Analysis of Linsker's application of Hebbian rules to Linear Networks

David J. C. MacKay
Department of Computation and Neural Systems
California Institute of Technology 164–30 CNS
Pasadena CA 91125
mackay@aurel.cns.caltech.edu

Kenneth D. Miller
Department of Physiology
University of California at San Francisco
San Francisco CA 94143 – 0444
ken@phyb.ucsf.edu

Abstract

Linsker has reported the development of structured receptive fields in simulations using a Hebb–type synaptic plasticity rule in a feed–forward linear network. The synapses develop under dynamics determined by a matrix that is closely related to the covariance matrix of input cell activities. We analyse the dynamics of the learning rule in terms of the eigenvectors of this matrix. These eigenvectors represent independently evolving weight structures. Some general theorems are presented regarding the properties of these eigenvectors and their eigenvalues. For a general covariance matrix four principal parameter regimes are predicted.

We concentrate on the gaussian covariances at layer $\mathcal{B} \rightarrow \mathcal{C}$ of Linsker's network. Analytic and numerical solutions for the eigenvectors at this layer are presented. Three eigenvectors dominate the dynamics: a DC eigenvector, in which all synapses have the same sign; a bi–lobed, oriented eigenvector; and a circularly symmetric, centre–surround eigenvector. Analysis of the circumstances in which each of these vectors dominates yields an explanation of the emergence of centre–surround structures and symmetry–breaking bi–lobed structures. Criteria are developed estimating the boundary of the parameter regime in which centre–surround structures emerge. The application of our analysis to Linsker's higher layers, at which the covariance functions were oscillatory, is briefly discussed.

Published as: *Network* **1**:257-298 (1990).

6/29/95: Present addresses of authors:
David MacKay
Radio Astronomy, Cavendish Laboratory, Madingley Road, Cambridge, CB3 0HE. U.K.
e-mail: mackay@mrao.cam.ac.uk; www: ftp://131.111.48.24/pub/mackay/homepage.html

Ken Miller
Dept. of Physiology, UCSF, SF, CA 94143-0444
e-mail: ken@phy.ucsf.edu; www: http://keck.ucsf.edu/ ken

Figure 1: **The first three layers of Linsker's network.**
Each neuron receives synapses only from neurons in the previous layer. There are no intralaminar or feedback connections. The density of synapses received by a neuron has a gaussian distribution, so that most synapses are received from neurons in a circular area directly below. This distribution of synapses is fixed, while the strengths of the synapses change according to equation (3). Cells in layer $\mathcal{A}$ are uncorrelated in their activities, and connect via excitatory synapses of fixed strength $w_{\max}$ to cells of layer $\mathcal{B}$. The overlap between the inputs to neurons in layer $\mathcal{B}$ produces gaussian correlations among the activities of those neurons. Cells of layer $\mathcal{B}$ in turn connect to layer $\mathcal{C}$ through synapses that may be positive or negative and which develop according to equation (3) from a random initial configuration.

1 Introduction

Linsker has studied by simulation the evolution of synaptic weight vectors in a feed–forward linear network [3, 4]. The network is shown in figure 1. Synaptic weight modification occured under a Hebbian rule that was linear up to saturating nonlinearities limiting the sizes of synaptic weights. Linsker found that in certain parameter regimes, "centre–surround" synaptic structures emerged at the third layer of the network (figure 2). Concatenation of several successive layers of such centre–surround cells and proper selection of parameters yielded a final regime of the Hebbian rule in which cells developed oriented receptive fields, consisting of alternating bars of positive and negative input synapses. These results are of interest for two reasons. First, the system studied is extremely simple: the network is linear and the connections develop under a simple teacherless learning rule. It seemed surprising that structured receptive fields arise in such a simple network. Secondly, because centre–surround and oriented receptive fields are found in the mammalian visual system, it seemed possible that these results might illustrate aspects of the dynamics underlying early neural development in the visual system.

The dynamical and biological bases for Linsker's results have not been clearly established. It has been pointed out that a Hebbian mechanism can perform certain types of principal component analysis [4, 14], but Linsker's results do not represent principal components of the input statistics, as we shall demonstrate. In this paper, we present an analysis of the dynamical mechanisms responsible for Linsker's results. We then comment briefly on the relationship of these results to biology.

1.1 Derivation of Linsker's Equations

We begin by reviewing the derivation of the equation that Linsker simulated. Consider a layer of cells, layer $\mathcal{L}$, connected to a single cell in the next layer, layer $\mathcal{M}$. Let $x_i^{\mathcal{L}}$ be the activity of cell i in layer $\mathcal{L}$, and let w_i be the strength of that input's connection to the cell in layer $\mathcal{M}$. Then the activity of the neuron in layer $\mathcal{M}$ is assumed to be a linear combination of the inputs:

$$x^{\mathcal{M}} = \sum_{i\in\mathcal{L}} w_i x_i^{\mathcal{L}} + a_1 \tag{1}$$

where a_1 is a constant.

A Hebbian rule for synaptic plasticity is one in which a synaptic strength is increased when pre– and post–synaptic firing are correlated, and possibly decreased when they are anticorrelated. The Hebb–type learning rule used by Linsker for modification of the connection strength w_i is:

$$\frac{d}{dt}w_i = (x^{\mathcal{M}} - c_2)(x_i^{\mathcal{L}} - c_1) + c_3 \tag{2}$$

c_1, c_2, and c_3 are constants. The dependence of this rule on the activity of the postsynaptic cell can be removed by substituting the value of $x^{\mathcal{M}}$ given by equation (1). This yields an equation for the evolution of the weights from layer $\mathcal{L}$ to the postsynaptic cell in terms of those weights and the activities of the cells in the presynaptic layer $\mathcal{L}$:

$$\frac{d}{dt}w_i = (\sum_{j\in\mathcal{L}} w_j x_j^{\mathcal{L}} + a_1 - c_2)(x_i^{\mathcal{L}} - c_1) + c_3$$

Assuming that $\mathbf{w}$ changes on a much longer time scale than the random variations in the inputs, we average over the ensemble of patterns $x_i^{\mathcal{L}}$ in layer $\mathcal{L}$. The mean rate of change of w_i is then given in terms of the mean activities $\bar{x}_i^{\mathcal{L}}$ and the covariance matrix $Q_{ij}^{\mathcal{L}} \equiv \langle (x_i^{\mathcal{L}} - \bar{x}_i^{\mathcal{L}})(x_j^{\mathcal{L}} - \bar{x}_j^{\mathcal{L}}) \rangle$ by:

$$\frac{d}{dt}w_i = \sum_{j\in\mathcal{L}} (Q_{ij}^{\mathcal{L}} + \bar{x}_j^{\mathcal{L}}(\bar{x}_i^{\mathcal{L}} - c_1))w_j + (a_1 - c_1)(\bar{x}_i^{\mathcal{L}} - c_1) + c_3$$

It is characteristic of Hebbian rules that synaptic strengths tend to increase without limit. To avoid this, Linsker added upper and lower bounds for all synaptic strengths:[1] $-w_{\max} \leq w_i \leq w_{\max}$.

The number of constants involved in this equation can be substantially reduced by assuming that the first-order statistics of the patterns in layer $\mathcal{L}$ are uniform, *i.e.* $\bar{x}_i^{\mathcal{L}} = \bar{x}^{\mathcal{L}}\ \forall i$. Then the equation for development of the synaptic weight vector becomes:

$$\frac{d}{dt}w_i = k_1 + \sum_{j\in\mathcal{L}} (Q_{ij}^{\mathcal{L}} + k_2)w_j \qquad \text{subject to } -w_{\max} \leq w_i \leq w_{\max} \tag{3}$$

where

$$\begin{aligned} k_1 &= (a_1 - c_2)(\bar{x}^{\mathcal{L}} - c_1) + c_3 \\ k_2 &= \bar{x}^{\mathcal{L}}(\bar{x}^{\mathcal{L}} - c_1) \end{aligned} \tag{4}$$

Linsker based his simulations on equation (3). Our definitions of the parameters differ slightly from Linsker's, in that no implicit dependence on the number of synapses is introduced.[2]

---

[1] Linsker allowed more general hard limits, $n_E - 1 \leq w_i \leq n_E$, $0 < n_E < 1$, which he implemented either directly or by allowing fractions n_E and $1 - n_E$ of the synapses to be excitatory and inhibitory respectively. Linsker reported no dependence of results upon n_E for the range that he studied, $0.35 \leq n_E \leq 0.65$. He concentrated on the case $n_E = 0.5$, and we have worked with this simplest case.

[2] Note that the transformation between equations (2) and (3) yields constants $\{k_1, k_2\}$ that depend on $\bar{x}^{\mathcal{L}}$ (equation (4)). So the generality of a set of constants does not carry over between the two equations. For example, a choice of $\{c_1, c_2, c_3\}$ that is used for several layers of a network does not necessarily map onto a set of constants $\{k_1, k_2\}$ that is the same for all layers.

Figure 2: **Linsker's results at layer C.**
Each circle represents the spatial pattern of inputs from layer B to a cell in layer C that resulted from a given choice of $k_1/|k_2|$, for k_2 large and negative. As the parameter k_1 was varied, synaptic structures ranging from saturated structures through centre–surround structures to bi–lobed oriented structures were reported.

1.2 Overview

The covariance matrix of activities of the inputs to the neuron, $\mathbf{Q}$, depends on two factors: the covariance function, which describes the dependence of the covariance of two input cells' activities on their separation in the input field; and the location of the synapses, which is determined by a synaptic density function.[3]

Depending on the covariance function, the synaptic density function, and the choice of the two parameters k_1 and k_2, different weight structures emerge. Linsker used a gaussian synaptic density function. Cells in his initial layer, layer $\mathcal{A}$, were taken to be uncorrelated in their activities (figure 1). Cells in that layer connect via excitatory synapses of fixed strength w_{max} to cells of layer $\mathcal{B}$. The overlap between the inputs to neurons in layer $\mathcal{B}$ causes the covariance in the activities of two layer $\mathcal{B}$ neurons to be a gaussian function of the separation between the neurons. Using these gaussian covariances, Linsker reported in his layer $\mathcal{B} \to \mathcal{C}$ connections the emergence of non–trivial weight structures: as the parameters were varied, these ranged from saturated structures through centre–surround structures to bi–lobed oriented structures (Figure 2). Given covariances that oscillate as a function of separation of the inputs (his layer $\mathcal{F} \to \mathcal{G}$), the development of cells with tri–lobed or multi–lobed oriented receptive fields was also reported.

The analysis in this paper examines the properties of equation (3). We concentrate on the class of covariance functions that are non–negative and monotonically decreasing, and in particular on the gaussian covariances in Linsker's layer $\mathcal{B} \to \mathcal{C}$ connections. We give an explanation of the occurrence of the structures shown in figure 2 and discuss criteria for the emergence of centre–surround weight structures. Several of the results are more general, applying to any covariance matrix $\mathbf{Q}$. Based on these general results, we comment briefly on the emergence of multilobed oriented cells at higher layers. We also briefly discuss the biological plausibility of the dynamical mechanisms found to underly Linsker's results. In appendix G the same methods are applied to a model one–dimensional network analogous to Linsker's two–dimensional network. Some of these results have been presented in briefer form elsewhere [6, 7].

2 Analysis in Terms of Eigenvectors

We write equation (3) as a first order differential equation for the weight vector $\mathbf{w}$:

[3] The synaptic density function and covariance function can be treated explicitly, as discussed in section 3.2 and appendix A.

Figure 3: **Dynamics of w relative to the fixed point.**
(a) Relative to the fixed point, the component of the weight vector in the direction of the eigenvector $\mathbf{e}^{(a)}$ with largest eigenvalue eventually grows fastest, if saturation limits are ignored. (b) Typical initial weight vectors resulting from the initial random distribution of synaptic strengths are indicated by the grey cloud at origin. If the fixed point is displaced from the origin a sufficient distance, relative to the size of a typical initial weight vector, an eigenvector with smaller eigenvalue can initially grow faster than the principal eigenvector. (c) The component in the direction of an eigenvector with negative eigenvalue decays. Ignoring saturation limits, the weight vector is eventually constrained to lie in the plane that is perpendicular to that eigenvector and that contains the fixed point.

$$\dot{\mathbf{w}} = (\mathbf{Q} + k_2\mathbf{J})\mathbf{w} + k_1\mathbf{n} \qquad\qquad \text{subject to } -w_{\max} \leq w_i \leq w_{\max} \tag{5}$$

where $\mathbf{J}$ is the matrix $J_{ij} = 1\ \forall i,j$, and $\mathbf{n}$ is the DC vector $n_i = 1\ \forall i$. This equation is linear, up to the hard limits on w_i. These hard limits define a hypercube in weight space within which the dynamics are confined. We make the following assumption:

Assumption 1 *The principal features of the dynamics are established before the hard limits are reached. When the hypercube is reached, it captures and preserves the existing weight structure with little subsequent change.*

The matrix $\mathbf{Q}+k_2\mathbf{J}$ is symmetric, so it has a complete orthonormal set of eigenvectors[4] $\mathbf{e}^{(a)}$ with real eigenvalues λ_a. Each eigenvector represents a weight configuration that evolves independently from the others. We now review how the linear dynamics within the hypercube can be fully characterised in terms of these eigenvectors and eigenvalues. Equation (5) has a fixed point at $\mathbf{w} = \mathbf{w}^{\mathrm{FP}}$ where $(\mathbf{Q} + k_2\mathbf{J})\mathbf{w}^{\mathrm{FP}} + k_1\mathbf{n} = 0$. The fixed point can be expressed explicitly in terms of the eigenvectors:

$$\mathbf{w}^{\mathrm{FP}} = -k_1(\mathbf{Q} + k_2\mathbf{J})^{-1}\mathbf{n} = -k_1 \sum_a \frac{\mathbf{e}^{(a)} \cdot \mathbf{n}}{\lambda_a}\mathbf{e}^{(a)} \tag{6}$$

Relative to the fixed point, the component of $\mathbf{w}$ in the direction of an eigenvector grows or decays exponentially at a rate proportional to the corresponding eigenvalue. The weight vector at time t can be written in the eigenvector basis as $\mathbf{w}(t) = \sum_a w_a(t)\mathbf{e}^{(a)}$, where $w_a(t) \equiv \mathbf{w}(t) \cdot \mathbf{e}^{(a)}$ is the

[4]The indices a and b will be used to denote the eigenvector basis for $\mathbf{w}$, while the indices i and j will be used for the synaptic basis.

component of $\mathbf{w}(t)$ in the direction of the eigenvector $\mathbf{e}^{(a)}$. Then, within the hypercube, equation (5) yields

$$w_a(t) - w_a^{\mathrm{FP}} = (w_a(0) - w_a^{\mathrm{FP}})e^{\lambda_a t}.$$

Suppose the typical initial weight vectors are distributed without bias around the origin. If the fixed point is near the origin then each eigenvector with $\lambda_a > 0$ reinforces itself and grows exponentially with rate λ_a. This means that after a short time the dynamics are typically dominated by the eigenvector with largest eigenvalue, which outgrows all the others (Figure 3a). But if one component of the fixed point w_b^{FP} is much larger than the typical size of the initial components, then the corresponding component w_b receives a substantial "head start" in its growth rate, and may initially outgrow eigenvectors with larger eigenvalue (Figure 3b). The component of $\mathbf{w}$ in the direction of an eigenvector $\mathbf{e}^{(c)}$ with negative eigenvalue decays towards the fixed point (Figure 3c), and the final weight vector $\mathbf{w}$ is constrained to lie in the hyperplane defined by $w_c = w_c^{\mathrm{FP}}$.

