LETTER **Communicated by Steven Zucker**

Hebbian Learning of Recurrent Connections: A Geometrical Perspective

Mathieu N. Galtier
m.galtier@jacobs-university.de
Olivier D. Faugeras
Olivier.Faugeras@inria.fr
NeuroMathComp Project Team, INRIA Sophia-Antipolis Méditerranée, 06902 Sophia Antipolis, France

Paul C. Bressloff
bresloff@math.utah.edu
Department of Mathematics, University of Utah, Salt Lake City, UT 84112, U.S.A., and Mathematical Institute, University of Oxford, Oxford OX1 3LB, U.K.

We show how a Hopfield network with modifiable recurrent connections undergoing slow Hebbian learning can extract the underlying geometry of an input space. First, we use a slow and fast analysis to derive an averaged system whose dynamics derives from an energy function and therefore always converges to equilibrium points. The equilibria reflect the correlation structure of the inputs, a global object extracted through local recurrent interactions only. Second, we use numerical methods to illustrate how learning extracts the hidden geometrical structure of the inputs. Indeed, multidimensional scaling methods make it possible to project the final connectivity matrix onto a Euclidean distance matrix in a high-dimensional space, with the neurons labeled by spatial position within this space. The resulting network structure turns out to be roughly convolutional. The residual of the projection defines the nonconvolutional part of the connectivity, which is minimized in the process. Finally, we show how restricting the dimension of the space where the neurons live gives rise to patterns similar to cortical maps. We motivate this using an energy efficiency argument based on wire length minimization. Finally, we show how this approach leads to the emergence of ocular dominance or orientation columns in primary visual cortex via the self-organization of recurrent rather than feedforward connections. In addition, we establish that the nonconvolutional (or long-range) connectivity is patchy and is co-aligned in the case of orientation learning.

 © 2012 Massachusetts Institute of Technology

1 Introduction

Activity-dependent synaptic plasticity is generally thought to be the basic cellular substrate underlying learning and memory in the brain. In Donald Hebb (1949) postulated that learning is based on the correlated activity of synaptically connected neurons: if both neurons A and B are active at the same time, then the synapses from A to B and B to A should be strengthened proportionally to the product of the activity of A and B. However, as it stands, Hebb's learning rule diverges. Therefore, various modifications of Hebb's rule have been developed, which basically take one of three forms (see Gerstner & Kistler, 2002, and Dayan & Abbott, 2001). First, a decay term can be added to the learning rule so that each synaptic weight is able to "forget" what it previously learned. Second, each synaptic modification can be normalized or projected on different subspaces. These constraint-based rules may be interpreted as implementing some form of competition for energy between dendrites and axons (for details, see Miller, 1996; Miller & MacKay, 1996; Ooyen, 2001). Third, a sliding threshold mechanism can be added to Hebbian learning. For instance, a postsynaptic threshold rule consists of multiplying the presynaptic activity and the subtraction of the average postsynaptic activity from its current value, which is referred to as covariance learning (see Sejnowski & Tesauro, 1989). Probably the best known of these rules is the BCM rule presented in Bienenstock, Cooper, and Munro (1982). It should be noted that history-based rules can also be defined without changing the qualitative dynamics of the system. Instead of considering the instantaneous value of the neurons' activity, these rules consider its weighted mean over a time window (see Földiák, 1991; Wallis & Baddeley, 1997). Recent experimental evidence suggests that learning may also depend on the precise timing of action potentials (see Bi & Poo, 2001). Contrary to most Hebbian rules that detect only correlations, these rules can also encode causal relationships in the patterns of neural activation. However, the mathematical treatment of these spike-timing-dependent rules is much more difficult than rate-based ones.

Hebbian-like learning rules have often been studied within the framework of unsupervised feedfoward neural networks (see Oja, 1982; Bienenstock et al., 1982; Miller & MacKay, 1996; Dayan & Abbott, 2001). They also form the basis of most weight-based models of cortical development, assuming fixed lateral connectivity (e.g., Mexican hat) and modifiable vertical connections (see the review of Swindale, 1996).[1] In these developmental models, the statistical structure of input correlations provides a mechanism for spontaneously breaking some underlying symmetry of the neuronal

[1] Only a few computational studies consider the joint development of lateral and vertical connections including Bartsch and Van Hemmen (2001) and Miikkulainen, Bednar, Choe, and Sirosh (2005).

receptive fields, leading to the emergence of feature selectivity. When such correlations are combined with fixed intracortical interactions, there is a simultaneous breaking of translation symmetry across the cortex, leading to the formation of a spatially periodic cortical feature map. A related mathematical formulation of cortical map formation has been developed in Takeuchi and Amari (1979) and Bressloff (2005) using the theory of self-organizing neural fields. Although very irregular, the two-dimensional cortical maps observed at a given stage of development, can be unfolded in higher dimensions to get smoother geometrical structures. Indeed, Bressloff, Cowan, Golubitsky, Thomas, and Wiener (2001) suggested that the network of orientation pinwheels in V1 is a direct product between a circle for orientation preference and a plane for position, based on a modification of the ice cube model of Hubel and Wiesel (1977). From a more abstract geometrical perspective, Petitot (2003) has associated such a structure to a 1-jet space and used this to develop some applications to computer vision. More recently, more complex geometrical structures such as spheres and hyperbolic surfaces that incorporate additional stimulus features such as spatial frequency and textures, were considered, respectively, in Bressloff and Cowan (2003) and Chossat and Faugeras (2009).

In this letter, we show how geometrical structures related to the distribution of inputs can emerge through unsupervised Hebbian learning applied to recurrent connections in a rate-based Hopfield network. Throughout this letter, the inputs are presented as an external nonautonomous forcing to the system and not an initial condition, as is often the case in Hopfield networks. It has previously been shown that in the case of a single fixed input, there exists an energy function that describes the joint gradient dynamics of the activity and weight variables (see Dong & Hopfield, 1992). This implies that the system converges to an equilibrium during learning. We use averaging theory to generalize the above result to the case of multiple inputs, under the adiabatic assumption that Hebbian learning occurs on a much slower timescale than both the activity dynamics and the sampling of the input distribution. We then show that the equilibrium distribution of weights, when embedded into $\mathbb{R}^k$ for sufficiently large integer k, encodes the geometrical structure of the inputs. Finally, we numerically show that the embedding of the weights in two dimensions ($k = 2$) gives rise to patterns that are qualitatively similar to experimentally observed cortical maps, with the emergence of features columns and patchy connectivity. In contrast to standard developmental models, cortical map formation arises via the self-organization of recurrent connections rather than feedforward connections from an input layer. Although the mathematical formalism we introduce here could be extended to most of the rate-based Hebbian rules in the literature, we present the theory for Hebbian learning with decay because of the simplicity of the resulting dynamics.

The use of geometrical objects to describe the emergence of connectivity patterns has previously been proposed by Amari in a different context.

Based on the theory of information geometry, Amari considers the geometry of the set of all the networks and defines learning as a trajectory on this manifold for perceptron networks in the framework of supervised learning (see Amari, 1998) or for unsupervised Boltzmann machines (see Amari, Kurata, & Nagaoka, 1992). He uses differential and Riemannian geometry to describe an object that is at a larger scale than the cortical maps considered here. Moreover, Zucker and colleagues are currently developing a nonlinear dimensionality-reduction approach to characterize the statistics of natural visual stimuli (see Zucker, Lawlor, & Holtmann-Rice, 2011; Coifman, Maggioni, Zucker, & Kevrekidis, 2005). Although they do not use learning neural networks and stay closer to the field of computer vision than this letter does, it turns out their approach is similar to the geometrical embedding approach we are using.

The structure of the letter is as follows. In section 2, we apply mathematical methods to analyze the behavior of a rate-based learning network. We formally introduce a nonautonomous model to be averaged in a second time. This then allows us to study the stability of the learning dynamics in the presence of multiple inputs by constructing an appropriate energy function. In section 3, we determine the geometrical structure of the equilibrium weight distribution and show how it reflects the structure of the inputs. We also relate this approach to the emergence of cortical maps. Finally, the results are discussed in section 4.

2 Analytical Treatment of a Hopfield-Type Learning Network

2.1 Model

2.1.1 Neural Network Evolution. A neural mass corresponds to a mesoscopic coherent group of neurons. It is convenient to consider them as building blocks for computational simplicity; for their direct relationship to macroscopic measurements of the brain (EEG, MEG, and optical imaging), which average over numerous neurons; and because one can functionally define coherent groups of neurons within cortical columns. For each neural mass $i \in \{1..N\}$ (which we abusively refer to as a neuron in the following), define the mean membrane potential $V_i(t)$ at time t. The instantaneous population firing rate $v_i(t)$ is linked to the membrane potential through the relation $v_i(t) = s(V_i(t))$, where s is a smooth sigmoid function. In the following, we choose

$$s(v) = \frac{S_m}{1 + \exp(-4S'_m(v - \phi))}, \tag{2.1}$$

where S_m, S'_m, and ϕ are, respectively, the maximal firing rate, the maximal slope, and the offset of the sigmoid.

Consider a Hopfield network of neural masses described by

$$\frac{dV_i}{dt}(t) = -\alpha V_i(t) + \sum_{j=1}^{N} W_{ij}(t)\, s(V_j(t)) + I_i(t). \tag{2.2}$$

The first term roughly corresponds to the intrinsic dynamics of the neural mass: it decays exponentially to zero at a rate α if it receives neither external inputs nor spikes from the other neural masses. We will fix the units of time by setting $\alpha = 1$. The second term corresponds to the rest of the network, sending information through spikes to the given neural mass i, with $W_{ij}(t)$ the effective synaptic weight from neural mass j. The synaptic weights are time dependent because they evolve according to a continuous time Hebbian learning rule (see below). The third term, $I_i(t)$, corresponds to an external input to neural mass i, such as information extracted by the retina or thalamocortical connections. We take the inputs to be piecewise constant in time; at regular time intervals, a new input is presented to the network. In this letter, we assume that the inputs are chosen by periodically cycling through a given set of M inputs. An alternative approach would be to randomly select each input from a given probability distribution (see Geman, 1979). It is convenient to introduce vector notation by representing the time-dependent membrane potentials by $V \in C^1(\mathbb{R}_+, \mathbb{R}^N)$, the time-dependent external inputs by $I \in C^0(\mathbb{R}_+, \mathbb{R}^N)$, and the time-dependent network weight matrix by $W \in C^1(\mathbb{R}_+, \mathbb{R}^{N\times N})$. We can then rewrite the above system of ordinary differential equations as a single vector-valued equation,

$$\frac{dV}{dt} = -V + W \cdot S(V) + I, \tag{2.3}$$

where $S : \mathbb{R}^N \to \mathbb{R}^N$ corrresponds to the term-by-term application of the sigmoid s: $S(V)_i = s(V_i)$.

2.1.2 Correlation-Based Hebbian Learning. The synaptic weights are assumed to evolve according to a correlation-based Hebbian learning rule of the form

$$\frac{dW_{ij}}{dt} = \epsilon(s(V_i)s(V_j) - \mu W_{ij}), \tag{2.4}$$

where ϵ is the learning rate, and we have included a decay term in order to stop the weights from diverging. In order to rewrite equation 2.4 in a more compact vector form, we introduce the tensor (or Kronecker) product $S(V) \otimes S(V)$ so that in component form,

$$[S(V) \otimes S(V)]_{ij} = S(V)_i S(V)_j, \tag{2.5}$$

where S is treated as a mapping from $\mathbb{R}^N$ to $\mathbb{R}^N$. The tensor product implements Hebb's rule that synaptic modifications involve the product of postynaptic and presynaptic firing rates. We can then rewrite the combined voltage and weight dynamics as the following nonautonomous (due to time-dependent inputs) dynamical system:

$$\Sigma : \quad \begin{cases} \dfrac{dV}{dt} = -V + W \cdot S(V) + I \\ \dfrac{dW}{dt} = \epsilon\big(S(V) \otimes S(V) - \mu W\big) \end{cases} . \tag{2.6}$$

Let us make few remarks about the existence and uniqueness of solutions. First, boundedness of S implies boundedness of the system Σ. More precisely, if I is bounded, the solutions are bounded. To prove this, note that the right-hand side of the equation for W is the sum of a bounded term and a linear decay term in W. Therefore, W is bounded, and, hence, the term $W \cdot S(V)$ is also bounded. The same reasoning applies to V. S being Lipschitz continuous implies that the right-hand side of the system is Lipschitz. This is sufficient to prove the existence and uniqueness of the solution by applying the Cauchy-Lipschitz theorem. In the following, we derive an averaged autonomous dynamical system Σ', which will be well defined for the same reasons.

2.2 Averaging the System. System Σ is a nonautonomous system that is difficult to analyze because the inputs are periodically changing. It has already been studied in the case of a single input (see Dong & Hopfield, 1992), but it remains to be analyzed in the case of multiple inputs. We show in section A.1 that this system can be approximated by an autonomous Cauchy problem, which will be much more convenient to handle. This averaging method makes the most of multiple timescales in the system.

