (Radulescu et al., 2009) we examined how crosstalk would affect the Oja model. We considered a learning network consisting of a single output neuron receiving, through a set of n input neurons, n signals $\mathbf{x}=(x_1,\ldots,x_n)^T$ drawn from a probability 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 was generated as the weighted sum of the inputs $y=\mathbf{x}^T\boldsymbol{\omega}$.

The synaptic weights ω_i were modified in accordance with Oja's rule of learning, by implementing first a Hebb-like strengthening of each ω_i proportionally with the product of x_i and y (with small constant of proportionality, or *learning rate*, γ), followed by an approximate "normalization" step (applicable for small γ and $\|\boldsymbol{\omega}\|$ close to one), maintaining the Euclidean norm of the weight vector $\boldsymbol{\omega}=(\omega_1,\ldots,\omega_n)^T$ close to one. We considered the long-term average of this Oja equation, using the input covariance matrix $\mathbf{C}=\langle\mathbf{x}\mathbf{x}^T\rangle$ as an appropriate characterization of the inputs, and studied the long-term behavior of the conditional expectation $\mathbf{w}(t)=\langle\boldsymbol{\omega}(t+1)|\boldsymbol{\omega}(t)\rangle$, given by the continuous time differential equation:

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

We then introduced inspecificity into the learning equation (Radulescu et al., 2009). We implemented this inspecificity by assuming that, on average, only a fraction q of the intended update reaches the appropriate connection, the remaining fraction $1-q$ being distributed amongst the other connections (following a rule which we defined according to plausible underlying biology). The quality factor q is analogous to a similar factor in molecular evolution theory that represents the fidelity of single-base copying (Swetina and Schuster, 1982). The actual update at a given connection thus includes contributions from erroneous or inaccurate updates from other connections. The erroneous updating process was formally described by an error matrix $\mathcal{E}$, independent of the inputs, whose elements, which depend on average on q, reflect at each time step t the fractional contribution that the activity through the connection with weight ω_i makes to the update of ω_j:

$$\omega_i(t+1)=\omega_i+\gamma y([\mathcal{E}\mathbf{x}]_i-y\omega_i)$$

Author's personal copy

The long-term, continuous-time statistics can be then written in matrix form as

$$\frac{d\mathbf{w}}{dt} = \gamma[\mathbf{ECw} - (\mathbf{w}^T\mathbf{Cw})\mathbf{w}] \tag{1}$$

where the average "error matrix" $\mathbf{E} = \langle \mathcal{E} \rangle$ is a symmetric matrix with positive entries, which equals the identity matrix $I \in \mathcal{M}_n(\mathbb{R})$ in case of perfect quality updates. Throughout the paper, we call this the (inspecific) Oja rule with continuous time updates, and we use the notation $f^{\mathbf{E}}(\mathbf{w}) = \gamma[\mathbf{ECw} - (\mathbf{w}^T\mathbf{Cw})\mathbf{w}]$. One may notice that the inspecific rule does not perform the same approximate normalization of $\mathbf{w}$ as the original continuous-time Oja rule; under Eq. (1), the norm of $\mathbf{w}$ remains bounded, without having to become eventually close to one (see details below). Clearly one can modify the rule by performing an exact normalization, and thus keep $\mathbf{w}$ of unit length and study one-dimensional dynamics on the circle. In this paper, we analyze extensively the first case (Section 2), but we also briefly study the case of exact normalization (Section 3).

We (Radulescu et al., 2009) and others (Botelho and Jamison, 2004) have studied the asymptotic behavior of the n-dimensional system defined by $f^{\mathbf{E}}(\mathbf{w})$. We started with a local linear analysis of the equilibria and their stability. Although this rule is nonlinear, the Hebbian update term is linear in the output, and we sometimes refer to this, and related, rules, as being "linear," in contrast to other Hebbian rules (Hyvärinen and Oja, 1998; Hyvärinen et al., 2001; Bell and Sejnowski, 1995; Olshausen et al., 1996; Elliott, 2003; Földiák, 1990; Cooper, 2004) which are nonlinear in the output.

Note that the symmetric, positive definite matrix $\mathbf{C} \in \mathcal{M}_n(\mathbb{R})$ defines a dot product between any two vectors $\mathbf{w}$ and $\mathbf{v}$ in $\mathbb{R}^n$ as $\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbf{C}} = \mathbf{v}^T\mathbf{Cw}$. 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: $\langle \mathbf{ECu}, \mathbf{v} \rangle_{\mathbf{C}} = (\mathbf{ECu})^T\mathbf{Cv} = \mathbf{u}^T\mathbf{C}^T\mathbf{E}^T\mathbf{Cv} = \mathbf{u}^T\mathbf{CECv} = \langle \mathbf{u}, \mathbf{ECv} \rangle_{\mathbf{C}}$, for all $\mathbf{u}$, $\mathbf{v} \in \mathbb{R}^n$. Hence $\mathbf{EC}$ has a basis of eigenvectors, orthogonal with respect to the dot product $\langle \cdot, \cdot \rangle_{\mathbf{C}}$. The following description of equilibria $\mathbf{w} = (w_1 \ldots w_n)^T$ such that $\mathbf{ECw} = (\mathbf{w}^T\mathbf{Cw})\mathbf{w}$ is immediate:

Description of equilibria: The equilibria of the system (1) are $\mathbf{w} = \mathbf{0}$ and all eigenvectors of $\mathbf{EC}$ (with corresponding eigenvalue $\lambda_{\mathbf{w}}$), normalized, with respect to the norm $\| \cdot \|_{\mathbf{C}} = \langle \cdot, \cdot \rangle_{\mathbf{C}}$, so that $\|\mathbf{w}\|_{\mathbf{C}} = \lambda_{\mathbf{w}}$.

At any equilibrium $\mathbf{w}$ of the system (1), the Jacobian matrix of $f^{\mathbf{E}}$ around $\mathbf{w}$ is (see Appendix A.1 for a complete calculation)

$$Df^{\mathbf{E}}_{\mathbf{w}} = \gamma[\mathbf{EC} - 2\mathbf{w}(\mathbf{Cw})^T - (\mathbf{w}^T\mathbf{Cw})\mathbf{I}] \tag{2}$$

Then we have the following (see Appendix A.1 for proof):

Stability criteria for equilibria: Suppose $\mathbf{EC}$ has a multiplicity one largest eigenvalue. A normalized eigenvector $\mathbf{w}$ is a local hyperbolic attracting equilibrium for (1) iff it corresponds to the maximal eigenvalue of $\mathbf{EC}$.

Such attractors always exist provided $\mathbf{EC}$ has a maximal eigenvalue of multiplicity one, a property which is generic for $\mathbf{EC}$ (Radulescu et al., 2009). When assuming a unique leading eigenvalue, the corresponding eigendirection is orthogonal in $\langle \cdot, \cdot \rangle_{\mathbf{C}}$ to all other eigenvectors of $\mathbf{EC}$. 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 $\|\mathbf{w}\|_{\mathbf{C}} = \lambda_{\mathbf{w}}$. It can be shown easily that these two attractors (the appropriately normalized eigenvectors corresponding to the maximal eigenvalue of $\mathbf{EC}$) can be the only attractors in the system (see Appendix B for proof).

In a previous paper (Radulescu et al., 2009), we further analyzed the sensitivity of the system under variations of parameters, for some biologically plausible forms of the covariance and error matrices:

$$\mathbf{C} = \begin{bmatrix} v+\delta_1 & c & \cdots & c \\ c & v+\delta_2 & \cdots & c \\ \vdots & & \ddots & \vdots \\ c & c & \cdots & v+\delta_n \end{bmatrix} \quad \text{and} \quad \mathbf{E} = \begin{bmatrix} q & \epsilon & \cdots & \epsilon \\ \epsilon & q & \cdots & \epsilon \\ \vdots & & \ddots & \vdots \\ \epsilon & \epsilon & \cdots & q \end{bmatrix}$$

where the input covariance matrix had uniform covariances $c > 0$ and variance biases $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_n$; the error matrix was defined such that $q > \epsilon > 0$, $q + (n-1)\epsilon = 1$. Our analysis of this system concluded that the effect of biologically realistic levels of crosstalk would typically only produce small gradual changes in the learning process, though when inputs carry very similar signals, the effects could be more dramatic. In this paper we explore this "very similar" scenario more thoroughly. In particular, we describe the effect of crosstalk in the special "unbiased" case, where the inputs have identical statistics.

1.3. Biased and unbiased inputs

Our previous analysis considered only distributions of inputs with a bias in the covariance matrix (we imposed the condition that $\mathbf{EC}$ has a leading eigenvalue of multiplicity one). While this case is mathematically generic, previous work using related models (without crosstalk, Miller et al., 1989) to study learning in the visual system, often assumed that the input statistics are "unbiased," or identical for each input (for example, because inputs from corresponding points in the left and right eyes represent the same point in visual space). It is well known that in the two-dimensional case, if the two inputs x_1 and x_2 are positively correlated (as one might anticipate for active vision), linear Hebbian learning does not predict the observed developmental segregation of visual afferents (Dayan and Abbott, 2002; Cooper, 2004; Swindale, 1996; Willshaw and Von Der Malsburg, 1976); negative correlations (or a nonlinear rule) are required. However, modifications in learning rules, for example subtractive normalization (Goodhill, 1993; Willshaw and Von Der Malsburg, 1976; Miller and MacKay, 1994; Linsker, 1986), a weight-dependent rule (Elliott and Shadbolt, 2002) or a BCM rule (Cooper, 2004), although not always originally developed to explain segregation, can overcome this difficulty. Subtractive rules lead to Hebbian learning driven by a modified version of the covariance matrix. In the current work, we examine the dynamics of Oja learning with crosstalk when inputs are unbiased, and how this changes when a slight bias is introduced.

We show here that in the unbiased negative correlation case, the system undergoes a bifurcation in dynamics at a critical crosstalk level. Related results have been obtained by Elliott (2012) and Cox and Adams (2009). While there is no true bifurcation in the near-unbiased case, the very dramatic change in learning that occurs over a small error range would be biologically indistinguishable from a true bifurcation. We discuss our results in relation to models of development and learning.

1.4. A reduced, two-dimensional model

In this paper, we will consider the continuous-time, two-dimensional nonlinear rule of Oja (i.e., for two input channels and one output), with covariance matrix $\mathbf{C}$ and error matrix $\mathbf{E}$ symmetric matrices having the forms

$$\mathbf{C} = \begin{pmatrix} v+\delta & c \\ c & v \end{pmatrix}$$

and

$$\mathbf{E} = \begin{pmatrix} q & 1-q \\ 1-q & q \end{pmatrix}$$

Author's personal copy

The parameters are such that $1/2 < q \leq 1$, $v > 0$ and $c < 0$, such that $v > |c|$, $v > |\delta|$ and $v(v+\delta) > c^2$ (i.e., $\det(\mathbf{C}) > 0$). The 2D system expands to

$$\dot{w}_1 = [q(v+\delta)+(1-q)c]w_1 + [qc+(1-q)v]w_2 - [vw_1^2 + 2cw_1w_2 + vw_2^2]w_1$$
$$\dot{w}_2 = [(1-q)(v+\delta)+qc]w_1 + [(1-q)c+qv]w_2 - [vw_1^2 + 2cw_1w_2 + vw_2^2]w_2 \quad (3)$$

The rest of the paper is centered around this 2-dimensional model. In Section 2, we establish the mathematical background of the model's behavior. We analyze some of its local and global dynamics, observe the dependence of these dynamics on parameters and discuss bifurcations. One of the phenomena central to our interest is how the behavior of the system changes when the bias parameter δ varies, in particular when it approaches zero (i.e., the inputs are very close to a perfectly unbiased state). In Section 3, we discuss an alternative, direct normalization, model. In Section 4 we explore stochastic versions of the model. In the Discussion, we embed the results in the context of visual modeling and ocular segregation of inputs.

2. Linear analysis of the 2D dynamics

2.1. Spectrum of the 2D system

We notice that the phase plane of the system is symmetric about the origin (i.e., if $w(t)$ is a solution curve for the system, then $-w(t)$ is as well). The trace, determinant and eigenvalues of **EC** can be obtained easily as expressions of the system parameters:

$$\det(\mathbf{EC}) = \det(\mathbf{E})\det(\mathbf{C}) = (2q-1)[v(v+\delta)-c^2] > 0$$
$$\mathrm{tr}(\mathbf{EC}) = 2(1-q)c + q(2v+\delta) > 0$$
(from the Cauchy–Schwartz inequality).

Lemma 2.1. *For all parameter values,* **EC** *has two real eigenvalues* $\mu_{1,2}$*, which are distinct unless the conditions* $\delta = 0$ *and* $q = q^* = v/(v-c)$ *are simultaneously satisfied. More precisely, when* $q \neq q^*$*, we have*

$$\mu_1 = \frac{2(1-q)c + q(2v+\delta) + \sqrt{\Delta}}{2}$$
larger eigenvalue, with eigenline of slope
$$z_1 = \frac{-q\delta + \sqrt{\Delta}}{2\beta}$$
$$\mu_2 = \frac{2(1-q)c + q(2v+\delta) - \sqrt{\Delta}}{2}$$
smaller eigenvalue, with eigenline of slope
$$z_2 = \frac{-q\delta - \sqrt{\Delta}}{2\beta}$$

where $\beta = qc + (1-q)v$ *and* $\Delta = [2qc + (1-q)(2v+\delta)]^2 + (2q-1)\delta^2$.

Proof. The calculation of eigenvalues and eigenvectors is immediate from the characteristic equation of **EC**: $X^2 - \mathrm{tr}(\mathbf{EC})X + \det(\mathbf{EC}) = 0$, with discriminant

$$\Delta = \mathrm{tr}(\mathbf{EC})^2 - 4\det(\mathbf{EC}) = [2qc + (1-q)(2v+\delta)]^2 + (2q-1)\delta^2$$

Notice that $\Delta \geq 0$, with equality $\Delta = 0$ (i.e., double eigenvalue for **EC**) iff both $\delta = 0$ and $\beta = 0$. The critical quality value (where the two eigenvalues are equal, producing a switch in the dynamics when $\delta = 0$) is $q^* = v/(v-c)$. Since $v > |c|$, this value occurs within the appropriate q range, $(1/2, 1]$ (see Fig. 1). □

2.2. Equilibria of the 2D system

Throughout this section, in addition to working in our generally specified parameter ranges, we will assume that **EC** has distinct eigenvalues (i.e., $\delta \neq 0$ or $q \neq q^*$). In this case, the system has as equilibria the origin $\mathbf{w} = \mathbf{0}$, and two pairs of opposite eigenvectors of **EC** normalized such that $\mathbf{w}^T\mathbf{C}\mathbf{w} = \mu$ (where μ is the respective eigenvalues of each pair).

The normalization condition can be written as

$$\mathbf{w}^T\mathbf{C}\mathbf{w} = (v+\delta)w_1^2 + 2cw_1w_2 + vw_2^2 = \mu$$

Using the same notation $z = w_2/w_1$, this can be rewritten as $vz^2 + 2cz + (v+\delta) = \mu/w_1^2$, so that

$$\|\mathbf{w}\| = \sqrt{\frac{\mu(z^2+1)}{vz^2+2cz+(v+\delta)}}$$

While a explicit formula of $\|\mathbf{w}\|$ in terms of parameters would be complicated, the above implicit expression suggests that the norm varies with both error and correlation (see Fig. 2).