Thus (while the weight vector is not in contact with the hard–limits) the dynamics are fully characterised by the eigenvalues and eigenvectors of $\mathbf{Q} + k_2\mathbf{J}$. The principal emergent features of the dynamics are determined by the following three factors:

1. The principal eigenvectors of $\mathbf{Q}+k_2\mathbf{J}$, that is, the eigenvectors with largest eigenvalues. These are the fastest growing weight configurations.

2. Eigenvectors of $\mathbf{Q} + k_2\mathbf{J}$ with negative eigenvalue. These are associated with attracting constraint surfaces.

3. The location of the fixed point of equation (3). This is important for two reasons:

   - It determines the location of the constraint surfaces. The constraint surfaces always contain the fixed point, so increasing the distance to the fixed point increases the distance of the constraint surfaces from the origin.
   - The fixed point gives a "head start" to the growth rate of eigenvectors that have a component in the direction of the fixed point.

3 Eigenvectors of Q

We first examine the eigenvectors and eigenvalues of the covariance matrix $\mathbf{Q}$ for Linsker's layer $\mathcal{B} \to \mathcal{C}$ connections. The principal eigenvector of $\mathbf{Q}$ dominates the dynamics of equation (5) for $k_1 = 0$, $k_2 = 0$. The subsequent eigenvectors of $\mathbf{Q}$ become important as k_1 and k_2 are varied.

3.1 Eigenvectors of non–negative covariance matrices

In general, where there are only positive correlations in the inputs, $\mathbf{Q}$ is a non-negative matrix, and the Perron–Frobenius Theorem [15, page 1] holds:

Theorem 1 *For a matrix whose entries are all non–negative, the components of the principal eigenvector all have the same sign.*

This means that where there are no anti–correlations in a cell's input field, and $k_1 = 0, k_2 = 0$, all the cell's synapses tend to reinforce each others' growth, and the final configuration is expected to be all–excitatory or all–inhibitory. This property applies to Linsker's layer $\mathcal{B} \to \mathcal{C}$ connections.

3.2 Continuum approximation

To make analytic results possible we go to the continuum limit, that is, the limit of an infinite number of synapses. Here the vector of synaptic strengths $\mathbf{w}$ is replaced by a weight function $v(\mathbf{r})$, with an associated synaptic density function $A(\mathbf{r})$. $v(\mathbf{r})$ represents the average strength of a synapse from position $\mathbf{r}$, while $A(\mathbf{r})$ represents the number of synapses per unit area from the region about $\mathbf{r}$. Details of the transformation between the v and $\mathbf{w}$ representations are presented in appendix A.

The covariance between synapses from locations $\mathbf{r}$ and $\mathbf{r}'$ is described by the covariance function $C(\mathbf{r}, \mathbf{r}')$. In the continuum limit, the matrix $\mathbf{Q} + k_2\mathbf{J}$ is replaced by an integral operator. The continuum version of equation (3) is then:

$$\frac{d}{dt}v(\mathbf{r}) = k_1 + \int (C(\mathbf{r}, \mathbf{r}') + k_2)\, A(r')v(\mathbf{r}')\, d^2\mathbf{r}' \qquad \text{subject to } -w_{\max} \leq v(\mathbf{r}) \leq w_{\max} \qquad (7)$$

At Linsker's layer $\mathcal{B} \to \mathcal{C}$, the covariance function is a gaussian $C(\mathbf{r}, \mathbf{r}') = e^{-(\mathbf{r}-\mathbf{r}')^2/2C}$, and the synaptic density function is another gaussian $A(\mathbf{r}) = e^{-\mathbf{r}^2/2A}$, where C and A denote the characteristic sizes of the covariance function and synaptic density (arbor) function respectively. Linsker used various values for the ratio C/A. Our analytic results leave this ratio as a free parameter; in the figures we have used $C/A = 2/3$, the value frequently used for layer $\mathcal{B} \to \mathcal{C}$ in [3].

We now investigate the eigenfunctions of the integral operator above for Linsker's layer $\mathcal{B} \to \mathcal{C}$. We will continue to refer to this integral operator as the matrix $\mathbf{Q}+k_2\mathbf{J}$; we will use the terms eigenfunction and eigenvector interchangeably. Appendix A briefly discusses the numerical calculation of eigenvectors.

3.3 Properties of circularly symmetric systems

If $\mathbf{Q}$ has a translation invariance or other symmetry property, then the eigenfunctions will have related symmetry properties. If in a two–dimensional system the covariance function between two inputs depends only on their separation, and if the synaptic density function has circular symmetry, then in the continuum limit the operator $\mathbf{Q}+k_2\mathbf{J}$ is unchanged under rotation of the system. So the eigenfunctions of $\mathbf{Q} + k_2\mathbf{J}$ can be written as simultaneous eigenfunctions of the rotation operator, and are therefore separable into the product of a radial function and one of the angular functions $\cos l\theta$, $\sin l\theta$, $l = 0, 1, 2...$ To describe these eigenfunctions we borrow from quantum mechanics the notation $n = 1, 2, 3...$ and $l = \mathrm{s}, \mathrm{p}, \mathrm{d}...$ to denote the function's total number of nodes $= 0, 1, 2...$ and number of angular nodes $= 0, 1, 2...$ respectively. For example, "2s" has one radial node and no angular nodes, and "2p" has one angular node and no radial nodes. This notation is illustrated in figure 4.

The eigenfunctions of Linsker's monotonic non–negative covariance operator and similar covariance operators are found to be ordered in eigenvalue by their numbers of radial and angular nodes in the tree structure shown in figure 4. A line between two eigenfunctions means that the upper eigenfunction has larger eigenvalue, that is, addition of a node always results in lowering of the eigenvalue. Thus, $\lambda_{1\mathrm{s}} > \lambda_{2\mathrm{s}} > (\lambda_{3\mathrm{s}}, \lambda_{3\mathrm{p}})$ and $\lambda_{1\mathrm{s}} > \lambda_{2\mathrm{p}} > (\lambda_{3\mathrm{d}}, \lambda_{3\mathrm{p}})$, but for example no relationship is asserted between $\lambda_{2\mathrm{s}}$ and either $\lambda_{2\mathrm{p}}$ or $\lambda_{3\mathrm{d}}$. We do not know how general this ordering property is. In one–dimensional systems, counterexamples have been found to the conjecture that all monotonic non–negative covariance operators have their eigenvectors ordered in eigenvalue by their number of nodes.

Figure 4: **Separability of eigenfunctions in a circularly symmetric system.** The eigenfunctions of an operator invariant under rotation are separable into the product of a radial function and one of the angular functions $\cos l\theta$, $\sin l\theta$, $l = 0, 1, 2...$ This figure illustrates our notation for the eigenfunctions. The family of 's–modes' 1s, 2s, 3s... are circularly symmetric, the 'p–modes' 2p, 3p... have one angular node, the 'd–modes' have two angular nodes, etc. The eigenfunctions of Linsker's and similar systems are found to be ordered in eigenvalue by the number of radial and angular nodes as shown: a line between two eigenfunctions means that the upper eigenfunction has larger eigenvalue.

3.4 Analytic Calculations for $k_2 = 0$

We have solved analytically for the first six eigenfunctions and eigenvalues of the covariance matrix for layer $\mathcal{B} \rightarrow \mathcal{C}$ of Linsker's network, in the continuum limit (Table 1; Appendix D). The relative sizes of the eigenvalues and the shapes of the eigenfunctions depend on a single free parameter, the ratio C/A of the size of the gaussian covariance function to the size of the gaussian synaptic density function. For all C/A, 1s, the function with no changes of sign, is the principal eigenfunction of $\mathbf{Q}$, as predicted by the Perron–Frobenius Theorem; 2p, the bilobed oriented function, is the second eigenfunction; and 2s, the centre–surround eigenfunction, is third.[5] Figure 5a shows the principal eigenfunctions for $C/A = 2/3$. Tang [16] has independently derived these analytic results, and also developed approximations for the eigenfunctions for nonzero k_2.

3.5 Summary for $k_1 = 0$, $k_2 = 0$

For $k_1 = 0$, $k_2 = 0$, the dynamics are dominated by the principal eigenfunction, 1s, in which all synapses have the same sign. The centre–surround eigenfunction 2s is third in line behind 2p, the bi–lobed function.

4 The effects of the parameters k_1 and k_2

Varying k_2 changes the eigenvectors and eigenvalues of the matrix $\mathbf{Q} + k_2\mathbf{J}$. Varying k_1 moves the fixed point of the dynamics with respect to the origin. We now analyse these two changes, and their effects on the dynamics. We will use the following definition:

[5] 2s is degenerate with 3d.

Figure 5: **Computed Eigenvectors for Linsker's $\mathcal{B} \to \mathcal{C}$ connections**
a) Eigenvectors of $\mathbf{Q}$. b) Eigenvectors of $\mathbf{Q} + k_2\mathbf{J}$ for $k_2 = -3$. In each row the eigenvectors have the same eigenvalue, with the largest eigenvalue at the top. The greyscale bar is linear in synaptic strength, with synaptic strength zero at the center. Each eigenvector has been individually normalised to use the full greyscale. Eigenvalues (in arbitrary units): a) $k_2 = 0$: 1s, 2.22; 2p, 1.0; 2s & 3d, 0.45. b) $k_2 = -3$: 2p, 1.0; 2s, 0.66; 1s, -17.8. Eigenvectors of the operator $(e^{-(\mathbf{r}-\mathbf{r}')^2/2C} + k_2)e^{-\mathbf{r}'^2/2A}$ were computed for $C/A = 2/3$ on a circle of radius 12.5 grid intervals, with $\sqrt{A} = 6.15$ grid intervals.

Table 1: **Analytic solutions for first six eigenfunctions of the operator $\mathbf{Q}(\mathbf{r}, \mathbf{r}')$**
The integral operator $\mathbf{Q}(\mathbf{r}, \mathbf{r}') = e^{-(\mathbf{r}-\mathbf{r}')^2/2C} e^{-\mathbf{r}'^2/2A}$, where C and A denote the characteristic sizes of the covariance function and synaptic density (arbor) function respectively. The eigenvalues λ are normalised by $N = 2\pi A$, the effective number of synapses.

Name	Number of nodes			Expression	Eigenvalue
	Radial	Angular	Total		λ/N
1s	0	0	0	$e^{-r^2/2R}$	LC/A
2p$_x$	0	1	1	$r\cos\theta\, e^{-r^2/2R}$	L^2C/A
2p$_y$	0	1	1	$r\sin\theta\, e^{-r^2/2R}$	L^2C/A
2s	1	0	1	$(1 - r^2/r_0^2)e^{-r^2/2R}$	L^3C/A
3d$_1$	0	2	2	$r^2\cos 2\theta\, e^{-r^2/2R}$	L^3C/A
3d$_2$	0	2	2	$r^2\sin 2\theta\, e^{-r^2/2R}$	L^3C/A

Constant	Value
R	$\frac{1}{2}C(1 + \sqrt{1 + 4A/C})$
L	$\frac{R-C}{R}$ $(0 < L < 1)$
r_0^2	$\frac{2A}{\sqrt{1+4A/C}}$

Definition

Let $\hat{\mathbf{n}}$ be the unit vector in the direction of the DC vector $\mathbf{n}$. We refer to $(\mathbf{w} \cdot \hat{\mathbf{n}})$ as the *DC component* of $\mathbf{w}$. The DC component is proportional to the sum of the synaptic strengths in a weight vector. For example, 2p has zero DC component, as do all the other eigenfunctions with angular antisymmetry. On the other hand, all the s–modes typically have a non–zero DC component.

4.1 General Theorem: the effect of k_2

The analysis of the previous section primarily referred to Linsker's layer $\mathcal{B} \to \mathcal{C}$, in which the covariance function is gaussian. We now characterise the effect of adding $k_2\mathbf{J}$ to *any* covariance matrix $\mathbf{Q}$.

