Linear analysis of auto-organization in Hebbian neural networks

JUAN CARLOS LETELIER and JORGE MPODOZIS

Departamento de Biología, Facultad de Ciencias, Universidad de Chile, Santiago, Chile

The self-organization of neurotopies where neural connections follow Hebbian dynamics is framed in terms of linear operator theory. A general and exact equation describing the time evolution of the overall synaptic strength connecting two neural laminae is derived. This linear matricial equation, which is similar to the equations used to describe oscillating systems in physics, is modified by the introduction of non-linear terms, in order to capture self-organizing (or auto-organizing) processes. The behavior of a simple and small system, that contains a non-linearity that mimics a metabolic constraint, is analyzed by computer simulations. The emergence of a simple "order" (or degree of organization) in this low-dimensionality model system is discussed.

***Key terms**: linear operator theory, linear systems, neurotopy, self-organization.*

INTRODUCTION

Artificial neural networks are a conceptual tool developed during the last half of the 20th century as a tool to build computational theories of brain function. In 1943, appeared what we can call "The first paper" in this field (McCulloch and Pitts, 1943). This work had an "algebraic" flavor as it tried to obtain results and theorems based on the point of view that neurons behave as elementary logical functionals like disjunction or negation. This fundamental paper, undoubtedly influenced by the work of Alan Turing, tried to deduce very general results about brain function. The pathway opened by this paper was important as it made clear that the understanding of brain function requires, beside the accumulation of experimental data, a theoretical framework. But neurons are not the simple "logical elements" envisaged by those authors, they have a complex biology in which the electrical phenomena at the level of the axon are the end result of complex biophysical interactions between a continuously changing set of postsynaptic currents that propagates in linear, and non linear, fashion between thousands of synapses and the cell body.

This more accurate picture of neuronal dynamics has been incorporated in neural networks models since the mid-fifties (see Anderson and Rosenfeld, 1988, for a review; Rochester *et al*, 1956, for an example of an early realistic simulation). An important simplification is usually assumed as many complex biophysical interactions are ignored and neurons are treated as simple linear elements that perform a weighted average of its inputs (Fig 1). Thus the output of a neuron is constructed as a linear combination of its inputs weighted by the synaptic strength (C_j) of each input:

$$\textbf{Output activity} = C_1 * (\text{Input}_1) + C_2 * (\text{Input}_2) + \ldots + C_n * (\text{Input}_n) = \sum_{j=1}^{n} C_j * (\text{Input}_j) \qquad \text{Eqn 0}$$

* **Correspondence to:** Dr Juan Carlos Letelier, Departamento de Biología, Facultad de Ciencias, Universidad de Chile, Casilla 653, Santiago, Chile. Fax: (56-2) 271-2983. E-mail: letelier@abello.seci.uchile.cl

In spite of the extreme character of this simplification, a complete field of research has been created around this view of neuronal function.

Another important factor incorporated into neural network models is the concept of "plasticity" or "learning" at the level of synaptic connections. The synaptic weights (the $\mathbf{C}_j$ of Fig 1) are thus assumed to be continuously varying over time. This variation changes the effective connectivity of the network; hence its computational properties. Two main mechanisms are used to define the direction and magnitude of the change for each $\mathbf{C}_j$. One type of mechanism depends on a global rule that somehow "knows" which are the "correct" values that the $\mathbf{C}_j$ must have and adjusts sequentially every synaptic connection in the network (Kohonen, 1982). Examples of this method of adjusting the $\mathbf{C}_j$ originated the "perceptron" (Block, 1962) and, more recently, the many versions of multi-layer neural networks constructed by the "back-propagation" method (Rumelhart *et al*, 1986). The second mechanism is more interesting from the point of view of neurobiology as the rule for changing each $\mathbf{C}_j$ is local, depending only on the input and output activity of neurons. This notion reflects our deep expectations about the biochemical steps that must happen at the pre and post synaptic levels. In effect, our current conception of synaptic plasticity demands that locally produced neuro-transmitters, factors or reverse-transmitters trigger the action of a complex enzymatic cascade that changes the biophysical properties of the synapse. This notion is not new, clear references can be found in writings of Cajal and Pavlov, but because Donald Hebb enunciated it in a particularly clear way in an influential book (Hebb, 1949), this type of plasticity is known today as "Hebbian". Hebb's rule can be translated in a particular simple mathematical form:

Change in synaptic strength = μ **(presynaptic activity) * (postsynaptic activity)**
For a particular synapse: $\Delta \mathbf{C}_j = \mu$ **(Input$_j$) * (Output activity)**

μ is a proportionality factor that specifies the rate of change. Because the change in strength depends on the simultaneous activation of the synapse and the postsynaptic cells this rule is also known as a "correlation" rule.