The position and stability of the four nonzero equilibria also vary with the parameters v, c, δ and q. If we aim to study the sensitivity of the system's dynamics under parameter perturbations, the next step should be establishing the linear stability of these equilibria. It is easy to show, using Proposition A1 (from Appendix A.1) that

Description and stability of equilibria: Suppose the matrix **EC** has distinct eigenvalues. The system (3) has five distinct equilibria, $\mathbf{w} = \mathbf{0}$ and four normalized eigenvectors of **EC**. The two (opposite) eigenvectors of the larger eigenvalue are hyperbolic attractors, and the two (opposite) eigenvectors corresponding to the lower eigenvalue are saddles. The origin is repelling.

More precisely, this means that if $\mu_{\mathbf{w}}$ is the larger eigenvalue of **EC**, the Jacobian matrix $Df^{\mathbf{E}}_{\mathbf{w}}$ has two negative eigenvalues, hence $\mathbf{w}$ is an attracting node. If instead $\mu_{\mathbf{w}}$ is the smaller eigenvalue of **EC**, then $Df^{\mathbf{E}}_{\mathbf{w}}$ has two real eigenvalues of opposite signs, and $\mathbf{w}$ is a saddle equilibrium.

We are particularly interested in the behavior near and at $\delta = 0$. The above characterization of equilibria applies when $\delta \neq 0$, but it breaks down in the parameter slice $\delta = 0$, at the critical point when **EC** has a double eigenvalue. In other words we expect that the system undergoes a bifurcation in the unbiased $\delta = 0$ slice, which does not exist in the other, $\delta \neq 0$ slices (i.e., when "bias" is present in the inputs), therefore we will study this case separately.

For the following paragraph (Section 2.2) we assume $\delta \neq 0$. The unbiased case $\delta = 0$ is discussed separately in Section 2.3. The results are integrated and concluded in Section 2.4.

2.3. Biased case: rotational dynamics and invariant lines

One way to describe the dynamics of the system, including the more global aspects and possibly cyclic behavior (which has not yet been excluded) is to follow the rotational direction of the solution trajectories in different regions of the (w_1, w_2) phase-plane under the velocity field $(\dot{w}_1, \dot{w}_2)$.

Consider the angle $\theta \in [-\pi/2, \pi/2]$ made by the direction (w_1, w_2) with the w_1 axis. As before, call $z = w_2/w_1 = \tan(\theta)$ and $\beta = qc + (1-q)v$. Then, along a trajectory in the (w_1, w_2) plane,

$$\dot{z} = \frac{d}{dt}\left(\frac{w_2}{w_1}\right) = \frac{\dot{w}_2w_1 - \dot{w}_1w_2}{w_1^2}$$
$$= [(1-q)(v+\delta)+qc] - q\delta z - [(1-q)v+qc]z^2$$
$$= -\beta z^2 - q\delta z + [\beta + (1-q)\delta]$$

We first want to establish if there are any values of z for which $\dot{z} = 0$. These are the slopes along which the rotational speed of the trajectories is zero; in other words, they would correspond to invariant lines in the phase-plane.

We consider the quadratic equation: $\dot{z} = -\beta z^2 - q\delta z + [\beta + (1-q)\delta] = 0$. The discriminant is the same as the one of the

Author's personal copy

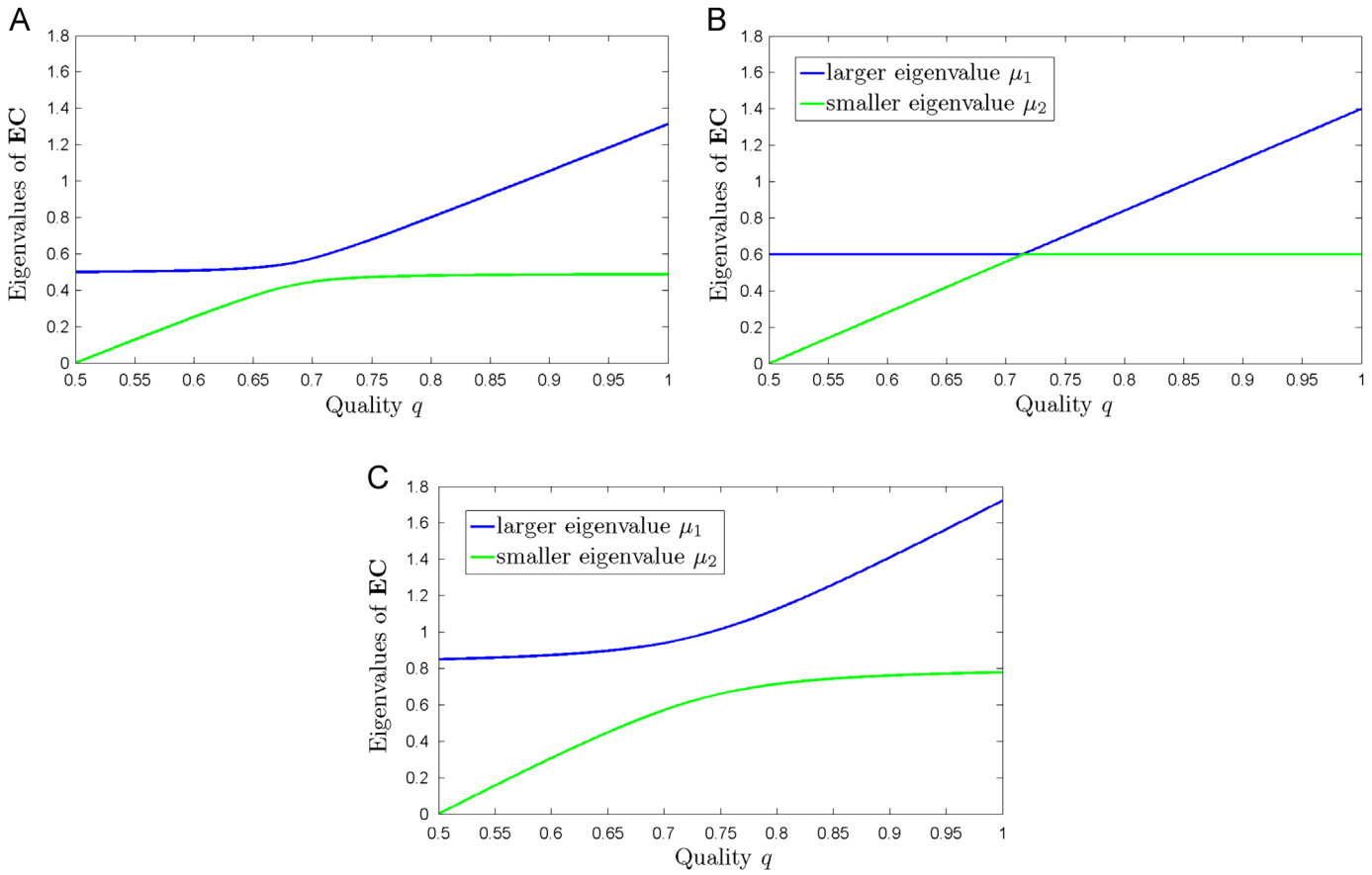

Fig. 1. Evolution of the eigenvalues as the quality q is varied, in three different δ slices. $\delta=-0.2$ (A), $\delta=0$ (B) and $\delta=0.5$ (C). Fixed parameters: $v=1$ and $c=-0.4$, hence $q^*=1/1.4–0.71$. When $\delta=0$, the eigenvalues μ_1 and μ_2 touch at $q=q^*$. For $\delta\neq 0$, the two curves avoid this crossing; the minimal distance between them occurs at $q=q^*$, but it is strictly positive.

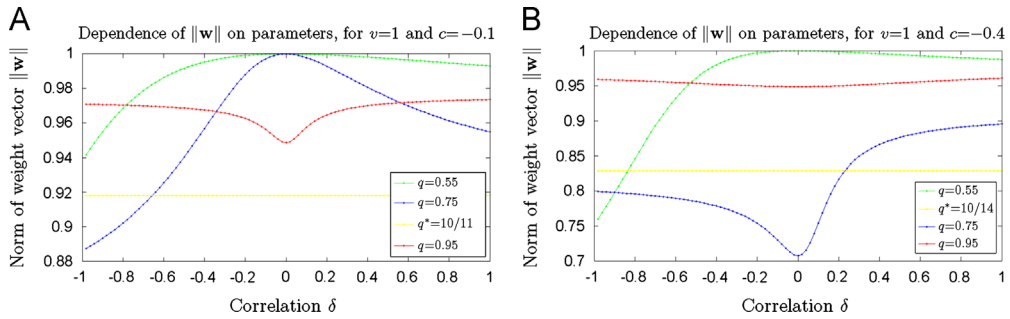

Fig. 2. The norm $\|\mathbf{w}\|$ varies with both quality q and correlation δ. (A) The panel illustrates $v=1$ and $c=-0.1$. Each color-coded plot represents how the norm $\|\mathbf{w}\|$ changes as a function of δ, for a different value of the quality q. (B) The panel illustrates $v=1$ and $c=-0.4$. Each color-coded plot represents how the norm $\|\mathbf{w}\|$ changes as a function of δ, for a different value of the quality q. In both panels, the colors refer respectively to $q=0.55$ (green), $q=0.75$ (blue), $q=0.95$ (red) and critical $q=q^*=v/(v-c)$ (yellow). (See also Appendix C, for a more analytical approach.) (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

characteristic equation of **EC**:

$$\Delta = q^2\delta^2 + 4\beta[\beta+(1-q)\delta]$$
$$= [2qc+(1-q)(2v_\delta)]^2+(2q-1)\delta^2$$

The solutions of the quadratic equation will the be exactly the slopes of the eigendirections of **EC**:

$$z_{1,2}=\frac{-q\delta\pm\sqrt{\Delta}}{2\beta}\in[-\infty,+\infty]$$

proving the following:

Lemma 2.2. *The eigendirections of the matrix* **EC** *represent invariant lines under the vector field of system* (1).

We want to better describe the phase-plane behavior *between* the invariant lines $z=z_1$ and $z=z_2$. For any fixed $q\in(1/2,q^*)\cup(q^*,1]$ (i.e., for $\beta\neq 0$), the rotational speed is given by the sign of the quadratic function $f(z)=-\beta z^2-q\delta z+[\beta+(1-q)\delta]$. In principle, we then have two situations:

i. $q\in(1/2,q^*)$ (i.e., $\beta>0$). Then $z_1>z_2$, with $z>0$ in (z_2,z_1) and $z<0$ on $(-\infty,z_2)\cup(z_1,\infty)$. The phase plane looks schematically as in Fig. 3A.
ii. $q\in(q^*,1]$ (i.e., $\beta<0$). Then $z_1<z_2$, with $z<0$ in (z_1,z_2) and $z>0$ on $(-\infty,z_1)\cup(z_2,\infty)$. The phase plane looks schematically as in Fig. 3B.

In other words, all trajectories move asymptotically towards the invariant line $z=z_1$.

Author's personal copy

Since the behavior of the system seems to be a large extent dictated by these invariant lines, we study how the positions of these lines change under variations of the quality parameter q. In other words, we want to study the monotonicity of $z_1 = z_1(q)$ and $z_2 = z_2(q)$. We get the following (for detailed proofs and limit-case behavior $\lim_{q \to q^*_\pm} z_{1,2}$, see Appendix C; for illustrations see Figs. 3 and 4):

Proposition 2.3. *If $\delta < 0$, then $dz_{1,2}/dq > 0$ and hence both z_1 and z_2 are increasing as $q \in (1/2, q^*) \cup (q^*, 1]$. In the system's phase plane, this corresponds to a continuous counter-clockwise rotation of the two invariant lines. If $\delta > 0$, then $dz_{1,2}/dq < 0$; hence both $z_{1,2}$ are in this case decreasing as $q \in (1/2, q^*) \cup (q^*, 1]$. In the phase plane, this corresponds to a clockwise rotation of the invariant lines.*

Proposition 2.4. *The angles $\theta_{1,2} \in [-\pi/2, \pi/2]$ between each invariant line and the w_1 abscissa are decreasing with respect to the parameter q in case $\delta > 0$, and are increasing with respect to the parameter q in case $\delta < 0$. Moreover, in both cases, the angular rate of change is finite, at all $q \in (1/2, 1]$.*

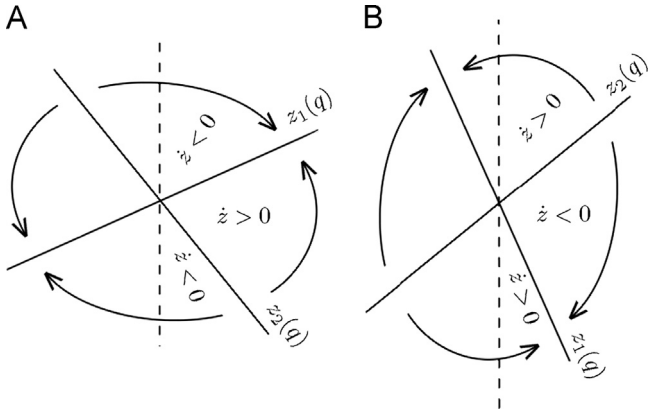

Fig. 3. Invariant lines and generic phase plane dynamics. The invariant lines are marked as $z_1(q)$ and $z_2(q)$. The arrows indicate the rotational direction of the vector field between the two invariant lines. This can be obtained in the right vertical half-plane (where we have defined our angle, $\theta \in [-\pi/2, \pi/2]$), then extended by symmetry in the opposite half-plane. For $q < q^*$ we have $z_1 > z_2$ (A). As q increases, the two invariant lines rotate: clockwise if $\delta > 0$ and anti-clockwise if $\delta < 0$. At $q = q^*$, one of the invariant lines goes through a vertical stage. For $\delta > 0$, θ_2 jumps from $-\pi/2$ to $\pi/2$, hence z_2 has a vertical asymptote at $q = q^*$, and jumps from $z_2 \to -\infty$ to $z_2 \to \infty$. For $\delta < 0$, θ_1 jumps from $\pi/2$ to $-\pi/2$, hence z_1 has a vertical asymptote at $q = q^*$, and jumps from $z_1 \to \infty$ to $z_1 \to -\infty$. In consequence, after this critical stage, for $q > q^*$, we have $z_1 < z_2$ (B) Although the rotation is continuous, either z_1 or z_2 has an infinite discontinuity, due to our definition (mod π) of the angles $\theta_1, 2$.

2.4. Unbiased case $\delta = 0$

For $\delta = 0$ the computations are simpler; however, as mentioned before, the system has an interesting critical transition which does not appear in the $\delta \neq 0$ slices (occurring from the "touching," or apparent crossing, of the two eigenvalues at $q = q^*$, as shown in Fig. 1).

Proposition 2.5. *Suppose $\delta = 0$. The phase plane of the system depends on the value of q as follows*:

i. *If $q < q^*$, then $\mu_1 = v + c$ is the larger eigenvalue, with eigendirection $z_1 = 1$ and norm of the corresponding attracting equilibria $\|w\| = 1$. $\mu_2 = (2q-1)(v-c)$ is the smaller eigenvalue, with eigendirection $z_2 = -1$ and norm of the corresponding saddle equilibria $\|w\| = \sqrt{q-1/2}$.*
ii. *If $q > q^*$, then $\mu_1 = (2q-1)(v-c)$ is the larger eigenvalue, with eigendirection $z_1 = -1$ and norm of the corresponding attracting equilibria $\|w\| = \sqrt{q-1/2}$. $\mu_2 = v + c$ is the smaller eigenvalue, with eigendirection $z_2 = 1$ and norm of the corresponding saddle equilibria $\|w\| = 1$.*
iii. *If $q = q^*$, the system contains an infinity of half-stable non-isolated equilibria (each direction will contain two opposite equilibria, describing overall an ellipse of equilibria around the origin).*

Proof. For $\delta = 0$, we have $\dot{z} = -\beta(z^2 - 1)$. The situation $q < q^*$ corresponds to $\beta > 0$, and $q > q^*$ corresponds to $\beta < 0$. Parts i and ii follow immediately. For $q = q^*$, $\dot{z} = 0$; all lines through the origin are invariant, and each contains two half-stable equilibria. In Appendix D, we show that the locus of these equilibria is an ellipse (see dotted curve in Fig. 4), and we describe its axes and foci. □

Remark 1. For $q = 1$, the attracting equilibria lay along the direction $z = -1$, so that $w_1 + w_2 = 0$. A simple way to quantify how far the stable equilibrium $w = (w_1, w_2)$ degrades from this error-free state as the quality q decreases, we can measure how much the sum $S(q) = |w_1 + w_2|$ deviates from zero, the outcome of perfect learning (Fig. 5).