Theorem 2 *For any covariance matrix* $\mathbf{Q}$*, the spectrum of eigenvectors and eigenvalues of* $\mathbf{Q}+k_2\mathbf{J}$ *obeys the following:*

1. *Eigenvectors of* $\mathbf{Q}$ *with no DC component, and their eigenvalues, are unaffected by* k_2*.*
2. *The other eigenvectors, with non–zero DC component, vary with* k_2*. Their eigenvalues increase continuously and monotonically with* k_2 *between asymptotic limits such that the upper limit of one eigenvalue is the lower limit of the eigenvalue above.*
3. *There is at most one negative eigenvalue.*
4. *All but one of the eigenvalues remain finite. In the limits* $k_2 \to \pm\infty$ *the eigenvector with eigenvalue of largest magnitude is the DC vector* $\hat{\mathbf{n}}$*, and it has eigenvalue* $\to k_2 N$*, where* N *is the dimensionality of the matrix* $\mathbf{Q}$ *(i.e. the number of synapses).*

The properties stated in this theorem, which is proved in appendix B, are summarised pictorially by the spectral structure shown in figure 6. These simple properties arise because $\mathbf{J}$ has only rank 1.

4.2 Implications

For circularly symmetric systems such as Linsker's, all the eigenfunctions with angular nodes have zero DC component and are thus independent of k_2. The eigenfunctions that vary with k_2 are the s–modes. Specifically, at Linsker's layer $\mathcal{B} \to \mathcal{C}$, the leading s–modes at $k_2 = 0$ are 1s, 2s; as k_2 is decreased to $-\infty$, these modes transform continuously into 2s, 3s, respectively, as shown by the annotations in figure 6. So 2s becomes the s–mode with largest eigenvalue. (The 2s eigenfunctions at $k_2 = 0$ and $k_2 = -\infty$ both have one radial node, but they are not identical functions. Figure 6 shows that the 2s mode at $k_2 = -\infty$ must have larger eigenvalue than the 2s mode at $k_2 = 0$.)

As $k_2 \to -\infty$, theorem 2 states that one eigenvector approaches the DC vector $\hat{\mathbf{n}}$. Because the eigenvectors are orthogonal, the other eigenvectors must have DC components that tend to zero. This means that the principal eigenvectors for $k_2 \to -\infty$ approach synaptic structures composed of equal sums of positive and negative synapses. These principal eigenvectors are therefore dramatically different in character from the principal eigenvector for $k_2 = 0$.

As discussed in section 2, an eigenvector $\mathbf{e}^{(c)}$ with negative eigenvalue enforces a constraint on the final synaptic structure, $\mathbf{w} \cdot \mathbf{e}^{(c)} = w_c^{\mathrm{FP}}$. In the limit of large negative k_2 the constraint enforced by the eigenvector $\hat{\mathbf{n}}$ determines the final average synaptic strength: $\bar{w} = \mathbf{w}^{\mathrm{FP}} \cdot \mathbf{n}/N$. Linsker showed [3] that all or all but one of the synapses in a stable final configuration have

Figure 6: **General spectrum of eigenvalues of $\mathbf{Q} + k_2\mathbf{J}$ as a function of k_2.**
A: Eigenvectors with non–zero DC component (solid lines) vary with k_2, and have eigenvalues that increase monotonically with k_2. B: Eigenvectors with zero DC component (thick dotted lines) are independent of k_2. C: Adjacent DC eigenvalues share a common asymptote (thin dotted lines). D: There is only one negative eigenvalue. The leading eigenvector at $k_2 = \infty$ and the negative eigenvector at $k_2 = -\infty$ are both equal to the DC vector, $\hat{\mathbf{n}}$.
The annotations in parentheses identify the eigenvectors of Linsker's system at layer $\mathcal{B} \to \mathcal{C}$. From [6].

synaptic strength $\pm w_{\max}$, so this constraint on the average synaptic strength effectively fixes the final percentages of positive synapses and of negative synapses.

4.3 What constitutes large k_2?

We say that k_2 is large and negative when the eigenfunctions of $\mathbf{Q} + k_2\mathbf{J}$ are well described by the $k_2 \to -\infty$ limit. This is the case if the largest eigenvalue of $-k_2\mathbf{J}$, $-k_2 N$, is much larger than any eigenvalue of $\mathbf{Q}$ (Appendix C). Using the eigenvalue of 1s in Table 1, this yields the requirement $|k_2| \gg \frac{C}{A} + \frac{C^2}{2A^2}(1 - \sqrt{1 + 4A/C})$.

In [3], Linsker used values of $\frac{C}{A}$ for layer $\mathcal{B} \to \mathcal{C}$ ranging between $\frac{2}{5}$ and $\frac{2}{3}$, for which the above requirement ranges from $k_2 \ll -0.22$ to $k_2 \ll -0.44$. Linsker used $k_2 = -3$ throughout that reference, so all his simulations used large negative k_2.

4.4 Numerical computations for Linsker's system

The computed eigenvectors at layer $\mathcal{B} \to \mathcal{C}$ are shown in figure 5b for $k_2 = -3$. Properties of these eigenvectors can also be estimated from the analytic solutions for $k_2 = 0$ via perturbation theory as shown in appendix C. As predicted, there is now a single eigenvector with large negative eigenvalue, 1s. The principal eigenvector is 2p. The centre–surround eigenvector, 2s, is not the principal eigenvector of $\mathbf{Q} + k_2\mathbf{J}$ for this system either at $k_2 = 0$ or at large negative k_2.[6] For large negative k_2, 2s is the principal *symmetric* eigenvector, but it still has smaller eigenvalue than 2p. So the regime $k_2 \to -\infty$, $k_1 = 0$ will be dominated by oriented bi–lobed weight structures, and the circular symmetry is broken.

[6] Tang [16] showed that there is an intermediate regime of small negative k_2 in which the principal eigenfunction has centre–surround structure. But this is not the regime in which Linsker's centre–surround cells emerged.

4.5 Effect of k_1

Varying k_1 changes the location of the fixed point of equation (5). As stated in equation (6), the component of the fixed point in the direction of an eigenvector $\mathbf{e}^{(a)}$ is proportional to $k_1\mathbf{e}^{(a)} \cdot \mathbf{n}$. So the fixed point is displaced from the origin only in the direction of eigenvectors that have a non–zero DC component, that is, only in the direction of the s–modes. This has two important effects, as discussed in section 2:

1. The s–modes are given a head start in growth rate when k_1 is increased. In particular, the principal s–mode, the centre–surround vector, may grow faster than the principal eigenvector 2p.

2. The constraint surface must pass through the fixed point, so its location is moved when k_1 is changed. For large negative k_2, the sum of synaptic strengths in the final weight vector is fixed on the constraint surface. To leading order in $1/k_2$, Linsker showed that the constraint is:
$$\sum w_j = k_1/|k_2|$$
To next order, this expression becomes $\sum w_j = k_1/|k_2 + \bar{q}|$, where $\bar{q} = \langle Q_{ij} \rangle$, the average covariance (averaged over i and j; appendix C).[7]

4.6 Summary of the effects of k_1 and k_2 in Linsker's system

We can now anticipate the explanation for the emergence of centre–surround cells: For $k_1 = 0$, $k_2 = 0$, the dynamics are dominated by 1s. The centre–surround eigenfunction 2s is third in line behind 2p, the bi–lobed function. Making k_2 large and negative removes 1s from the lead. As $k_2 \to \infty$, 1s takes on large negative eigenvalue and tends to the DC vector $\mathbf{n}$, so it enforces a constraint on the final number of positive and negative synapses. 2p becomes the principal eigenfunction and dominates the dynamics for $k_1 = 0$. Finally, increasing $k_1/|k_2|$ gives a head start to the principal s–mode, the centre–surround function 2s.

Increasing $k_1/|k_2|$ also increases the final average synaptic strength, so large $k_1/|k_2|$ not only gives 2s a large head start, it also produces a large DC bias. Centre–surround structures emerge when $k_1/|k_2|$ is large enough that 2s dominates over 2p, and small enough that the DC bias does not obscure the centre–surround structure. Therefore an on–centre centre–surround regime lies sandwiched between a 2p–dominated regime and an all–excitatory regime, and an off–centre centre–surround regime lies between the 2p–dominated regime and an all–inhibitory regime. In section 6 we will estimate the boundaries of these parameter regime in which centre-surround structures emerge.

5 Analysis in Terms of Constrained Dynamics

A complementary conceptual framework for understanding the emergence of centre–surround structures can be obtained by considering the dynamics on the constraint surface.

For large negative k_2, the constraint surface is the hyperplane $\sum w_i = k_1/|k_2|$. On the constraint surface, we can divide the weight vector into $\mathbf{w}(t) = \mathbf{w}_{\mathrm{AC}}(t) + \mathbf{w}_{\mathrm{DC}}$, where $\mathbf{w}_{\mathrm{DC}} = \frac{k_1}{k_2 N}\mathbf{n}$ is the

[7] The additional term largely resolves the discrepancy between Linsker's g and k_1/k_2 in [3]. In the continuum limit, $\bar{q} = \frac{1}{1+2A/C}$, using the notation of table 1. In the example on p. 7511 of [3], $A/C = 1.5$, $k_1 = 0.45$, $k_2 = -3$. Hence $k_1/|k_2| = 0.15$ and $k_1/|k_2 + \bar{q}| = 0.164$, while the observed value of g was 0.166 ± 0.002. In the example on p. 8391, $A/C = 2.5$, $k_1 = 0.35$, $k_2 = -3$, $k_1/|k_2| = 0.117$, $k_1/|k_2 + \bar{q}| = 0.124$, and the observed value of g was 0.126 ± 0.001.

time–independent DC part of $\mathbf{w}$ and $\mathbf{w}_{\mathrm{AC}}(t)$ is a time–varying AC vector, that is, a vector with zero DC component. In appendix E we show that the dynamics on the constraint surface in the limit of large negative k_2 are described by

$$\dot{\mathbf{w}}_{\mathrm{AC}} = \mathbf{PQPw}_{\mathrm{AC}} + \frac{k_1}{|k_2|N}\mathbf{PQn} \tag{8}$$

where $\mathbf{P}$ is the orthogonal projection operator onto the surface $\sum w_j = 0$. $\mathbf{P}$ subtracts out the DC part of a vector, leaving the AC part.[8] Apart from the DC eigenvector, the eigenvectors and eigenvalues of the matrix $\mathbf{PQP}$ are identical to those of the matrix $\mathbf{Q} + k_2\mathbf{J}$ in the limit $k_2 \to -\infty$.

The two terms in equation (8) represent the growth of two sorts of AC pattern. The first term represents the contribution of the time–varying, AC part of $\mathbf{w}$ to $\dot{\mathbf{w}}$. In terms of the eigenvectors of $\mathbf{Q} + k_2\mathbf{J}$, the component in the direction of 2p grows faster than that of 2s under this term due to its larger eigenvalue. The second term, $\mathbf{PQn}$, represents the contribution of the time–invariant, DC part of $\mathbf{w}$ to $\dot{\mathbf{w}}$. This term corresponds to the head start in growth rate. Because $\mathbf{n}$ is circularly symmetric, $\mathbf{PQn}$ is circularly symmetric. Therefore, it is a superposition of s–modes, dominated by the centre–surround mode 2s. This centre–surround structure is continually "spewed out" at a rate proportional to the DC bias $\frac{k_1}{|k_2|N}$. The first term amplifies this centre–surround structure, and also other AC structures present in the initial fluctuations. In order for a symmetric centre–surround structure to emerge, the static second term in equation (8) must compensate for the advantage in eigenvalue of asymmetric over symmetric structures in the first term.

This alternative representation of the dynamics is useful for speeding up numerical simulations of equation (3) for large negative k_2. The matrix $\mathbf{Q} + k_2\mathbf{J}$ has one very large negative eigenvalue $\simeq k_2 N$, so very small steps have to made in iterations of equation (3) to avoid unstable oscillations. In equation (8), on the other hand, the matrix $\mathbf{PQP}$ has no extreme or negative eigenvalues, since the DC constraint is enforced separately.

6 Criteria for The Emergence of Centre–Surround Cells

We now wish to estimate the boundaries of the parameter regime in which the head start is sufficient for centre–surround cells to emerge. We use two approaches to determine the DC bias at which 2s and 2p are equally favoured. This gives an estimate for the boundary between the regimes dominated by 2s and 2p. If 2s dominates for DC bias sufficiently small that the surround has a significant size, then centre-surround structures will emerge in the corresponding parameter regime.

1. **Energy Criterion:** We first estimate the level of DC bias at which the weight vector composed of (2s plus DC bias) and the weight vector composed of (2p plus DC bias) are energetically equally favoured. This gives an estimate of the level of DC bias above which 2s will dominate under simulated annealing, which explores the entire space of possible weight configurations.

2. **Time Development Criterion:** Second, we estimate the level of DC bias above which 2s will dominate over 2p under simulations of time development of equation (3). We estimate the relationship between the parameters such that, starting from a typical random distribution of initial weights, the 2s mode reaches the saturating hypercube at the same time as the 2p mode.

[8] It should be noted that this projection operator maintains the constraint *subtractively.* Very different dynamics would result if the constraint were enforced multiplicatively. This subject will be developed in a future publication.

We shall consider two eigenvectors, $\mathbf{e}^{(1)}$ and $\mathbf{e}^{(2)}$, with eigenvalues $\lambda_1 > \lambda_2$. $\mathbf{e}^{(1)}$ corresponds to 2p, and $\mathbf{e}^{(2)}$ corresponds to 2s. Let the fixed point have component w_2^{FP} in the direction of $\mathbf{e}^{(2)}$. $\mathbf{e}^{(1)}$ is an AC vector, so $w_1^{\mathrm{FP}} = 0$. Both criteria for $\mathbf{e}^{(2)}$ to dominate over $\mathbf{e}^{(1)}$ will depend on an estimate of the effect of the weight limits $-w_{\max} \leq w_i \leq w_{\max}$. (Without this hypercube of saturation constraints, $\mathbf{e}^{(1)}$ will always dominate the dynamics of equation (3) after a sufficiently long time.) For the growth of a typical eigenvector, the hypercube represents a complex constraint, as it is possible for further growth in the direction of the vector to take place after the largest components of $\mathbf{w}$ reach saturation. This makes it difficult to assess the exact effect of the hypercube on the dynamics. To estimate this effect, we make the following assumptions additional to assumption 1:

Assumption 2 *In the abscence of constraints, we estimate that the hypercube has been 'reached' by a typical vector* $\mathbf{w}$ *when* $w_{\mathrm{rms}} = w_{\max}$.