Indeed, it is natural to consider that learning occurs on a much slower timescale than the evolution of the membrane potentials (as determined by α):

$$\epsilon \ll 1. \tag{2.7}$$

Second, an additional timescale arises from the rate at which the inputs are sampled by the network. That is, the network cycles periodically through M fixed inputs, with the period of cycling given by T. It follows that I is T-periodic, piecewise constant. We assume that the sampling rate is also much slower than the evolution of the membrane potentials:

$$\frac{M}{T} \ll 1. \tag{2.8}$$

Finally, we assume that the period T is small compared to the timescale of the learning dynamics,

$$\epsilon \ll \frac{1}{T}. \tag{2.9}$$

In section A.1, we simplify the system Σ by applying Tikhonov's theorem for slow and fast systems and then classical averaging methods for periodic systems. It leads to defining another system Σ', which will be a good approximation of Σ in the aymptotic regime.

In order to define the averaged system Σ', we need to introduce some additional notation. Let us label the M inputs by $I^{(a)}, a = 1, \ldots, M$ and denote by $V^{(a)}$ the fixed-point solution of the equation $V^{(a)} = W \cdot S(V^{(a)}) + I^{(a)}$. If we now introduce the $N \times M$ matrices $\mathcal{V}$ and $\mathcal{I}$ with components $\mathcal{V}_{ia} = V_i^{(a)}$ and $\mathcal{I}_{ia} = I_i^{(a)}$, then we define

$$\Sigma' : \quad \begin{cases} \dfrac{d\mathcal{V}}{dt} = -\mathcal{V} + W \cdot S(\mathcal{V}) + \mathcal{I} \\ \dfrac{dW}{dt} = \epsilon\left(\dfrac{1}{M} S(\mathcal{V}) \cdot S(\mathcal{V})^T - \mu W(t)\right) \end{cases}. \tag{2.10}$$

To illustrate this approximation, we simulate a simple network with both exact (i.e., Σ) and averaged (i.e., Σ') evolution equations. For these simulations, the network consists of $N = 10$ fully connected neurons and is presented with $M = 10$ different random inputs taken uniformly in the intervals $[0, 1]^N$. For this simulation, we use $s(x) = \frac{1}{1+e^{-4(x-1)}}$, and $\mu = 10$. Figure 1 (left) shows the percentage of error between final connectivities for different values of ϵ and T/M. Figure 1 (right) shows the temporal evolution of the norm of the connectivity for both the exact and averaged system for $T = 10^3$ and $\epsilon = 10^{-3}$.

In the remainder of the letter, we focus on the system Σ' whose solutions are close to those of the original system Σ provided condition A.2 in the appendix is satisfied, that is, the network is weakly connected. Finally, note that it is straighforward to extend our approach to time-functional rules (e.g., sliding threshold or BCM rules as described in Bienenstock et al., 1982), which, in this new framework, would be approximated by simple ordinary differential equations (as opposed to time-functional differential equations) provided S is redefined appropriately.

2.3 Stability

2.3.1 Lyapunov Function. In the case of a single fixed input ($M = 1$), systems Σ and Σ' are equivalent and reduce to the neural network with adapting synapses previously analyzed by Dong and Hopfield (1992).

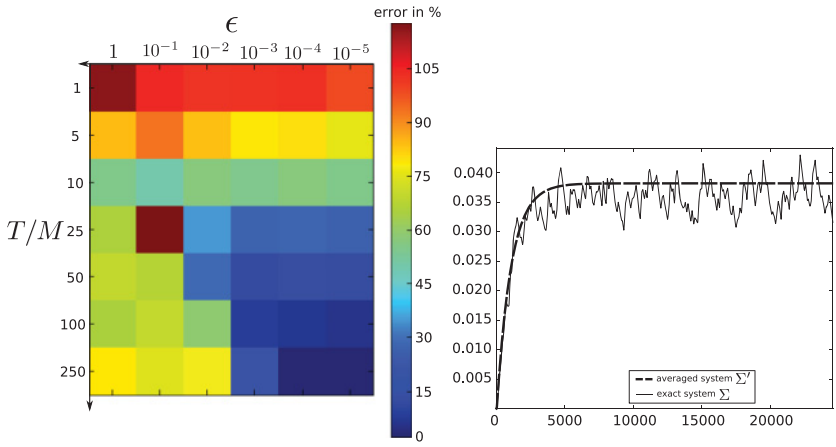


Figure 1: Comparison of exact and averaged systems. (Left) Percentage of error between final connectivities for the exact and averaged system. (Right) Temporal evolution of the norm of the connectivities of the exact system Σ and averaged system Σ'.

Under the additional constraint that the weights are symmetric ($W_{ij} = W_{ji}$), these authors showed that the simultaneous evolution of the neuronal activity variables and the synaptic weights can be reexpressed as a gradient dynamical system that minimizes a Lyapunov or energy function of state. We can generalize their analysis to the case of multiple inputs ($M > 1$) and nonsymmetric weights using the averaged system Σ'. That is, following along lines similar to Dong and Hopfield (1992), we introduce the energy function

$$E(\mathcal{U}, W) = -\frac{1}{2}\langle \mathcal{U}, W \cdot \mathcal{U}\rangle - \langle \mathcal{I}, \mathcal{U}\rangle + \langle 1, \overline{S^{-1}}(\mathcal{U})\rangle + \frac{M\mu}{2}\|W\|^2 \qquad (2.11)$$

where $\mathcal{U} = S(\mathcal{V}), \quad \|W\|^2 = \langle W, W\rangle = \sum_{i,j} W_{ij}^2,$

$$\langle \mathcal{U}, W \cdot \mathcal{U}\rangle = \sum_{a=1}^{M}\sum_{i=1}^{N} U_i^{(a)} W_{ij} U_j^{(a)}, \quad \langle \mathcal{I}, \mathcal{U}\rangle = \sum_{a=1}^{M}\sum_{i=1}^{N} I_i^{(a)} U_i^{(a)} \qquad (2.12)$$

and

$$\langle 1, \overline{S^{-1}}(\mathcal{U})\rangle = \sum_{a=1}^{M}\sum_{i=1}^{N} \int_0^{U_i^{(a)}} S^{-1}(\xi)\, d\xi. \qquad (2.13)$$

In contrast to Dong and Hopfield (1992), we do not require a priori that the weight matrix is symmetric. However, it can be shown that the system always converges to a symmetric connectivity pattern. More precisely, $\mathcal{A} = \{(\mathcal{V}, W) \in \mathbb{R}^{N\times M} \times \mathbb{R}^{N\times N} : W = W^T\}$ is an attractor of the system Σ'. A proof can be found in section A.2. It can then be shown that on $\mathcal{A}$ (symmetric weights), E is a Lyapunov function of the dynamical system Σ', that is,

$$\frac{dE}{dt} \leq 0, \quad \text{and} \quad \frac{dE}{dt} = 0 \Longrightarrow \frac{d\mathcal{Y}}{dt} = 0, \quad \mathcal{Y} = (\mathcal{V}, W)^T.$$

The boundedness of E and the Krasovskii-LaSalle invariance principle then implies that the system converges to an equilibrium (see Khalil & Grizzle, 1996). We thus have

Theorem 1. *The initial value problem for the system Σ' with $\mathcal{Y}(0) \in \mathcal{H}$, converges to an equilibrium state.*

Proof. See section A.3.

It follows that neither oscillatory nor chaotic attractor dynamics can occur.

2.3.2 Linear Stability. Although we have shown that there are stable fixed points, not all of the fixed points are stable. However, we can apply a linear stability analysis on the system Σ' to derive a simple sufficient condition for a fixed point to be stable. The method we use in the proof could be extended to more complex rules.

Theorem 2. *The equilibria of system Σ' satisfy*

$$\begin{cases} \mathcal{V}^* = \dfrac{1}{\mu M} S(\mathcal{V}^*) \cdot S(\mathcal{V}^*)^T \cdot S(\mathcal{V}^*) + \mathcal{I}, \\ W^* = \dfrac{1}{\mu M} S(\mathcal{V}^*) \cdot S(\mathcal{V}^*)^T, \end{cases} \tag{2.14}$$

and a sufficient condition for stability is

$$3S'_m \|W^*\| < 1, \tag{2.15}$$

provided $1 > \epsilon\mu$, which is probably the case since $\epsilon \ll 1$.

Proof. See section A.4.

This condition is strikingly similar to that derived in Faugeras, Grimbert, and Slotine (2008) (in fact, it is stronger than the contracting condition they find). It says the network may converge to a weakly connected situation.

It also says that the dynamics of $\mathcal{V}$ is likely (because the condition is only sufficient) to be contracting and therefore subject to no bifurcations: a fully recurrent learning neural network is likely to have a simple dynamics.

2.4 Equilibrium Points. It follows from equation 2.14 that the equilibrium weight matrix W^* is given by the correlation matrix of the firing rates. Moreover, in the case of sufficiently large inputs, the matrix of equilibrium membrane potentials satisfies $\mathcal{V}^* \approx \mathcal{I}$. More precisely, if $|S(I_i^{(a)})| \ll |I_i^{(a)}|$ for all $a = 1, \ldots, M$ and $i = 1, \ldots, N$, then we can generate an iterative solution for $\mathcal{V}^*$ of the form

$$\mathcal{V}^* = \mathcal{I} + \frac{1}{\mu} S(\mathcal{I}) \cdot S(\mathcal{I})^T \cdot S(\mathcal{I}) + \text{h.o.t.}$$

If the inputs are comparable in size to the synaptic weights, then there is no explicit solution for $\mathcal{V}^*$. If no input is presented to the network ($I = 0$), then $S(0) \neq 0$ implies that the activity is nonzero, that is, there is spontaneous activity. Combining these observations, we see that the network roughly extracts and stores the correlation matrix of the strongest inputs within the weights of the network.

3 Geometrical Structure of Equilibrium Points

3.1 From a Symmetric Connectivity Matrix to a Convolutional Network. So far, neurons have been identified by a label $i \in \{1..N\}$; there is no notion of geometry or space in the preceding results. However, as we show below, the inputs may contain a spatial structure that can be encoded by the final connectivity. In this section, we show that the network behaves as a convolutional network on this geometrical structure. The idea is to interpret the final connectivity as a matrix describing the distance between the neurons living in a k-dimensional space. This is quite natural since W^* is symmetric and has positive coefficients, properties shared with a Euclidean distance matrix. More specifically, we want to find an integer $k \in \mathbb{N}$ and N points in $\mathbb{R}^k$, denoted by x_i, $i \in \{1, \ldots, N\}$, so that the connectivity can roughly be written as $W_{ij}^* \simeq g(\|x_i - x_j\|^2)$, where g is a positive decreasing real function. If we manage to do so, then the interaction term in system Σ becomes

$$\{W \cdot S(V)\}_i \simeq \sum_{j=1}^{N} g(\|x_i - x_j\|^2) S(V(x_j)), \tag{3.1}$$

where we redefine the variable V as a field such that $V(x_j) = V_j$. This equation says that the network is convolutional with respect to the variables x_i, $i = 1, .., N$ and the associated convolution kernel is $g(\|x\|^2)$.

In practice, it is not always possible to find a geometry for which the connectivity is a distance matrix. Therefore, we project the appropriate matrix on the set of Euclidean distance matrices. This is the set of matrices M such that $M_{ij} = \|x_i - x_j\|^2$ with $x_i \in \mathbb{R}^k$. More precisely, we define $D = g^{-1}(W^*)$, where g^{-1} is applied to the coefficients of W^*. We then search for the distance matrix $D_\perp$ such that $\|D_{||}\|^2 = \|D - D_\perp\|^2$ is minimal. In this letter, we consider an L_2-norm. Although the choice of an L_p-norm will be motivated by the wiring length minimization argument in section 3.3, note that the choice of a L_2 norm is somewhat arbitrary and corresponds to penalizing long-distance connections. The minimization turns out to be a least square minimization whose parameters are the $x_i \in \mathbb{R}^k$. This can be implemented by a set of methods known as multidimensional scaling, which are reviewed in Borg and Groenen (2005). In particular, we use the stress majorization or SMACOF algorithm for the stress1 cost function throughout this letter. This leads to writing $D = D_\perp + D_{||}$ and therefore $W^*_{ij} = g(D_{\perp ij} + D_{||ij})$.

We now consider two choices of the function g:

1. If $g(x) = a\left(1 - \frac{x}{\lambda^2}\right)$ with $a > 1$ and $\lambda \in \mathbb{R}_+$, then one can always write

$$W^*_{ij} = W^*(x_i, x_j) = M(x_i, x_j) + g(\|x_i - x_j\|^2) \tag{3.2}$$

   such that $M(x_i, x_j) = -\frac{a}{\lambda^2} D_{||ij}$.
2. If $g(x) = ae^{-\frac{x}{\lambda^2}}$ with $a, \lambda \in \mathbb{R}_+$, then one can always write

$$W^*_{ij} = W^*(x_i, x_j) = M(x_i, x_j)\, g(\|x_i - x_j\|^2) \tag{3.3}$$

   such that $M(x_i, x_j) = e^{-\frac{D_{||ij}}{\lambda^2}}$,

where W^* is also redefined as a function over the x_i: $W^*(x_i, x_j) = W^*_{ij}$. For obvious reasons, M is called the nonconvolutional connectivity. It is the role of multidimensional scaling methods to minimize the role of the undetermined function M in the previous equations, that is, ideally having $M \equiv 0$ (resp. $M \equiv 1$) for the first (resp. second) assumption above. The ideal case of a fully convolutional connectivity can alway be obtained if k is large enough. Indeed, proposition 1 shows that $D = g^{-1}(W^*)$ satisfies the triangular inequality for matrices (i.e., $D_{ij} \leq (\sqrt{D_{ik}} + \sqrt{D_{kj}})^2$) for both g under some plausible assumptions. Therefore, it has all the properties of a distance matrix (symmetric, positive coefficients, and triangular inequality), and one can find points in $\mathbb{R}^k$ such that it is the distance matrix of these points iff $k \leq N - 1$ is large enough. In this case, the connectivity on the space defined by these points is fully convolutional; equation 3.1 is exactly verified.