In Fig. 1B, the inputs are unbiased ($\delta = 0$), and in the absence of crosstalk ($q = 1$) the inputs segregate completely. As crosstalk

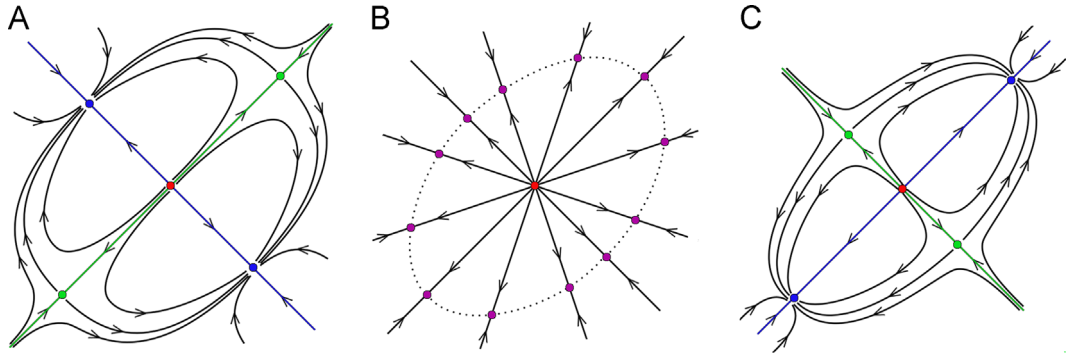

Fig. 4. Transitions of the phase plane and bifurcation at $q = q^*$, in the slice $\delta = 0$. (A) When $q > q^*$, the stable equilibria are the two vectors of norm $\sqrt{q-1/2}$ (blue dots) along the invariant line of slope $z_1 = -1$; the saddle equilibria are the two eigenvectors of norm 1 (green dots) along the invariant line of slope $z_2 = 1$. As q decreases from $q = 1$ towards $q = q^*$, the saddles remain unchanged, but the attractors gradually approach the origin (their norm $\sqrt{q-1/2}$ decreases). (B) When $q = q^*$, the system traverses a bifurcation state, characterized by an infinite number (an entire ellipse) of neutrally stable equilibria. This critical state permits the swap of stability between the two invariant lines. (C) When $q < q^*$, the stable equilibria are now the two vectors of norm 1 (blue dots) along the invariant line of slope $z_1 = 1$, while the saddle equilibria swapped to the two eigenvectors of norm $\sqrt{q-1/2}$ (green dots) along the invariant line of slope $z_2 = -1$. As q continues to decrease from $q = q^*$ towards $q = 1/2$, the attractors remain unchanged, and the saddles approach the origin (collapsing into the origin in the limit of $q \to 1/2$). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Author's personal copy

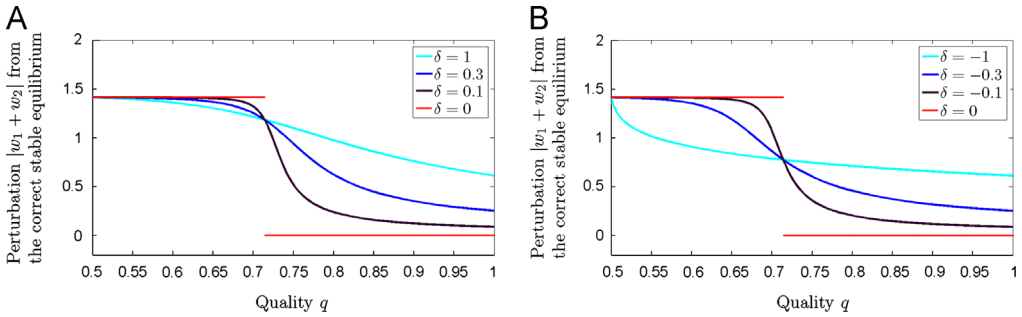

Fig. 5. $S(q)=|w_1+w_2|$ as a measure of the increasing inspecificity of the stable equilibrium, compared to its ideal state $S(1)=0$, as q decays from $q=1$. For $v=1$, $c=-0.4$, we plotted $S(q)$. (A) For $\delta \geq 0$: $\delta=1$ (cyan); $\delta=0.3$ (blue); $\delta=0.1$ (purple); $\delta=0$ (red). (B) For $\delta \leq 0$: $\delta=-1$ (cyan); $\delta=-0.3$ (blue); $\delta=-0.1$ (purple); $\delta=0$ (red). In both panels, all continuous curves for $\delta \neq 0$ concur at one point, which corresponds to the fact that, for both $\delta>0$ and $\delta<0$, the stable equilibrium at $q=q^*$ is independent on the magnitude of δ. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

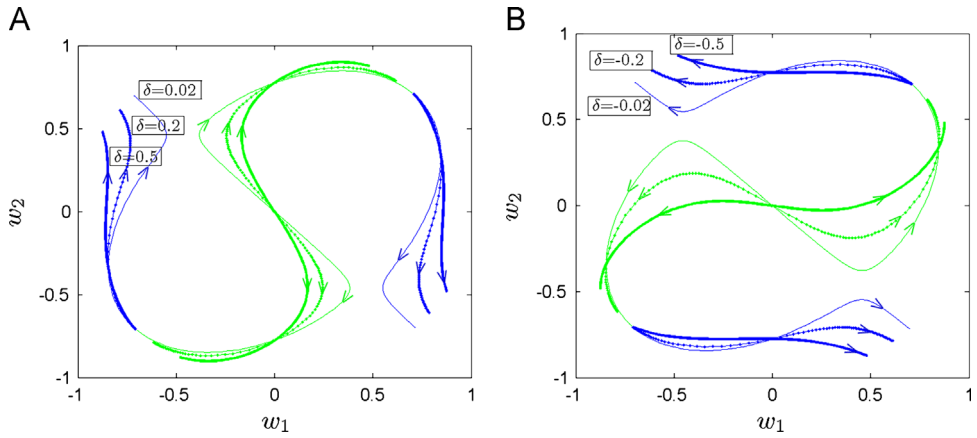

Fig. 6. Equilibria curves in the phase plane, as q changes. The blue curves represent the stable equilibrium locus, and the green curves the saddle equilibrium. (A) Plots for a few representative positive δ values: $\delta=0.02$ (thin curves), $\delta=0.2$ (thin dotted curves) and $\delta=0.5$ (thick curves). All green saddle curves concur at one point (on the vertical axis), and all blue stable curves also concur at a point, corresponding to the fact that the position of the two equilibria is independent on the magnitude of $\delta>0$. (B) Plots for a few representative negative δ values: $\delta=-0.02$ (thin curves), $\delta=-0.2$ (thin dotted curves) and $\delta=-0.5$ (thick curves). All green saddle curves concur at one point, and all blue stable curves also concur at a point (on the vertical axis), corresponding to the fact that the position of the two equilibria is independent on the magnitude of $\delta<0$. The arrows along the curves indicate the direction of increasing q. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

increases, the separation between the eigenvalues at first decreases, though the inputs remain completely segregated. However, as crosstalk increases further, the two eigenvalues equalize at the critical quality value $q^*=v/(v-c)$. With further increases in crosstalk, the inputs become completely unsegregated, and the eigenvalues now move apart. This qualitative change at q^* is a bifurcation. Note that although the qualitative behavior only changes at q^*, there is a biologically less important quantitative change: the two symmetric equilibrium weight vectors decrease continuously in length as $\sqrt{q-1/2}$, as q decreases from $q=1$ until the bifurcation at $q=q^*$, then remain of unit length for $q<q^*$.

In the slightly biased cases $\delta=-0.2$ and 0.5 (Fig. 1A and C), this overall behavior persists, although the eigenvalues always remain distinct, and there is no true bifurcation. Thus in part A, as crosstalk increases, the eigenvalues at first approach each other, and the solution remains almost segregated. At the "pseudocritical", value of $q=((2v+\delta)(2v+\delta-2c)-\delta^2)/(2v+\delta-2c)$, the eigenvalues reach their closest value (in an "avoided crossing") and then start to separate as crosstalk increases further; significantly beyond this pseudocritical value, the outcome is almost unsegregated (see Figs. 5 and 6). Of course, for q values very close to this pseudocritical value, desegregation is very rapidly increasing with increases in crosstalk (see Figs. 5 and 6), especially with very small values of δ. Thus even with slight input bias, the overall behavior, switching from segregation to unsegregation at a critical crosstalk value, resembles that seen in the unbiased situation.

2.5. Conclusions: mathematical behavior of the 2D system

Corollary 2.6. *For any combination of parameters, the phase-plane of the system (1) contains no cycles. Moreover, the system has only two (opposite) attracting equilibria, with attraction basins two open half-planes.*

Remark. The result holds more generally for an n-dimensional system, as shown in Appendix A.1.

Since we are looking at a 2-dimensional system, this means, according to the Poincaré–Bendixon theorem that the only attracting sets can be attracting equilibria. The two attracting equilibria of the system (by Proposition 2.3) lie along the invariant line corresponding to the largest eigenvalue of the covariance matrix **C**, hence their position (direction and distance to origin) depends on the values of the parameters (in particular on the quality q and bias factor δ). Fig. 6 illustrates the evolution of these points in the phase plane for a fixed $\delta \neq 0$, as q increases. (We used Matcont continuation algorithms (Dhooge et al., 2003) to numerically estimate the equilibria and draw the equilibrium curves.)

The following two paragraphs summarize the conclusions obtained throughout the previous sections:

Biased dynamics: When the system is biased (i.e., $\delta \neq 0$) the two eigenvalues of the modified input covariance matrix **EC** are always separated. The phase plane has two pairs of nonzero opposing

Author's personal copy

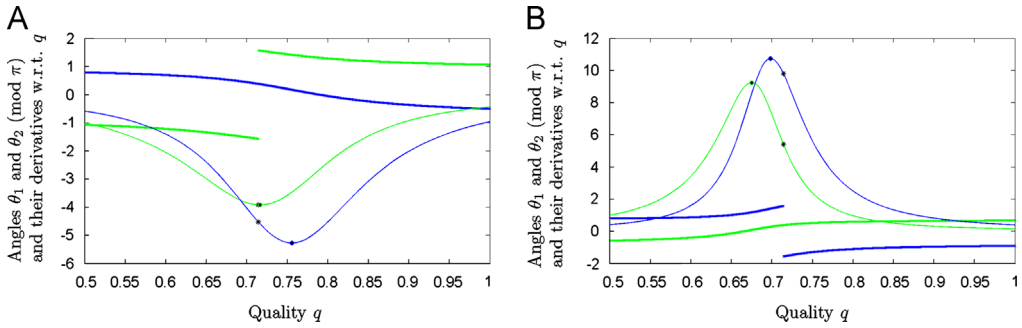

Fig. 7. Illustration of the evolution of the angles $\theta_{1,2}$ of the invariant lines with the abscissa, as q increases. In both panels, $v=1$ and $c=-0.4$. (A) $\delta=0.5$; (B) $\delta=-0.2$. The graphs of the functions are shown in thick lines, θ_1 in blue and θ_2 in green. The graphs of the derivatives are plotted in thin lines, with $d\theta_1/dq$ in blue and $d\theta_2/dq$ in green. On the graphs of the derivatives, we marked with a black star the points corresponding to $q=q^*$, and with a bullet the points of extremum (the inflection points for $\theta_{1,2}$, where the rotational speed is maximal). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

equilibria, each situated on one of the two distinct invariant lines through the origin (i.e., the two eigendirections of **EC**). The invariant line of slope z_1 corresponding to the higher eigenvalue μ_1 of **EC** contains the pair of opposing attracting equilibria; the invariant line of slope z_2 corresponding to the lower eigenvalue μ_2 of **EC** separates their two basins of attraction and also contains the pair of opposing saddles. As the parameter q increases, the invariant lines rotate (clockwise if $\delta>0$ and counter-clockwise if $\delta<0$) in a continuously differentiable manner, with an angular speed that depends on q. This rotation gets arbitrarily fast (e.g., at its point of maximal rotational speed) as $\delta\to 0$.

Unbiased dynamics: When the system is unbiased (i.e., $\delta=0$) the two eigenvalues of the modified input covariance matrix **EC** collide at the critical value of the quality parameter $q=q^*$. For any $q\neq q^*$, the phase plane has two pairs of nonzero opposing equilibria, each situated on one of the two distinct invariant lines $\binom{1}{\pm 1}$ through the origin. The invariant line corresponding to the higher eigenvalue of **EC** contains the pair of opposing attracting equilibria; the invariant line corresponding to the lower eigenvalue of **EC** separates their two basins of attraction and also contains the pair of opposing saddles. As the parameter q increases, the invariant lines remain unchanged, until they swap instantaneously as q traverses the critical state $q=q^*$ (stability-swapping *bifurcation*). At the bifurcation point, the phase plane has an entire ellipse of half-stable equilibria.

Remark. The *codimension 2 bifurcation* that occurs at $q=q^*$ in the slice $\delta=0$ can be considered a limit case of the phase–plane transition sequence obtained when increasing q, when making $\delta\to 0$ in the biased case. The rotational speed blows up to ∞ as $\delta\to 0$, and, in the $\delta=0$ slice, the rotation becomes instantaneous via what appears to be the bifurcation's "swap" of eigendirections. The evolution of the rotation speed with respect to q as $\delta\to 0$ is further illustrated in Fig. 7.

3. Alternative models: Euclidean normalization of weights versus the Oja model

The Oja rule is an elegant and classical solution to the well-known problem that unconstrained Hebbian learning is unstable (Dayan and Abbott, 2002; Oja, 1982). It has the biologically appealing feature that it is local, although it does require that the "normalizing" adjustment is proportional to the current weight. We have shown that it is still useful when some crosstalk is present, although the stable norm, and the exact direction of the learned weight vector, changes. One can imagine various other ways, possibly involving "homeostasis" or "synaptic scaling" (Turrigiano et al., 1998; Turrigiano and Nelson, 2004) of promoting stability, and some studies invoke various combinations of these mechanisms. A less biologically plausible, nonlocal, but extremely simple and highly effective method, which might capture features of any more plausible scheme and which works even for nonlinear rules, is to impose a specific norm after each weight vector update. Here we examine how crosstalk affects such "explicit" or "brute" normalization.

As before, Hebb's rule lies at the basis of the weight updates: $\Delta\mathbf{w}=\gamma y\mathbf{x}$, with $y=\mathbf{w}^T\mathbf{x}=\mathbf{x}^T\mathbf{w}$.