This means that the hypercube is 'reached' when the component of $\mathbf{w}$ in the direction of a typical normalised vector is $h = \sqrt{N} w_{\max}$.

We introduce $g = k_1/(|k_2| N w_{\max})$ as a measure of the average synaptic strength induced by the DC constraint $\sum w_i = k_1/|k_2|$, as a fraction of $w_{\max}$. Thus $g = 1$ corresponds to all synapses saturated at $+w_{\max}$, and $g = 0$ corresponds to equal numbers of positive and negative synapses.[9] Now as the DC level g is increased, the amount of growth possible before the hypercube is reached decreases. We estimate the amount of growth possible in the direction of a typical AC vector as a function of the DC bias g. When $g = 1$, all synapses are saturated, and no growth in any direction is possible. We interpolate linearly between $h = \sqrt{N} w_{\max}$ at $g = 0$ (Assumption 2) and $h = 0$ at $g = 1$, assuming:

Assumption 3 *When the DC level is constrained to be* g*, the component* $h(g)$ *in the direction of a typical unit AC vector at which the hypercube constraint is 'reached' is* $h(g) = \sqrt{N} w_{\max}(1 - g)$.

Assumptions 1–3 may not adequately characterize the effects of the hypercube on the dynamics, so the numerical estimates of the precise locations of the boundaries between the regions may be in error. However, the qualitative picture of the division of parameter regimes that they present is informative.

6.1 Energy Criterion

Linsker [3, 4] suggested analysis of equation (3) in terms of the energy function on which the dynamics perform constrained gradient descent:

$$E = -\frac{1}{2}\mathbf{w}^{\mathrm{T}}(\mathbf{Q} + k_2\mathbf{J})\mathbf{w} - k_1 \mathbf{w} \cdot \mathbf{n}$$

Neglecting the initial conditions, we can examine the result of minimising this energy subject to the hard limit constraints on w_i. Expressing the weight vector in terms of the eigenvectors of $\mathbf{Q} + k_2\mathbf{J}$, the energy of a configuration $\mathbf{w} = \sum w_a \mathbf{e}^{(a)}$ is:

$$E = -\frac{1}{2}\sum_a \lambda_a w_a^2 - \sqrt{N} k_1 \sum_a w_a n_a$$

where n_a is the DC component of eigenvector $\mathbf{e}^{(a)}$. Without loss of generality we assume that $n_2 > 0$.

[9] This is equal to twice Linsker's g, since he did not include normalization by $w_{\max}$, and he used $w_{\max} = 0.5$.

Figure 7: **Schematic diagram illustrating the criteria for $\mathbf{e}^{(2)}$ to dominate.** The polygon of size $h(g)$ represents the hypercube. **Energy criterion:** The points marked E_1 and E_2 show the locations at which the energy estimates were made. **Time development criterion:** The grey cloud surrounding the origin represents the distribution of initial weight vectors. If $w_1(0)$ is sufficiently small compared with w_2^{FP}, and if the hypercube is sufficiently close, then the weight vector hits the hypercube in the direction of $\mathbf{e}^{(2)}$ before w_1 has grown appreciably.

We consider two configurations, one with w_1 equal to its maximum value $h(g)$ and $w_2 = 0$, and one with $w_2 = h(g)$ and $w_1 = 0$. The DC component $\mathbf{w} \cdot \hat{\mathbf{n}}$ is the same in both cases. All the other components are assumed to be small and to contribute no bias in energy between the two configurations. The energies E_1 and E_2 of these configurations will be our estimates of the energies of saturated configurations obtained by saturating $\mathbf{e}^{(1)}$ and $\mathbf{e}^{(2)}$ respectively, subject to the constraint on $\mathbf{w} \cdot \hat{\mathbf{n}}$. We will compare these two energies and find the DC level $g = g^{\mathrm{E}}$ at which E_1 and E_2 are equal. This g will be a coarse estimate of the boundary between the regimes in which simulated annealing gives centre–surround and symmetry–breaking structures.

Neglecting additive constants,

$$E_1 = -\frac{1}{2}\lambda_1 h(g)^2$$

and

$$E_2 = -\frac{1}{2}\lambda_2 h(g)^2 - \sqrt{N} k_1 h(g) n_2$$

It is the last term $-\sqrt{N} k_1 h(g) n_2$ that gives an energy advantage to $\mathbf{e}^{(2)}$ as g is increased.

To find an estimate of the critical value of g for simulated annealing we set $E_1 = E_2$. Substituting $h(g) = \sqrt{N}(1-g)w_{\max}$ and $k_1 = g|k_2|N w_{\max}$ and rearranging, we obtain as our estimate for the boundary between the $\mathbf{e}^{(1)}$–dominated regime and the $\mathbf{e}^{(2)}$–dominated regime:[10]

$$g^{\mathrm{E}} = \frac{1}{1 + 2\frac{n_2|k_2|}{(\lambda_1 - \lambda_2)/N}}$$

We can take the analytic results for $k_2 = 0$ (section 3.4) and use perturbation theory as outlined in appendix C to estimate the parameters $n_2 k_2$, λ_2/N and λ_1/N for large k_2. Using these estimates, we plot a graph of g^{E} against the one remaining degree of freedom, the ratio C/A of the spatial sizes of the covariance function and the synaptic density function (Figure 8). This is the value of

[10] λ/N is written as a single entity because $\lambda \propto N$. Also $n_2 k_2$ tends to a constant as $k_2 \to \infty$ (Appendix C).

Figure 8: **Energy criterion: g^E versus C/A.**
Above the line $g^E(C/A)$, centre–surround structures are estimated to be energetically favoured; below the line, bilobed structures are favoured. The perturbation theory used to estimate g^E becomes invalid for $C/A <\sim 0.2$. Linsker used $C/A = 2/3$ or $2/5$ for layer $\mathcal{B} \to \mathcal{C}$.

g at which we estimate that 2s and 2p are energetically equally favoured. It must be emphasised that this estimate depends on a substantial number of assumptions and approximations, so it is not at all exact. For Linsker's choice of $C/A = 2/3$, g^{E} is 0.16.

6.2 Time Development Criterion

The energy criterion does not take into account the initial conditions from which equation (3) starts. We now derive a second criterion that attempts to do this.

If the initial random component in the direction of 2p, $w_{\mathrm{2p}}(0)$, is sufficiently small compared to the head start that 2s has, $w_{\mathrm{2s}}^{\mathrm{FP}}$, then w_{2p} may never start growing appreciably before the growth of w_{2s} saturates (Figure 7). The initial component $w_{\mathrm{2p}}(0)$ is a random quantity whose typical magnitude can be estimated statistically from the weight initialisation parameters. The head start for w_{2s}, $w_{\mathrm{2s}}^{\mathrm{FP}}$, can be evaluated in terms of the parameters λ_{2s} and the DC component of 2s. These two quantities, $w_{\mathrm{2p}}(0)$ and $w_{\mathrm{2s}}^{\mathrm{FP}}$, scale differently with the number of synapses N. $w_{\mathrm{2p}}(0)$ is a zero–mean quantity related to N random variables; $w_{\mathrm{2s}}^{\mathrm{FP}}$ is related to N non–random variables. As suggested by the law of large numbers, the typical magnitude of $w_{\mathrm{2p}}(0)$ scales as $1/\sqrt{N}$ relative to $w_{\mathrm{2s}}^{\mathrm{FP}}$. Hence the initial relative magnitude of w_{2p} can be made arbitrarily small by increasing N, and the emergence of centre–surround structures may be achieved at any g by using an N sufficiently large to suppress the initial symmetry breaking fluctuations. We estimate the boundary between the regimes dominated by 2s and 2p by finding the relationship between N and g such that $w_1(t)$ and $w_2(t)$ hit the hypercube at the same time.

Before saturation occurs, the components $w_a(t)$ in the eigenvector basis can be written in terms of the initial components $w_a(0)$:

$$w_1(t) = w_1(0)e^{\lambda_1 t}$$

$$w_2(t) = w_2(0)e^{\lambda_2 t} - (e^{\lambda_2 t} - 1)w_2^{\mathrm{FP}}$$

We estimate statistically the typical starting component $w_1(0)$ once $\mathbf{w}$ has been projected into the constraint surface of fixed average w_i; when the initial weights are randomly generated between

Figure 9: **Boundaries estimated by the two criteria for $C/A = 2/3$**
To the left of the line labelled g^{E}, the energy criterion predicts that 2p is favoured; to the right, 2s is favoured. Above and below the line $N^*(g)$, the time development criterion estimates that 2s and 2p respectively will dominate equation (3). The regions A, B, C, D denote the four regions of parameter space separated by these two criteria. The regions X, Y, mark the regimes studied by Linsker: X) $N = 300 - 600$, $g = 0.3 - 0.6$: the region in which Linsker reported robust centre–surround structures; Y) $N = 300-600$, $g <\sim 0.2$: asymmetric centre–surround structures and (near $g = 0$) bi–lobed structures. The estimates of g^{E} and $N^*(g)$ are both obtained from the analytic results using perturbation theory (Appendices C, D). From [6].

$-w_{\mathrm{max}}$ and w_{max}, we obtain

$$w_1(0)_{\mathrm{rms}} = \sigma(g) w_{\mathrm{max}} \sqrt{d} \tag{9}$$

where $\sigma(g)$ is a dimensionless standard deviation derived in appendix F and d is the degeneracy of $\mathbf{e}^{(1)}$. We evaluate the size of the component w_1 at the time t_2 at which w_2 reaches the hypercube. Without loss of generality we assume $w_2^{\mathrm{FP}} < 0$ so that w_2 increases with time.[11] Then by assumption 3, $w_2(t_2) = h(g) = \sqrt{N} w_{\mathrm{max}}(1 - g)$, so

$$t_2 = \frac{1}{\lambda_2} \ln \left(1 + \frac{\sqrt{N} w_{\mathrm{max}}(1 - g)}{|w_2^{\mathrm{FP}}|} \right)$$

Now letting $w_1(t_2) = h(g)$ we obtain an implicit equation for N^*, the number of synapses above which w_2 dominates, in terms of g:

$$w_1(0) e^{\lambda_1 t_2} = \sqrt{N^*} w_{\mathrm{max}} (1 - g)$$

Substituting $|w_2^{\mathrm{FP}}| = n_2 \sqrt{N} k_1 / \lambda_2 = n_2 \sqrt{N} g |k_2| w_{\mathrm{max}} / (\lambda_2 / N)$, and $w_1(0) = \sigma(g) w_{\mathrm{max}} \sqrt{d}$, we obtain the following expression for N^*:

$$\sqrt{N^*} = \tfrac{\sigma(g)\sqrt{d}}{(1-g)} \left(1 + \tfrac{1-g}{g} \tfrac{\lambda_2/N}{|n_2 k_2|} \right)^{\lambda_1/\lambda_2} \tag{10}$$

To make a detailed estimate of N^*, we would need to consider other sources of symmetry–breaking fluctuations such as asymmetries in the synaptic locations.

[11] We set $w_2(0) = 0$, neglecting its fluctuations, which for large $g\sqrt{N}$ (*i.e.* $g\sqrt{N} \gg \frac{\sigma(g)\lambda_2/N}{|n_2 k_2|}$) are negligible compared with w_2^{FP}.

6.3 Discussion of the two criteria

For Linsker's system, figure 9 shows N^* as a function of g, the boundary estimated by the time development criterion, and g^{E}, the boundary estimated by the energy criterion. Figure 9 also shows the regions in the parameter space at which Linsker made the simulations he reported. The two criteria give different boundaries. In regime A, 2p is estimated to both be energetically favoured, and to emerge under equation (3). Similarly, in regime C, 2s is estimated to be energetically favoured, and to dominate equation (3). In regime D the initial fluctuations are so big that although 2s is energetically favoured, symmetry breaking structures can dominate equation (3). (Note that large fluctuations always help the growth of 2p but hinder the growth of 2s half of the time, because half of the time the initial component of 2s will be *towards* the fixed point, so that the 2s component must first shrink to zero before it can grow in the favoured direction.) Lastly, in regime B, although 2p is energetically favoured, 2s will reach saturation first because N is sufficiently large that the symmetry breaking fluctuations are suppressed. Whether this saturated 2s structure will be stable, or whether it might gradually destabilise into a 2p–like structure, is not predicted by our analysis.[12]

The possible difference between simulated annealing and the evolution of equation (3) makes it clear that if initial conditions are important (regimes B and D), the use of simulated annealing on the energy function as a quick way of finding the outcome of equation (3) may give erroneous results.[13]

7 Parameter regimes for general Q and for Linsker's system

For a general $\mathbf{Q}$ in equation (3), we predict up to four main parameter regimes for varying k_1 and k_2.[14] These regimes, shown in figure 10a, are dominated by the following weight structures:

Regime 1	$k_2 = 0$, $k_1 = 0$:	The principal eigenvector of $\mathbf{Q}$.
Regime 2	$k_2 =$ large positive and/or $k_1 =$ large	The flat DC weight vector.
Regime 3	$k_2 =$ large negative, $k_1 \simeq 0$	The principal eigenvector of $\mathbf{Q} + k_2\mathbf{J}$ for $k_2 \to -\infty$.
Regime 4	$k_2 =$ large negative, $k_1 =$ intermediate	The principal eigenvector with non–zero DC component of $\mathbf{Q} + k_2\mathbf{J}$ for $k_2 \to -\infty$. This vector is given a head start in growth rate. This regime may not exist if the head start is too small.

For non–negative $\mathbf{Q}$, *i.e.* no anti–correlations in the inputs, the principal component of $\mathbf{Q}$ has no zero–crossings and hence regime 1 produces saturated structures similar or identical to the DC weight structure. This leaves regimes 3 and 4 as the only parameter regimes in which alternative weight structures might arise.