These two ideas (*i.e.*, that neurons produce a weighted average of their inputs, and that synaptic strength depends on the correlation between pre and post synaptic activities) are at the very center of most modern models of neural activity. A review of this enormous field is not our goal. Instead this paper explores how, when these two ideas are taken together, an equation for the time evolution of the $\mathbf{C}_j$ can be found. Furthermore, this evolution equation is similar to equations found in the physics of oscillating systems, such as strings and membranes, and is related with the algebraic analysis of systems of linear differential equations and linear operator theory.

MATHEMATICAL FORMULATION AND COMPUTER SIMULATION

To clarify ideas we consider the following model of a "neural network". Let **Q** and **P** represent two one-dimensional layers with **q** and **p** neurons respectively (Fig 2). The **Q** layer has no lateral interaction between its neurons, while the **P** layer has lateral interactions represented by *intralaminar* synaptic weights **W** and receives input fibers from **Q** represented by $\mathbf{C}_s$. **Q** layer neurons are indexed by Greek symbols (α, β, ... γ) while **P** neurons are indexed by Latin symbols (x, y, ... z). The *interlaminar* synaptic weight between **Q** layer neuron at position α and a **P** layer neuron at position **x**, at time **t**, is $\mathbf{C}_{x\alpha}(\mathbf{t})$. In the **P** layer, the influence that cell at position **y** has over a neuron located at position **x** is represented by $\mathbf{W}_{xy}$ and it is considered time-invariant. The activities of **Q** and **P** neurons are represented by **I** and **A**, respectively. The **Q** layer could be thought as the "input" layer while the **P** layer would be the "processing" layer.

Our first task is to combine the notions described by equation 0, figure 1 and figure 2 to obtain an expression for the rate of change of the interlaminar synaptic weights $\mathbf{C}_{x\alpha}(\mathbf{t})$. It is important to notice that the complete set of $\mathbf{C}_{x\alpha}(\mathbf{t})$ forms a rectangular **q** x **p**

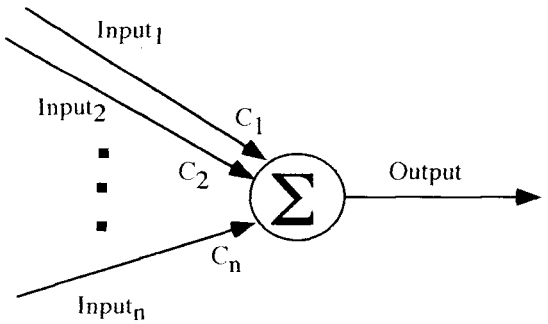


Fig 1. *The neuron as a weighted averager.* The neuronal inputs (Inputx) affect the post-synaptic cell through coupling variables (C_x) that reflect the synaptic strength. The post-synaptic output ("Output") is simply the summation of all the terms (C_x * $Input_x$). This formal model of a neuron is very simple as it neglects saturation effects.

matrix **C(t)**. This calculation would be developed with care as its deduction illuminates important aspects of the mathematics behind neural networks.

The activity of a **P** neuron, at position **x**, is equal to:

$$A_x(t) = \{W_{x1}A_1(t) + W_{x2}A_2(t) + ... + W_{xp}A_p(t)\} + \{C_{x1}(t)I_1(t) + C_{x2}(t) + ... + C_{xq}(t)I_q(t)\}$$