Proposition 1. *If the neurons are equally excited on average (i.e., $\|S(\mathcal{V}_i)\| = c \in \mathbb{R}_+$), and*

1. *If* $g(x) = a\left(1 - \frac{x}{\lambda^2}\right)$ *with* $a, \lambda \in \mathbb{R}_+$, *then* $D = g^{-1}(W^*)$ *satisfies the triangular inequality.*
2. *If* $g(x) = ae^{-\frac{x}{\lambda^2}}$ *with* $a, \lambda \in \mathbb{R}_+$, *then* $D = g^{-1}(W^*)$ *satisfies the triangular inequality if the following assumption is satisfied:*

$$\arcsin\left(S(0)\right) - \arcsin\left(\sqrt{a^3 - \sqrt{a^6 - a^3}}\right) \geq \frac{\pi}{8}. \tag{3.4}$$

Proof. See section A.5.

3.2 Unveiling the Geometrical Structure of the Inputs. We hypothesize that the space defined by the x_i reflects the underlying geometrical structure of the inputs. We have not found a way to prove this, so we provide numerical examples that illustrate this claim. Therefore, the following is only a (numerical) proof of concept. For each example, we feed the network with inputs having a defined geometrical structure and then show how this structure can be extracted from the connectivity by the method outlined in section 3.1. In particular, we assume that the inputs are uniformly distributed over a manifold Ω with fixed geometry. This strong assumption amounts to considering that the feedforward connectivity (which we do not consider here) has already properly filtered the information coming from the sensory organs. More precisely, define the set of inputs by the matrix $\mathcal{I} \in \mathbb{R}^{N \times M}$ such that $I_i^{(a)} = f(\|y_i - z_a\|_\Omega)$ where the z_a are uniformly distributed points over Ω, the y_i are the positions on Ω that label the ith neuron, and f is a decreasing function on $\mathbb{R}_+$. The norm $\|.\|_\Omega$ is the natural norm defined over the manifold Ω. For simplicity, assume $f(x) = f_\sigma(x) = Ae^{-\frac{x^2}{\sigma^2}}$ so that the inputs are localized bell-shaped bumps on the shape Ω.

3.2.1 Planar Retinotopy. First, we consider a set of spatial gaussian inputs uniformly distributed over a two-dimensional plane: $\Omega = [0, 1] \times [0, 1]$. For simplicity, we take $N = M = K^2$ and set $z_a = y_i$ for $i = a, a \in \{1, \ldots, M\}$. (The numerical results show an identical structure for the final connectivity when the y_j correspond to random points, but the analysis is harder.) In the simpler case of one-dimensional gaussians with $N = M = K$, the input matrix takes the form $\mathcal{I} = T_{f_\sigma}$, where T_f is a symmetric Toeplitz matrix:

$$T_f = \begin{pmatrix} f(0) & f(1) & f(2) & \cdots & \cdots & f(K) \\ f(1) & f(0) & f(1) & f(2) & \cdots & f(K-1) \\ f(2) & f(1) & f(0) & f(1) & \cdots & f(K-2) \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ f(K) & f(K-1) & f(K-2) & \cdots & \cdots & f(0) \end{pmatrix}. \tag{3.5}$$

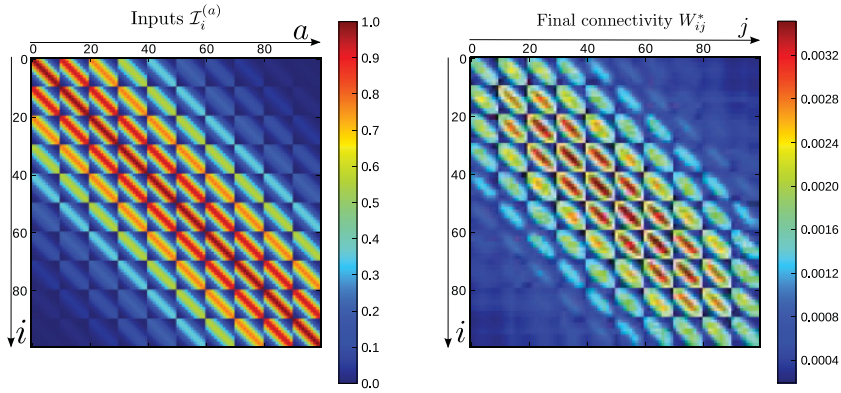


Figure 2: Plot of planar retinotopic inputs on $\Omega = [0,1] \times [0,1]$ (left) and final connectivity matrix of the system Σ' (right). The parameters used for this simulation are $s(x) = \frac{1}{1+e^{-4(x-1)}}$, $l = 1$, $\mu = 10$, $N = M = 100$, $\sigma = 4$. (Left) These inputs correspond to $I_m = 1$.

In the two-dimensional case, we set $y = (u, v) \in \Omega$ and introduce the labeling $y_{k+(l-1)K} = (u_k, v_l)$ for $k, l = 1, \ldots K$. It follows that $I_i^{(a)} \sim \exp(-\frac{(u_k - u_{k'})^2}{\sigma^2}) \exp(-\frac{(v_l - v_{l'})^2}{\sigma^2})$ for $i = k + (l-1)K$ and $a = k' + (l'-1)K$. Hence, we can write $\mathcal{I} = T_{f_\sigma} \otimes T_{f_\sigma}$, where $\otimes$ is the Kronecker product; the Kronecker product is responsible for the $K \times K$ substructure we can observe in Figure 2, with $K = 10$. Note that if we were interested in a n-dimensional retinotopy, then the input matrix could be written as a Kronecker product between n Toeplitz matrices. As previously mentioned, the final connectivity matrix roughly corresponds to the correlation matrix of the input matrix. It turns out that the correlation matrix of $\mathcal{I}$ is also a Kronecker product of two Toeplitz matrixes generated by a single gaussian (with a different standard deviation). Thus, the connectivity matrix has the same basic form as the input matrix when $z_a = y_i$ for $i = a$. The inputs and stable equilibrium points of the simulated system are shown in Figure 2. The positions x_i of the neurons after multidimensional scaling are shown in Figure 3 for different parameters. Note that we find no significant change in the position x_i of the neurons when the convolutional kernel g varies (as will also be shown in section 3.3.1). Thus, we show results for only one of these kernels, $g(x) = e^{-x}$.

If the standard deviation of the inputs is properly chosen as in Figure 3b, we observe that the neurons are distributed on a regular grid, which is retinotopically organized. In other words, the network has learned the geometric shape of the inputs. This can also be observed in Figure 3d, which corresponds to the same connectivity matrix as in Figure 3b but is represented in three dimensions. The neurons self-organize on a two-dimensional

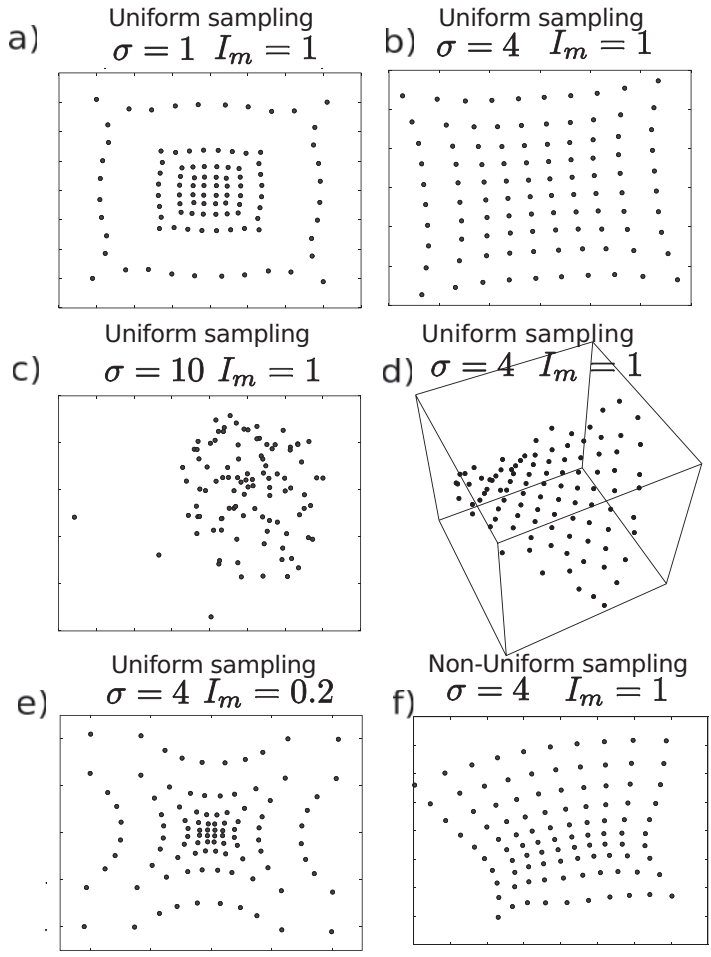


Figure 3: Positions x_i of the neurons after having applied multidimensional scaling to the equilibrium connectivity matrix of a learning network of $N = 100$ neurons driven by planar retinotopic inputs as described in Figure 2. In all figures, the convolution kernel $g(x) = e^{-x}$; this choice has virtually no impact on the shape of the figures. (a) Uniformly sampled inputs with $M = 100$, $\sigma = 1$, $I_m = 1$, and $k = 2$. (b) Uniformly sampled inputs with $M = 100$, $\sigma = 4$, $I_m = 1$, and $k = 2$. (c) Uniformly sampled inputs with $M = 100$, $\sigma = 10$, $I_m = 1$, and $k = 2$. (d) Uniformly sampled inputs with $M = 100$, $\sigma = 4$, $I_m = 0.2$, and $k = 2$. (e) Same as panel b, but in three dimensions: $k = 3$. (f) Nonuniformly sampled inputs with $M = 150$, $\sigma = 4$, $I_m = 1$, and $k = 2$. The first 100 inputs are as in panel B, but 50 more inputs of the same type are presented to half of the visual field. This corresponds to the denser part of the picture.

saddle shape that accounts for the border distortions that can be observed in two dimensions (which we discuss in the next paragraph). If σ is too large, as can be observed in Figure 3c, the final result is poor. Indeed, the inputs are no longer local and cover most of the visual field. Therefore, the neurons saturate, $S(V_i) \simeq S_m$, for all the inputs, and no structure can be read in the activity variable. If σ is small, the neurons still seem to self-organize (as long as the inputs are not completely localized on single neurons) but with significant border effects.

There are several reasons that we observe border distortions in Figure 3. We believe the most important is due to an unequal average excitation of the neurons. Indeed, the neurons corresponding to the border of the "visual field" are less excited than the others. For example, consider a neuron on the left border of the visual field. It has no neighbors on its left and therefore is less likely to be excited by its neighbors and less excited on average. The consequence is that it is less connected to the rest of the network (see, e.g., the top line of the right side of Figure 2) because their connection depends on their level of excitment through the correlation of the activity. Therefore, it is farther away from the other neurons, which is what we observe. When the inputs are localized, the border neurons are even less excited on average and thus are farther away, as shown in Figure 3a.

Another way to get distortions in the positions x_i is to reduce or increase excessively the amplitude $I_m = \max_{i,a} |I_i^{(a)}|$ of the inputs. Indeed, if it is small, the equilibrium actitivity described by equation 2.14 is also small and likely to be the flat part of the sigmoid. In this case, neurons tend to be more homogeneously excited and less sensitive to the particular shape of the inputs. Therefore, the network loses some information about the underlying structures of the inputs. Actually the neurons become relatively more sensitive to the neighborhood structure of the network, and the border neurons have a different behavior from the rest of the network, as shown in Figure 3e. The parameter μ has much less impact on the final shape since it corresponds only to a homogeneous scaling of the final connectivity.

So far, we have assumed that the inputs were uniformly spread on the manifold Ω. If this assumption is broken, the final position of the neurons will be affected. As shown in Figure 3f, where 50 inputs were added to the case of Figure 3b in only half of the visual field, the neurons that code for this area tend to be closer. Indeed, they tend not to be equally excited on average (as supposed in proposition 1), and a distortion effect occurs. This means that a proper understanding of the role of the vertical connectivity would be needed to complete this geometrical picture of the functioning of the network. This is, however, beyond the scope of this letter.