For Linsker's $\mathcal{B} \to \mathcal{C}$ connections (Figure 10b), the principal eigenvector of $\mathbf{Q}+k_2\mathbf{J}$ as $k_2 \to -\infty$ is the symmetry–breaking oriented function, 2p. This eigenvector dominates regime 3. 2s, the centre–surround structure, dominates in regime 4. In principle there could be systems with non–

[12]In the one–dimensional model system of appendix G we have found that such energetically unfavourable saturated weight vectors may be stable or unstable, depending sensitively on the parameters.

[13]As noted by Linsker, simulated annealing may of course be a more appropriate strategy if equation (3) is viewed as being subjected to random noise.

[14]not counting the symmetric regimes $(k_1, k_2) \leftrightarrow (-k_1, k_2)$ in which all the weight structures are inverted in sign.

Figure 10: **Four principal parameter regimes**
a) The four regimes for a general covariance matrix $\mathbf{Q}$. b) The regimes for Linsker's layer $\mathcal{B} \to \mathcal{C}$ connections. See text for explanation of the regimes. When k_2 is large and negative, the DC constraint is approximately constant along the radial lines of constant $k_1/(k_2 + \bar{q})$, so each of the parameter regimes with large negative k_2 is wedge shaped. From [6].

negative $\mathbf{Q}$ in which 2s is the principal eigenvector at $k_2 = -\infty$: then in both regimes 3 and 4, 2s would dominate. This never occurs for any choice of parameters in Linsker's network, at least not in the range of parameters where our perturbation theory is valid (section D.3).

8 Further applications of the analysis

8.1 Higher layers in Linsker's network

The analysis in terms of the eigenvectors of $\mathbf{Q} + k_2\mathbf{J}$ should apply to the higher layers. Four principal regimes are expected as described in section 7, with the proviso that it is possible that there may be some near degeneracy of eigenvectors, so that some regimes may be dominated by a combination of several eigenvectors. We have not studied these higher layers in detail; we offer the following brief comments.

The covariance function for the higher layers is no longer non–negative, but instead oscillates with distance. The leading eigenvectors are expected to have the same characteristic spatial frequency as the covariance function. If the spatial period is comparable to the diameter of the synaptic density function, as in Linsker's layers $\mathcal{C} \to \mathcal{D}$ through $\mathcal{E} \to \mathcal{F}$, the leading eigenvectors of $\mathbf{Q} + k_2\mathbf{J}$ for $k_2 \in (0, -\infty)$ are 2p and 2s. Either 2p or 2s dominates in regimes 1 and 3, depending on the parameters; the 2s eigenvector dominates the dynamics in regime 4 by the same mechanism as in layer $\mathcal{B} \to \mathcal{C}$. Regime 4 was the one studied by Linsker in these layers.

At layer $\mathcal{F} \to \mathcal{G}$, Linsker used larger input arbors, so that several oscillations of the covariance function were contained in one cell's arbor. This means that the leading eigenvectors are expected also to have several oscillations. Linsker then reported a "bubble" in the $(g, A_\mathcal{G}/A_\mathcal{F})$ parameter space in which tri– or multi–lobed oriented cells arose.[15] The formation of these cells may be partially understood in terms of the mixing of the principal eigenvectors of $\mathbf{Q}+k_2\mathbf{J}$. In computations using several covariance functions similar to that used by Linsker, we have found that the three

[15]Linsker termed cells with a central excitatory stripe and two inhibitory flanking stripes 'bi–lobed cells': we call them 'tri–lobed' and reserve 'bi–lobed' for 2p–like functions.

principal eigenvectors are, in order, an s, p, and d vector. The leading s–mode has negligible DC component at $k_2 = 0$ for the range of arbor widths used at layer $\mathcal{F} \to \mathcal{G}$, so that the principal eigenvectors and eigenvalues are virtually unaltered as k_2 varies from 0 to $-\infty$. The number of radial nodes in these vectors increases as the arbors broaden, that is, as the parameter $A_{\mathcal{G}}/A_{\mathcal{F}}$ increases. The principal eigenvectors range from 2s, 3p, 3d for the narrowest arbors used by Linsker at layer $\mathcal{F} \to \mathcal{G}$, to 4s, 4p, 5d for the broadest arbors. As the arbor diameter increases, the eigenvalues of the three vectors become more nearly degenerate.

The mixed vector [Ns + (N+1)d] gives tri– or multi–lobed oriented arrangements of synapses. However, this does not explain why oriented, grating–like structures *should* occur. Other combinations of the principal eigenvectors that lack grating–like regularity would be equally likely to emerge from considerations of linear dynamics. Such other combinations appear to account for the various structures Linsker found as he varied g for each value of $A_{\mathcal{G}}/A_{\mathcal{F}}$. Variation of g should not significantly alter the linear dynamics, because the leading modes have negligible DC component. Since the multi-lobed outcome occurs only for a narrow range of g, this suggests that the multi-lobed outcome is determined in the nonlinear regime. Indeed, within the parameter–space "bubble" in which Linsker obtained oriented cells, the only published example of time development [5] shows that a 3s structure develops initially. Only in the nonlinear regime, when many synapses were saturated, does this 3s structure convert to a tri-lobed, oriented structure. Thus, what appears to be robust in the linear dynamics is that a receptive field should emerge with a characteristic spatial frequency of oscillation between positive and negative inputs, as seen in the leading eigenvectors. This frequency is determined by the frequency of oscillation of the covariance function within an arbor, i.e. by $A_{\mathcal{G}}/A_{\mathcal{F}}$. The precise form of the receptive field depends on the choice of the parameter g, and appears to depend upon the nonlinearities that limit synaptic growth.

8.2 A one–dimensional model system

In appendices G and H, analytic solutions are presented for a model one–dimensional network analogous to Linsker's two–dimensional network. This model system shares all the properties of Linsker's layers $\mathcal{A}$ through $\mathcal{C}$, except that a centre–surround structure is never the global minimum of the energy function. Rather, asymmetric structures (the analogue of 2p–dominated structures) are always the energy–minimising configurations, that is $g^{\mathrm{E}} = 1$. This exact result enables us to examine the performance of the estimate of g^{E} given by the energy criterion (section 6.1). In this extreme case of high g^{E}, the estimate is low by nearly a factor of two.

In simulations of time development under equation (3), centre–surround structures initially dominate the dynamics for large N, when symmetry–breaking fluctuations are suppressed. However the energetically unfavourable saturated centre–surround structures that result may be unstable under equation (3), transforming synapse by synapse into an asymmetric structure, or they may be stable. This depends very sensitively on the parameter values.

9 Discussion

We have analysed Linsker's equation (3) by examining the eigenvectors of the matrix that drives the dynamics. Our analysis depends on assumption 1 that the principal features of the dynamics are determined before the saturating limits on synaptic strength become important. Justification for this assumption at layer $\mathcal{B} \to \mathcal{C}$ is provided by Linsker's account of the simulated development of centre–surround cells, in which centre–surround structures emerge before any saturation occurs [3, page 7512]. In certain parameter regimes (regions B and D of figure 9), the initial conditions may give preference to energetically unfavored weight configurations. In these regimes, assumption

1 may be violated, because the nonlinear dynamics do not always preserve saturated structures that are energetically unfavourable. This is demonstrated in the the model system of appendix G. The development of Linsker's oriented cells at layer $\mathcal{F} \to \mathcal{G}$ also may violate assumption 1. This development appears to depend on the details of the non–linear dynamics when synapses are saturated, as discussed in section 8.1. Nonetheless, our analysis of the dynamics in the linear regime gives insight into the results at layer $\mathcal{F} \to \mathcal{G}$, by identifying the principal weight structures that appear to contribute to the final outcome.

For the layer $\mathcal{B} \to \mathcal{C}$ connections, where covariances are purely positive, all synapses reinforce one another if $k_1 = k_2 = 0$, leading to a synaptic structure in which all synapses have the same sign. However, for large negative k_2, the parameters k_1 and k_2 enforce a constraint fixing the average synaptic strength, so that growth of some synapses requires others to become negative. Then three different weight structures dominate the dynamics. The fastest–growing weight structure is a bi–lobed oriented eigenvector that breaks the circular symmetry of the covariance and synaptic density functions. This structure dominates the dynamics for small $k_1/|k_2|$. A centre–surround structure is the fastest-growing circularly symmetric eigenvector. This structure dominates the dynamics for intermediate values of $k_1/|k_2|$ due to its 'head start' in growth rate. The flat DC eigenvector has large negative eigenvalue, enforcing the constraint on the average synaptic strength.

We conjecture that the qualitative properties of these leading eigenvectors at layer $\mathcal{B} \to \mathcal{C}$ are robust, and would be unchanged if the gaussian covariance and synaptic density function were replaced by any other monotonic non–negative covariance and synaptic density functions. Linsker suggested that the emergence of centre–surround structures may depend on the peaked synaptic density function that he used [3, page 7512]. However, with a flat 'pill–box' density function, we have calculated that the eigenfunctions are qualitatively unchanged. Therefore we conjecture that centre–surround structures should emerge by the same 'head start' mechanism with a pill–box density function, though the locations of the parameter regime boundaries will be different.

Weight structures that include both positive and negative synapses can be obtained even in the absence of the constraints enforced by k_1 and k_2 if covariances oscillate in sign significantly within the arbor, as in Linsker's higher layers. The positive and negative regions of the receptive field then oscillate with a spatial frequency roughly determined by the frequency of oscillation of the covariance function.

We have shown that simulated annealing, used by Linsker to speed up simulations of equation (3) in higher layers, may give different parameter regime boundaries from simulations of the time development of equation (3). If the initial symmetry breaking fluctuations are sufficiently small, the centre–surround function may dominate the dynamics for arbitrarily small non–zero k_1. The details of this difference will depend on whether or not energetically unfavourable saturated weight structures are stabilised by the saturation constraints.

It should be noted that because of the terms k_1 and k_2, the outcome of equation (3) under either time development or simulated annealing is different from the outcomes of both principal components analysis (PCA) and information maximisation. When applied to a single output cell, both PCA and information maximisation maximise the quadratic form $\mathbf{w}^{\mathrm{T}}\mathbf{Q}\mathbf{w}$ subject to the constraint $\sum w_i^2 = \text{const}$. Equation (3) also maximises $\mathbf{w}^{\mathrm{T}}\mathbf{Q}\mathbf{w}$, but does so subject to the very different constraints $\sum w_i = \text{const}$ and $|w_i| \leq w_{\max}$.

Biological discussion

The development of centre–surround synaptic structures in Linsker's system depends on two features that are biologically problematic: the use of synaptic strengths that may take either positive or negative values; and the constraints, enforced by the terms k_1 and k_2, that fix the final percentage of positive and of negative synapses onto each postsynaptic cell. In the absence of either of these features, only all–excitatory or all–inhibitory synaptic structures would develop. We discuss these features briefly here. A thorough discussion of correlation–driven learning rules as models of biological systems can be found in [10].

In biology, all the synapses from a single neuron are either exclusively positive or exclusively negative. Synaptic strengths that may take either positive or negative values can be used to describe the summed strength of two separate populations, one of exclusively positive synapses and one of exclusively negative synapses [3]. However, for such a sum to be described by a simple equation like equation (3), the following conditions must hold [11]: the rules of cortical activation and of Hebbian plasticity must be linear; the two populations must be statistically indistinguishable in their connectivities and patterns of activity, so that a single covariance function describes both the covariance within each population and the covariance between the populations; and the positive and negative synapses must obey identical Hebbian learning rules. However, many feedforward projections, such as the retinogeniculate and geniculocortical projections in the mammalian visual system, are exclusively excitatory. Furthermore, where excitatory and inhibitory projections do coexist, they are not likely to be equivalent (discussed in [11]): excitatory and inhibitory populations often have distinct patterns of connectivity and of activation; and there is currently no evidence that inhibitory synapses can be modified by Hebbian rules.

One can derive a linear Hebb rule like Linsker's without the use of negative synapses by studying the *difference* between the innervation strengths of two equivalent excitatory projections [11]. Biological examples include ON–centre and OFF–centre inputs [9] and left–eye and right–eye inputs [12] in the mammalian visual system. In this case, however, the constants k_1 and k_2 disappear from the equation for the development of the difference of synaptic strengths because these constants take on equal values for each of the two equivalent populations. Therefore, such a biologically motivated linear Hebb model has $k_1 = k_2 = 0$, and lacks the constraints on which many of Linsker's results depend. Such a model can nonetheless develop orientation–selective receptive field structures if oscillations exist in the covariance functions of the input layer and if lateral interactions are introduced in the output layer [9]. In this case, orientation–selective receptive fields develop in the early, linear regime of development. Linsker's constraints, applied to a model with ON– and OFF–centre inputs, would fix the final percentages of ON and of OFF inputs in the receptive field. This in turn may determine the spatial phase of the resulting receptive field: for example, if a majority of synapses are ON, the central lobe of a tri–lobed cell must be composed of ON synapses. In the absence of constraints, receptive fields may vary their spatial phase. This leads to very different predictions for the organization of orientation selectivity across the cortex from those made by Linsker in [3] ([9]). This will be discussed in more detail in a future publication.

A How to treat the synaptic density and covariance functions explicitly

For the purposes of simulating equation (3) or computing the eigenvectors of the matrix $\mathbf{Q} + k_2\mathbf{J}$, it can be convenient to use an alternative representation of the synaptic strengths. Instead of having a label for every individual synapse w_i (so that the synaptic density function is implicit), a set of equally spaced representative synapses v_j is used, with positions $\mathbf{r}_j$. The number of synapses represented by v_j is given by the local synaptic density $A_j = A(\mathbf{r}_j)$, where $A(\mathbf{r})$ is the synaptic density function. v_j represents the average value of these synapses. Then the activity of the postsynaptic cell, for example, is $\sum_j v_j A_j x_j$, where x_j are the representative pre–synaptic activities.