In other words: $\mathbf{w}(n+1)=\mathbf{w}+\gamma y\mathbf{x}$. As in the Oja model, we can think of Hebbian inspecificity being formalized as a stochastic error matrix $\boldsymbol{\epsilon}$, so that, at each time step:

$$\mathbf{w}\to\mathbf{w}+\gamma y\boldsymbol{\epsilon}x$$

Taking expectation of both sides and re-naming $\mathbf{w}=\langle\mathbf{w}\rangle$ (the long-term average of the weight vector), $\mathbf{C}=\langle\mathbf{x}\mathbf{x}^T\rangle$ (the covariance matrix of the input distribution) and $\mathbf{E}=\langle\boldsymbol{\epsilon}\rangle$ (the average error matrix), we obtain the iteration: $\mathbf{w}\to\mathbf{w}+\gamma\langle\boldsymbol{\epsilon}\mathbf{x}\mathbf{x}^T\rangle\mathbf{w}=\mathbf{w}+\gamma\mathbf{E}\mathbf{C}\mathbf{w}$.

We normalize to keep $\|\mathbf{w}\|=1$, and make no further approximations to implement this normalization biologically. We get the new iteration function that describes the average iterative process, with errors, becomes

$$f^{\mathbf{E}}(\mathbf{w})=\frac{\mathbf{w}+\gamma\mathbf{E}\mathbf{C}\mathbf{w}}{\|\mathbf{w}+\gamma\mathbf{E}\mathbf{C}\mathbf{w}\|}$$

where the "modified" covariance matrix is as before **EC**; unlike in the Oja case, **EC** is now involved in the normalization step as well. Notice that, since **EC** has positive eigenvalues, the matrix $\mathbf{I}+\gamma\mathbf{E}\mathbf{C}$ is nonsingular, hence $f^{\mathbf{E}}$ is defined for all $\mathbf{w}\in\mathbb{R}^n\backslash\{0\}$. This direct normalization confines the trajectories to the unit circle. It is easy to show that the fixed points $\mathbf{w}$ of the system $\mathbf{w}\to f^{\mathbf{E}}(\mathbf{w})$ are all normalized eigenvectors of the matrix **EC**.

The Jacobian matrix of the system around a fixed point $\mathbf{w}$ is (see Appendix A.2)

$$Df^{\mathbf{E}}_{\mathbf{w}}=\frac{(\mathbf{I}-\mathbf{w}\mathbf{w}^T)(\gamma\mathbf{A}+\mathbf{I})}{\|\mathbf{w}+\gamma\mathbf{E}\mathbf{C}\mathbf{w}\|} \tag{4}$$

Then we have the following (see Appendix A.2 for proof):

Description and stable fixed points: The system has stable fixed points iff the modified correlation matrix **EC** has a maximal eigenvalue of multiplicity one. Then, a point $\mathbf{w}$ is a stable fixed point of the system iff it is a unit eigenvector of **EC** corresponding to the unique maximal eigenvalue of **EC**.

It is clear that the phase space of this system, although not dynamically equivalent to the phase space of the corresponding Oja model, is very similar. Disregarding the origin (which is not in the domain of one, but is a repelling fixed point for the other), the other fixed points have the same qualitative behavior (stability) for both systems, if assuming γ sufficiently small. Moreover, the stability transitions occur at the same bifurcation points (where

Author's personal copy

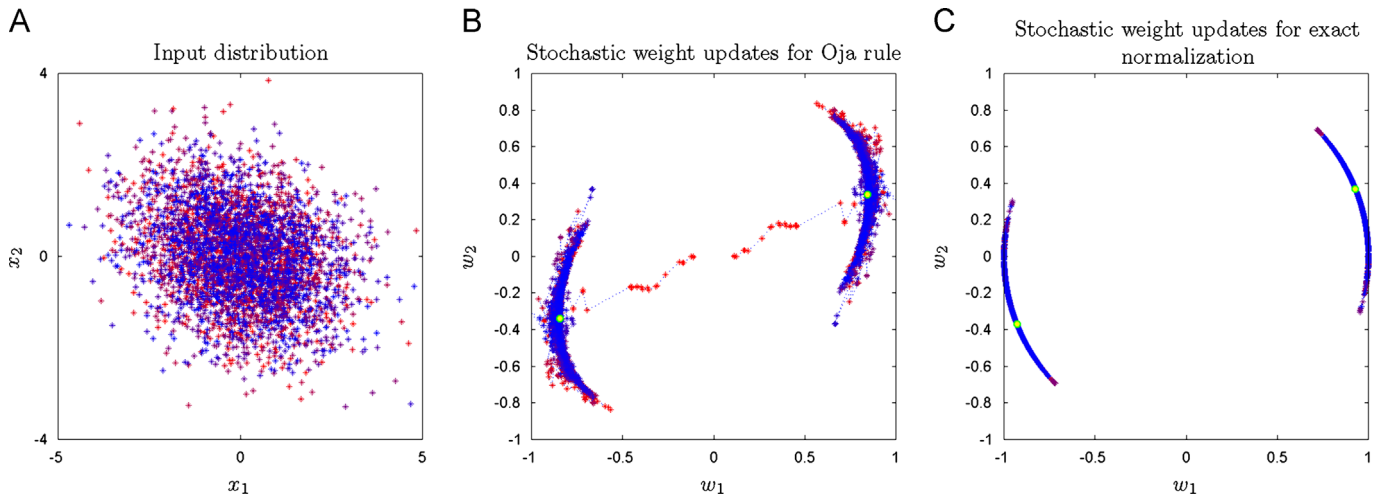

Fig. 8. Behavior of stochastic weight updates for a biased input distribution $\delta \neq 0$. (A) A discrete input sample ($N=4000$) was drawn out of a Gaussian input distribution with $\nu=1$, $c=-0.4$, $\delta=1$, and used to update the weights. (B) Depending on their initial state, the weight vector stabilizes towards small stochastic fluctuations around either one of the attracting equilibria (the pair of appropriately normalized eigenvectors corresponding to the larger eigenvalue of **EC**). (C) The corresponding iterations are shown in the case of exact normalization at each step (fewer iterations are shown in this case, since more weights, all living on the unit circle, would obstruct the clarity of the figure.) In all three panels, the points were colored update-chronologically from red to blue. We used $q=\nu/(\nu-c)\sim 0.71$; the corresponding equilibria for the deterministic, continuous-time system are marked with yellow dots in both phase spaces. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

the eigenvalues of **EC** collide with each other), and the bifurcation phase-planes are themselves similar.

In the case of the two-dimensional model discussed in this paper, the eigenvalue swap occurs as before when $q=q^*=\nu/(\nu-c)$. For example, in the unbiased case $\delta=0$, the bifurcation phase plane at $q=q^*$ again exhibits a ring attractor. Indeed, at the codimension 2 bifurcation corresponding to $\delta=0$ and $q=q^*$, the iteration function becomes: $f(\mathbf{w})=\mathbf{w}/\|\mathbf{w}\|$, which maps radially any $\mathbf{w}$ in the plane to the unit circle, and keeps it fixed thereafter.

In Section 4 we further investigate whether the long-term evolution of $\mathbf{w}$ predicted by this model is comparable with the behavior in the more realistic situation of stochastic weight updates, in discrete time and at finite learning rate. We study both the case of an explicit Euclidean normalization, and the case of a "subtractive" Taylor approximation of it (see also Radulescu et al., 2009).

4. Stochastic models. Simulations and predictions

Here we briefly study the more biologically realistic situation in which the weights update stochastically with a small finite learning rate γ. More precisely: while assuming a fixed (average) error matrix **E** throughout the process, the weight updates are driven by individual inputs (taken in our simulations to be normally distributed), rather than by the mean statistics in the negligible learning rate limit. For both types of normalization (online Oja, as described in Section 2, and Euclidean, as described in Section 3), we study in particular whether the convergence to eigenvector equilibria, and the transitions in dynamics between different values of the parameter q, still occur as in the deterministic model.

Our numerical simulations show, as expected, that convergence is conserved, in the following sense: when a pair of attracting equilibria exist for the deterministic system (i.e., **EC** has distinct eigenvalues), the discrete sequence of updating $\boldsymbol{\omega}$ eventually stabilizes to small, stochastic fluctuations around one of these two equilibria (which are, as we recall, the appropriately normalized eigenvectors corresponding to the larger eigenvalue of **EC**). This is illustrated in Figs. 8 and 9A and C. In Fig. 8, $\mathbf{x}$ is drawn out of a biased, Gaussian distribution of inputs (shown on the left), for which the two eigenvalues of **EC** are warranted to be distinct for any value of q, in particular for the value chosen here ($q=q^*=1/1.4$). In Fig. 9, the inputs are unbiased, so the same remark applies only if $q\neq q^*$. In Fig. 9A, we illustrate the case $q>q^*$, in which the attracting vectors are $\pm\sqrt{q-1/2}\binom{1}{-1}$; in Fig. 9C, we illustrate the case $q<q^*$, in which the attracting vectors are $\pm(1/\sqrt{2})\binom{1}{1}$. In both cases, the stochastic update settles to fluctuations about either one of these vectors, depending on the initial conditions.

Fig. 9B illustrates the unbiased case corresponding to the codimension 2 bifurcation in the deterministic dynamics; that is, when $q=q^*$. Recall that, in the deterministic phase-plane, this case was characterized by an ellipse of neutrally stable equilibria, so that each initial condition would converge radially towards a unique nonisolated equilibrium on this curve. This situation changes in the model driven by stochastic updates. An initial weight vector $\boldsymbol{\omega}$ will quickly be attracted towards the ellipse; however, the orbit does not fluctuate around a particular point on the curve, but rather diffuses along the curve, eventually covering densely the entire ellipse.

5. Discussion

We have proposed (Cox and Adams, 2009; Adams and Cox, 2002b) that a central problem for biological learning is that the activity-dependent processes that lead to connection strength adjustments cannot be completely synapse specific, because they must obey the laws of physics. This truism provides a new viewpoint: it raises the possibility that sophisticated learning, such as presumably occurs in the neocortex, is enabled as much by special machinery for enhancing specificity, as by special algorithms (Adams and Cox, 2006). We have suggested (Adams, 1998; Cox and Adams, 2009) that these plasticity errors are analogous to mutations, and that cortical circuitry might reduce such errors, just as "proofreading" reduces DNA copying mistakes (Kornberg and Baker, 1992). The idea of "synaptic mutation" comes in at least two forms. First, and most simply, neural activity of a connection might not only affect its strength, in a Hebbian manner, but might also affect the strength of other, existing connections ("crosstalk"). Second, synaptogenesis, perhaps initiated by spine formation, could result in the creation of a new connection (Le Bé

Author's personal copy

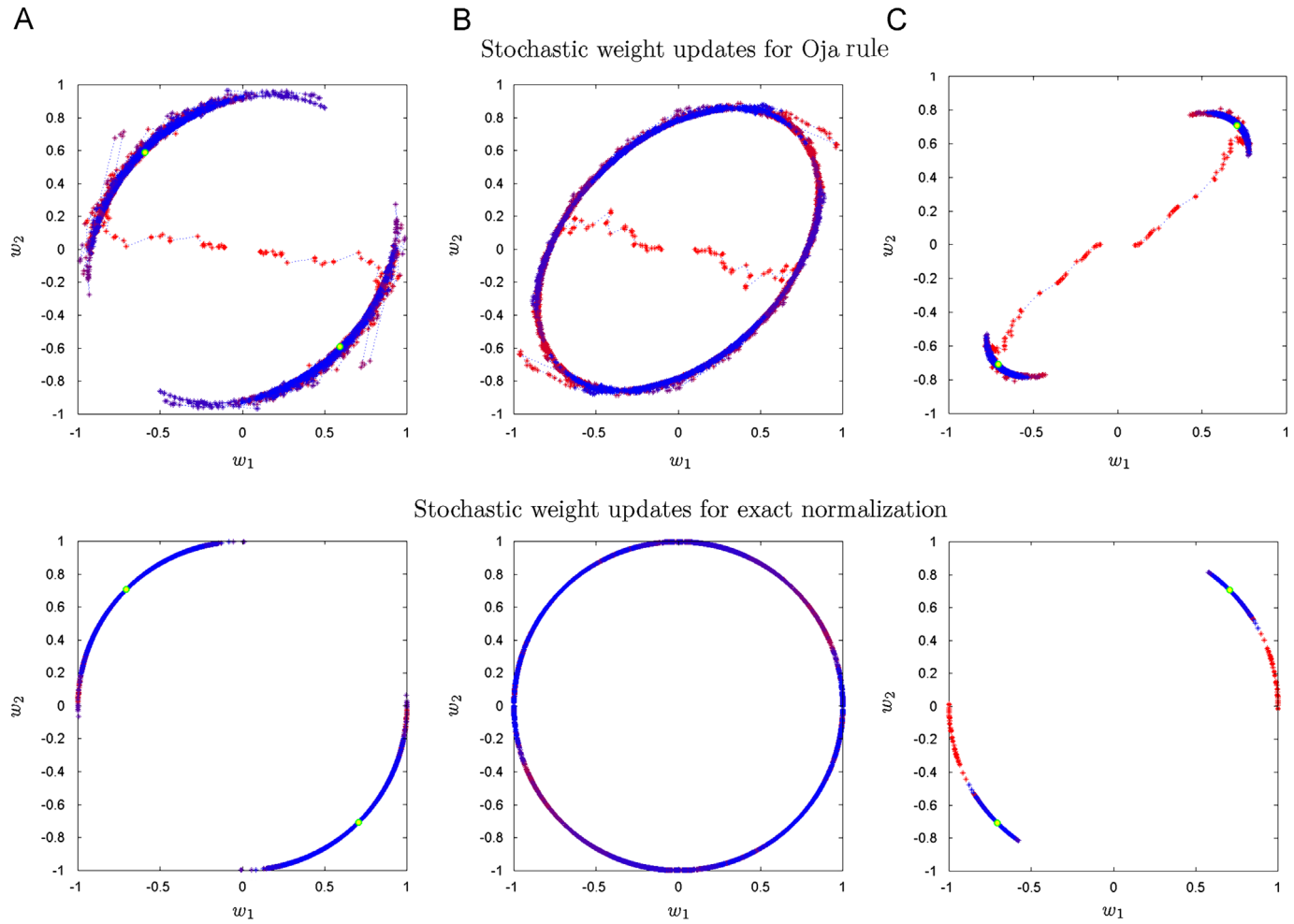

Fig. 9. Differences in stochastic behavior when q is varied, in the unbiased input case $\delta = 0$ (compare with Fig. 4). A discrete number of input vectors $\mathbf{x}(t) = (x_1(t), x_2(t))$ are drawn from a distribution with covariance matrix $\mathbf{C}$, with $v = 1$, $c = -0.4$ (so that the critical quality value $q^* = 1/1.4 \sim 0.71$). The weights $\boldsymbol{\omega} = (\omega_1(t), \omega_2(t))$, adjusting with a small learning rate $\gamma = 0.1$, are plotted in the (ω_1, ω_2) plane, with the color of the points changing chronologically from red to blue. The top panels show the behavior of the Oja model, while the bottom panels, for corresponding parameters, show the behavior for exact normalization of weights at each step. (A) For good transmission quality $q = 0.85 > q^*$, $\boldsymbol{\omega}$ is converging in the long term to a state of small fluctuations around either $\pm\sqrt{q-1/2}\,(1-1)^T$ (marked with yellow dots), depending on the initial state. The plot illustrates the trajectories for two initial states, each stabilizing around one of these opposite eigenvectors. (B) For critical transmission quality $q = q^*$, $\boldsymbol{\omega}$ converges to fluctuations around the ellipse of neutrally attracting equilibria, but will perpetually drift around, filling the ellipse, driven by input fluctuations from the mean statistics, without remaining asymptotically near any particular equilibrium state. (C) For poor transmission quality $q = 0.6 < q^*$, $\boldsymbol{\omega}$ is converging in the long term to a state of small fluctuations around $\pm(1/\sqrt{2})(11)^T$ (marked with yellow dots), depending on the initial state. The plot illustrates the trajectories for two initial states, each stabilizing around one of these opposite eigenvectors. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

and Markram, 2006; Stepanyants and Chklovskii, 2005; Zito et al., 2009). In our earliest work (Adams, 1998; Adams and Cox, 2002a), we conflated these two ideas by postulating that synapse strengthening (weakening) involved the formation (removal) of new, functional, synapses; "synaptic mutation" would occur if these new synapses are (erroneously) made or removed at inactive (or even currently non-existent) connections. However, it now seems more likely that new synapses are initially silent (Isaac et al., 1995; Liao et al., 1995; Montgomery et al., 2001), and that Hebbian LTP involves strengthening of either these silent synapses, or of existing functional synapses, in a progressive, albeit quantized manner (Petersen et al., 1998; O'Connor et al., 2005). We now prefer to focus on the clean and core issue of the accuracy of the adjustment of existing connections, with crosstalk being analogous to the base-copying error (or mutation) rate in Eigen-type molecular evolution models. Synaptogenesis (which might be activity-regulated Kwon and Sabatini, 2011) is a separate, though linked issue.