3.2.2 Toroïdal Retinotopy. We now assume that the inputs are uniformly distributed over a two-dimensional torus, $\Omega = \mathbb{T}^2$. That is, the input labels z_a are randomly distributed on the torus. The neuron labels y_i are regularly

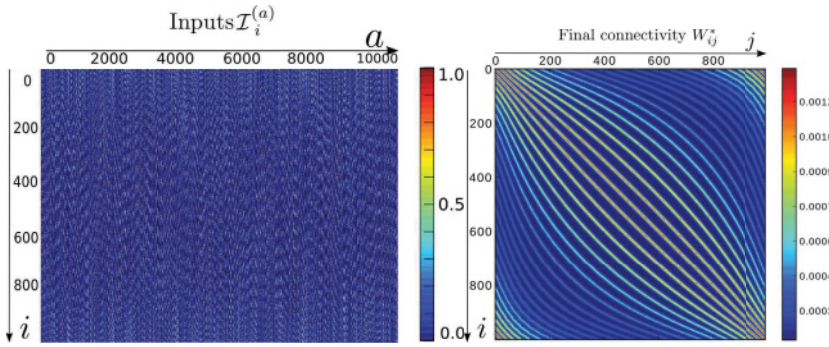


Figure 4: Plot of retinotopic inputs on $\Omega = \mathbb{T}^2$ (left) and the final connectivity matrix (right) for the system Σ'. The parameters used for this simulation are $s(x) = \frac{1}{1+e^{-4(x-1)}}$, $l = 1$, $\mu = 10$, $N = 1000$, $M = 10,000$, $\sigma = 2$.

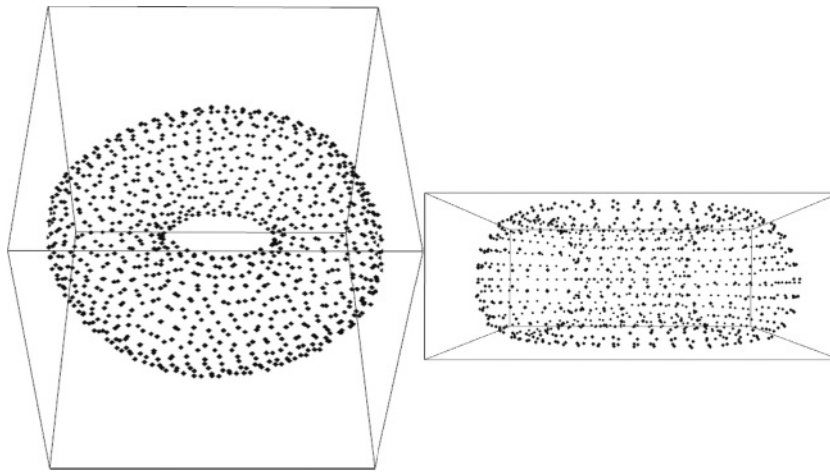


Figure 5: Positions x_i of the neurons for $k = 3$ after having applied multidimensional scaling methods presented in section 3.1 to the final connectivity matrix shown in Figure 4.

and uniformly distributed on the torus. The inputs and final stable weight matrix of the simulated system are shown in Figure 4. The positions x_i of the neurons after multidimensional scaling for $k = 3$ are shown in Figure 5 and appear to form a cloud of points distributed on a torus. In contrast to the previous example, there are no distortions now because there are no borders on the torus. In fact, the neurons are equally excited on average in this case which makes proposition 1 valid.

3.3 Links with Neuroanatomy. The brain is subject to energy constraints, which are completely neglected in the above formulation. These constraints most likely have a significant impact on the positions of real neurons in the brain. Indeed, it seems reasonable to assume that the positions and connections of neurons reflect a trade-off between the energy costs of biological tissue and their need to process information effectively. For instance, it has been suggested that a principle of wire length minimization may occur in the brain (see Swindale, 1996; Chklovskii, Schikorski, & Stevens, 2002). In our neural mass framework, one may consider that the stronger two neural masses are connected, the larger the number of real axons linking the neurons together. Therefore, minimizing axonal length can be read as that the stronger the connection, the closer, which is consistent with the convolutional part of the weight matrix. However, the underlying geometry of natural inputs is likely to be very high-dimensional, whereas the brain lies in a three-dimensional world. In fact, the cortex is so flat that it is effectively two-dimensional. Hence, the positions of real neurons are different from the positions $x_i \in \mathbb{R}^k$ in a high-dimensional vector space; since the cortex is roughly two-dimensional, the positions could be realized physically only if $k = 2$. Therefore, the three-dimensional toric geometry or any higher-dimensional structure could not be perfectly implemented in the cortex without the help of nonconvolutional long-range connections. Indeed, we suggest that the cortical connectivity is made of two parts: (1) a local convolutional connectivity corresponding to the convolutional term g in equations 3.2 and 3.3, which is consistent with the requirements of energy efficiency, and (2) a nonconvolutional connectivity corresponding to the factor M in equations 3.2 and 3.3, which is required in order to represent various stimulus features. If the cortex were higher-dimensional ($k \gg 2$), then there would no nonconvolutionnal connectivity M, that is, $M \equiv 0$ for the linear convolutional kernel or $M \equiv 1$ for the exponential one.

We illustrate this claim by considering two examples based on the functional anatomy of the primary visual cortex: the emergence of ocular dominance columns and orientation columns, respectively. We proceed by returning to the case of planar retinotopy (see section 5.3.1) but now with additional input structure. In the first case, the inputs are taken to be binocular and isotropic, whereas in the second case, they are taken to be monocular and anisotropic. The details are presented below. Given a set of prescribed inputs, the network evolves according to equation 2.10, and the lateral connections converge to a stable equilibrium. The resulting weight matrix is then projected onto the set of distance matrices for $k = 2$ (as described in section 3.2.1) using the stress majorization or SMACOF algorithm for the stress1 cost function as described in Borg and Groenen (2005). We thus assign a position $x_i \in \mathbb{R}^2$ to the ith neuron, $i = 1, \ldots, N$. (Note that the position x_i extracted from the weights using multidimensional scaling is distinct from the "physical" position y_i of the neuron in the retinocortical plane; the latter determines the center of its receptive field.) The convolutional

connectivity (g in equations 3.2 and 3.3) is therefore completely defined. On the planar map of points x_i, neurons are isotropically connected to their neighbors; the closer the neurons are, the stronger is their convolutional connection. Moreover, since the stimulus feature preferences (orientation, ocular dominance) of each neuron i, $i = 1, \ldots, N$, are prescribed, we can superimpose these feature preferences on the planar map of points x_i. In both examples, we find that neurons with the same ocular or orientation selectivity tend to cluster together: interpolating these clusters then generates corresponding feature columns. It is important to emphasize that the retinocortical positions y_i do not have any columnar structure, that is, they do not form clusters with similar feature preferences. Thus, in contrast to standard developmental models of vertical connections, the columnar structure emerges from the recurrent weights after learning, which are intrepreted as a Euclidean distances. It follows that neurons coding for the same feature tend to be strongly connected; indeed, the multidimensional scaling algorithm has the property that it positions strongly connected neurons close together. Equations 3.2 and 3.3 also suggest that the connectivity has a nonconvolutional part, M, which is a consequence of the low dimensionality ($k = 2$). In order to illustrate the structure of the nonconvolutional connectivity, we select a neuron i in the plane and draw a link from it at position x_i to the neurons at position x_j for which $M(x_i, x_j)$ is maximal in Figures 7, 8, and 9. We find that M tends to be patchy; it connects neurons having the same feature preferences. In the case of orientation, M also tends to be coaligned, that is, connecting neurons with similar orientation preference along a vector in the plane of the same orientation.

Note that the proposed method to get cortical maps is artificial. First, the networks learn its connectivity and, second, the equilibrium connectivity matrix is used to give a position to the neurons. In biological tissues, the cortical maps are likely to emerge slowly during learning. Therefore, a more biological way of addressing the emergence of cortical maps may consist in repositioning the neurons at each time step, taking the position of the neurons at time t as the initial condition of the algorithm at time $t + 1$. This would correspond better to the real effect of the energetic constraints (leading to wire length minimization), which occur not only after learning but also at each time step. However, when learning is finished, the energetic landscape corresponding to the minimization of the wire length would still be the same. In fact, repositioning the neurons successively during learning changes only the local minimum where the algorithm converges: it corresponds to a different initialization of the neurons' positions. Yet we do not think this corresponds to a more relevant minimum since the positions of the points at time $t = 0$ are still randomly chosen. Although not biological, the a posteriori method is not significantly worse than the application of algorithm at each time step and gives an intuition of the shapes of the local minima of the functional leading to the minimization of the wire length.

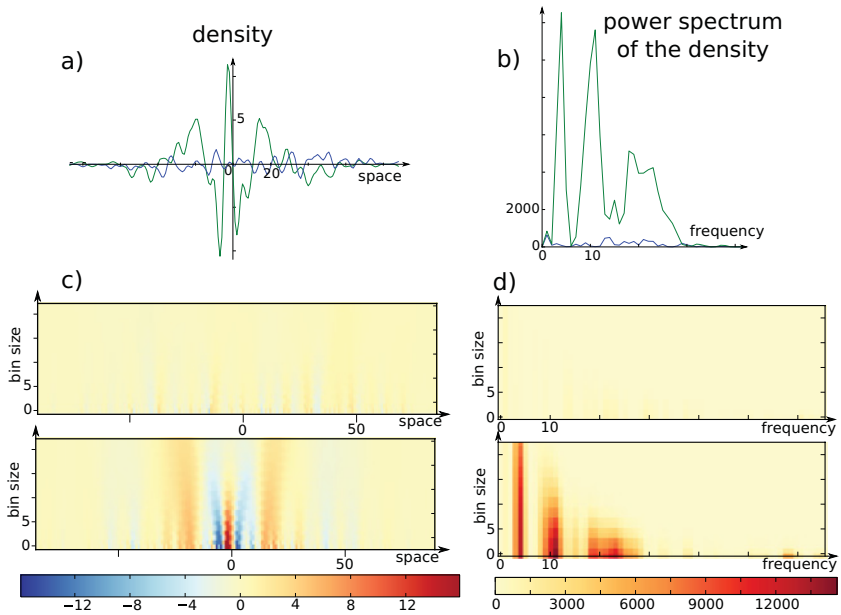


Figure 6: Emergence of ocular dominance columns. Analysis of the equilibrium connectivity of a network of $N = 1000$ neurons exposed to $M = 3000$ inputs as described in equation 3.6 with $\sigma' = 10$. The parameters used for this simulation are $I_m = 1$, $\mu = 10$ and $s(x) = \frac{1}{1+e^{-4(x-1)}}$. (a) Relative density of the network assuming that the "weights" of the left neurons are $+1$ and the "weights" of the right eye neuron are -1. Thus, a positive (resp. negative) lobe corresponds to a higher number of left neurons (resp. right neurons), and the presence of oscillations implies the existence of ocular dominance columns. The size of the bin to compute the density is 5. The blue (resp. green) curve corresponds to $\gamma = 0$ (resp. $\gamma = 1$). It can be seen that the case $\gamma = 1$ exhibits significant oscillations consistent with the formation of ocular dominance columns. (b) Power spectra of curves plotted in panel a. The dependence of the density and power spectrum on bin size is shown in panels c and d, respectively. The top pictures correspond to the blue curves (i.e., no binocular disparity), and the bottom pictures correspond to the green curves $\gamma = 1$ (i.e., a higher binocular disparity).

3.3.1 Ocular Dominance Columns and Patchy Connectivity. In order to construct binocular inputs, we partition the N neurons into two sets $i \in \{1, \ldots, N/2\}$ and $i \in \{N/2 + 1, \ldots, N\}$ that code for the left and right eyes, respectively. The ith neuron is then given a retinocortical position y_i, with the y_i uniformly distributed across the line for Figure 6 and across the plane for Figures 7 and 8. We do not assume a priori that there exist any ocular dominance columns, that is, neurons with similar retinocortical positions y_i

do not form clusters of cells coding for the same eye. We then take the ath input to the network to be of the form

$$\begin{aligned} &\text{Left eye} \quad I_i^{(a)} = (1+\gamma(a))e^{-\frac{(y_i - z_a)^2}{\sigma'^2}}, \quad i = 1, \ldots, N/2, \\ &\text{Right eye} \quad I_i^{(a)} = (1-\gamma(a))e^{-\frac{(y_i - z_a)^2}{\sigma'^2}}, \quad i = N/2+1, \ldots, N, \end{aligned} \tag{3.6}$$

where the z_a are randomly generated from $[0, 1]$ in the one-dimensional case and $[0, 1]^2$ in the two-dimensional case. For each input a, $\gamma(a)$ is randomly and uniformly taken in $[-\gamma, \gamma]$ with $\gamma \in [0, 1]$ (see Bressloff, 2005). Thus, if $\gamma(a) > 0$ ($\gamma(a) < 0$), then the corresponding input is predominantly from the left (right) eye.

First, we illustrate the results of ocular dominance simulations in one dimension in Figure 6. Although it is not biologically realistic, taking the visual field to be one-dimensional makes it possible to visualize the emergence of ocular dominance columns more easily. Indeed, in Figure 6 we analyze the role of the binocular disparity of the network, that is, we change the value of γ. If $\gamma = 0$ (the blue curves in Figures 6a and 6b and top pictures in Figures 6c and 6d), there are virtually no differences between left and right eyes, and we observe much less segregation than in the case $\gamma = 1$ (the green curves in Figures 6a and 6b and the bottom pictures in Figures 6c and 6d). Increasing the binocular disparity between the two eyes results in the emergence of ocular dominance columns. Yet there does not seem to be any spatial scale associated with these columns: they form on various scales, as shown in Figure 6d.