- When equation (3) is transformed to the $\mathbf{v}$ representation we obtain

$$\dot{\mathbf{v}} = (\mathbf{C} + k_2\mathbf{J}^{\mathbf{v}})\mathbf{A}\mathbf{v} + k_1\mathbf{n} \qquad \text{subject to } -w_{\max} \leq v_j \leq w_{\max}.$$

  Here $\mathbf{A} = \mathrm{diag}\{A_j\}$, and $\mathbf{J}^{\mathbf{v}}$ is the appropriate size matrix of all 1's. $\mathbf{C}$ is the matrix constructed from the covariance function: $C_{jk} = C(\mathbf{r}_j, \mathbf{r}_k)$.

- In this representation the constraint $\sum w_i = k_1/|k_2|$ becomes $\sum v_j A_j = k_1/|k_2|$.

- For each eigenvector $\mathbf{e}^{\mathbf{v}}$ in the $\mathbf{v}$ representation, there is an equivalent eigenvector $\mathbf{e}^{\mathbf{w}}$ in the $\mathbf{w}$ representation with approximately identical eigenvalue. As long as the function $C(\mathbf{r}_j, \mathbf{r}_k)$ varies slowly on the scale of the spacing between the representative synapses, the additional eigenvectors of $\mathbf{Q} + k_2\mathbf{J}$ will have approximately zero eigenvalue.[16] Thus, the principal eigenvectors of $\mathbf{Q} + k_2\mathbf{J}$ can be found by computing the eigenvectors of $(\mathbf{C} + k_2\mathbf{J}^{\mathbf{v}})\mathbf{A}$.

- Eigenvectors of $(\mathbf{C} + k_2\mathbf{J}^{\mathbf{v}})\mathbf{A}$ can be found using the transformation $t_j = A_j^{\frac{1}{2}} v_j$ which symmetrises the matrix. In the continuum limit, we can represent the matrix $(\mathbf{C} + k_2\mathbf{J}^{\mathbf{v}})\mathbf{A}$ by the integral operator with kernel $(C(\mathbf{r}, \mathbf{r}') + k_2)A(\mathbf{r}')$.

- Let $\mathbf{n}$ be the DC vector in the $\mathbf{w}$ representation. Then the DC component of $\hat{\mathbf{e}}^{\mathbf{w}}$ is given in terms of $\mathbf{e}^{\mathbf{v}}$ by:

$$\hat{\mathbf{e}}^{\mathbf{w}} \cdot \hat{\mathbf{n}} \equiv \frac{\mathbf{e}^{\mathbf{w}} \cdot \mathbf{n}}{|\mathbf{n}||\mathbf{e}^{\mathbf{w}}|} = \frac{\sum_j e_j^{\mathbf{v}} A_j}{(\sum_j A_j)^{\frac{1}{2}} (\sum_j e_j^{\mathbf{v}2} A_j)^{\frac{1}{2}}}. \tag{11}$$

B Proof of Theorem 2

To prove theorem 2 we will construct the eigenvalue spectrum of $(\mathbf{Q} + k_2\mathbf{J})$ for a general covariance matrix $\mathbf{Q}$.

Lemma 1 *At $k_2 = 0$ all the eigenvalues are positive, for any covariance matrix $\mathbf{Q}$.*

Proof. The quadratic form $\mathbf{w}^{\mathrm{T}}\mathbf{Q}\mathbf{w}$ is always positive for $\mathbf{Q}$ a covariance:

$$\mathbf{w}^{\mathrm{T}}\mathbf{Q}\mathbf{w} = \mathbf{w}^{\mathrm{T}}\langle(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x}^{\mathrm{T}} - \bar{\mathbf{x}}^{\mathrm{T}})\rangle\mathbf{w} = \langle(\mathbf{w}^{\mathrm{T}}(\mathbf{x} - \bar{\mathbf{x}}))^2\rangle$$

Therefore the eigenvalues of $\mathbf{Q}$ must all be positive. •

[16] If for all $\mathbf{r}_j$ and $\mathbf{r}_k$, all covariances between synapses represented at $\mathbf{r}_j$ and $\mathbf{r}_k$ were exactly equal to $C(\mathbf{r}_j - \mathbf{r}_k)$, then the eigenvalue of an eigenvector in the v representation would exactly equal the eigenvalue of the equivalent eigenvector in the w representation, and the additional eigenvectors of $\mathbf{Q} + k_2\mathbf{J}$ would have eigenvalue exactly zero.

Lemma 2 *If* $\mathbf{Q}$ *has an eigenvector* $\mathbf{e}^{\mathrm{AC}}$ *that has no DC component then* $\mathbf{e}^{\mathrm{AC}}$ *is an eigenvector of* $(\mathbf{Q}+k_2\mathbf{J})$ *for all* k_2*, and its eigenvalue is independent of* k_2*.*

Proof. Let $\mathbf{Q}\mathbf{e}^{\mathrm{AC}}=\lambda_{\mathrm{AC}}\mathbf{e}^{\mathrm{AC}}$, where $\mathbf{e}^{\mathrm{AC}}\cdot\mathbf{n}=0$. $\mathbf{J}=\mathbf{n}\mathbf{n}^{\mathrm{T}}$, so

$$(\mathbf{Q}+k_2\mathbf{J})\mathbf{e}^{\mathrm{AC}}=\lambda_{\mathrm{AC}}\mathbf{e}^{\mathrm{AC}}+k_2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{e}^{\mathrm{AC}})=\lambda_{\mathrm{AC}}\mathbf{e}^{\mathrm{AC}}$$ •

So we can divide the eigenvectors into a set that are independent of k_2 (possibly empty), and a set that vary with k_2. The first set have no DC component so will be referred to as the AC set. The others will be called the DC–mixed set. The dependence of the DC–mixed eigenvalues on k_2 is addressed by the remainder of this theorem.

Lemma 3 *In the limit* $k_2\to\pm\infty$*, only one eigenvalue diverges to* $\pm\infty$*. Its eigenvector is the DC vector.*

Lemma 4 *The two sets of eigenvectors, in the two limits* $k_2\to\pm\infty$*, are identical.*

Lemma 5 *The two sets of eigenvalues, in the two limits* $k_2\to\pm\infty$*, are identical, except for the DC eigenvector's eigenvalues (Lemma 3).*

Proof. In the limit $k_2\to\pm\infty$, $\mathbf{Q}$ represents a negligible contribution to $(\mathbf{Q}+k_2\mathbf{J})$ and can be treated as a first-order perturbation to $k_2\mathbf{J}$ (see Appendix C). To first order, the eigenvectors of $(\mathbf{Q}+k_2\mathbf{J})$ in each non–degenerate subspace of $\mathbf{J}$ are the eigenvectors of $k_2\mathbf{J}$. $\mathbf{J}$ has only one eigenvector with non–zero eigenvalue, namely the DC vector $\hat{\mathbf{n}}$, which has eigenvalue N, the dimension of the matrices $\mathbf{Q}$ and $\mathbf{J}$. Thus to first order, the DC vector is an eigenvector of $(\mathbf{Q}+k_2\mathbf{J})$ with eigenvalue $k_2N+\langle\hat{\mathbf{n}}|\mathbf{Q}|\hat{\mathbf{n}}\rangle=N(k_2+\bar{q})$, where $\bar{q}=\langle Q_{ij}\rangle$, the average covariance (averaged over pairs of synapses i,j). To zero order, the other eigenvectors $\hat{\mathbf{e}}$ of $\mathbf{Q}+k_2\mathbf{J}$ are the eigenvectors of $\mathbf{Q}$ in the subspace orthogonal to $\mathbf{n}$.[17] The corresponding eigenvalues are $\langle\hat{\mathbf{e}}|\mathbf{Q}|\hat{\mathbf{e}}\rangle$, which is always finite. The degenerate subspace is identical for the two matrices $k_2\mathbf{J}, k_2=\pm\infty$, so the sets of eigenvectors and eigenvalues in the degenerate subspace are identical in the two cases. •

Note that lemma 3 means that in the limit $k_2\to\pm\infty$, since the eigenvectors are orthogonal, all the other eigenvectors must have DC component of order k_2^{-1} or smaller.

Lemma 6 *If* $\mathbf{A}(t)$ *is a differentiable Hermitian matrix function with positive semi–definite derivative* $\frac{d\mathbf{A}}{dt}$ *then the eigenvalues* $\lambda_a(t)$ *of* $\mathbf{A}$ *are non-decreasing functions of* t*.*

For proof see Theorem V.2.3 in [1, p. 459ff.].

Now $d/dk_2(\mathbf{Q}+k_2\mathbf{J})=\mathbf{J}$, and $\mathbf{J}$ is positive semi–definite, so:

Lemma 7 *The DC–mixed eigenvalues of* $(\mathbf{Q}+k_2\mathbf{J})$ *are monotonically increasing continuous functions of* k_2*.*

Now we are ready to construct the spectrum, by "joining the dots". Subject to the constraints of monotonicity and continuity, we have to join the three sets of points at $k_2=0,\pm\infty$ (dots in figure 11). There is only one way in which this can be done, shown by the lines in figure 11.

Result 1 *There is at most one eigenvector with negative eigenvalue. It tends to a flat DC function in the limit of large negative* k_2*.*

Result 2 *All eigenvalues but one remain finite for all* k_2*.*

Result 3 *The eigenvalues vary monotonically and continuously between asymptotes, such that each eigenvalue has a limited range as a function of* k_2*, and these ranges touch without overlapping.*

[17] The definition of an eigenvector $\mathbf{e}$ of $\mathbf{Q}$ in a given subspace is: $\mathbf{e}$ lies in the subspace and is an eigenvector of $\mathbf{PQP}$ where $\mathbf{P}$ is the orthogonal projection operator into the subspace.

Figure 11: **Combining the lemmas.**
The lemmas uniquely determine the form of the spectrum of eigenvalues of the DC–mixed eigenvectors as a function of k_2. The AC eigenvalues are horizontal lines, not shown in this figure. (a) Lemma 1 establishes that all the intersections with the λ axis are positive (black dots at $k_2 = 0$). Lemmas 3, 4 and 5 establish the relationship between the eigenvalues at $\pm\infty$ (black dots at $k_2 = \pm\infty$). (b) Lemma 7 establishes that these points must be joined continuously and monotonically. The lines show the only way of joining the dots to satisfy the lemmas.

C Perturbation theory

We derive perturbation theory [8] approximations for some of the eigenvalues $\lambda_a^{k_2}$ and DC components $n_a^{k_2}$ of the eigenvectors of $\mathbf{Q} + k_2\mathbf{J}$ for large k_2. These are expressed in terms in terms of the eigenvalues λ_a and DC components n_a of the eigenvectors of $\mathbf{Q}$, $\{\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, ...\}$. We use odd subscripts to denote the AC eigenvectors and even subscripts to denote the eigenvectors with non–zero DC component.

For large k_2, consider $\mathbf{Q}$ as a first order perturbation to the matrix $k_2\mathbf{J}$. This is valid iff all the eigenvalues of $\mathbf{Q}$ are much smaller than the non–zero eigenvalue of $k_2\mathbf{J}$, *i.e.* if $\lambda_0 \ll k_2 N$. The orthonormal eigenvectors of $k_2\mathbf{J}$ are the DC vector $\hat{\mathbf{n}}$, which has eigenvalue Nk_2, and any set of AC eigenvectors orthogonal to $\hat{\mathbf{n}}$, which have eigenvalue zero. The perturbing matrix $\mathbf{Q}$ breaks the degeneracy of the AC subspace. So the eigenvectors of $k_2\mathbf{J} + \mathbf{Q}$ for $k_2 \to \pm\infty$ are, to zero order, $\hat{\mathbf{n}}$ and the eigenvectors of $\mathbf{Q}$ in the AC subspace (*i.e.* the eigenvectors of $\mathbf{PQP}$ where $\mathbf{P} = \mathbf{I} - \mathbf{n}\mathbf{n}^{\mathrm{T}}$ is the orthogonal projection operator onto the AC subspace). We denote these eigenvectors by $\{\hat{\mathbf{n}}, \mathbf{e}_1^\infty, \mathbf{e}_2^\infty, \mathbf{e}_3^\infty, ...\} = \{\hat{\mathbf{n}}, \mathbf{e}_1, \mathbf{e}_2^\infty, \mathbf{e}_3, ...\}$. We use these eigenvectors as the basis in which we perform first order perturbation theory. In this basis, $k_2\mathbf{J}$ is diagonal, and $\mathbf{Q}$ is diagonal except for cross terms in the first row and column:

$$k_2\mathbf{J} = \begin{pmatrix} Nk_2 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{12}$$

$$\mathbf{Q} = \begin{pmatrix} \langle \hat{\mathbf{n}}|\mathbf{Q}|\hat{\mathbf{n}}\rangle & 0 & \langle \hat{\mathbf{n}}|\mathbf{Q}|\mathbf{e}_2^\infty\rangle & 0 & \cdots \\ 0 & \langle \mathbf{e}_1|\mathbf{Q}|\mathbf{e}_1\rangle & 0 & 0 & \cdots \\ \langle \mathbf{e}_2^\infty|\mathbf{Q}|\hat{\mathbf{n}}\rangle & 0 & \langle \mathbf{e}_2^\infty|\mathbf{Q}|\mathbf{e}_2^\infty\rangle & 0 & \cdots \\ 0 & 0 & 0 & \langle \mathbf{e}_3|\mathbf{Q}|\mathbf{e}_3\rangle & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{13}$$

To first order, the perturbed eigenvalue of the principal eigenvector $\hat{\mathbf{n}}$ is $\lambda_0^{k_2} = Nk_2 + \langle \hat{\mathbf{n}}|\mathbf{Q}|\hat{\mathbf{n}}\rangle = N(k_2 + \bar{q})$, where $\bar{q} = \langle \hat{\mathbf{n}}|\mathbf{Q}|\hat{\mathbf{n}}\rangle / N$, the average covariance.