This equation establishes that the activity of every **P** cell is controlled by two contributions. One is derived from outside the **P** lamina and is modulated by the set of inter-laminar weights **C** (right parenthesis), and the other is produced inside the **P** layer and modulated by the synaptic weights **W** that represent lateral interactions (left parenthesis). This equation shows that the activity of any given neuron of the **P** layer depends on all the other cells of that same layer. The solution, for the activities $A_x(t)$, can be easily obtained, as this equation is formally equivalent to a system of (linear) **p** linked equations. In fact, the set of equations describing the activity of all **P** layer cells is:

$$(1 - w_{11})A_1(t) - w_{12}A_2(t) - ... - w_{1p}A_p(t) = \sum_{j=1}^{q} C_{1j}(t)I_j$$

$$-w_{x1}A_1(t) - w_{x2}A_2(t) - ... - (1 - w_{xx})A_x(t) - ... - w_{xp}A_p(t) = \sum_{j=1}^{q} C_{xj}(t)I_j$$

$$-w_{p1}A_1(t) - w_{p2}A_2(t) - ... - (1 - w_{pp})A_p(t) = \sum_{j=1}^{q} C_{pj}(t)I_j$$

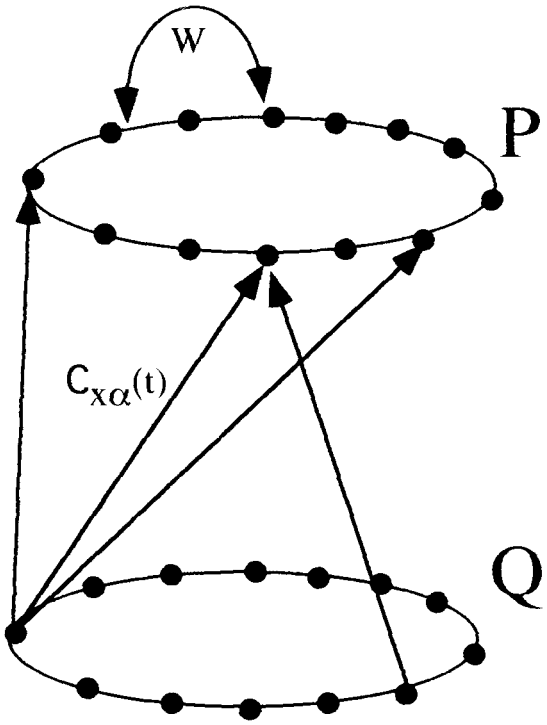


Fig 2. *Geometrical model of the neural network.* Two layers of neurons (P and Q) are connected via plastic synaptic connections ($C_{x\alpha}(t)$ that change over time according to Hebb's rule (arrows). Each P layer neuron receives inputs from all Q layer neurons and from all P layer neurons via an invariant network of lateral interactions (depicted here by **W**) that are assumed to decrease with distance. In order to avoid "border" effects, and to maintain strict translational invariance, each layer is assumed to be connected as a circle, thus every neuron has a "left" and "right" neighbors. This architecture is easily applied to bi-dimensional layers.

This linear system of equations can be written in matricial form as:

$$\mathbf{(Id-W)A(t) = C(t)I(t)}$$

with **Id** denoting the identity matrix of order **p**.

Thus, if **(Id-W)** is invertible, it has the following solution:

$$\mathbf{A(t) = (Id-W)^{-1}C(t)I(t) = \widetilde{W}C(t)I(t)}$$

with $\widetilde{W} = (Id-W)^{-1}$

Thus, given a set of input activities **I**, in layer **Q**, the set of activities **A** induced in layer **P**, at position **x**, is:

$$A_x(t) = \sum_{y=1}^{p} \widetilde{w}_{xy} \sum_{\beta=1}^{q} C_{y\beta}(t) I_{\beta}(t) \qquad \text{Eqn I}$$

Using this last equation we can calculate the rate of change of the matrix **C(t)** over time. As stated above, if each single inter-laminar connection follows a Hebbian dynamics it instantaneous rate of change is:

$$\Delta C_{x\alpha}(t) = \mu\, A_x(t)\, I_{\alpha}(t) \qquad \text{Eqn II}$$

This basic equation relates the strength of a single inter-laminar synaptic weight with the activities of both layers, and we would like to use it as a starting point to derive an "evolution" equation relating **C(t)** with structural parameters of layers **Q** and **P**.