From this point of view, it seems possible that the key to overcoming the curse of dimensionality that underlies difficult, and apparently almost intractable, learning problems lies not just in finding good approximations, architectures and techniques, but also in perfecting the relevant biological plasticity apparatus. Indeed, it seems possible that problems of survival and reproduction are so diverse that no single algorithm can solve them all, so that no "universal" or "canonical" cortical circuit would be expected. In these circumstances, as Rutherford once said about physics (Birks and Segrè, 1963), neuroscience would become a type of stamp collecting. However, if every specialized algorithm relies on extraordinarily specific synaptic weight adjustment, then finding machinery that allows such specificity would be tantamount to discovering new neurobiological general principles, somewhat along the lines that established the main framework for modern biology (Darwinian evolution, Mendelian genetics, DNA structure and function, replication mechanisms etc). We have speculated that an important part of such machinery, at least in the neocortex, might lie outside the synapse itself, in the form of complex circuitry performing a proofreading operation analogous to that procuring accuracy for polynucleotide copying (Adams and Cox, 2002b, 2006, 2012). However, such machinery would be less necessary if update inaccuracy merely degraded learning, rather than preventing it. In particular even if temporarily unfavorable

Author's personal copy

(e.g., "noisy") input statistics led to imperfect learning because of Hebbian inspecificity, the degraded weights might still be a useful starting point for better learning when input statistics improve. On the other hand, if inspecificity completely prevented even partial learning, then rapid and successful learning from newly favorable statistics might be impossible. These considerations have impelled us to examine the effect of Hebbian "crosstalk" in various classical models of unsupervised learning, using both linear (Radulescu et al., 2009) and nonlinear rules (Cox and Adams, 2009) (see also Elliott, 2012).

5.1. Separate but equal: segregation without bias

In this paper we extended our previous study (Radulescu et al., 2009) of the effect of crosstalk on the simple linear Hebbian model of Oja to situations approaching the "unbiased" case where all inputs have the same statistical distribution. This case has often been invoked in discussions of the emergence of ocular dominance wiring and other forms of neural development, but it might also apply to any situation in which sets of inputs disconnect completely, or "segregate," to form pruned wiring patterns that are then "sculpted" by a more subtle synaptic learning process. (In the present model weights and activities can be negative; one would interpret negative weights as disconnections. Other updating schemes, restricting the weights to be positive, could of course lead to different long-term dynamics.) For the case of visual input, it seems likely that statistics would be similar, and positively correlated, for the two eyes, which look at the same world, and it is well known (Dayan and Abbott, 2002) that a linear Hebb rule with unbiased inputs, under either implicit or explicit normalization, leads to the symmetric, equal-weight, and thus apparently unbiological, outcome. A possible solution to this is to use a "subtractive" normalization scheme, although this also requires imposing weight limits (Miller and MacKay, 1994). It has been shown that a wide variety of nonlinear rules (Elliott, 2003), including the BCM rule (Bienenstock et al., 1982) and STDP (Elliott, 2008) can lead to ocular segregation under unbiased statistics. The key point is that segregated states can be created by typically nonlinear, "symmetry-breaking" mechanisms even when the inputs themselves do not favor particular segregated outcomes.

A natural question would be: if such segregated outcomes are an important part of normal development (which then constrain subsequent, more detailed, "refining," plasticity processes, including learning), how could the determining "unbiased" statistics arise, and conversely, how would plasticity errors, such as crosstalk, or other alterations in the form of the rule, affect the outcome? In particular, we show here that, unsurprisingly, crosstalk tends to prevent segregation, especially when the inputs are close to unbiased. This might set a limit to the use of symmetry breaking to generate specific wiring, or require special specificity-enhancing circuitry, such as "proofreading," even during development. At the very least it suggests that internally generated patterns deriving segregation, such as negative correlations induced by mutual inhibition, might have to be quite strong to overcome the desegregating effect of inevitable crosstalk.

Before exploring this further, we comment briefly about "unbias" in relation to Hebbian learning. Although here we focus on lack of bias in the second order statistics, one can also postulate unbias at all order, an assumption which greatly simplifies the study of nonlinear Hebbian plasticity, essentially eliminating the possibility of learning and restricting analysis to development. To what would unbiased high-order statistics correspond? It seems that they correspond to the radially symmetric distributions recently considered by Lyu and Simoncelli (Lyu and Simoncelli, 2009), where the joint pdf equal density contour lines are nested hyperspheres with nonGaussian spacings. One might expect that with completely unbiased (spherical) input statistics no particular direction in weight space would be favored and therefore the outcomes would be either symmetric (equal weights), or broken symmetric (various combinations of opposite but equal magnitude weights); the particular set of outcomes would be determined by the higher-order correlations, and could be quite complicated. Indeed, Elliott (2003) finds that segregated outcomes are quite typical of nonlinear Hebbian rules with unbiased statistics and shows that crosstalk can induce bifurcations in these cases (Elliott, 2012).

Recently, it has been suggested that the Oja rule (even without crosstalk) and Eigen's replication/mutation equation might be "isomorphic" (Fernando and Szathmáry, 2009; Fernando et al., 2010). Indeed both equations describe normalized growth processes. However, our work shows that the Oja equation only shows a bifurcation at a critical crosstalk value in very narrow conditions. We suggest that the important analogy lies less in detailed mathematical equivalencies, and more in the fundamental need for accuracy in elementary biological adaptational processes. In particular, it is clear that superaccurate polynucleotide copying underlies Darwinian evolution, and similarly superaccurate Hebbian plasticity might be needed for neural learning.

5.2. Error matrix and effect of crosstalk

The analysis reported here essentially shows that the well known bifurcation that occurs in linear Hebbian learning as unbiased negative correlations become positive (from segregated to unsegregated states) still occurs in the presence of crosstalk, but at a new, crosstalk-dependent critical negative correlation level. This effect is quite intuitive: crosstalk favors the unsegregated state, and therefore allows the switch to occur at negative correlation, rather than at zero correlation. Of course this situation changes dramatically as soon as any degree of bias is introduced, since now the eigenvalues of $\mathbf{C}$ become distinct, and our previous analysis (Radulescu et al., 2009) applies: crosstalk produces a smooth change in the direction of the learned weights (the dominant eigenvector of $\mathbf{EC}$). Our present analysis attempts to characterize the relation between these two regimes. In particular, we show that the smooth change can be very rapid when bias is weak.

The change in the normalization produced by crosstalk in the Oja model is largely irrelevant, and indeed one still sees the same behavior with explicit normalization (Section 3). Our analysis also gives insight into the codimension two bifurcation that occurs for unbiased inputs at the critical quality q^*, via an ellipse of half-stable equilibria. For very small input bias, the motion "towards the ellipse" becomes extremely rapid when approaching q^* (Fig. 6 and Appendix C), which permits the exchange of stability between the two invariant lines at this critical state (Fig. 3). Although a true bifurcation is only seen for unbiased inputs and for negligible learning rates, the behavior remains practically indistinguishable from a bifurcation even with slightly biased inputs and bounded learning rates. An example was already discussed in our previous paper (see Fig. 4 in Radulescu et al., 2009).

A similar situation occurs with models of phase transitions: a true bifurcation of the dynamics only occurs in the "thermodynamic limit," but this is effectively established even for quite small systems (Sollich, 1994). We have previously called attention to the analogy between Hebbian learning and molecular evolution (Adams and Cox, 2002a; Adams, 1998), with crosstalk playing the role of mutation. In Eigen's evolution model (Eigen, 1971), the transition from the ordered, living, state to the disordered, chemical, state is quite sharp even for polynucleotide lengths ~ 50, though a true phase transition (identical to that of the

Author's personal copy

surface of the 2-dimensional Ising model) is only seen with unlimited chains (Saakian and Hu, 2006). Interestingly, the model becomes easiest to analyze in this limit, and the relevant dimensionless control parameter is simply the product of the mutation rate and the (binary) chain length (for binary strings). Although we analyzed here the $n=2$ case, we assume that weights are specified with unlimited bit resolution (i.e., reals). In this case the dimensionless control parameter, equivalent to that in the thermodynamic limit of the Eigen model, is q. In the standard Eigen model, the mutation rate is the same at all chain positions.

In our earlier paper (Radulescu et al., 2009) we illustrated our general mathematical results using the particular case of $\mathbf{E}$ with all diagonal elements equal and all off-diagonal elements equal (i.e., we assumed that the Hebbian adjustment of any weight was equally affected by error, and does not depend either on the strength of that weight or its identity). Such "isotropicity" seems a reasonable first assumption, like neglecting bumps on an inclined plane in mechanics. However, experimentally crosstalk has been, for technical reasons, mostly documented between anatomically neighboring synapses (Harvey and Svoboda, 2007; Bi, 2002; Engert and Bonhoeffer, 1997). In that paper we justified isotropicity based on the finding that individual cortical connections are composed of multiple synapses which are scattered over the dendritic tree (Radulescu et al., 2009; Varga et al., 2011; Chen et al., 2011; Jia et al., 2010). To some extent, crosstalk locality could be captured using different, nonisotropic, forms for $\mathbf{E}$ without affecting our main conclusions. However, in the present paper we present result only for the $n=2$ input case, where the distinction between local and global crosstalk does not arise.

5.3. Relevance to ocular dominance and general developmental mechanisms

A useful though rather fuzzy distinction can be drawn between developmental mechanisms which generate sets of connections ("circuits"), or, perhaps, "incipient" or "potential" connections (Adams and Cox, 2002b; Stepanyants and Chklovskii, 2005) that can be made actual without axo-dendritic rewiring merely by adding postsynaptic spines or presynaptic "drumsticks" (Anderson and Martin, 2001; Sherman and Guillery, 2001), and "learning," which refines (perhaps in crucial ways) the overall framework established by development. This distinction is related to that between "Nature" and "Nurture," or, in the context of Chomskyan linguisitics, "principles" and "parameters." The Oja model encapsulates this distinction in minimal form: by definition, when the inputs are unbiased there can be no learning about the external world, and only two outcomes are possible, which we call segregated or unsegregated. The classic biological example is that in many species early in development a geniculate axon diffusely innervates a patch of layer IV of cortex (though it does not necessarily contact all the neurons whose dendrites ramify in that patch), but then retracts from stripes within that patch that become selectively innervated by axons carrying signals from the other eye. Cells within a stripe then become largely monocular, although they develop different selectivities for different stimulus features such as orientation. In the Oja model, segregation appears in response to unbiased (or, effectively, nearly unbiased) inputs at a critical level of negative correlation, which depends on the degree of crosstalk. In real animals, segregation appears before the onset of visual experience, and is thought to be driven by unbiased inputs generated by spontaneous firing. While one might expect crosstalk to hinder segregation (since it tends to equalize weights), our results show this is not quite correct in the Oja model: it merely shifts the critical degree of (unbiased) correlation required. Various proposals exist for how such inputs can induce segregation even when correlations are positive (Miller and MacKay, 1994; Elliott, 2003) and it's likely that crosstalk will also have the same weak effect here. Indeed, Elliott (2012) has shown that while crosstalk can induce a bifurcation from segregation to unsegregation in a weight-dependent model, the critical crosstalk value (his equation (3.8)) can be shifted by changing correlation. Thus, the endogenous developmental machinery that creates circuits probably does not require great Hebbian accuracy (and might not require Hebbian machinery at all Crowley and Katz, 2000; Paik and Ringach, 2011). If the aforementioned postulated layer VI proofreading circuit (Adams and Cox, 2006, 2012) underlies accuracy, it would not be needed until learning begins, consistent with evidence that the final stages of layer VI circuitry (for example, feedback to relay cells) is late to develop. Indeed, much of the initial pruning that takes place in development might serve to improve the accuracy of proofreading circuitry essential for true learning.

Once detailed, and biased, sensory input occurs, it can drive quantitative adjustments in the already correctly segregated circuits, involving both synapse-strength change and stabilization and un-silencing of new spines (and removal of weak synapses). However, even in the highly simplified Oja model, appropriate adjustment now requires great accuracy, and therefore presumably "proofreading," especially when correlation bias is weak.

6. Conclusion

The inspecific Oja rule does not generically show bifurcations with variation in the crosstalk parameter. In this paper we analyze an interesting special case which does show a bifurcation: when the input statistics are unbiased. We also describe the behavior in the vicinity of this special case, which is practically indistinguishable from a bifurcation. Essentially in this region "learning" changes rather abruptly from being dominated by second-order input statistics (at sufficiently low crosstalk) to being dominated by the internal pattern of crosstalk itself. However, we regard this behavior as being biologically rather uninteresting, since synaptic mechanisms are usually accurate enough that it never occurs. The one exception would be during development, where near-unbiased statistics might be used by the brain to induce initial selective wiring. Our results suggest that even in this case, high Hebbian accuracy might be required. However, extreme accuracy is probably most essential for nonlinear learning from higher-order statistics (Cox and Adams, 2009; Elliott, 2012). We suggest that the unsupervised learning "environment" should be construed as the complete set of input correlations to the plastic neuron or network. In evolution models, the fitness of each possible sequence must be specified (and could be unique for each sequence). Input correlations, at all orders, would similarly define growth rates for every possible weight vector. However, in both cases, update errors would set sharp limits to the acquisition of information by adaptation. Perhaps the necessary extraordinary accuracy is achieved in analogous ways in both biological and neural adaptation.