In Figures 7 and 8, we plot the results of ocular dominance simulations in two dimensions. In particular, we illustrate the role of changing the binocular disparity γ, changing the standard deviation of the inputs σ', and using different convolutional kernels g. We plot the points x_i obtained by performing multidimensional scaling on the final connectivity matrix for $k = 2$ and superimposing on this the ocular dominance map obtained by interpolating between clusters of neurons with the same eye preference. The convolutional connectivity (g in equations 3.2 and 3.3) is implicitly described by the position of the neurons: the closer the neurons, the stronger their connections. We also illustrate the nonconvolutional connectivity (M in equations 3.2 and 3.3) by linking one selected neuron to the neurons it is most strongly connected to. The color of the link refers to the color of the target neuron. The multidimensional scaling algorithm was applied for each set of parameters with different initial conditions, and the best final solution (i.e., with the smallest nonconvolutional part) was kept and plotted. The initial conditions were random distributions of neurons or artificially created ocular dominance stripes with different numbers of neurons per stripe. It turns out the algorithm performed better on the latter. (The number of tunable parameters was too high for the system to converge to

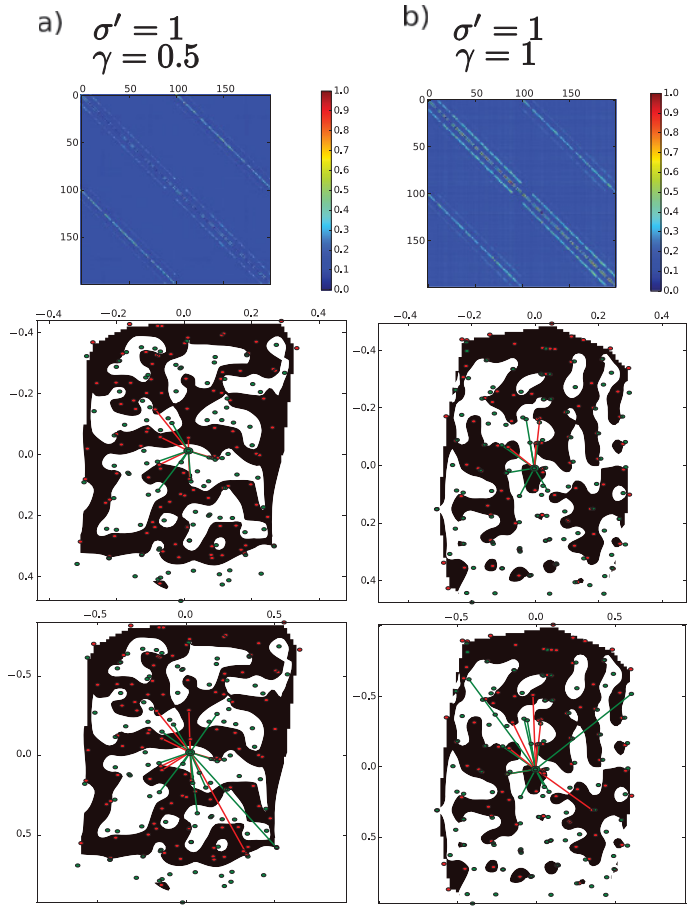


Figure 7: Analysis of the equilibirum connectivity of a modifiable recurrent network driven by two-dimensional binocular inputs. This figure and Figure 8 correspond to particular values of the disparity γ and standard deviation σ'. Each cell shows the profile of the inputs (top), the position of the neurons for a linear convolutional kernel (middle), and an exponential kernel (bottom). The parameters of the kernel (a and λ) were automatically chosen to minimize the nonconvolutional part of the connectivity. It can be seen that the choice of the convolutional kernel has little impact on the position of the neurons. (a, c) these panels correspond to $\gamma = 0.5$, which means there is little binocular disparity. Therefore, the nonconvolutional connectivity connects neurons of opposite eye preference more than for $\gamma = 1$, as shown in panels b and d. The inputs for panels a and b have a smaller standard deviation than for panels c and d. It can be seen that the neurons coding for the same eye tend to be closer when σ' is larger. The other parameters used for these simulations are $s(x) = 1/(1 + e^{-4(x-1)})$, $l = 1$, $\mu = 10$, $N = M = 200$.

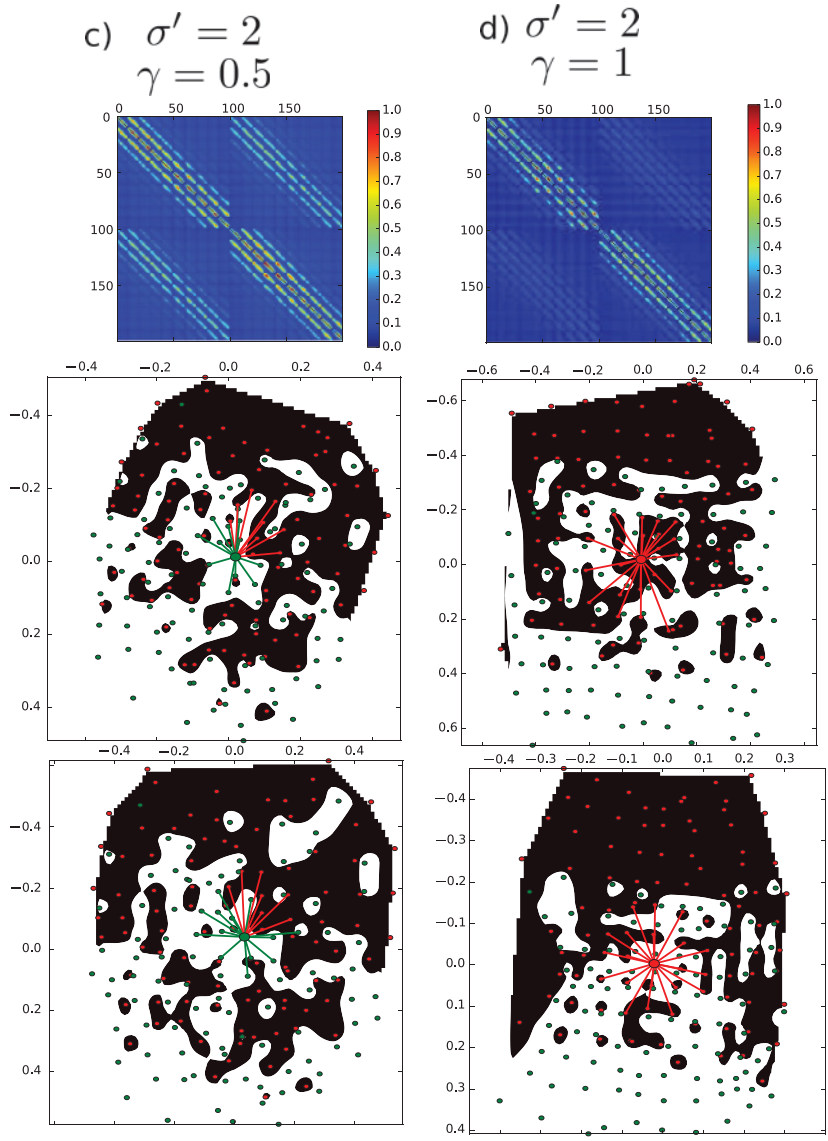


Figure 8: See Figure 7 for the commentaries.

a global equilibrium for a random initial condition.) Our results show that nonconvolutional or long-range connections tend to link cells with the same ocular dominance provided the inputs are sufficiently strong and different for each eye.

3.3.2 Orientation Columns and Collinear Connectivity. In order to construct oriented inputs, we partition the N neurons into four groups Σ_θ corresponding to different orientation preferences $\theta = \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$. Thus, if neuron $i \in \Sigma_\theta$, then its orientation preference is $\theta_i = \theta$. For each group, the neurons are randomly assigned a retinocortical position $y_i \in [0,1] \times [0,1]$. Again, we do not assume a priori that there exist any orientation columns, that is, neurons with similar retinocortical positions y_i do not form clusters of cells coding for the same orientation preference. Each cortical input $I_i^{(a)}$ is generated by convolving a thalamic input consisting of an oriented gaussian with a Gabor-like receptive field (as in Miikkulainen et al., 2005). Let $\mathcal{R}_\theta$ denote a two-dimensional rigid body rotation in the plane with $\theta \in [0, 2\pi)$. Then

$$I_i^{(a)} = \int G_i(\xi - y_i) I_a(\xi - z_a) d\xi, \tag{3.7}$$

where

$$G_i(\xi) = G_0(\mathcal{R}_{\theta_i}\xi) \tag{3.8}$$

and $G_0(\xi)$ is the Gabor-like function,

$$G_0(\xi) = A_+ e^{-\xi^T.\Lambda^{-1}.\xi} - A_- e^{-(\xi - e_0)^T.\Lambda^{-1}.(\xi - e_0)} - A_- e^{-(\xi + e_0)^T.\Lambda^{-1}.(\xi + e_0)},$$

with $e_0 = (0,1)$ and

$$\Lambda = \begin{pmatrix} \sigma_{large} & 0 \\ 0 & \sigma_{small} \end{pmatrix}.$$

The amplitudes A_+, A_- are chosen so that $\int G_0(\xi) d\xi = 0$. Similarly, the thalamic input $I_a(\xi) = I(\mathcal{R}_{\theta'_a}\xi)$ with $I(\xi)$ the anisotropic gaussian

$$I(\xi) = e^{-\xi^T.\Lambda'^{-1}.\xi}, \qquad \Lambda' = \begin{pmatrix} \sigma'_{large} & 0 \\ 0 & \sigma'_{small} \end{pmatrix}.$$

The input parameters θ'_a and z_a are randomly generated from $[0, \pi)$ and $[0,1]^2$, respectively. In our simulations, we take $\sigma_{large} = 0.133\ldots$, $\sigma'_{large} = 0.266\ldots$, and $\sigma_{small} = \sigma'_{small} = 0.0333\ldots$. The results of our simulations are shown in the Figure 9 (left). In particular, we plot the points x_i obtained by performing multidimensional scaling on the final connectivity matrix for $k = 2$, and superimposing on this the orientation preference map obtained by interpolating between clusters of neurons with the same orientation preference. To avoid border problems, we have zoomed on the

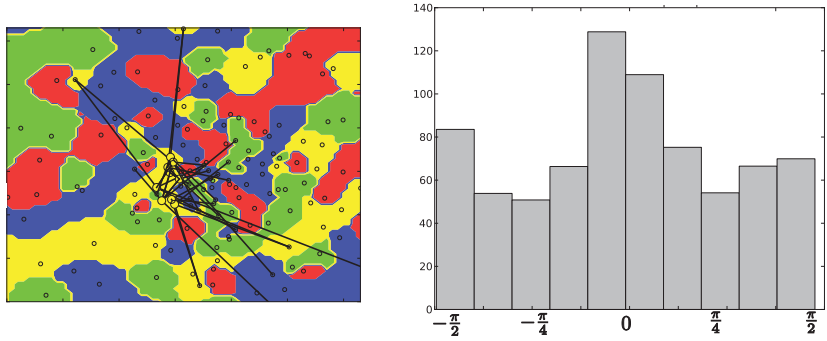


Figure 9: Emergence of orientation columns. (Left) Plot of the positions x_i of neurons for $k = 2$ obtained by multidimensional scaling of the weight matrix. Neurons are clustered in orientation columns represented by the colored areas, which are computed by interpolation. The strongest components of the nonconvolutional connectivity (M in equations 3.2 and 3.3) from a particular neuron in the yellow area are illustrated by drawing black links from this neuron to the target neurons. Since yellow corresponds to an orientation of $\frac{3\pi}{4}$, the nonconvolutional connectivity shows the existence of a colinear connectivity as exposed in Bosking, Zhang, Schofield, and Fitzpatrick (1997). The parameters used for this simulation are $s(x) = \frac{1}{1+e^{-4(x-1)}}$, $l = 1$, $\mu = 10$, $N = 900$, $M = 9000$. (Right) Histogram of the five largest components of the nonconvolutional connectivity for 80 neurons randomly chosen among those shown in the left panel. The abscissa corresponds to the difference in radian between the direction preference of the neuron and the direction of the links between the neuron and the target neurons. This histogram is weighted by the strength of the nonconvolutional connectivity. It shows a preference for coaligned neurons but also a slight preference for perpendicularly aligned neurons (e.g., neurons of the same orientation but parallel to each other). We chose 80 neurons in the center of the picture because the border effects shown in Figure 3 do not arise and it roughly corresponds to the number of neurons in the left panel.

center on the map. We also illustrate the nonconvolutional connectivity by linking a group of neurons gathered in an orientation column to all other neurons for which M is maximal. The patchy, anisotropic nature of the long-range connections is clear. The anisotropic nature of the connections is further quantified in the histogram of Figure 9.