The mean rate of change of $C_{x\alpha}(t)$ can be approximated as:

$$\Theta_m(t) = C_{x\alpha}(t+k) - Cx\alpha(t)$$

which in term of the instantaneous change is:

$$\Theta_m(t) = \sum_{i=1}^{m} \Delta C_{x\alpha}(t+i) = \mu \sum_{i=1}^{m} A_x(t+i) I_{\alpha}(t+i)$$

using the expression for $A_x(t)$:

$$\Theta_m(t) = \mu \sum_{i=1}^{m} \left\{ \sum_{y=1}^{p} \widetilde{w}_{xy} \sum_{\beta=1}^{q} C_{y\beta}(t+i) I_{\beta}(t+i) \right\} I_{\alpha}(t+1)$$

Because the intra-laminar interactions **W** do not change with time and we are only interested in the mean variations of **C** -thus effectively stating that $C_{y\beta}(t+k) \approx C_{y\beta}(t)\ \forall y, \beta$ (see Fig 3)- we can approximate the time-evolution of each $C_{y\beta}$ by two components. One component represents the average *trend* of the variation of $C_{y\beta}$, while the other component is an unpredictable "noise". Thus the trend is obtained by taking the average of **Cy_** during time intervals that are "long" with respect to the fast and noisy transitions. With this simplification the last equation can be written as:

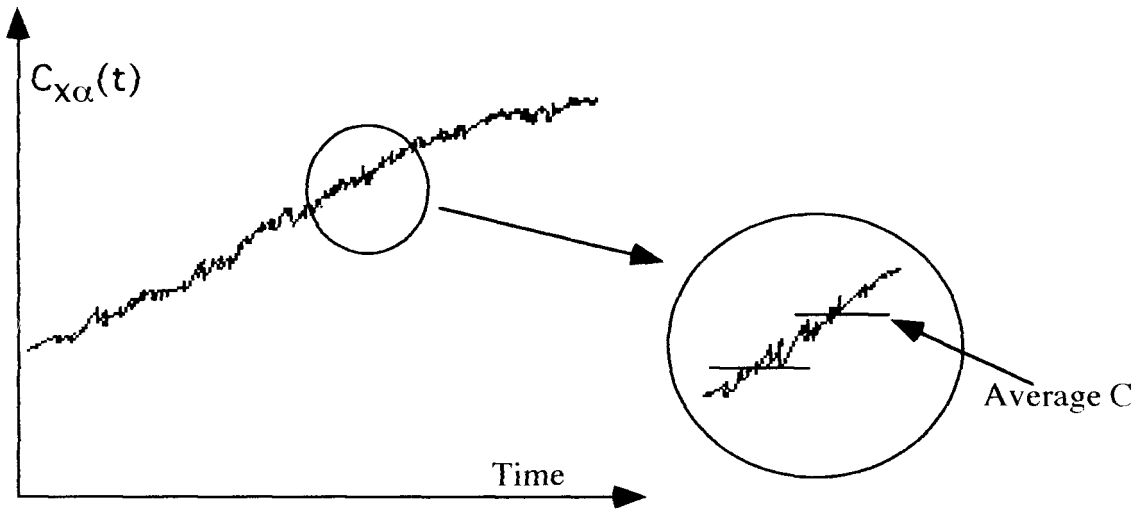


Fig 3. *A crucial approximation.* The dynamics that governs the time evolution of each ($C_{x\alpha}(t)$ can be subdivided into fast and slow components. The fast component can be thought as a random "noise" (jagged line) superimposed to a general trend. Thus, locally, the ($C_{x\alpha}(t)$ profile is approximated by a step function that replaces ($C_{x\alpha}(t)$) by a local average of the recent past. Without this approximation equation III can not be deduced. This prediction helps us to obtain an evolution equation about the average value of each connection.

$$\Theta_m(t) = \mu \sum_{y=1}^{p} \sum_{\beta=1}^{q} \tilde{w}_{xy} C_{y\beta}(t) \left\{ \sum_{i=1}^{m} I_\alpha(t+i)\ I_\beta(t+1) \right\} \text{ with } \left\{ \sum_{i=1}^{m} I_\alpha(t+i)\ I_\beta(t+i) \right\} = m\, J_{\alpha\beta}\ \forall t$$