Appendix A. Stability of equilibria

A.1. Stability of equilibria for the inspecific Oja rule

Consider the inspecific Oja system $d\mathbf{w}/dt = f^{\mathbf{E}}(\mathbf{w})$, with $f^{\mathbf{E}}(\mathbf{w}) = \gamma[\mathbf{ECw}-(\mathbf{w}^T\mathbf{Cw})\mathbf{w}]$. Then:

Lemma A1. *The Jacobian of the system around an equilibrium w* **is**:

$$Df^{\mathbf{E}}_{\mathbf{w}} = \gamma[\mathbf{EC}-2\mathbf{w}(\mathbf{Cw})^T-(\mathbf{w}^T\mathbf{Cw})\mathbf{I}] \quad (5)$$

Author's personal copy

Proof. Call $g(\mathbf{w})=(\mathbf{w}^T\mathbf{C}\mathbf{w})\mathbf{w}$, so $f^{\mathbf{E}}(\mathbf{w})=\gamma[\mathbf{E}\mathbf{C}\mathbf{w}-g(\mathbf{w})]$

$g_i(\mathbf{w})=(\mathbf{w}^T\mathbf{C}\mathbf{w})w_i$

If $i\neq j$:

$$\frac{\partial g_i}{\partial w_j}(\mathbf{w})=\frac{\partial}{\partial w_j}\left(\sum_{k,l}C_{kl}w_kw_l\right)w_i=2\left(\sum_k C_{kj}w_k\right)w_i=2[\mathbf{C}\mathbf{w}]_jw_i$$

If $i=j$:

$$\frac{\partial g_i}{\partial w_i}(\mathbf{w})=\frac{\partial}{\partial w_i}\left(\sum_{k,l}C_{kl}w_kw_l\right)w_i+\sum_{k,l}C_{kl}w_kw_l=2\left(\sum_k C_{ki}w_k\right)w_i+\mathbf{w}^T\mathbf{C}\mathbf{w}=2[\mathbf{C}\mathbf{w}]_iw_i+\mathbf{w}^T\mathbf{C}\mathbf{w}$$

So:

$Dg_{\mathbf{w}}=2\mathbf{w}(\mathbf{C}\mathbf{w})^T+(\mathbf{w}^T\mathbf{C}\mathbf{w})\mathbf{I}$ □

Proposition A1. *Suppose* **EC** *has a multiplicity one largest eigenvalue. An equilibrium* **w** *is a local hyperbolic attractor for the system iff it is an eigenvector corresponding to the maximal eigenvalue of* **EC**.

Proof. Recall that a vector **w** is an equilibrium for the system if either $\mathbf{w}=\mathbf{0}$ or if **w** is an eigenvector of **EC** with eigenvalue $\lambda_{\mathbf{w}}$, normalized so that $\|w\|_{\mathbf{C}}=\lambda_{\mathbf{w}}$. Clearly, $Df^{\mathbf{E}}_{\mathbf{w}=\mathbf{0}}=\gamma\mathbf{E}\mathbf{C}$, which has at least one eigenvalue >0; hence $\mathbf{w}=\mathbf{0}$ is unstable.

Fix now an eigenvector $\mathbf{w}\neq\mathbf{0}$ of **EC**, with $\mathbf{E}\mathbf{C}\mathbf{w}=\lambda_{\mathbf{w}}\mathbf{w}$. Then:

$$Df^{\mathbf{E}}_{\mathbf{w}}\mathbf{w}=\gamma[\mathbf{E}\mathbf{C}\mathbf{w}-2\mathbf{w}(\mathbf{C}\mathbf{w})^T\mathbf{w}-(\mathbf{w}^T\mathbf{C}\mathbf{w})\mathbf{w}] \tag{6}$$

$$Df^{\mathbf{E}}_{\mathbf{w}}\mathbf{w}=\gamma[-2\mathbf{w}\mathbf{w}^T\mathbf{C}\mathbf{w}]=-2\gamma\lambda_{\mathbf{w}}\mathbf{w} \tag{7}$$

Recall that the vector **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 $\mathbf{E}\mathbf{C}\mathbf{v}=\lambda_{\mathbf{v}}\mathbf{v}$, and $\langle\mathbf{w},\mathbf{v}\rangle_{\mathbf{C}}=\mathbf{w}^T\mathbf{C}\mathbf{v}=0$. We calculate:

$$Df^{\mathbf{E}}_{\mathbf{w}}\mathbf{v}=\gamma[\mathbf{E}\mathbf{C}\mathbf{v}-2\mathbf{w}\mathbf{w}^T\mathbf{C}\mathbf{v}-\lambda_{\mathbf{w}}\mathbf{v}] \tag{8}$$

$$Df^{\mathbf{E}}_{\mathbf{w}}\mathbf{v}=\gamma[(\lambda_{\mathbf{v}}-\lambda_{\mathbf{w}})\mathbf{v}-2\langle\mathbf{w},\mathbf{v}\rangle_{\mathbf{C}}\mathbf{w}]=-\gamma[\lambda_{\mathbf{w}}-\lambda_{\mathbf{v}}]\mathbf{v} \tag{9}$$

So $\mathcal{B}$ is also a basis of eigenvectors for $Df^{\mathbf{E}}_{\mathbf{w}}$. The corresponding eigenvalues are $-2\gamma\lambda_{\mathbf{w}}$ (for the eigenvector **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 **w** to be a hyperbolic attractor for the system is that all the eigenvalues of $Df^{\mathbf{E}}_{\mathbf{w}}$ are <0. Since the learning rate γ and the eigenvalue $\lambda_{\mathbf{w}}$ are both >0, this condition is further 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 **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, or in other words if **w** is in the direction of the principal eigenvector of **EC**). □

A.2. Stability of equilibria for the exact normalization rule

Consider the direct normalization system $\Delta\mathbf{w}=f^{\mathbf{E}}(\mathbf{w})$, with $f^{\mathbf{E}}(\mathbf{w})=\gamma[\mathbf{E}\mathbf{C}\mathbf{w}-(\mathbf{w}^T\mathbf{C}\mathbf{w})\mathbf{w}]$. Then:

Lemma A2. *The Jacobian of the system around a fixed point* **w** *is*:

$$Df^{\mathbf{E}}_{\mathbf{w}}=\frac{(\mathbf{I}-\mathbf{w}\mathbf{w}^T)(\gamma\mathbf{A}+\mathbf{I})}{\|\mathbf{w}+\gamma\mathbf{E}\mathbf{C}\mathbf{w}\|} \tag{10}$$

Proof. In order to slightly simplify the notation, we call $\mathbf{A}=\mathbf{E}\mathbf{C}$, $\mathbf{u}=\mathbf{w}+\gamma\mathbf{E}\mathbf{C}\mathbf{w}$ and $a=\|\mathbf{u}\|$, notation which we will use whenever it is convenient. The vector **w** is a fixed point of $f(\mathbf{w})$ iff $\mathbf{w}+\gamma\mathbf{A}\mathbf{w}=a\mathbf{w}$, i.e. **w** is a unit eigenvector of **A** (with the Euclidean norm). To establish the stability, we compute the Jacobian matrix of $f^{\mathbf{E}}$ at each fixed point.

Fix $j\in\overline{1,n}$. Then, for any $i\neq j$:

$$\frac{\partial u_i}{\partial w_j}=\frac{\partial}{\partial w_j}(w_i+\gamma[\mathbf{A}\mathbf{w}]_i)=\gamma A_{ij}$$

When $i=j$, we have similarly:

$$\frac{\partial u_j}{\partial w_j}=\frac{\partial}{\partial w_j}(w_j+\gamma[\mathbf{A}\mathbf{w}]_j)=1+\gamma A_{jj}$$

Hence, overall:

$$\frac{\partial}{\partial w_j}\|\mathbf{u}\|^2=2u_j(1+\gamma A_{jj})+\sum_{i\neq j}2u_i\gamma A_{ij}=2u_j+2\gamma\sum_i u_iA_{ij}=2u_j+2\gamma[\mathbf{A}^T\mathbf{u}]_j \tag{11}$$

In matrix form:

$$\frac{\partial}{\partial\mathbf{w}}\|\mathbf{u}\|^2=2\gamma\mathbf{A}^T\mathbf{u}+2\mathbf{u} \tag{12}$$

Now, fix $i\in\overline{1,n}$. For $j\neq i$, we have:

$$\frac{\partial f^{\mathbf{E}}_i}{\partial w_j}=\frac{\gamma A_{ij}\|u\|-u_i\|\mathbf{u}\|^{-1}[\gamma\mathbf{A}^T\mathbf{u}+\mathbf{u}]_j}{\|\mathbf{u}\|^2}$$

For $j=i$, we have:

$$\frac{\partial f^{\mathbf{E}}_i}{\partial w_i}=\frac{(1+\gamma A_{ii})\|\mathbf{u}\|-u_i\|\mathbf{u}\|^{-1}[\gamma\mathbf{A}^T\mathbf{u}+\mathbf{u}]_i}{\|\mathbf{u}\|^2}$$

Rewritten in matrix form:

$$\frac{\partial f^{\mathbf{E}}}{\partial\mathbf{w}}=\frac{\gamma}{a}\mathbf{A}-\frac{1}{a^3}(\gamma\mathbf{u}\mathbf{u}^T\mathbf{A}+\mathbf{u}\mathbf{u}^T)+\frac{1}{a}\mathbf{I}=\frac{1}{a}\left(\mathbf{I}-\frac{1}{a^2}\mathbf{u}\mathbf{u}^T\right)(\gamma\mathbf{A}+\mathbf{I}) \tag{13}$$

where **I** is the appropriate size identity matrix.

At any fixed point **w**, for which automatically $\|\mathbf{w}\|=1$ and $\mathbf{A}\mathbf{w}=\lambda_{\mathbf{w}}\mathbf{w}$, where $1+\lambda_{\mathbf{w}}\gamma=a$, we have that:

$$\mathbf{u}\mathbf{u}^T=(\mathbf{w}+\gamma\mathbf{A}\mathbf{w})(\mathbf{w}+\gamma\mathbf{A}\mathbf{w})^T=(1+2\gamma\lambda_{\mathbf{w}}+\gamma^2\lambda_{\mathbf{w}}^2)\mathbf{w}\mathbf{w}^T=(1+\lambda_{\mathbf{w}}\gamma)^2\mathbf{w}\mathbf{w}^T \tag{14}$$

The Jacobian at a fixed point **w** can be then simplified to:

$$\frac{\partial f^{\mathbf{E}}}{\partial\mathbf{w}}=\frac{1}{a}(\mathbf{I}-\mathbf{w}\mathbf{w}^T)(\gamma\mathbf{A}+\mathbf{I}) \qquad \square \tag{15}$$

Proposition A2. *Suppose* **EC** *has a multiplicity one largest eigenvalue. A fixed point* **w** (*i.e., a unit eigenvector of* **EC**) *is attracting iff it is an eigenvector corresponding to the maximal eigenvalue of* **EC**.

Proof. We calculate:

$$\frac{\partial f^{\mathbf{E}}}{\partial\mathbf{w}}(\mathbf{w})=\frac{1}{a}(\mathbf{I}-\mathbf{w}\mathbf{w}^T)(\gamma\lambda_{\mathbf{w}}+1)\mathbf{w}=\frac{1}{a}(\mathbf{w}-\mathbf{w}(\mathbf{w}^T\mathbf{w}))(\gamma\lambda_{\mathbf{w}}+1)=0 \tag{16}$$

Complete **w** to a basis of eigenvectors of **A** (not necessarily mutually orthogonal). Let $\mathbf{v}\neq\mathbf{w}$ any of the vectors in this basis (with eigenvalue $\lambda_{\mathbf{v}}$), and consider $\mathbf{z}=\mathbf{v}-[\mathbf{w}^T\mathbf{v}]\mathbf{w}$ the projection of **v** on the orthogonal complement of **w**. Then:

$$(\gamma\mathbf{A}+\mathbf{I})(\mathbf{z})=(\gamma\lambda_{\mathbf{v}}+1)\mathbf{v}-(\gamma\lambda_{\mathbf{w}}+1)[\mathbf{w}^T\mathbf{v}]\mathbf{w} \tag{17}$$

Hence

$$\frac{\partial f^{\mathbf{E}}}{\partial\mathbf{w}}(\mathbf{z})=\frac{1}{a}(\gamma\lambda_{\mathbf{v}}+1)(\mathbf{v}-[\mathbf{w}^T\mathbf{v}]\mathbf{w})=\frac{1}{a}(\gamma\lambda_{\mathbf{v}}+1)\mathbf{z}=\frac{\gamma\lambda_{\mathbf{v}}+1}{\gamma\lambda_{\mathbf{w}}+1}\mathbf{z} \tag{18}$$

A normalized eigenvector **w** of **EC** is stable as a fixed point of the system if all the eigenvalues of the Jacobian $\partial f^{\mathbf{E}}/\partial\mathbf{w}$ at **w** are less

Author's personal copy

than one in absolute value:

$$\left|\frac{\gamma\lambda_{\mathbf{v}}+1}{\gamma\lambda_{\mathbf{w}}+1}\right|<1$$

Since all eigenvalues of **EC** are positive (recall that **EC** is diagonalizable with the dot product $\langle\cdot,\cdot\rangle_{\mathbf{C}}$), this is equivalent to $(\gamma\lambda_{\mathbf{v}}+1)/(\gamma\lambda_{\mathbf{w}}+1)<1$, and thus to $\lambda_{\mathbf{w}}>\lambda_{\mathbf{v}}$ for every $\mathbf{v}\neq\mathbf{w}$. □

Appendix B. An extension to higher dimensions

Theorem B1. *Suppose the modified covariance matrix* **EC** *has a unique maximal eigenvalue* λ_1. *Then the two eigenvectors* $\pm\mathbf{w}_{\mathbf{EC}}$ *corresponding to* λ_1, *normalized such that* $\|\mathbf{w}\|_{\mathbf{C}}=\lambda_1$, *are the only two attractors of the system. More precisely, the phase space is divided into two basins of attraction, of* $\mathbf{w}_{\mathbf{EC}}$ *and* $-\mathbf{w}_{\mathbf{EC}}$ *respectively, separated by the subspace* $\langle\mathbf{w},\mathbf{w}_{\mathbf{EC}}\rangle=0$.

Proof. We perform the change of variable $\mathbf{u}=\sqrt{\mathbf{C}}\mathbf{w}$, so that $\mathbf{u}^T\mathbf{u}=\mathbf{w}^T\mathbf{C}\mathbf{u}$. Notice that $\sqrt{\mathbf{C}}$ is also a symmetric matrix, and that $\mathbf{w}=\sqrt{\mathbf{C}}^{-1}\mathbf{u}$; the system then becomes:

$$\sqrt{\mathbf{C}}^{-1}\dot{\mathbf{u}}=\mathbf{E}\mathbf{C}\sqrt{\mathbf{C}}^{-1}\mathbf{u}-(\mathbf{u}^T\sqrt{\mathbf{C}}^{-1}\mathbf{C}\sqrt{\mathbf{C}}^{-1}\mathbf{u})\sqrt{\mathbf{C}}^{-1}\mathbf{u}=\mathbf{E}\sqrt{\mathbf{C}}\mathbf{u}-(\mathbf{u}^T\mathbf{u})\sqrt{\mathbf{C}}^{-1}\mathbf{u}$$

or equivalently:

$$\dot{\mathbf{u}}=\sqrt{\mathbf{C}}\mathbf{E}\sqrt{\mathbf{C}}\mathbf{u}-(\mathbf{u}^T\mathbf{u})\mathbf{u}=\mathbf{A}\mathbf{u}-(\mathbf{u}^T\mathbf{u})\mathbf{u} \tag{19}$$

where, of course, we defined $\mathbf{A}=\sqrt{\mathbf{C}}\mathbf{E}\sqrt{\mathbf{C}}$. Clearly, **A** a symmetric matrix, having the same eigenvalues as **EC**. More precisely, **w** is an eigenvector of **EC** with eigenvalue μ iff $\sqrt{\mathbf{C}}\mathbf{v}$ is an eigenvector of **A** with eigenvalue μ. Moreover: any two distinct eigenvectors $\mathbf{v}\neq\mathbf{w}$ of **EC** are known to be orthogonal, hence any two distinct eigenvectors of **A** are orthogonal in the regular Euclidean dot product: $(\sqrt{\mathbf{C}}\mathbf{v})^T(\sqrt{\mathbf{C}}\mathbf{w})=\mathbf{v}^T\sqrt{\mathbf{C}}\sqrt{\mathbf{C}}\mathbf{w}=\mathbf{v}^T\mathbf{C}\mathbf{w}=\mathbf{0}$.

Consider then **v** to be the principal component of **A** (i.e., the eigenvector corresponding to its maximal eigenvalue), and let $\mathbf{u}=\mathbf{u}(t)$ be a trajectory of the system (19). We want to observe the evolution in time of the angle between the variable vector **u** and the fixed vector **v**.