4 Discussion

In this letter, we have shown how a neural network can learn the underlying geometry of a set of inputs. We have considered a fully recurrent neural network whose dynamics is described by a simple nonlinear rate equation, together with unsupervised Hebbian learning with decay that

occurs on a much slower timescale. Although several inputs are periodically presented to the network, so that the resulting dynamical system is nonautonomous, we have shown that such a system has a fairly simple dynamics: the network connectivity matrix always converges to an equilibrium point. We then demonstrated how this connectivity matrix can be expressed as a distance matrix in $\mathbb{R}^k$ for sufficiently large k, which can be related to the underlying geometrical structure of the inputs. If the connectivity matrix is embedded in a lower two-dimensional space ($k = 2$), then the emerging patterns are qualitatively similar to experimentally observed cortical feature maps, although we have considered simplistic stimuli and the feedforward connectivity as fixed. That is, neurons with the same feature preferences tend to cluster together, forming cortical columns within the embedding space. Moreover, the recurrent weights decompose into a local isotropic convolutional part, which is consistent with the requirements of energy efficiency and a longer-range nonconvolutional part that is patchy. This suggest a new interpretation of the cortical maps: they correspond to two-dimensional embeddings of the underlying geometry of the inputs.

Geometric diffusion methods (see Coifman et al., 2005) are also an efficient way to reveal the underlying geometry of sets of inputs. There are differences with our approach, although both share the same philosophy. The main difference is that geometric harmonics deals with the probability of co-occurence, whereas our approach is more focused on wiring length, which is indirectly linked to the inputs through Hebbian coincidence. From an algoritmic point of view, our method is concerned with the position of N neurons and not M inputs, which can be a clear advantage in certain regimes. Indeed, we deal with matrices of size $N \times N$, whereas the total size of the inputs is $N \times M$, which is potentially much higher. Finally, this letter is devoted to decomposing the connectivity between a convolutional and nonconvolutional part, and this is why we focus not only on the spatial structure but also on the shape of the activity equation on this structure. These two results come together when decomposing the connectivity. Actually, this focus on the connectivity was necessary to use the energy minization argument of section 2.3.1 and compute the cortical maps in section 3.3.

One of the limitations of applying simple Hebbian learning to recurrent cortical connections is that it takes into account only excitatory connections, whereas 20% of cortical neurons are inhibitory. Indeed, in most developmental models of feedforward connections, it is assumed that the local and convolutional connections in cortex have a Mexican hat shape with negative (inhibitory) lobes for neurons that are sufficiently far from each other. From a computational perspective, it is possible to obtain such a weight distribution by replacing Hebbian learning with some form of covariance learning (see Sejnowski & Tesauro, 1989). However, it is difficult to prove convergence to a fixed point in the case of the covariance learning rule, and the multidimensional scaling method cannot be applied directly

unless the Mexican hat function is truncated so that it is invertible. Another limitation of rate-based Hebbian learning is that it does not take into account causality, in contrast to more biologically detailed mechanisms such as spike-timing-dependent plasticity.

The approach taken here is very different from standard treatments of cortical development (as in Miller, Keller, & Stryker, 1989; Swindale, 1996), in which the recurrent connections are assumed to be fixed and of convolutional Mexican hat form while the feedforward vertical connections undergo some form of correlation-based Hebbian learning. In the latter case, cortical feature maps form in the physical space of retinocortoical coordinates y_i, rather than in the more abstract planar space of points x_i obtained by applying multidimensional scaling to recurrent weights undergoing Hebbian learning in the presence of fixed vertical connections. A particular feature of cortical maps formed by modifiable feedforward connections is that the mean size of a column is determined by a Turing-like pattern forming instability and depends on the length scales of the Mexican hat weight function and the two-point input correlations (see Miller et al., 1989; Swindale, 1996). No such Turing mechanism exists in our approach, so the resulting cortical maps tend to be more fractal-like (many length scales) compared to real cortical maps. Nevertheless, we have established that the geometrical structure of cortical feature maps can also be encoded by modifiable recurrent connections. This should have interesting consequences for models that consider the joint development of feedforward and recurrent cortical connections. One possibility is that the embedding space of points x_i arising from multidimensional scaling of the weights becomes identified with the physical space of retinocortical positions y_i. The emergence of local convolutional structures, together with sparser long-range connections, would then be consistent with energy efficiency constraints in physical space.

Our letter also draws a direct link between the recurrent connectivity of the network and the positions of neurons in some vector space such as $\mathbb{R}^2$. In other words, learning corresponds to moving neurons or nodes so that their final position will match the inputs' geometrical structure. Similarly, the Kohonen algorithm detailed in Kohonen (1990) describes a way to move nodes according to the inputs presented to the network. It also converges toward the underlying geometry of the set of inputs. Although these approaches are not formally equivalent, it seems that both have the same qualitative behavior. However, our method is more general in the sense that no neighborhood structure is assumed a priori; such a structure emerges via the embedding into $\mathbb{R}^k$.

Finally, note that we have used a discrete formalism based on a finite number of neurons. However, the resulting convolutional structure obtained by expressing the weight matrix as a distance matrix in $\mathbb{R}^k$ (see equations 3.2 and 3.3) allows us to take an appropriate continuum limit. This then generates a continuous neural field model in the form of an

integro-differential equation whose integral kernel is given by the underlying weight distribution. Neural fields have been used increasingly to study large-scale cortical dynamics (see Coombes, 2005 for a review). Our geometrical learning theory provides a developmental mechanism for the formation of these neural fields. One of the useful features of neural fields from a mathematical perspective is that many of the methods of partial differential equations can be carried over. Indeed, for a general class of connectivity functions defined over continuous neural fields, a reaction-diffusion equation can be derived whose solution approximates the firing rate of the associated neural field (see Degond & Mas-Gallic, 1989; Cottet, 1995; Edwards, 1996). It appears that the necessary connectivity functions are precisely those that can be written in the form 3.2. This suggests that a network that has been trained on a set of inputs with an appropriate geometrical structure behaves as a diffusion equation in a high-dimensional space together with a reaction term corresponding to the inputs.

Appendix

A.1 Proof of System's Σ Averaging. Here, we show that system Σ can be approximated by the autonomous Cauchy problem Σ′. Indeed, we can simplify system Σ by applying Tikhonov's theorem for slow and fast systems and then classical averaging methods for periodic systems.

A.1.1 Tikhonov's Theorem. Tikhonov's theorem (see Tikhonov, 1952, and Verhulst, 2007 for a clear introduction) deals with slow and fast systems. It says the following:

Theorem 3. *Consider the initial value problem,*

$$\dot{x} = f(x, y, t), \quad x(0) = x_0, \quad x \in \mathbb{R}^n, t \in \mathbb{R}_+,$$
$$\epsilon\dot{y} = g(x, y, t), \quad y(0) = y_0, \quad y \in \mathbb{R}^m.$$

Assume that:

1. *A unique solution of the initial value problem exists, and, we suppose, this holds also for the reduced problem,*

$$\dot{x} = f(x, y, t), \; x(0) = x_0,$$
$$0 = g(x, y, t)$$

   with solutions $\bar{x}(t)$, $\bar{y}(t)$.
2. *The equation* $0 = g(x, y, t)$ *is solved by* $\bar{y}(t) = \phi(x, t)$, *where* $\phi(x, t)$ *is a continuous function and an isolated root. Also suppose that* $\bar{y}(t) = \phi(x, t)$

is an asymptotically stable solution of the equation $\frac{dy}{d\tau} = g(x, y, \tau)$ *that is uniform in the parameters* $x \in \mathbb{R}^n$ *and* $t \in \mathbb{R}_+$.

3. $y(0)$ *is contained in an interior subset of the domain of attraction of* $\bar{y}$.

Then we have

$$\lim_{\epsilon \to 0} x_\epsilon(t) = \bar{x}(t), \ 0 \le t \le L,$$
$$\lim_{\epsilon \to 0} y_\epsilon(t) = \bar{y}(t), \ 0 \le d \le t \le L,$$

with d and L constants independent of ϵ.

In order to apply Tikhonov's theorem directly to system Σ, we first need to rescale time according to $t \to \epsilon t$. This gives

$$\epsilon \frac{dV}{dt} = -V + W \cdot S(V) + I,$$
$$\frac{dW}{dt} = S(V) \otimes S(V) - \mu W.$$

Tikhonov's theorem then implies that solutions of Σ are close to solutions of the reduced system (in the unscaled time variable)

$$\begin{cases} V(t) = W \cdot S(V(t)) + I(t), \\ \dot{W} = \epsilon (S(V) \otimes S(V) - \mu W), \end{cases} \tag{A.1}$$

provided that the dynamical systems Σ in equation 3.6, and equation A.1 are well defined. It is easy to show that both systems are Lipschitz because of the properties of S. Following Faugeras et al. (2008), we know that if

$$S'_m \|W\| < 1, \tag{A.2}$$

there exists an isolated root $\bar{V} : \mathbb{R}_+ \to \mathbb{R}^N$ of the equation $V = W \cdot S(V) + I$ and $\bar{V}$ is asymptotically stable. Equation A.2 corresponds to the weakly connected case. Moreover, the initial condition belongs to the basin of attraction of this single fixed point. Note that we require $\frac{M}{T} \ll 1$ so that the membrane potentials have sufficient time to approach the equilibrium associated with a given input before the next input is presented to the network. In fact, this assumption make it reasonable to neglect the transient activity dynamics due to the switching between inputs.

A.1.2 Periodic Averaging. The system given by equation A.1 corresponds to a differential equation for W with T-periodic forcing due to the presence of V on the right-hand side. Since $T \ll \epsilon^{-1}$, we can use classical averaging methods to show that solutions of equation A.1 are close to solutions of

the following autonomous system on the time interval $[0, \frac{1}{\epsilon}]$ (which we suppose is large because $\epsilon \ll 1$):

$$\Sigma_0 : \quad \begin{cases} V(t) = W \cdot S(V(t)) + I(t) \\ \dfrac{dW}{dt} = \epsilon \left(\dfrac{1}{T} \displaystyle\int_0^T S(V(s)) \otimes S(V(s)) ds - \mu W(t) \right) \end{cases}. \tag{A.3}$$

It follows that solutions of Σ are also close to solutions of Σ_0. Finding the explicit solution $V(t)$ for each input $I(t)$ is difficult and requires fixed-point methods (e.g., a Picard algorithm). Therefore, we consider yet another system Σ' whose solutions are also close to Σ_0 and hence Σ. In order to construct Σ', we need to introduce some additional notation.

Let us label the M inputs by $I^{(a)}$, $a = 1, \ldots, M$ and denote by $V^{(a)}$ the fixed-point solution of the equation $V^{(a)} = W \cdot S(V^{(a)}) + I^{(a)}$. Given the periodic sampling of the inputs, we can rewrite equation A.3 as

$$\begin{aligned} V^{(a)} &= W \cdot S(V^{(a)}) + I^{(a)}, \\ \frac{dW}{dt} &= \epsilon \left(\frac{1}{M} \sum_{a=1}^{M} S(V^{(a)}) \otimes S(V^{(a)}) - \mu W(t) \right). \end{aligned} \tag{A.4}$$

If we now introduce the $N \times M$ matrices $\mathcal{V}$ and $\mathcal{I}$ with components $\mathcal{V}_{ia} = V_i^{(a)}$ and $\mathcal{I}_{ia} = I_i^{(a)}$, we can eliminate the tensor product and simply write equation A.4 in the matrix form

$$\begin{aligned} \mathcal{V} &= W \cdot S(\mathcal{V}) + \mathcal{I}, \\ \frac{dW}{dt} &= \epsilon \left(\frac{1}{M} S(\mathcal{V}) \cdot S(\mathcal{V})^T - \mu W(t) \right), \end{aligned} \tag{A.5}$$

where $S(\mathcal{V}) \in \mathbb{R}^{N \times M}$ such that $[S(\mathcal{V})]_{ia} = s(V_i^{(a)})$. A second application of Tikhonov's theorem (in the reverse direction) then establishes that solutions of the system Σ_0 (written in the matrix form A.5) are close to solutions of the matrix system

$$\Sigma' : \quad \begin{cases} \dfrac{d\mathcal{V}}{dt} = -\mathcal{V} + W \cdot S(\mathcal{V}) + \mathcal{I} \\ \dfrac{dW}{dt} = \epsilon \left(\dfrac{1}{M} S(\mathcal{V}) \cdot S(\mathcal{V})^T - \mu W(t) \right) \end{cases}. \tag{A.6}$$

In this letter, we focus on system Σ' whose solutions are close to those of the original system Σ provided condition A.2 is satisfied, that is, the network is weakly connected. Clearly the fixed points (V^*, W^*) of system Σ satisfy $\|W^*\| \leq \frac{S^2_m}{\mu}$. Therefore, equation A.2 says that if $\frac{S^2_m S'_m}{\mu} < 1$, then Tikhonov's theorem can be applied, and systems Σ and Σ' can be reasonably considered as good approximations of each other.

A.2 Proof of the Convergence to the Symmetric Attractor $\mathcal{A}$. We need to prove the two points: (1) $\mathcal{A}$ is an invariant set and (2) for all $\mathcal{Y}(0) \in \mathbb{R}^{N\times M} \times \mathbb{R}^{N\times N}$, $\mathcal{Y}(t)$ converges to $\mathcal{A}$ as $t \to +\infty$. Since $\mathbb{R}^{N\times N}$ is the direct sum of the set of symmetric connectivities and the set of antisymmetric connectivies, we write $W(t) = W_S(t) + W_A(t),\ \forall t \in \mathbb{R}_+$, where W_S is symetric and W_A is antisymetric.