The expression inside the brackets can be thought of terms of the spatial auto-correlation of the activity of layer **Q**. If we assume that this layer is analogous to a sensory lamina (like the retina), its neural activity **I** must have the same spatial correlations **J** as the spatial correlations of the set of stimuli impinging of that lamina. Thus, the instantaneous rate of change of the mean is:

$$\Delta C_{x\alpha}(t) = \frac{C_{x\alpha}(t+k) - C_{x\alpha}(t)}{m} = \mu \sum_{y,\beta}^{p,q} \tilde{w}_{xy} C_{y\beta}(t) J_{\alpha\beta}$$

In matrix form, the following evolution equation can be obtained:

$$\dot{\mathbf{C}} = \mu\, \tilde{\mathbf{W}}\, \mathbf{C}\, \mathbf{J} \qquad \text{Eqn III}$$

This equation, which is many times asserted without demonstration, serves as a foundation of all modern studies concerning auto-organization (Malsburg, 1973; Linsker, 1986; Miller *et al*, 1989; Goodhill, 1993), and has many important consequences. First, it is a linear differential equation where a linear operator $H(X) = \tilde{W}XC$ appears spontaneously (thus we are looking to solutions of $X' = \mu H(X)$). Second, the operator H reflects two "structural" facts: a) the connectivity of layer **P** and b) the correlation of the set of inputs.

Interestingly, equations of this type have been intensively studied by modern functional analysis. The solutions can be found by studying the eigenvalues and eigenvectors of operator H and they have the following general form:

$$\mathbf{C(t)} = \psi_{ml}\, e^{\lambda_{ml} t}$$

where ψ_{ml} is an eigenvector of **H** and λ_{ml} its associated eigenvalue.

However, a simple glimpse shows that this mathematical model is incomplete to simulate the auto-organization of neural connections. In effect, the solutions of equation III are all exponential functions that grow (or decay) towards infinity (or to zero), thus collapsing any order in the connectivity pattern. In the literature it is common to see renormalization techniques, like to divide each $\mathbf{C}_{x\alpha}(t)$ by the maximal value of **C(t)** at that moment to tackle this problem. This *ad-hoc* technique solves the mathematical problem of uncontrolled growth, but it is an unsatisfactory method as it does not map easily into a physiological process.

To mathematically capture an *auto-organizing agent*, equation II should be modified with extra terms that reflect mechanisms of synaptic modification not based in correlation (*i.e.* "extra Hebbian mechanisms") and are coherent with cellular mechanisms. The first modification is to incorporate competition, among synapses, for the presence of a scarce metabolite required for changing the synaptic apparatus. In this context, each synapse in the network is under the action of two "forces": a positive growth due to a Hebbian mechanism and a decay due to the level of "maturation" of the network. Initially this (secondary) effect is small, but it grows non-linearly and it affects all synapses equally. Equation III is then transformed to:

$$\dot{\mathbf{C}} = \mu\, \tilde{\mathbf{W}}\, \mathbf{C}\, \mathbf{J} - \| \mathbf{C} \|^2$$

The above equation reflects how a competition mechanism is incorporated into the language of linear operators. It can not be solved analytically as it contains two "non-linearities": the quadratic term and the fact that the **C**'s can not be negative (we are only modeling excitatory synapses). In the absence of an analytical solution the equation can be numerically solved; however, the framework of linear systems can help us to understand the simulations and to predict, quantitatively, the properties of the end state. To clarify ideas we are going to present examples of one-dimensional networks with