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

We differentiate and obtain:

$$\begin{aligned}-\|\mathbf{v}\|\sin(\theta)\dot{\theta}&=\frac{1}{\|\mathbf{u}\|^2}\left[\langle\mathbf{v},\dot{\mathbf{u}}\rangle\cdot\|\mathbf{u}\|-\langle\mathbf{v},\mathbf{u}\rangle\frac{\langle\mathbf{u},\dot{\mathbf{u}}\rangle}{\|\mathbf{u}\|}\right]\\&=\frac{(\mathbf{v}^T\dot{\mathbf{u}})\|\mathbf{u}\|^2-(\mathbf{v}^T\mathbf{u})(\mathbf{u}^T\mathbf{u})}{\|\mathbf{u}\|^3}\end{aligned} \tag{20}$$

The numerator of this expression

$$\begin{aligned}h(\mathbf{u})&=(\mathbf{v}^T\dot{\mathbf{u}})(\mathbf{u}^T\mathbf{u})-(\mathbf{v}^T\mathbf{u})(\mathbf{u}^T\dot{\mathbf{u}})\\&=(\mathbf{u}^T\mathbf{u})(\mathbf{v}^T[\mathbf{A}\mathbf{u}-(\mathbf{u}^T\mathbf{u})\mathbf{u}])+(\mathbf{u}^T\mathbf{v})-(\mathbf{v}^T\mathbf{u})(\mathbf{u}^T-[\mathbf{A}\mathbf{u}-(\mathbf{u}^T\mathbf{u})\mathbf{u}])\\&=(\mathbf{u}^T\mathbf{u})(\mathbf{v}^T\mathbf{A}\mathbf{u})-(\mathbf{v}^T\mathbf{u})(\mathbf{u}^T\mathbf{A}\mathbf{u})\end{aligned}$$

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:

$$\begin{aligned}h(\mathbf{u})&=[(\mathbf{P}\mathbf{u}^T)(\mathbf{P}\mathbf{u})][(\mathbf{P}\mathbf{v}^T)\mathbf{D}(\mathbf{P}\mathbf{u})]-[(\mathbf{P}\mathbf{v}^T)(\mathbf{P}\mathbf{u})][(\mathbf{P}\mathbf{u}^T)\mathbf{D}(\mathbf{P}\mathbf{u})]\\&=(\mathbf{z}^T\mathbf{z})(\mathbf{y}^T\mathbf{D}\mathbf{z})-(\mathbf{y}^T\mathbf{z})(\mathbf{z}^T\mathbf{D}\mathbf{z})\end{aligned}$$

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 $\lambda_1>\lambda_2\geq\ldots\geq\lambda_n$ is the largest eigenvalue of **EC**, assumed to have multiplicity one. Hence:

$$\begin{aligned}h(\mathbf{u})&=(\mathbf{z}^T\mathbf{z})(\mathbf{y}^T\mathbf{D}\mathbf{z})-(\mathbf{y}^T\mathbf{z})(\mathbf{z}^T\mathbf{D}\mathbf{z})\\&=(\mathbf{z}^T\mathbf{z})\lambda_1(\mathbf{y}^T\mathbf{z})-(\mathbf{y}^T\mathbf{z})(\mathbf{z}^T\mathbf{D}\mathbf{z})\\&=(\mathbf{y}^T\mathbf{z})[\lambda_1(\mathbf{y}^T\mathbf{z})-\mathbf{z}^T\mathbf{D}\mathbf{z}]\\&=(\mathbf{y}^T\mathbf{z})\left[\lambda_1\textstyle\sum z_j^2-\sum\lambda_j z_j^2\right]=(\mathbf{y}^T\mathbf{z})\left[\sum(\lambda_1-\lambda_j)z_j^2\right]\end{aligned}$$

Hence, if $\mathbf{y}^T\mathbf{z}>0$, then $h(\mathbf{u})>0$. In other words: if $\mathbf{v}^T\mathbf{u}>0$ then $-\|\mathbf{v}\|\sin(\theta)\dot{\theta}>0$, hence that $\dot{\theta}<0$. 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 towards the principal eigenvector $\mathbf{w}_{\mathbf{EC}}$ of the matrix **EC**. □

Appendix C. Sensitivity analysis

This is a technical section, in which we calculate how the invariant directions $z_{1,2}$ change when varying q.

Remark. In order to simplify further computations, we rewrite

$$\begin{aligned}z_{1,2}&=\frac{-q\delta\pm\sqrt{\Delta}}{2\beta}=\frac{-q\delta\pm\sqrt{q^2\delta^2+4\beta^2+4\beta\delta(1-q)}}{2\beta}\\&=-\frac{1}{2}\left(\frac{q\delta}{\beta}\right)\pm\frac{1}{2}\,\mathrm{sign}(\beta)\sqrt{\left(\frac{q\delta}{\beta}\right)^2+\frac{4[\beta+(1-q)\delta]}{\beta}}\end{aligned}$$

Call $\gamma=q\delta/\beta=q\delta/(cq+(1-q)v)$. Then $(1-q)\delta/\beta=(\delta-c\gamma)/v$, and hence

$$z_{1,2}=-\frac{1}{2}\gamma\pm\frac{1}{2}\,\mathrm{sign}(\beta)\sqrt{\eta}$$

where

$$\eta=\frac{\Delta}{\beta^2}=\gamma^2+4\left[1+\frac{\delta-c\gamma}{v}\right]$$

Then we can use the chain rule to express $dz_{1,2}/dq=(dz_{1,2}/d\gamma)\cdot(d\gamma/dq)$.

Lemma. *The derivative* $d\gamma/dq=\delta v/\beta^2$. *Also, for* $q\in(1/2,q^*)\cup(q^*,1]$ (*i.e., where* $\beta\neq0$), *we have*

$$\frac{dz_{1,2}}{d\gamma}=\frac{1}{2}\left[\frac{\pm\left(q\delta-\frac{2c\beta}{v}\right)}{\sqrt{\Delta}}-1\right]<0$$

Proof.

$$\frac{d\gamma}{dq}=\frac{d}{dq}\left(\frac{q\delta}{\beta}\right)=\frac{\delta[\beta-q\dot{\beta}]}{\beta^2}=\frac{\delta[\beta-q(c-v)]}{\beta^2}=\frac{\delta v}{\beta^2}$$

For $q\in(1/2,q^*)\cup(q^*,1]$, we also have directly that

$$\frac{dz_{1,2}}{d\gamma}=-\frac{1}{2}\pm\frac{1}{2}\frac{\mathrm{sign}(\beta)\frac{d\eta}{d\gamma}}{\sqrt{\eta}}=-\frac{1}{2}\pm\frac{1}{2}\,\mathrm{sign}(\beta)\frac{\gamma-\frac{2c}{v}}{\sqrt{\gamma^2+4\left[1+\frac{\delta-c\gamma}{v}\right]}} \tag{21}$$

$$\frac{dz_{1,2}}{d\gamma}=\frac{1}{2}\left[\pm\frac{\beta\left(\gamma-\frac{2c\beta}{v}\right)}{\sqrt{\Delta}/|\beta|}-1\right]=\frac{1}{2}\left[\frac{\pm\left(q\delta-\frac{2c\beta}{v}\right)}{\sqrt{\Delta}}-1\right] \tag{22}$$

Since

$$\left(\gamma^2+4\left[1+\frac{\delta-c\gamma}{v}\right]\right)-\left(\gamma-\frac{2c}{v}\right)^2=\frac{4[v(v+\delta)-c^2]}{v^2}>0,$$

it follows that

Author's personal copy

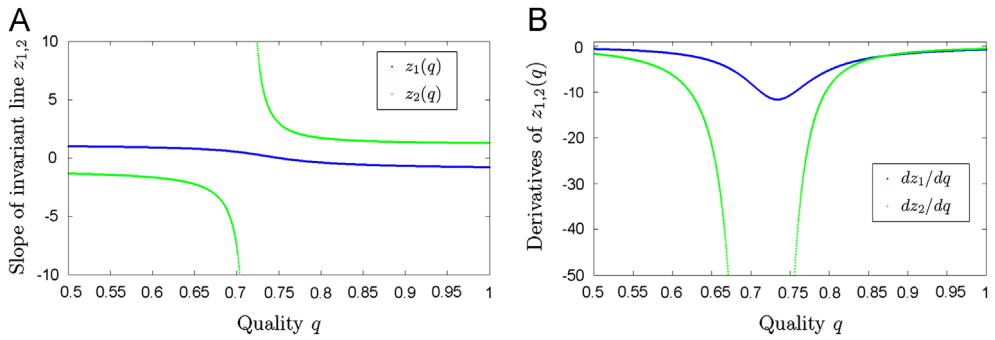

Fig. A1. Slopes of invariant lines (A) and their change as q is varied (B). In both panels, the other parameters values were fixed to $\nu=1$, $c=-0.4$ and $\delta=0.2$. Notice that, in accordance with Proposition C.1, z_1 and its derivative dz_1/dq are continuous (blue curves) on $[1/2,1]$, while z_2 and its derivative dz_2/dq (green curves) have vertical asymptotes at $q=q^*=\nu/(\nu-c)\sim 0.71$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

$$\sqrt{\gamma^2+4\left[1+\frac{\delta-c\gamma}{\nu}\right]} > |\gamma-\frac{2c}{\nu}| \geq \pm\left(\gamma-\frac{2c}{\nu}\right)$$

and hence

$$\frac{|\mathrm{sign}(\beta)\left(\gamma-\frac{2c}{\nu}\right)|}{\sqrt{\gamma^2+4\left[1+\frac{\delta-c\gamma}{\nu}\right]}}<1$$

It immediately follows from (2) that $dz_{1,2}/d\gamma<0$. □

Corollary. *The slope of the invariant lines changes with respect to q according to*

$$\frac{dz_{1,2}}{dq}=\frac{\delta\nu}{2\beta^2}\left[\frac{\pm\left(q\delta-\frac{2c\beta}{\nu}\right)}{\sqrt{\Delta}}-1\right]$$

hence $\mathrm{sign}(dz_{1,2}/dq)=-\mathrm{sign}(\delta)$ *for* $q\in(1/2,q^*)\cup(q^*,1]$.

Proof. The conclusion follows directly from the chain rule that $dz_{1,2}/dq=(\delta\nu/\beta^2)\cdot(dz_{1,2}/d\gamma)$. □

At this stage, we can distinguish two cases: $\delta<0$ and $\delta>0$. We analyze in detail the case $\delta>0$. The other is very similar (although not symmetric about $\delta=0$), and we will only state the results, and show some graphic illustrations.

Proposition C1. *If* $\delta>0$, *then* $dz_{1,2}/dq<0$; *hence both* $z_{1,2}$ *are decreasing as* $q\in(1/2,q^*)\cup(q^*,1]$. *Furthermore, the monotonicity, asymptotes and end behavior of the functions* $z_{1,2}(q)$ *are sketched in the following table*:

q	$1/2$		q^*		1
z_1	1	$\searrow$	$\frac{1-q^*}{q^*}$	$\searrow$	$\frac{\delta-\sqrt{4c^2+\delta^2}}{-2c}$
z_2	$-1-\frac{\delta}{2(\nu+c)}$	$\searrow$	$-\infty\|^{\infty}$	$\searrow$	$\frac{\delta+\sqrt{4c^2+\delta^2}}{-2c}$

Remark C1. In the system's phase plane, this corresponds to a continuous clockwise rotation of the two invariant lines (the vertical asymptote at q^* corresponds to the z_2 line going through a the vertical position). A phase–plane sketch of this process is shown in Fig. 2, and the graphs of the actual functions $z_{1,2}(q)$ and of their derivatives $dz_{1,2}/dq$, for some fixed values of the parameters $\nu,c,\delta>0$, are shown in Fig. A1.

Remark C2. Clearly from the table, it is shown that the angular position of the two equilibria at $q=q^*$ does not depend on the bias δ. It can be easily shown that the norm of these points is also independent on δ. For example, the norm of the stable equilibrium is

$$\|\mathbf{w}\|^2=\frac{\mu_1(z_1^2+1)}{\nu z_1^2+2cz_1+(\nu+\delta)}=\frac{(1-q^*)c+q^*(\nu+\delta)}{\nu(1-q^*)^2+2cq^*(1-q^*)+(\nu+\delta)q^{*2}}\times q^{*2}(z_1^2+1) \tag{23}$$

$$\|\mathbf{w}\|^2=\frac{(1-q^*)c+q^*(\nu+\delta)}{q^*[(1-q^*)c+q^*(\nu+\delta)]}\cdot q^{*2}(z_1^2+1)=\frac{1-2q^*+2q^{*2}}{q^*} \tag{24}$$

Hence the position of the two equilibria at critical quality is the same for all bias values $\delta>0$.

Proof. The monotonicity follows from the Corollary. The limit values follow from direct computation. For example:

$$\lim_{q\to q^*_+} z_2=\lim_{q\to q^*_+}\frac{-q\delta-\sqrt{\Delta}}{2\beta}=\lim_{q\to q^*_+}\frac{-q^*\delta}{0^-}=+\infty$$

$$\lim_{q\to q^*_-} z_2=\lim_{q\to q^*_-}\frac{-q^*\delta}{0^+}=-\infty$$

$$\lim_{q\to q^*} z_1=\lim_{q\to q^*}\frac{-q\delta+\sqrt{\Delta}}{2\beta}\cdot\frac{-q\delta-\sqrt{\Delta}}{-q\delta-\sqrt{\Delta}}=\lim_{q\to q^*}\frac{\beta+(1-q)\delta}{q\delta+\sqrt{\Delta}}=\frac{1-q^*}{q^*}$$ □

Remark C3. If $\delta<0$, then $dz_{1,2}/dq>0$ and hence both z_1 and z_2 are increasing as $q\in(1/2,q^*)\cup(q^*,1]$. In the system's phase plane, this corresponds to a continuous counter-clockwise rotation of the two invariant lines.

Proposition C2. *For* $\delta>0$, *the angle* $\theta_{1,2}\in[-\pi/2,\pi/2]$ *between each invariant line and the* w_1 *abscissa is decreasing with respect to the parameter q. Moreover, the angular rate of change is finite, at all* $q\in(1/2,1]$.

q	$\frac{1}{2}$		q^*		1
$d\theta_1/dq$		$(-)$	$-\frac{\det(C)}{\nu\delta(1-2*q+2q^2)}$	$(-)$	
$d\theta_2/dq$		$(-)$	$-\frac{\nu}{q^2\delta}$	$(-)$	

Proof. The relation between the slope z and the actual angle θ is given by: $z=\tan\theta$ (we will avoid indices wherever there is no danger of confusion). Hence, for $q\in(1/2,q^*)\cup(q^*,1]$, we have

$$\cos^2(\theta)\cdot\frac{d\theta}{d\gamma}\Longrightarrow\frac{d\theta}{d\gamma}=\frac{dz}{d\gamma}\cdot\frac{dz}{d\gamma}\cdot\frac{1}{z^2+1}.$$

Author's personal copy

So

$$\frac{d\theta}{dq}=\frac{d\gamma}{dq}\cdot\frac{d\theta}{d\gamma}=\frac{\delta v}{2\beta^2}\cdot\frac{dz}{d\gamma}\cdot\frac{1}{z^2+1} \tag{25}$$

hence $\text{sign}(d\theta/dq)=-\text{sign}(\delta)$, for all $q\in(1/2,q^*)\cup(q^*,1]$.