1. In equation 2.10, the right-hand side of the equation for $\dot{W}$ is symmetric. Therefore, if $\exists t_1 \in R_+$ such that $W_A(t_1) = 0$, then W remains in $\mathcal{A}$ for $t \geq t_1$.
2. Projecting the expression for $\dot{W}$ in equation 2.10 on to the antisymmetric component leads to

$$\frac{dW_A}{dt} = -\epsilon\mu W_A(t), \tag{A.7}$$

whose solution is $W_A(t) = W_A(0)\exp(-\epsilon\mu t), \forall t \in \mathbb{R}_+$. Therefore, $\lim_{t\to+\infty} W_A(t) = 0$. The system converges exponentially to $\mathcal{A}$.

A.3 Proof of Theorem 1. Consider the following Lyapunov function (see equation 2.10),

$$E(\mathcal{U}, W) = -\frac{1}{2}\langle \mathcal{U}, W\cdot\mathcal{U}\rangle - \langle \mathcal{I}, \mathcal{U}\rangle + \langle 1, \overline{S^{-1}}(\mathcal{U})\rangle + \frac{\tilde{\mu}}{2}\|W\|^2, \tag{A.8}$$

where $\tilde{\mu} = \mu M$, such that if $W = W_S + W_A$, where W_S is symmetric and W_A is antisymmetric:

$$-\nabla E(\mathcal{U}, W) = \begin{pmatrix} W_S\cdot\mathcal{U} + I - S^{-1}(\mathcal{U}) \\ \mathcal{U}\cdot\mathcal{U}^T - \mu W \end{pmatrix}. \tag{A.9}$$

Therefore, writing the system Σ', equation 2.10, as

$$\frac{d\mathcal{Y}}{dt} = \gamma \begin{pmatrix} W_S\cdot S(\mathcal{V}) + I - S^{-1}(S(\mathcal{V})) \\ S(\mathcal{V})\cdot S(\mathcal{V})^T - \tilde{\mu}W \end{pmatrix} + \gamma \begin{pmatrix} W_A.S(\mathcal{V}) \\ 0 \end{pmatrix},$$

where $\mathcal{Y} = (\mathcal{V}, W)^T$, we see that

$$\frac{d\mathcal{Y}}{dt} = -\gamma(\nabla E(\sigma(\mathcal{V}, W))) + \Gamma(t), \tag{A.10}$$

where $\gamma(\mathcal{V}, W)^T = (\mathcal{V}, \epsilon W/M)^T$, $\sigma(\mathcal{V}, W) = (S(\mathcal{V}), W)$, and $\Gamma : \mathbb{R}_+ \to \mathcal{H}$ such that $\|\Gamma\|_{t\to+\infty} \to 0$ exponentially (because the system converges to $\mathcal{A}$). It follows that the time derivative of $\tilde{E} = E \circ \sigma$ along trajectories is given by

$$\frac{d\tilde{E}}{dt} = \left\langle \nabla\tilde{E}, \frac{d\mathcal{Y}}{dt} \right\rangle = \left\langle \nabla_{\mathcal{V}}\tilde{E}, \frac{d\mathcal{V}}{dt} \right\rangle + \left\langle \nabla_W \tilde{E}, \frac{dW}{dt} \right\rangle. \tag{A.11}$$

Substituting equation A.10 then yields

$$\begin{aligned}\frac{d\tilde{E}}{dt} &= -\langle \nabla\tilde{E}, \gamma(\nabla E \circ \sigma)\rangle + \underbrace{\langle \nabla\tilde{E}, \Gamma(t)\rangle}_{\tilde{\Gamma}(t)} \\ &= -\langle S'(\mathcal{V})\nabla_{\mathcal{U}} E \circ \sigma, \nabla_{\mathcal{U}} E \circ \sigma\rangle - \frac{\epsilon}{M}\langle \nabla_W E \circ \sigma, \nabla_W E \circ \sigma\rangle + \tilde{\Gamma}(t).\end{aligned} \tag{A.12}$$

We have used the chain rule of differentiation whereby

$$\nabla_V(\tilde{E}) = \nabla_V(E \circ \sigma) = S'(\mathcal{V})\nabla_{\mathcal{U}} E \circ \sigma,$$

and $S'(\mathcal{V})\nabla_{\mathcal{U}} E$ (without dots) denotes the Hadamard (term-by-term) product, that is,

$$[S'(\mathcal{V})\nabla_{\mathcal{U}} E]_{ia} = s'(V_i^{(a)})\frac{\partial E}{\partial U_i^{(a)}}.$$

Note that $|\tilde{\Gamma}|_{t\to+\infty} \to 0$ exponentially because $\nabla\tilde{E}$ is bounded, and $S'(\mathcal{V}) > 0$ because the trajectories are bounded. Thus, there exists $t_1 \in \mathbb{R}_+$ such that $\forall t > t_1$, $\exists k \in \mathbb{R}_+^*$ such that

$$\frac{d\tilde{E}}{dt} \le -k\|\nabla E \circ \sigma\|^2 \le 0. \tag{A.13}$$

As in Cohen and Grossberg (1983) and Dong and Hopfield (1992), we apply the Krasovskii-LaSalle invariance principle detailed in Khalil and Grizzle (1996). We check that:

- $\tilde{E}$ is lower-bounded. Indeed, $\mathcal{V}$ and W are bounded. Given that $\mathcal{I}$ and S are also bounded, it is clear that $\tilde{E}$ is bounded.
- $\frac{d\tilde{E}}{dt}$ is negative semidefinite on the trajectories as shown in equation A.13.

Then the invariance principle tells us that the solutions of system Σ' approach the set $M = \{\mathcal{Y} \in \mathcal{H} : \frac{d\tilde{E}}{dt}(\mathcal{Y}) = 0\}$. Equation A.13 implies that $M = \{Y \in \mathcal{H} : \nabla E \circ \sigma = 0\}$. Since $\frac{d\mathcal{Y}}{dt} = -\gamma(\nabla E \circ \sigma)$ and $\gamma \neq 0$ everywhere, M consists of the equilibrium points of the system. This completes the proof.

A.4 Proof of Theorem 2. Denote the right-hand side of system Σ', equation 2.10, by

$$F(\mathcal{V}, W) = \begin{cases} -\mathcal{V} + W \cdot S(\mathcal{V}) + I \\ \frac{\epsilon}{M}(S(\mathcal{V}).S(\mathcal{V})^T - \mu M W) \end{cases}.$$

The fixed points satisfy condition $F(\mathcal{V}, W) = 0$, which immediately leads to equations 2.14. Let us now check the linear stability of this system. The differential of F at $\mathcal{V}^*, W^*$ is

$$dF_{(\mathcal{V}^*, W^*)}(\mathcal{Z}, J) = \begin{pmatrix} -\mathcal{Z} + W^* \cdot (S'(\mathcal{V}^*)\mathcal{Z}) + J \cdot S(\mathcal{V}^*) \\ \frac{\epsilon}{M}((S'(\mathcal{V}^*)\mathcal{Z}) \cdot S(\mathcal{V}^*)^T + S(\mathcal{V}^*) \cdot (S'(\mathcal{V}^*)\mathcal{Z})^T - \mu M J), \end{pmatrix},$$

where $S'(\mathcal{V}^*)\mathcal{Z}$ denotes a Hadamard product, that is, $[S'(\mathcal{V}^*)\mathcal{Z}]_{ia} = s'(V_i^{*(a)})Z_i^{(a)}$. Assume that there exist $\lambda \in \mathbb{C}^*$, $(\mathcal{Z}, J) \in \mathcal{H}$ such that $dF_{(V^*, W^*)}(\begin{smallmatrix}\mathcal{Z} \\ J\end{smallmatrix}) = \lambda(\begin{smallmatrix}\mathcal{Z} \\ J\end{smallmatrix})$. Taking the second component of this equation and computing the dot product with $S(\mathcal{V}^*)$ leads to

$$(\lambda + \epsilon\mu)J \cdot S = \frac{\epsilon}{M}((S'\mathcal{Z}) \cdot S^T \cdot S + S \cdot (S'\mathcal{Z})^T \cdot S),$$

where $S = S(\mathcal{V}^*)$, $S' = S'(\mathcal{V}^*)$. Substituting this expression in the first equation leads to

$$M(\lambda + \epsilon\mu)(\lambda + 1)\mathcal{Z} = \left(\frac{\lambda}{\mu} + \epsilon\right) S \cdot S^T \cdot (S'\mathcal{Z}) + \epsilon(S'\mathcal{Z}) \cdot S^T \cdot S + \epsilon S \cdot (S'\mathcal{Z})^T \cdot S. \quad \text{(A.14)}$$

Observe that setting $\epsilon = 0$ in the previous equation leads to an eigenvalue equation for the membrane potential only:

$$(\lambda + 1)\mathcal{Z} = \frac{1}{\mu M} S \cdot S^T \cdot (S'\mathcal{Z}).$$

Since $W^* = \frac{1}{\mu M}(S \cdot S^T)$, this equation implies that $\lambda + 1$ is an eigenvalue of the operator $X \mapsto W^*.(S'X)$. The magnitudes of the eigenvalues are always smaller than the norm of the operator. Therefore, we can say that if $1 > \|W^*\| S'_m$, then all the possible eigenvalues λ must have a negative real part. This sufficient condition for stability is the same as in Faugeras et al. (2008). It says that fixed points sufficiently close to the origin are always stable.

Let us now consider the case $\epsilon \neq 0$. Recall that $\mathcal{Z}$ is a matrix. We now "flatten" $\mathcal{Z}$ by storing its rows in a vector called $\mathcal{Z}_{row}$. We use the following result in Brewer (1978): the matrix notation of operator $X \mapsto A \cdot X \cdot B$ is $A \otimes B^T$, where $\otimes$ is the Kronecker product. In this formalism, the previous equation becomes

$$\begin{aligned} &M(\lambda + \epsilon\mu)(\lambda + l)\mathcal{Z}_{row} \\ &= \left(\left(\frac{\lambda}{\mu} + \epsilon\right) S \cdot S^T \otimes I_d + \epsilon I_d \otimes S^T \cdot S + \epsilon S \otimes S^T\right) \cdot (S'Z)_{\text{row}}, \end{aligned} \tag{A.15}$$

where we assume that the Kronecker product has the priority over the dot product. We focus on the linear operator $\mathcal{O}$ defined by the right-hand side and bound its norm. Note that we use the following norm $\|W\|_\infty = \sup_X \frac{\|W.X\|}{\|X\|}$, which is equal to the largest magnitude of the eigenvalues of W:

$$\begin{aligned} \|\mathcal{O}\|_\infty \leq S'_m \Bigg(&\left|\frac{\lambda}{\mu}\right| \|S \cdot S^T \otimes I_d\|_\infty + \epsilon \|S \cdot S^T \otimes I_d\|_\infty \\ &+ \epsilon \|I_d \otimes S^T \cdot S\|_\infty + \epsilon \|S \otimes S^T\|_\infty \Bigg). \end{aligned} \tag{A.16}$$

Define ν_m to be the magnitude of the largest eigenvalue of $W^* = \frac{1}{\mu M}(S \cdot S^T)$. First, note that $S \cdot S^T$ and $S^T \cdot S$ have the same eigenvalues $(\mu M)\nu_i$ but different eigenvectors denoted by u_i for $S \cdot S^T$ and v_i for $S^T \cdot S$. In the basis set spanned by the $u_i \otimes v_j$, we find that $S \cdot S^T \otimes I_d$ and $I_d \otimes S^T \cdot S$ are diagonal with $(\mu M)\nu_i$ as eigenvalues. Therefore, $\|S \cdot S^T \otimes I_d\|_\infty = (\mu M)\nu_m$ and $\|I_d \otimes S^T \cdot S\|_\infty = (\mu M)\nu_m$. Moreover, observe that

$$\begin{aligned} (S^T \otimes S)^T \cdot (S^T \otimes S) \cdot (u_i \otimes v_j) &= (S \cdot S^T \cdot u_i) \otimes (S^T \cdot S \cdot v_j) \\ &= (\mu M)^2 \nu_i \nu_j \, u_i \otimes v_j. \end{aligned} \tag{A.17}$$

Therefore, $(S^T \otimes S)^T \cdot (S^T \otimes S) = (\mu M)^2 \mathrm{diag}(v_i v_j)$. In other words, $S^T \otimes S$ is the composition of an orthogonal operator (i.e., an isometry) and a diagonal matrix. Immediately, it follows that $\|S^T \otimes S\| \leq (\mu M) v_m$.