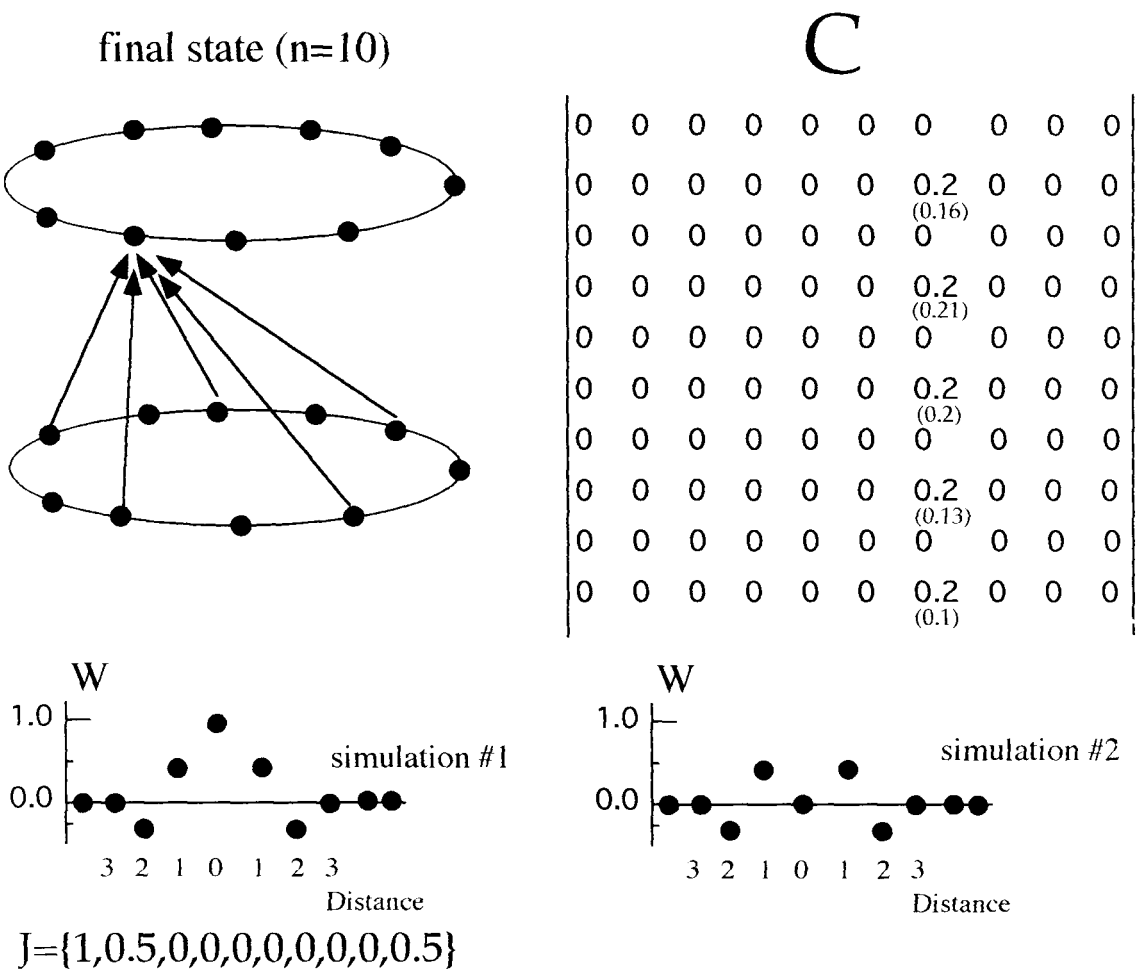
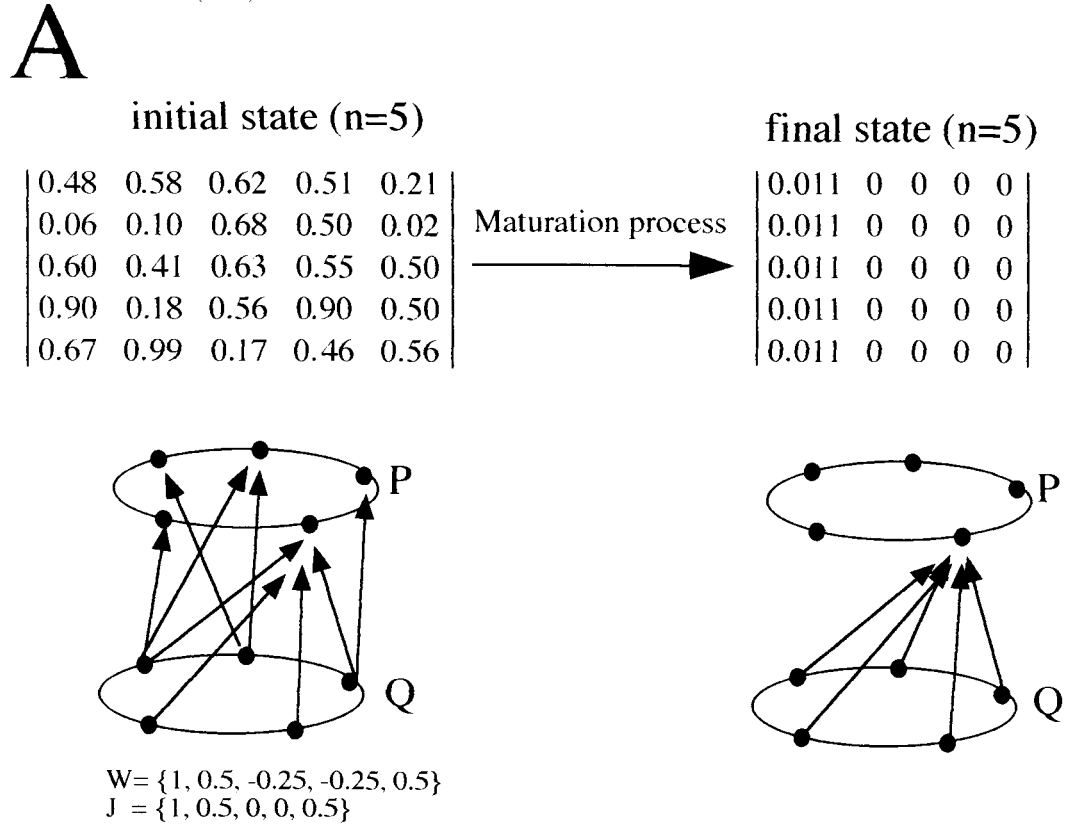