We yet have to check that the rate of change $d\gamma/dq$ remains finite (i.e., does not blow up to $-\infty$) as $q\to q^*$. Elaborating on (25) we have, for $q\in(1/2,q^*)\cup(q^*,1]$:

$$\begin{aligned}\lim_{q\to q^*}\frac{d\theta_2}{dq}&=\lim_{q\to q^*}\frac{\delta v}{2\beta^2}\cdot\frac{-\left(q\delta-\frac{2c\beta}{v}\right)-\sqrt{\Delta}}{\sqrt{\Delta}}\cdot\frac{4\beta^2}{\Delta+4\beta^2-2q\delta\sqrt{\Delta}}\\&=\frac{2\delta v}{q^*\delta}\cdot\frac{-q^*\delta-q^*\delta}{2q^{*2}\delta^2+2q^{*2}\delta^{*2}}=-\frac{v}{q^{*2}\delta}\end{aligned} \tag{26}$$

We also notice that

$$\lim_{q\to q^*}z_1=\frac{1-q^*}{q^*}\Longrightarrow\lim_{q\to q^*}(z_1^2+1)=\frac{1-2q^*+2q^{*2}}{q^{*2}} \tag{27}$$

and that

$$\begin{aligned}\lim_{q\to q^*}\frac{dz_1}{dq}&=\lim_{q\to q^*}\frac{\delta v}{2\beta^2}\cdot\frac{\left(q\delta-\frac{2c\beta}{v}\right)-\sqrt{\Delta}}{\sqrt{\Delta}}=\lim_{q\to q^*}\frac{\delta v}{2\beta^2}\cdot\frac{\left(q\delta-\frac{2c\beta}{v}\right)-\sqrt{\Delta}}{\sqrt{\Delta}}\\&\quad\times\frac{\left(q\delta-\frac{2c\beta}{v}\right)+\sqrt{\Delta}}{\left(q\delta-\frac{2c\beta}{v}\right)+\sqrt{\Delta}}\\&=\lim_{q\to q^*}\frac{\delta v}{2\beta^2\sqrt{\Delta}}\cdot\frac{\left(q\delta-\frac{2c\beta}{v}\right)^2-\Delta}{\left(q\delta-\frac{2c\beta}{v}\right)+\sqrt{\Delta}}\\&=\lim_{q\to q^*}\frac{\delta v}{2\beta^2\sqrt{\Delta}}\cdot\frac{\frac{-4\beta^2}{v^2}(c^2-v^2-v\delta)}{\left(q\delta-\frac{2c\beta}{v}\right)+\sqrt{\Delta}}\\&=\lim_{q\to q^*}\frac{\delta v}{2\beta^2\sqrt{\Delta}}\cdot\frac{-4\beta^2\det(\mathbf{C})}{v^2}\cdot\frac{1}{\left(q\delta-\frac{2c\beta}{v}\right)+\sqrt{\Delta}}\\&=\frac{\delta v}{2\beta^2q^*\delta}\cdot\frac{-4\beta^2\det(\mathbf{C})}{v^2}\cdot\frac{1}{2q^*\delta}=\frac{-\det(\mathbf{C})}{v\delta q^{*2}}\end{aligned} \tag{28}$$

Combining (26) and (28), we have

$$\lim_{q\to q^*}\frac{d\theta_2}{dq}=\frac{dz_1}{dq}\frac{1}{z_1^2+1}=\frac{-\det(\mathbf{C})}{v\delta q^{*2}}\cdot\frac{q^{*2}}{1-2q^*+2q^{*2}}=\frac{-\det(\mathbf{C})}{v\delta(1-2q^*+2q^{*2})}\quad\square \tag{29}$$

Appendix D. Description of the ellipse attractor

For unbiased inputs $\delta=0$ and critical quality $q=q^*$, **EC** has a double eigenvalue $\mu=v+c=(2q^*-1)(v-c)$. The eigenspace of **EC** is $\mathbb{R}^2$, hence each direction (i.e. slope $z=\tan\theta\in[-\infty,+\infty]$) produces two equilibria, normalized as follows:

$$\begin{aligned}\|\mathbf{w}\|^2&=\frac{\mu(z^2+1)}{vz^2+2cz+v}=\frac{(v+c)[\tan^2\theta+1]}{v\tan^2\theta+2c\tan\theta+v}\\&=\frac{v+c}{v\sin^2\theta+2c\sin\theta\cos\theta+v\cos^2\theta}\\&=\frac{v+c}{v+c\sin(2\theta)}\end{aligned} \tag{30}$$

We show that this is the polar equation of an ellipse with foci along the first diagonal $\theta=\pi/4$. Indeed, under a clockwise rotation by $-\pi/4$, Eq. (30) becomes

$$\begin{aligned}\rho^2&=\frac{v+c}{v+c\sin\left(2\left[\theta-\frac{\pi}{4}\right]\right)}=\frac{v+c}{v+c\cos(2\theta)}\\&==\frac{v+c}{v[\cos^2\theta+\sin^2\theta]+c[\cos^2\theta-\sin^2\theta]}\\&=\frac{v+c}{(v+c)\cos^2\theta+(v-c)\sin^2\theta}\\&==\frac{v+c}{\sqrt{v^2-c^2}}\cdot\frac{\sqrt{v^2-c^2}}{(v+c)\cos^2\theta+(v-c)\sin^2\theta}\end{aligned} \tag{31}$$

In polar coordinates, this is the equation of an ellipse

$$\rho^2=\frac{a^2b^2}{a^2\cos^2\theta+b^2\sin^2\theta}$$

with radial coordinate $\rho=\|\mathbf{w}\|(\sqrt{v^2-c^2}/(v+c))$ and angular coordinate θ, semi-major radius $a=\sqrt{v+c}$ and semi-minor radius $b=\sqrt{v-c}$.

References

Adams, P., 1998. Hebb and Darwin. Journal of Theoretical Biology 195, 419–438.

Adams, P., Cox, K., 2002a. Synaptic Darwinism and neocortical function. Neurocomputing 42, 197–214.

Adams, P., Cox, K., 2002b. A new interpretation of thalamocortical circuitry. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 357, 1767–1779.

Adams, P., Cox, K., 2006. A neurobiological perspective on building intelligent devices. Neuromorphic Engineering 3, 2–8.

Adams, P., Cox, K., 2012. From life to mind: two prosaic miracles. In: Simeonov, P., Smith, A., Ehresmann, A. (Eds.), Integral Biomathics, vol. 67. Springer, pp. 147–154.

Anderson, J., Martin, K., 2001. Does bouton morphology optimize axon length? Nature Neuroscience 4, 1166–1167.

Atick, J.J., Redlich, A.N., 1992. What does the retina know about natural scenes? Neural Computation 4, 196–210.

Baldwin, J., 1909. The influence of Darwin on theory of knowledge and philosophy. Psychological Review 16, 207.

Bell, A., Sejnowski, T., 1995. An information-maximization approach to blind separation and blind deconvolution. Neural Computation 7, 1129–1159.

Bi, G., 2002. Spatiotemporal specificity of synaptic plasticity: cellular rules and mechanisms. Biological Cybernetics 87, 319–332.

Bienenstock, E., Cooper, L., Munro, P., 1982. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. The Journal of Neuroscience 2, 32–48.

Birks, J., Segrè, E., 1963. Rutherford at Manchester. Physics Today 16, 71.

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

Calvin, W., 1996. The Cerebral Code: Thinking a Thought in the Mosaics of the Mind. The MIT Press.

Changeux, J., Courrège, P., Danchin, A., 1973. A theory of the epigenesis of neuronal networks by selective stabilization of synapses. Proceedings of the National Academy of Sciences 70, 2974.

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

Cooper, L., 2004. Theory of Cortical Plasticity. World Scientific Pub Co Inc.

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

Crowley, J., Katz, L., 2000. Early development of ocular dominance columns. Science 290, 1321.

Dayan, P., Abbott, L., 2002. Theoretical neuroscience: computational and mathematical modeling of neural systems. Philosophical Psychology 15.

Dhooge, A., Govaerts, W., Kuznetsov, Y.A., 2003. Matcont: a matlab package for numerical bifurcation analysis of odes. ACM Transactions on Mathematical Software (TOMS) 29, 141–164.

Edelman, G., 1987. Neural Darwinism: The Theory of Neuronal Group Selection. Basic Books.

Eigen, M., 1971. Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58, 465–523.

Elliott, T., 2003. An analysis of synaptic normalization in a general class of Hebbian models. Neural Computation 15, 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, 2253–2307.

Elliott, T., 2012. Cross-talk induces bifurcations in nonlinear models of synaptic plasticity. Neural Computation, 1–68.

Author's personal copy

Elliott, T., Shadbolt, N., 2002. Multiplicative synaptic normalization and a nonlinear Hebb rule underlie a neurotrophic model of competitive synaptic plasticity. Neural Computation 14, 1311–1322.
Engert, F., Bonhoeffer, T., 1997. Synapse specificity of long-term potentiation breaks down at short distances. Nature 388, 279–284.
Fernando, C., Szathmáry, E., 2009. Chemical, Neuronal and Linguistic Replicators. Towards an Extended Evolutionary Synthesis Cambridge. MIT Press, MA.
Fernando, C., Goldstein, R., Szathmáry, E., 2010. The neuronal replicator hypothesis. Neural Computation 22, 2809–2857.
Földiák, P., 1990. Forming sparse representations by local anti-Hebbian learning. Biological Cybernetics 64, 165–170.
Goodhill, G., 1993. Topography and ocular dominance: a model exploring positive correlations. Biological Cybernetics 69, 109–118.
Harvey, C., Svoboda, K., 2007. Locally dynamic synaptic learning rules in pyramidal neuron dendrites. Nature 450, 1195–1200.
Hinton, G., Nowlan, S., 1987. How learning can guide evolution. Complex Systems 1, 495–502.
Hyvärinen, A., Oja, E., 1998. Independent component analysis by general nonlinear Hebbian-like learning rules. Signal Processing 64, 301–313.
Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. Wiley-Interscience, vol. 26.
Isaac, J.T., Nicoll, R.A., Malenka, R.C., 1995. Evidence for silent synapses: implications for the expression of ltp. Neuron 15, 427–434.
Jerne, N., 1994. Antibodies and learning: selection versus instruction. Biology and Computation: A Physicist's Choice 2, 278.
Jia, H., Rochefort, N.L., Chen, X., Konnerth, A., 2010. Dendritic organization of sensory input to cortical neurons in vivo. Nature 464, 1307–1312.
Kornberg, A., Baker, T.A., 1992. DNA Replication. WH Freeman, New York, vol. 5.
Kwon, H.B., Sabatini, B.L., 2011. Glutamate induces de novo growth of functional spines in developing cortex. Nature 474, 100–104.
Le Bé, J.V., Markram, H., 2006. Spontaneous and evoked synaptic rewiring in the neonatal neocortex. Proceedings of the National Academy of Sciences of the United States of America 103, 13214–13219.
Liao, D., Hessler, N.A., Malinow, R., et al., 1995. Activation of postsynaptically silent synapses during pairing-induced ltp in ca1 region of hippocampal slice. Nature 375, 400–403.
Linsker, R., 1986. From basic network principles to neural architecture: emergence of orientation columns. Proceedings of the National Academy of Sciences of the United States of America 83, 8779.
Lyu, S., Simoncelli, E., 2009. Nonlinear extraction of independent components of natural images using radial Gaussianization. Neural Computation 21, 1485–1519.
Malsburg, C., 1973. Self-organization of orientation sensitive cells in the striate cortex. Biological Cybernetics 14, 85–100.
Miller, K., MacKay, D., 1994. The role of constraints in Hebbian learning. Neural Computation 6, 100–126.
Miller, K., Keller, J., Stryker, M., 1989. Ocular dominance column development: analysis and simulation. Science 245, 605.
Montgomery, J.M., Pavlidis, P., Madison, D.V., 2001. Pair recordings reveal all-silent synaptic connections and the postsynaptic expression of long-term potentiation. Neuron 29, 691–701.
O'Connor, D.H., Wittenberg, G.M., Wang, S.S.H., 2005. Graded bidirectional synaptic plasticity is composed of switch-like unitary events. Proceedings of the National Academy of Sciences of the United States of America 102, 9679–9684.
Oja, E., 1982. Simplified neuron model as a principal component analyzer. Journal of Mathematical Biology 15, 267–273.
Olshausen, B., et al., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609.
Otto, S., 2009. The evolutionary enigma of sex. The American Naturalist 174, S1–S14.
Paik, S.B., Ringach, D.L., 2011. Retinal origin of orientation maps in visual cortex. Nature Neuroscience 14, 919–925.
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 of the United States of America 95, 4732–4737.
Radulescu, A., Cox, K., Adams, P., 2009. Hebbian errors in learning: an analysis using the Oja model. Journal of Theoretical Biology 258, 489–501.
Ridley, M., 2001. Mendel's Demon: Gene Justice and Complexity of Life.
Saakian, D., Hu, C., 2004. Eigen model as a quantum spin chain: Exact dynamics. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics 69 021913–1.
Saakian, D., Hu, C., 2006. Exact solution of the Eigen model with general fitness functions and degradation rates. Proceedings of the National Academy of Sciences of the United States of America 103, 4935.
Schuster, P., Swetina, J., 1988. Stationary mutant distributions and evolutionary optimization. Bulletin of Mathematical Biology 50, 635–660.
Sherman, S., Guillery, R., 2001. Exploring the Thalamus. Academic Press.
Smith, J., Szathmary, E., 1997. The Major Transitions in Evolution. OUP Oxford.
Sollich, P., 1994. Finite-size effects in learning and generalization in linear perceptrons. Journal of Physics A: Mathematical and General 27, 7771.
Srinivasan, M., Laughlin, S., Dubs, A., 1982. Predictive coding: a fresh view of inhibition in the retina. Proceedings of the Royal Society of London Series B Biological Sciences 216, 427–459.
Stepanyants, A., Chklovskii, D., 2005. Neurogeometry and potential synaptic connectivity. Trends in Neurosciences 28, 387–394.
Swetina, J., Schuster, P., 1982. Self-replication with errors: a model for polynucleotide replication. Biophysical Chemistry 16, 329–345.
Swindale, N., 1996. The development of topography in the visual cortex: a review of models. Network: Computation in Neural Systems 7, 161–247.
Turrigiano, G., Nelson, S., 2004. Homeostatic plasticity in the developing nervous system. Nature Reviews Neuroscience 5, 97–107.
Turrigiano, G., Leslie, K., Desai, N., Rutherford, L., Nelson, S., 1998. Activity-dependent scaling of quantal amplitude in neocortical neurons. Nature, 892–895.
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 of the United States of America 108, 15420–15425.
Willshaw, D., Von Der Malsburg, C., 1976. How patterned neural connections can be set up by self-organization. Proceedings of the Royal Society of London Series B Biological Sciences 194, 431–445.
Wimbauer, S., Wenisch, O., Miller, K.D., van Hemmen, L., 1997a. Development of spatiotemporal receptive fields of simple cells: I. Model formulation Biological Cybernetics 77, 453–461.
Wimbauer, S., Wenisch, O., van Hemmen, L., Miller, K.D., 1997b. Development of spatiotemporal receptive fields of simple cells: II. Simulation and Analysis Biological Cybernetics 77, 463–477.
Wimbauer, S., Gerstner, W., van Hemmen, L., 1998. Analysis of a correlation-based model for the development of orientation-selective receptive fields in the visual cortex. Network: Computation in Neural Systems 9, 449–466.
Young, J., 1979. Learning as a process of selection and amplification. Journal of the Royal Society of Medicine 72, 801.
Zito, K., Scheuss, V., Knott, G., Hill, T., Svoboda, K., 2009. Rapid functional maturation of nascent dendritic spines. Neuron 61, 247–258.