Compute the norm of equation A.15:

$$|(\lambda + \epsilon\mu)(\lambda + 1)| \leq S'_m(|\lambda| + 3\epsilon\mu)v_m. \tag{A.18}$$

Define $f_\epsilon : \mathbb{C} \to \mathbb{R}$ such that $f_\epsilon(\lambda) = |(\lambda + \epsilon\mu)||(\lambda + 1)| - (|\lambda| + 3\epsilon\mu)S'_m v_m$. We want to find a condition such that $f_\epsilon(\mathbb{C}_+) > 0$, where $\mathbb{C}_+$ is the right half complex plane. This condition on ϵ, μ, v_m, and S'_m will be a sufficient condition for linear stability. Indeed, under this condition, we can show that only eigenvalues with a negative real part can meet the necessary condition, equation A.18. A complex number of the right half-plane cannot be eigenvalues, and thus the system is stable. The case $\epsilon = 0$ tells us that $f_0(\mathbb{C}_+) > 0$ if $1 > S'_m v_m$; compute

$$\frac{\partial f_\epsilon}{\partial \epsilon}(\lambda) = \mu(\Re(\lambda) + \mu\epsilon)\frac{|(\lambda + 1)|}{|(\lambda + \epsilon\mu)|} - 3\mu S'_m v_m.$$

If $1 \geq \epsilon\mu$, which is probably true given that $\epsilon \ll 1$, then $\frac{|(\lambda+1)|}{|(\lambda+\epsilon\mu)|} \geq 1$. Assuming $\lambda \in \mathbb{C}_+$ leads to

$$\frac{\partial f_\epsilon}{\partial \epsilon}(\lambda) \geq \mu(\mu\epsilon - 3S'_m v_m) \geq \mu(1 - 3S'_m v_m).$$

Therefore, the condition $3S'_m v_m < 1$, which implies $S'_m v_m < 1$ and leads to $f_\epsilon(\mathbb{C}_+) > 0$.

A.5 Proof of Proposition 1. The uniform excitment of neurons on average reads, for all $i = 1, .., N$, $\|S(\mathcal{V}_i)\| = c \in \mathbb{R}_+$. For simplicity and without loss of generality, we can assume $c = 1$ and $a \geq 1$ (we can play with a and λ to generalize this to any $c, a \in \mathbb{R}_+$ as long as $a > c$). The triangular inequality we want to prove can therefore be written as

$$\sqrt{g^{-1}(S(\mathcal{V}_i).S(\mathcal{V}_j)^T)} \leq \sqrt{g^{-1}(S(\mathcal{V}_i).S(\mathcal{V}_k)^T)} + \sqrt{g^{-1}(S(\mathcal{V}_j).S(\mathcal{V}_k)^T)}. \tag{A.19}$$

For readability we rewrite $u = S(\mathcal{V}_i)$, $v = S(\mathcal{V}_j)$, and $w = S(\mathcal{V}_k)$. These three vectors are on the unity sphere and have only positive coefficients. Actually, there is a distance between these vectors that consists of computing the geodesic angle between them. In other words, consider the intersection of *vect*(*u*,*v*) and the *M*-dimensional sphere. This is a circle where u and v are located. The angle between the two points is written $\theta_{u,v}$. It is a distance on

the sphere and thus satisfies the triangular inequality:

$$\theta_{u,v} \leq \theta_{u,w} + \theta_{w,v}. \tag{A.20}$$

Actually, all these angles belong to $[0, \frac{\pi}{2}[$ because u, v, and w only have positive coefficients.

Observe that $g^{-1}(u.v^T) = g^{-1}\big(\cos(\theta_{u,v})\big)$ and separate the two cases for the choice of function g:

1. If $g(x) = a\big(1 - \frac{x}{\lambda^2}\big)$ with $a \geq 1$, then $g^{-1}(x) = \lambda^2\big(1 - \frac{x}{a}\big)$. Therefore, define $h_1 : x \mapsto \lambda\sqrt{1 - \frac{\cos(x)}{a}}$. We now want to apply h_1 to equation A.20, but h_1 is monotonic only if $x \leq \frac{\pi}{2}$. Therefore, divide equation A.20 by 2 and apply h_1 on both sides to get

$$h_1\left(\frac{\theta_{u,v}}{2}\right) \leq h_1\left(\frac{1}{2}\theta_{u,w} + \frac{1}{2}\theta_{w,v}\right).$$

   Now consider the function $\eta_a(x, y) = h_1(x + y) - h_1(x) - h_1(y)$ for $x, y \in [0, \frac{\pi}{4}[$. Because, h_1 is increasing, it is clear that $\frac{\partial \eta_a}{\partial x}(x, y) \leq 0$ (and similarly for y), such that η_a reaches its maximum for $x = y = \frac{\pi}{4}$. Besides, $\eta_a(\frac{\pi}{4}, \frac{\pi}{4}) \leq \eta_1(\frac{\pi}{4}, \frac{\pi}{4}) < 0$. This proves that $2h_1(\frac{1}{2}\theta_{u,w} + \frac{1}{2}\theta_{w,v}) \leq h_1(\theta_{u,w}) + h_1(\theta_{w,v})$. Moreover, it is easy to observe that $2h_1(\frac{x}{2}) \geq h_1(x)$ for all $a > 1$. This concludes the proof for $g(x) = a\big(1 - \frac{x}{\lambda^2}\big)$.
2. If $g(x) = ae^{-x}$, then $g^{-1}(x) = \lambda^2 ln(\frac{a}{x})$. As before, define $h_2 : x \mapsto \lambda\sqrt{ln(\frac{a}{\cos(x)})}$. We still want to apply h_2 to equation A.20, but h_2 is not defined for $x > \frac{\pi}{2}$, which is likely for the right-hand side of equation A.20. Therefore, we apply h_2 to equation A.20, divided by two, and use the fact that h_2 is increasing on $[0, \frac{\pi}{2}[$. This leads to $h_2(\frac{\theta_{u,v}}{2}) \leq h_2(\frac{1}{2}\theta_{u,w} + \frac{1}{2}\theta_{w,v})$. First, we use the convexity of h_2 to get $2h_2(\frac{\theta_{u,v}}{2}) \leq h_2(\theta_{u,w}) + h_2(\theta_{w,v})$, and then we use the fact that $2h_2(\frac{x}{2}) \geq h_2(x)$ for $x \in [0, \delta[$, with $\delta < \frac{\pi}{2}$. This would conclude the proof, but we have to make sure the angles remain in $[0, \delta[$. Actually, we can compute δ, which verifies $2h_2(\frac{\delta}{2}) = h_2(\delta)$. This leads to $\delta = 2\arccos(\sqrt{a^3 - \sqrt{a^6 - a^3}})$. In fact, the coefficients of u, v, and w are strictly positive and larger than $S(0)$. Therefore, the angles between them are strictly smaller than $\frac{\pi}{2}$. More precisely, $\theta_{u,v} \in [0, \frac{\pi}{2} - 2\arcsin\big(S(0)\big)[$. Therefore, a necessary condition for the result to be true is $2\arccos(\sqrt{a^3 - \sqrt{a^6 - a^3}}) \geq \frac{\pi}{2} - 2\arcsin\big(S(0)\big)$, using the fact that $\arccos(x) = \frac{\pi}{2} - \arcsin(x)$ leads to $\arcsin\big(S(0)\big) - \arcsin(\sqrt{a^3 - \sqrt{a^6 - a^3}}) \geq \frac{\pi}{8}$.

Acknowledgments

M.N.G. and O.D.F. were partially funded by ERC advanced grant NerVi nb 227747. M.N.G. was partially funded by the région PACA, France. This letter

was based on work supported in part by the National Science Foundation (DMS-0813677) and by Award KUK-C1-013-4 made by King Abdullah University of Science and Technology. P.C.B. was also partially supported by the Royal Society Wolfson Foundation. This work was partially supported by the IP project BrainScales No. 269921.

References

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, *10*(2), 251–276.

Amari, S., Kurata, K., & Nagaoka, H. (1992). Information geometry of Boltzmann machines. *IEEE Transactions on Neural Networks*, *3*(2), 260–271.

Bartsch, A., & Van Hemmen, J. (2001). Combined Hebbian development of geniculocortical and lateral connectivity in a model of primary visual cortex. *Biological Cybernetics*, *84*(1), 41–55.

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

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

Borg, I., & Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications*. New York: Springer-Verlag.

Bosking, W., Zhang, Y., Schofield, B., & Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *Journal of Neuroscience*, *17*(6), 2112–2127.

Bressloff, P. (2005). Spontaneous symmetry breaking in self-organizing neural fields. *Biological Cybernetics*, *93*(4), 256–274.

Bressloff, P. C., & Cowan, J. D. (2003). A spherical model for orientation and spatial frequency tuning in a cortical hypercolumn. *Philosophical Transactions of the Royal Society B*, *358*, 1643–1667.

Bressloff, P., Cowan, J., Golubitsky, M., Thomas, P., & Wiener, M. (2001). Geometric visual hallucinations, Euclidean symmetry and the functional architecture of striate cortex. *Phil. Trans. R. Soc. Lond. B*, *306*(1407), 299–330.

Brewer, J. (1978). Kronecker products and matrix calculus in system theory. *IEEE Transactions on Circuits and Systems*, *25*(9), 772–781.

Chklovskii, D., Schikorski, T., & Stevens, C. (2002). Wiring optimization in cortical circuits. *Neuron*, *34*(3), 341–347.

Chossat, P., & Faugeras, O. (2009). Hyperbolic planforms in relation to visual edges and textures perception. *PLoS Computational Biology*, *5*(12), 367–375.

Cohen, M., & Grossberg, S. (1983). Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, SMC*, *13*, 815–826.

Coifman, R., Maggioni, M., Zucker, S., & Kevrekidis, I. (2005). Geometric diffusions for the analysis of data from sensor networks. *Current Opinion in Neurobiology*, *15*(5), 576–584.

Coombes, S. (2005). Waves, bumps, and patterns in neural field theories. *Biological Cybernetics*, *93*(2), 91–108.

Cottet, G. (1995). Neural networks: Continuous approach and applications to image processing. *Journal of Biological Systems*, *3*, 1131–1139.

Dayan, P., & Abbott, L. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

Degond, P., & Mas-Gallic, S. (1989). The weighted particle method For convection-diffusion equations: Part 1: The case of an isotropic viscosity. *Mathematics of Computation*, *53*, 485–507.

Dong, D., & Hopfield, J. (1992). Dynamic properties of neural networks with adapting synapses. *Network: Computation in Neural Systems*, *3*(3), 267–283.

Edwards, R. (1996). Approximation of neural network dynamics by reaction-diffusion equations. *Mathematical Methods in the Applied Sciences*, *19*(8), 651–677.

Faugeras, O., Grimbert, F., & Slotine, J.-J. (2008). Abolute stability and complete synchronization in a class of neural fields models. *SIAM J. Appl. Math*, *61*(1), 205–250.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*(2), 194–200.

Geman, S. (1979). Some averaging and stability results for random differential equations. *SIAM J. Appl. Math*, *36*(1), 86–105.

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

Hebb, D. (1949). *The organization of behavior: A neuropsychological theory*. Hoboken, NJ: Wiley.

Hubel, D. H., & Wiesel, T. N. (1977). Functional architecture of macaque monkey visual cortex. *Proc. Roy. Soc. B*, *198*, 1–59.

Khalil, H., & Grizzle, J. (1996). *Nonlinear systems*. Upper Saddle River, NJ: Prentice Hall.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Miikkulainen, R., Bednar, J., Choe, Y., & Sirosh, J. (2005). *Computational maps in the visual cortex*. New York: Springer.

Miller, K. (1996). Synaptic economics: Competition and cooperation in synaptic plasticity. *Neuron*, *17*, 371–374.

Miller, K. D., Keller, J. B., & Stryker, M. P. (1989). Ocular dominance column development: Analysis and simulation. *Science*, *245*, 605–615.

Miller, K., & MacKay, D. (1996). The role of constraints in Hebbian learning. *Neural Comp*, *6*, 100–126.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J. Math. Biology*, *15*, 267–273.

Ooyen, A. (2001). Competition in the development of nerve connections: A review of models. *Network: Computation in Neural Systems*, *12*(1), 1–47.

Petitot, J. (2003). The neurogeometry of pinwheels as a sub-Riemannian contact structure. *Journal of Physiology–Paris*, *97*(2–3), 265–309.

Sejnowski, T., & Tesauro, G. (1989). The Hebb rule for synaptic plasticity: Algorithms and implementations. In J. H. Byrne & W. O. Berry (Eds.), *Neural models of plasticity: Experimental and theoretical approaches* (pp. 94–103). Orlando, FL: Academic Press.

Swindale, N. (1996). The development of topography in the visual cortex: A review of models. *Network: Computation in Neural Systems*, *7*(2), 161–247.

Takeuchi, A., & Amari, S. (1979). Formation of topographic maps and columnar microstructures in nerve fields. *Biological Cybernetics*, *35*(2), 63–72.
Tikhonov, A. (1952). Systems of differential equations with small parameters multiplying the derivatives. *Matem. sb*, *31*(3), 575–586.
Verhulst, F. (2007). Singular perturbation methods for slow-fast dynamics. *Nonlinear Dynamics*, *50*(4), 747–753.
Wallis, G., & Baddeley, R. (1997). Optimal, unsupervised learning in invariant object recognition. *Neural Computation*, *9*(4), 883–894.
Zucker, S., Lawlor, M., & Holtmann-Rice, D. (2011). Third order edge statistics reveal curvature dependency. *Journal of Vision*, *11*, 1073.

Received February 1, 2011; accepted March 1, 2012.