Fig 4. *Auto-organization in a small network (n = 10).* Two examples of the final state when n is even. The two simulations differ in the pattern of lateral interactions, in simulation #1 the autoexcitation value $W_0 = 1.0$; in simulation #2, $W_0 = 0$. This difference, which radically changes matrix $\widetilde{W}$, does not qualitatively affect the final state. A single P layer neuron receives all the inputs from alternate cells in layer Q. The final results of simulation #2 are given in parentheses in the final matrix **C**. In both cases the **J** matrix was built by shifting the vector **J**.

small numbers of neurons (n = 5, ... 10). In these examples, the network topology is assumed to be a cylinder and the pattern of lateral interactions, as well as the spatial correlation of the set of stimuli, are invariant under translation. Thus, matrices **W** and **J** are circular matrices. This geometry is artificial, as it avoids border effects, but enables the use of linear analysis and clearly shows the emergence of order in a spatially homogenous system. Under these restrictions linear analysis of equation II predicts that: a) the final steady state is reached in exponential time, b) the final connections must contain symmetries, c) some of these symmetries must reflect the periodic solutions proper of the homogeneous (linear) system $X' = \mu H(X)$.

Simulation for n = 6, 8, 10

The cylindrical topology, when **n** is even, implies that a periodic function (*i.e.*, the natural eigenfunctions of the homogeneous problem) can be fitted exactly. Numerical simulations of these cases show an striking degree of final auto-organization (Fig 4), with only one **P** neuron receiving inputs from the **Q** layer. Most of the original connections ($n^2 = 36$, 64 or 100) disappear and only n/2 connections survive. Interestingly, the surviving connections are symmetrically separated by connections of strength 0 and this pattern does not depend on a specific configuration of the intensity of lateral interaction (**W**) in layer P. This alternation