To estimate the eigenvalues and DC components of $\mathbf{e}_2^\infty$, we need to make a further assumption: writing $\hat{\mathbf{n}}$ in terms of the eigenvectors of $\mathbf{Q}$,

$$\hat{\mathbf{n}} = n_0\mathbf{e}_0 + n_1\mathbf{e}_1 + n_2\mathbf{e}_2 + n_3\mathbf{e}_3 + \ldots \tag{14}$$

where n_a is the DC component of $\mathbf{e}_a$, we now assume that $\mathbf{e}_0$ is close to the DC vector $\hat{\mathbf{n}}$, and that $|n_0| \gg |n_2| \gg |n_4| > |n_a|\ \forall a > 4$. So we are assuming that $\mathbf{e}_0$ accounts for most of $\hat{\mathbf{n}}$ in equation (14), and that $\mathbf{e}_2$ accounts for most of the remainder. Then we can approximate $\mathbf{e}_2^\infty$ as follows:

$$\mathbf{e}_2^\infty = \frac{n_0\mathbf{e}_2 - n_2\mathbf{e}_0}{\sqrt{n_0^2 + n_2^2}} \tag{15}$$

The corrections to this expression are of order n_4. The assumption that $\mathbf{e}_0$ is close to the DC vector $\hat{\mathbf{n}}$ can be motivated for non–negative matrices $\mathbf{Q}$ by the Frobenius–Perron theorem (section 3.1).

So using first order perturbation theory and equations (13) and (15), we can derive the eigenvalue λ_2^∞ and the DC component $n_2^{k_2}$:

$$\lambda_2^\infty \simeq \langle \mathbf{e}_2^\infty|\mathbf{Q}|\mathbf{e}_2^\infty\rangle = \frac{n_0^2\lambda_2 + n_2^2\lambda_0}{n_0^2 + n_2^2}$$

$$n_2^{k_2} \simeq \frac{\langle \hat{\mathbf{n}}|\mathbf{Q}|\mathbf{e}_2^\infty\rangle}{\lambda_2^{k_2} - \lambda_0^{k_2}} \simeq \frac{(\lambda_0 - \lambda_2)n_0n_2}{Nk_2\sqrt{n_0^2 + n_2^2}}$$

Note that this shows for $k_2 \to -\infty$ that $n_2^{k_2}$ and n_2 have opposite signs, and:

$$n_2^{k_2}k_2 \simeq \frac{(\lambda_0 - \lambda_2)n_0n_2}{N\sqrt{n_0^2 + n_2^2}}$$

The accuracy of these perturbation theory approximations depends on two factors. 1) The second and higher order corrections due to the use of finite k_2 are negligible if $\frac{\lambda_0/N}{|k_2|} \ll 1$. This condition is satisfied for Linsker's parameters as already discussed in section 4.3. 2) Corrections to equation (15) are small if $1 - (n_0^2 + n_2^2) \ll 1$;[18] this condition is increasingly poorly satisfied as C/A decreases and breaks down for $C/A <\sim 0.2$. For $C/A = 2/3$ we have compared the above perturbation theory approximations of eigenvalues and DC components with the results of the numerical calculations described in the caption of figure 5; the two agree to within 15%.

[18]Strictly, corrections are small if $n_4\frac{\lambda_0-\lambda_2}{\lambda_2^\infty-\lambda_4^\infty} \ll 1$. The condition in the text only assures $n_4 \ll 1$.

D Analytic results for Linsker's system

D.1 Solution for eigenfunctions and eigenvalues

The covariance operator is $\mathbf{Q} = e^{-r^2/2C} * e^{-r^2/2A}\times$, *i.e.* multiplication by a gaussian of size $A = r_A^2$ followed by convolution with a gaussian of size $C = r_C^2$, where r_A and r_C are the characteristic arbor radius and the characteristic covariance radius (Appendix A).

Because of the circular symmetry, the eigenfunctions of this operator can be written as the product of a radial function $f(r)$ with one of the angular functions $\cos l\theta, \sin l\theta$.

We can find the radial functions for a given l by the use of guesswork and l-th order Hankel transforms, which are the Fourier transforms for cylindrical systems.

D.1.1 Hankel transforms

The two–dimensional Fourier transform of $\cos(l\theta)f(r)$ is $2\pi i^l \cos(l\phi)F_l(k)$, where

$$F_l(k) = \int_0^\infty r f(r) J_l(kr) dr$$

is the l-th order Hankel transform of $f(r)$, and $J_l(x)$ is a Bessel function

$$J_l(x) = \frac{1}{\pi}\int_0^\pi cos(l\theta - x sin\theta)d\theta$$

The inverse transform is symmetric:

$$f(r) = \int_0^\infty k F_l(k) J_l(kr) dk$$

We use the following Hankel transforms [13, 2]:[19]

$f(r)$	$F_0(k)$
$e^{-r^2/2B}$	$Be^{-Bk^2/2}$
$r^2e^{-r^2/2B}$	$B^2(2 - Bk^2)e^{-Bk^2/2}$
$f(r) * g(r)$	$2\pi F_0(k)G_0(k)$

$f(r)$	$F_l(k)$
$r^l e^{-r^2/2B}$	$B^{l+1}k^l e^{-Bk^2/2}$
$f(ar)$	$a^{-2}F_l(k/a)$

D.1.2 Eigenfunctions

We guess radial functions with the form of a gaussian $e^{-r^2/2R}$ multiplied by a polynomial, then find the parameters that make the functions into eigenfunctions by solving the eigenfunction equation $\mathbf{Q}(\mathbf{r},\mathbf{r}')cos(l\theta)f(r') = \lambda cos(l\theta)f(r)$. The way to evaluate $\mathbf{Q}(\mathbf{r},\mathbf{r}')cos(l\theta)f(r')$ is to first multiply $f(r')$ by $e^{-r'^2/2A}$ then convolve it with $e^{-r^2/2C}$ by going into Fourier space, multiplying by the transform $e^{-Ck^2/2}$, and returning from Fourier space. The first six eigenfunctions have been given in table 1. We note in addition the expression for the eigenfunctions with l angular nodes and no radial nodes: eigenfunction $= r^m \cos(m\theta)e^{-r^2/2R}$, $\lambda/N = L^{l+1}C/A$, using the notation of table 1. Note that all the eigenvalues differ by a ratio of $L = (R - C)/R$ raised to some power.

[19] $*$ denotes two–dimensional convolution

Figure 12: **Eigenvalue ratio and DC components as a function of C/A:** The ratio $\lambda_{2p}/\lambda_{2s}$, and the DC components n_{1s} and n_{2s} are shown exactly for $k_2 = 0$. The ratio $\lambda_{2p}/\lambda_{2s}^{\infty}$ is estimated by perturbation theory for $k_2 \to -\infty$ (this perturbation estimate becomes invalid for $C/A <\sim 0.2$). Linsker used $C/A = 2/3$ or $2/5$ for layer $\mathcal{B} \to \mathcal{C}$.

D.2 Derivations

Equipped with these expressions for the eigenfunctions and eigenvalues, we can derive or estimate several properties of the system analytically as a function of the gaussian parameter ratio C/A. We can evaluate exactly the DC components n_{1s} and n_{2s} of the 1s and 2s functions at $k_2 = 0$ using equation (11):

$$n_a = \frac{\int f_a(r) e^{-r^2/2A} r\, dr}{\left(\int f_a^2(r) e^{-r^2/2A} r\, dr\right)^{1/2} \left(\int e^{-r^2/2A} r\, dr\right)^{1/2}}$$

Adopting the convention that the 1s function is positive, and the 2s function has a positive center, this gives

$$n_{1s} = \frac{u}{A^{1/2} v^{1/2}}$$

and

$$n_{2s} = \frac{u(1 - 2u/r_0^2)}{\sqrt{Av(1 - 4v/r_0^2 + 8v^2/r_0^4)}}$$

where $u = RA/(R + A)$ and $v = RA/(R + 2A)$. n_{1s} and n_{2s} are plotted against C/A in figure 12. n_{2s} is always negative, and $|n_{2s}|$ attains its maximum value at $C/A \simeq 1/8$.

We can also evaluate $\bar{q} = 1/(1 + 2A/C)$, and the effective number of synapses $N = 2\pi A$.

D.3 Perturbation theory estimates

Using the results of appendix C, we can estimate the eigenvalue and DC component of 2s as $k_2 \to -\infty$. In the interval of C/A where the perturbation approximations are valid, the eigenvalue of 2s at $k_2 = -\infty$ is greater than the eigenvalue of 2s at $k_2 = 0$, but it never exceeds the eigenvalue of 2p (Figure 12).

These approximations and the expressions in section 6 were used to estimate the critical DC level g^{E} (Figure 8) and the critical number of synapses N^* (Figure 9).

D.4 The sign of the centre of the centre–surround structures

Linsker noted empirically that the centre of the centre–surround structure always has the same sign as the DC bias g. Our analysis can explain this. We again adopt the convention that the 1s function is positive and the 2s function has a positive centre at $k_2 = 0$. We have just shown (appendix D.2) that $n_{2\mathrm{s}}(k_2 = 0)$ is negative for any C/A. Thus, the 2s function for $k_2 \to -\infty$ defined in equation (15) also has a positive center. Perturbation theory (Appendix C) at $k_2 = 0$ shows that $n_{2\mathrm{s}}(k_2 \to -\infty)$ has opposite sign to $n_{2\mathrm{s}}(k_2 = 0)$, so $n_{2\mathrm{s}}$ is positive for the 2s eigenvector as $k_2 \to -\infty$. Now $w_{2\mathrm{s}}^{\mathrm{FP}} = -n_{2\mathrm{s}}|k_2|\sqrt{N} g w_{\max}/(\lambda_{2\mathrm{s}}/N)$, so $w_{2\mathrm{s}}^{\mathrm{FP}}$ has the opposite sign from g. $w_{2\mathrm{s}}$ starts near the origin and grows in the direction opposite to the fixed point, so a positive bias will cause a centre–surround structure with positive centre to emerge, and a negative bias causes a negative centre.

E Derivation of Constrained Dynamics

Let $\mathbf{P}$ be the projection operator onto the surface $\sum w_j = 0$, $\mathbf{P} = (\mathbf{I} - \hat{\mathbf{n}}\hat{\mathbf{n}}^{\mathrm{T}})$, where $\mathbf{I}$ is the identity matrix. We start from equation (5):

$$\dot{\mathbf{w}} = (\mathbf{Q} + k_2\mathbf{J})\mathbf{w} + k_1\mathbf{n}$$

Writing $\mathbf{w} = \mathbf{Pw} + \frac{1}{N}(\mathbf{w}\cdot\mathbf{n})\mathbf{n}$ we examine the evolution of the DC component and the AC part of $\mathbf{w}$:

$$\begin{aligned}
\frac{d}{dt}(\mathbf{w}\cdot\mathbf{n}) &= \mathbf{n}\cdot(\mathbf{Q} + k_2\mathbf{J})\mathbf{w} + k_1\mathbf{n}\cdot\mathbf{n} \\
&= \mathbf{n}\cdot(\mathbf{Q} + k_2\mathbf{J})(\mathbf{Pw} + \frac{1}{N}(\mathbf{w}\cdot\mathbf{n})\mathbf{n}) + Nk_1 \\
&= \mathbf{nQPw} + \frac{1}{N}(\mathbf{w}\cdot\mathbf{n})\mathbf{nQn} + \frac{k_2}{N}(\mathbf{w}\cdot\mathbf{n})\mathbf{nJn} + Nk_1 \\
&= \mathbf{nQPw} + (\mathbf{w}\cdot\mathbf{n})(\bar{q}N + k_2N) + Nk_1
\end{aligned}$$

where $\bar{q} = \langle Q_{ij}\rangle$, the average covariance. For large k_1 and k_2 we can neglect the first two terms and we have

$$\frac{d}{dt}(\mathbf{w}\cdot\mathbf{n}) = N((\mathbf{w}\cdot\mathbf{n})k_2 + k_1)$$

So if k_2 is negative, $(\mathbf{w}\cdot\mathbf{n})$ decays towards an equilibrium value of $\frac{k_1}{|k_2|}$.

Now the evolution of the AC component is derived:

$$\begin{aligned}
(\dot{\mathbf{Pw}}) &= \mathbf{P}(\mathbf{Q} + k_2\mathbf{J})\mathbf{w} + k_1\mathbf{Pn} \\
&= \mathbf{PQ}(\mathbf{Pw} + \frac{1}{N}(\mathbf{w}\cdot\mathbf{n})\mathbf{n})
\end{aligned}$$

since $\mathbf{PJ} = 0$ and $\mathbf{Pn} = 0$. So once $\mathbf{w}$ lies in the constraint surface,

$$\dot{\mathbf{w}} = \mathbf{PQPw} + \frac{k_1}{|k_2|N}\mathbf{PQn} \tag{16}$$

F Derivation of $w_1(0)_{\mathrm{rms}}$ (section 6.2)

We estimate the typical magnitude of $w_1(0)$ with the root mean square value of $\mathbf{e}_1\cdot\mathbf{w}(0)$:

$$\mathrm{var}(\mathbf{e}_1\cdot\mathbf{w}(0)) = \mathrm{var}(\sum_i e_i^{(1)} w_i(0)) = \sum_i (e_i^{(1)})^2 \mathrm{var}(w_i(0))$$

Figure 13: **The model system studied**
a) The network is made up of one–dimensional layers of neurons. Each neuron receives synapses with uniform density from a row of n neurons directly below. b) The number of inputs common to two layer $\mathcal{B}$ neurons with overlapping arbors is $n - |j-k|$, where j and k label the neurons sequentially. c) Covariance function in layer $\mathcal{B}$ as a function of separation: assuming no correlations in layer $\mathcal{A}$, the covariance is proportional to the overlap.