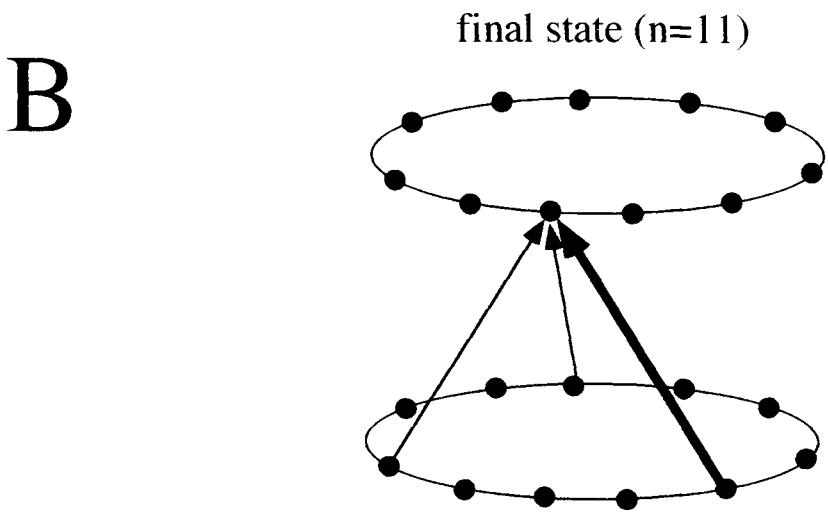


Fig 5. *Auto-organization in a small network (n = 5 and 11).* Two examples of auto-organization when n is odd. **A)** For small n (5) every Q cell layer projects to the P layer equally. Thus all arrows depicted in the final state have the same value. To give an idea of the initial state the initial matrix **C** is explicitly given, some (9) of the initial "arrows" connecting P to Q are also given. The final state is independent of the values found in the initial matrix **C. B)** For larger n (9, 11, 13) the final state resembles the final situation found for n = even. The final connections are not equal (the wider and darker arrow represents a final connection having approximately the double value that the other two) and they are not equally spaced in the network. The **W** and **J** matrices, for both simulations, were built using the $\widetilde{\mathbf{W}}$ and **J** vectors shown in A.

reflects the competition between adjacent connections mediated by the lateral interactions in layer **P**. Because of the strong symmetry derived from translational invariance the **P** neuron that receives all the surviving inputs change from simulation to simulation, but the overall pattern remains unchanged.

Simulation for n = 5, 7, 9, 11

Networks with an odd number of neurons behave rather differently because the eigenfunctions of the homogenous problem do not fit exactly into the network. Thus, as the final solution must respect the fundamental symmetry introduced by translational invariance, a qualitative new form is reached. The final state again has most of its connections equal to 0, and only a single **P** layer neuron receives all the connections. The difference, with respect to the case **n = 2 * k**, arises in the number and identity of the neurons from the input layer that project to the **P** layer (Fig 5). The periodic pattern found for even cases is replaced by a more irregular one, in which not all surviving connections have the same final strength (Fig 5A) and where the strict periodic alternation is replaced by an approximate alternation, or every **Q** layer neuron sends inputs to a single **P** layer cell (Fig 5B). Also, the final state is reached much more slowly demanding 5-10 times more iterations.

CONCLUSION

The language of linear operators, and especially the notion of eigenfunctions, is particularly suited to formalize auto-organizing neural systems that follow Hebb's rule with modifications that reflect metabolic constraints. Although current mathematical theory can not solve exactly the non-linear problems that characterize self-organization in such complex systems, it can give qualitative arguments about the asymptotic behavior of solutions. The study of small systems in search of semi-analytical tools is extremely important as pure numerical (*i.e.*, computer) simulations become more and more complex. In effect, in current computer simulations it is not clear which self-organizing features are the consequences of *ad-hoc* elaborated numerical techniques. Finally, we must add that the examples shown in this paper are particular examples of a more general phenomenon: the appearance of order through the recursive application of local rules among components having the same set of properties.

REFERENCES

ANDERSON JA, ROSENFELD E (eds) (1988) Neurocomputing: Foundations of Research. Cambridge, MA: MIT Press

BLOCK HD (1962) The perceptron: a model for brain functioning. I. Rev Mod Phys 34: 123-135

GOODHILL GJ (1993) Topography and ocular dominance: a model exploring positive correlations. Biol Cybern 69: 109-118

McCULLOCH WS, PITTS W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5: 115-133

HEBB DO (1949) The Organization of Behavior. New York, Wiley

KOHONEN T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43: 59-69

LINSKER R (1986) From basic network principles to neural architecture: Emergence of spatial-opponent cells. Proc Natl Acad Sci USA 83: 7508-7512

MALSBURG C von der (1973) Self-organization of orientation sensitive cells in the striate cortex. Kybernetik 14: 85-100

MILLER KD, KELLER JB, STRYKER MP (1989) Ocular dominance column development: analysis and simulation. Science 245: 605-615

ROCHESTER N, HOLLAND JH, HAIBT LH, DUDA WL (1956) Test on a cell assembly theory of the action of the brain using a large digital computer. IRE Trans Inform Th IT 2: 80-93

RUMELHART DE, HINTON GE, WILLIAMS RJ (1986) Learning representations by back-propagating errors. Nature 323: 533-536