Because $\mathbf{e}_1$ is normalised, $\sum_i (e_i^{(1)})^2 = 1$, so

$$\text{var}(\mathbf{e}_1 \cdot \mathbf{w}(0)) = \text{var}(w_i(0))$$

When $g = 0$, if the $w_i(0)$ are uniformly distributed on $[-w_{\text{max}}, w_{\text{max}}]$, $\text{var}(w_i(0)) = \frac{1}{3}w_{\text{max}}^2$. Thus $w_1(0)_{\text{rms}} = \frac{1}{\sqrt{3}}w_{\text{max}}$. When $g > 0$, the weight vector is initially projected into the hyperplane $\bar{w}(0) = gw_{\text{max}}$. The initial fluctuations in $w_i(0)$ are then diminished because a certain fraction of the weights get squashed against the upper hard limit and are no longer random variables. After this projection, the distribution of weights has a delta function of size $u/2$ at $w(0) = w_{\text{max}}$ and is uniform on the interval $(-(1-u)w_{\text{max}}, w_{\text{max}})$, where $u = 2(1-\sqrt{1-g})$. This distribution has $\bar{w}(0) = gw_{\text{max}}$, and $\sqrt{\text{var}(w_i(0))} = \sigma(g)w_{\text{max}}$ where $\sigma(g) = \sqrt{\frac{1}{6}(2 - 3u^2 + 2u^3 - \frac{3}{8}u^4)}$. Thus,

$$w_1(0)_{\text{rms}} = \sigma(g)w_{\text{max}}.$$

If the principal eigenfunction has a degeneracy d (for example 2p has $d = 2$), then the effective fluctuations are larger by a factor of $\sqrt{d}$, so:

$$w_1(0)_{\text{rms}} = \sigma(g)w_{\text{max}}\sqrt{d}$$

G Solution for a Model System

A complete solution has been obtained for equation (3) for a model one-dimensional system. It was hoped that the emergence of centre–surround receptive fields in two–dimensional systems should be replicable in a network with one–dimensional layers (Figure 13a), and would be easier to understand there. This model system does share all the properties of Linsker's system, except for a subtle twist in the robustness of the centre–surround structures.

Let each neuron receive connections from a finite arbor of width n with uniform synaptic density. By analogy with Linsker's network, the first layer of weights are all set to the positive hard–limit. This generates correlations between the activities of the units in the second layer $\mathcal{B}$. These correlations will drive the production of non–trivial connections to layer $\mathcal{C}$.

If the noise in the first layer is uncorrelated, the covariance between two units in layer $\mathcal{B}$ is just proportional to the amount of overlap in their inputs (Figure 13b,c):[20]

$$Q_{jk}^{\mathcal{B}} = \begin{cases} n - |j-k| & |j-k| < n \\ 0 & |j-k| > n \end{cases} \tag{17}$$

where j and k label the neurons in layer $\mathcal{B}$ sequentially. So the covariance matrix for a layer $\mathcal{C}$ neuron with a row of n inputs is:

$$\mathbf{Q}^{\mathcal{B}} = \begin{pmatrix} n & n-1 & n-2 & \cdots & 1 \\ n-1 & n & n-1 & \cdots & 2 \\ n-2 & n-1 & n & \cdots & 3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 3 & \cdots & n \end{pmatrix}$$

Let us go to the continuum limit and consider a single neuron in layer $\mathcal{C}$ that receives a continuum of inputs from the interval $-n/2 \leq x \leq n/2$. The synaptic strengths are now represented by a weight function on the interval $(-n/2, n/2)$ rather than an N–dimensional vector $\mathbf{w}$. The covariance function is (Figure 13):

$$Q(x,y) = n - |x-y| \qquad\qquad |x|,|y| \leq n/2$$

where x and y label the distances of input cells from the centre of the arbor. Now we want to know:

- What are the eigenfunctions and eigenvalues of the operator $(\mathbf{Q} + k_2\mathbf{J})$?
- How can the fixed point of the dynamics be characterised?
- Can centre–surround weight functions dominate the dynamics of this system?

G.1 Eigenfunctions

The function $Q(x,y) + k_2$ is symmetric under interchange of x and y and also under the transformation $(x,y) \to (-x,-y)$. (In terms of matrices, the matrix $\mathbf{Q} + k_2\mathbf{J}$ is symmetric about both diagonals.) This means that not only are the eigenfunctions real and orthogonal, but they must also be eigenstates of the transformation $x \to -x$, *i.e.*

Property 1 *The eigenfunctions can be divided into a set of symmetric and a set of antisymmetric functions.*

The symmetric eigenfunctions are the analog of the s–modes in Linsker's 2D system. Now since $\mathbf{J} = \mathbf{n}\mathbf{n}^{\mathrm{T}}$, $\mathbf{J} \cdot \mathbf{w}^A = 0$ for all antisymmetric $\mathbf{w}^A$ (and for any $\mathbf{w}$ with no DC component). So

Property 2 *The antisymmetric eigenfunctions and eigenvalues are independent of k_2.*

The eigenfunctions are defined by:

$$\int_{-n/2}^{n/2} (Q(x,y) + k_2) w(y) dy = \lambda w(x) \tag{18}$$

Such eigenfunction equations cannot generally be solved analytically. This one can be solved, however, because the integral operator is a function of $(x-y)$ (*i.e.* is Toeplitz), and has a simple piecewise linear form. As shown in Appendix H.1,

[20] Here we use $Q_{jk}^{A} = \delta_{jk}$ and $w_{\max} = 1$.

Figure 14: **Eigenfunctions and eigenvalues as a function of k_2**
This schematic diagram shows the largest eigenvalues as a function of k_2. The horizontal lines are the eigenvalues of the antisymmetric functions, which are independent of k_2. The lines with positive gradient are the eigenvalues of the symmetric eigenfunctions. The little boxes show schematically the structure of the eigenfunction associated with each eigenvalue.

Property 3 *The antisymmetric eigenfunctions are* sin *functions and the symmetric eigenfunctions are* cos *functions with frequency ω related to λ by $\omega^2 = 2/\lambda$. For sufficiently negative k_2 there is also a* cosh *solution with negative eigenvalue.*

The possible frequencies are discretised and depend on k_2. The discretisation condition is derived in Appendix H.2. The eigenfunctions and their eigenvalues are shown schematically as a function of k_2 in figure 14.

G.2 Discussion of eigenfunctions

As we vary k_2, the properties of the eigenfunctions distinguish two regimes:

- $k_2 > -n/2$:

  Here, the eigenfunctions are strictly ordered in eigenvalue by the number of changes of sign in each eigenfunction, and there are no eigenfunctions with negative eigenvalue. The eigenfunction[21] $w^{(0)}(x)$ with no zero crossings dominates the dynamics, since it has the largest eigenvalue.

  Property 4 *When $k_2 > -n/2$, the principal eigenfunction is symmetric and has no changes of sign.*

  The centre–surround eigenfunction with two zero crossings, $w^{(2)}(x)$, is the eigenfunction with third largest eigenvalue, so it has smaller typical growth rate.

- $k_2 < -n/2$:

  Here, the principal eigenfunction is the antisymmetric function $w^{(1)}(x)$;

  Property 5 *When $k_2 < -n/2$, the principal eigenfunction is antisymmetric and has one node.*

[21] The eigenfunctions will be labelled with a superscript denoting their number of zero–crossings.

$w^{(0)}(x)$ changes from a cos wave with no zero crossings to a cosh solution with negative eigenvalue. $w^{(0)}(x)$ is now associated with a constraint surface since it has a negative eigenvalue. In the limit $k_2 \to -\infty$, $w^{(0)}(x)$ tends to the flat DC function, with eigenvalue $\lambda \to (k_2+2n/3)n$.[22]

Property 6 *When $k_2 < -n/2$, there is a DC eigenfunction with large negative eigenvalue.*

The dynamics attract the trajectory to a plane perpendicular to this direction, so

Property 7 *When $k_2 < -n/2$, the weight vector is constrained to have a certain total DC component.*

Equivalently, the average synaptic strength is constrained to be a constant g, given in terms of the parameters by:[23]

$$g = \mathbf{w} \cdot \mathbf{n}/n = \frac{k_1}{n|k_2 + 2n/3|}$$

Within this constraint hyperplane, if $k_1 = 0$ (so that the fixed point is at the origin), the antisymmetric eigenfunction $w^{(1)}(x)$ with one zero crossing dominates the dynamics, since it is now the principal eigenfunction.

So for $k_1 = 0$, there are two parameter regimes, $k_2 > -n/2$ and $k_2 < -n/2$, in which $w^{(0)}(x)$ and $w^{(1)}(x)$ dominate respectively.

Property 8 *The centre–surround eigenfunction is never the principal eigenfunction. It is the principal* symmetric *eigenfunction for $k_2 < -n/2$.*

G.3 Fixed Point

Substituting into equation (6), the component of the fixed point in the direction of an eigenfunction is $k_1 \mathbf{e}^{(a)} \cdot \mathbf{n}/\lambda_a$, *i.e.* it is proportional to the DC component of the eigenfunction. So

Property 9 *The fixed point has no component in the direction of any antisymmetric eigenfunction.*

In the limit of large k_2, the components of the fixed point in the direction of each eigenvector can be found analytically. As shown in Appendix H.4, the components in the direction of the symmetric eigenfunctions are all of equal order. Thus

Property 10 *The fixed point gives a velocity advantage to the growth of symmetric eigenfunctions.*

G.4 Discussion of fixed point

The only chance for centre–surround weight functions to be obtained with high probability is if it is possible for the fixed point to give a "head start" to the centre–surround function $w^{(2)}(x)$. The fixed point only has non–zero components in the direction of symmetric eigenfunctions. So having chosen $k_2 \ll -n/2$ to kill the dominant eigenfunction $w^{(0)}(x)$, it seems possible that k_1 might be chosen to give the second–in–line $w^{(2)}(x)$ a sufficient advantage in growth rate for it to dominate over the leading antisymmetric eigenfunction $w^{(1)}(x)$. However increasing k_1 has two effects. The distance of the fixed point from the origin is proportional to k_1, so

[22] see Appendix H.3

[23] see Appendix H.4

Property 11 *Increasing k_1*

1. *gives a velocity advantage to the growth of symmetric eigenfunctions.*
2. *increases the size of DC component that the weight function is constrained to have.*

We can now refer to the criteria of section 6 to estimate whether centre–surround structures will dominate the dynamics.

Energy criterion

For this simple system it is possible to evaluate the energies of the appropriate fully saturated structures exactly, and we are able to see how badly our approximate estimates perform. The energies of a saturated symmetric centre–surround structure and a saturated asymmetric structure are (to within an additive constant):

$$E_{\text{symm}} = -\frac{1}{48}n^3(1-g^2)(1+3g) \quad \text{and} \quad E_{\text{asymm}} = -\frac{1}{12}n^3(1-g^2)$$

So $E_{\text{symm}} = E_{\text{asymm}}$ implies $g = 1$, and a saturated centre–surround structure is never the energy–minimising configuration for this system. This exact result $g^{\text{E}} = 1$ can be compared with the estimate of g^{E} (using $\lambda_1 = 2n^2/\pi^2$, $\lambda_2 = n^2/2\pi^2$, $n_2k_2 = \sqrt{2}\lambda_2/n$), $\hat{g}^{\text{E}} = 1/(1+2\sqrt{2}/3) = 0.51$. We can see that for the extreme case $g^{\text{E}} = 1$ our estimator performs rather poorly.

So in this model system, there are only two energetically robust parameter regimes, dominated by $w^{(0)}$ and $w^{(1)}$; this is a difference from Linsker's system, in which the centre–surround structure becomes energetically favoured before the DC level swamps the visibility of AC structures.

Property 12 *Centre–surround structures are never the energy–minimising structures.*

Time Development Criterion

N^* can be evaluated also:

$$N^* = \frac{\sigma^2(g)\left((\sqrt{2}-1)g+1\right)^8}{16g^8(1-g)^2}$$

and with a sufficient number of synapses, centre–surround structures do indeed dominate the initial dynamics; but the subtle twist is that (in contradiction to assumption 1) the resulting saturated centre–surround structures are not necessarily stable under equation (3): saturated centre–surround structures can collapse synapse by synapse into the more energetically favourable asymmetric state. Whether this collapse occurs depends sensitively on the parameters. For a neuron with N equally spaced inputs it can be shown that there are $N/2+1$ regions in which saturated centre–surround is stable and $N/2$ regions in which it is not (for N even; similar results hold for N odd). We state without proof the boundaries between these regions: for large negative k_2 there is a transition between a stable and unstable centre–surround region at every $g = \pm(2I+1)/N$, where I is an integer $\in [0, N/2-1]$. A set of $(N+1)$ zebra stripes in parameter space result. Thus the final twist in this model system is that although centre–surround structures *can* dominate the initial dynamics if symmetry breaking fluctuations are suppressed, their final stability is not robust to small variations in the parameters.

Property 13 *Centre–surround structures do not necessarily occur robustly.*

It is not known whether this instability that is so sensitive to the parameters is a special property of this system. Perhaps the extremely ordered form of the network may be responsible for the strange behaviour, and a randomised form of the same network would behave differently. Another possibility is that the stability of domain boundaries to distortions in one– and two–dimensional systems might be different.

G.5 Summary

The important features that we wanted to know to characterise the dynamics were the principal components of $\mathbf{Q} + k_2\mathbf{J}$, its negative eigenfunctions, and a characterisation of the fixed point.

There are two regimes.

- When $k_2 \simeq 0$, the principal component is a cos function with no changes of sign. There are no negative eigenvalues.
- When $k_2 \ll 0$:
  - The principal component is an antisymmetric sin function with one change of sign.
  - The principal symmetric eigenfunction is the centre–surround eigenfunction.
  - There is a single negative eigenvalue corresponding to a flat eigenfunction, which enforces a constraint on the total synaptic sum.
  - The location of the fixed point fixes the level of the constraint, and it favours the growth of symmetric eigenfunctions. By energy considerations, centre–surround structures are never favoured over asymmetric structures. If symmetry–breaking fluctuations are suppressed, centre–surround may dominate the initial dynamics, but their stability on the hypercube depends crazily on the precise values of the parameters.

So this model system has properties analogous to all the properties of Linsker's system, with the exception that there is no regime in which centre–surround structures are robustly favoured.

H Derivations for model system

H.1 Eigenfunctions

Starting from equation (18), the eigenfunctions satisfy:

$$\int_{-m}^{m} (n + k_2 - |x - y|)w(y)dy = \lambda w(x) \tag{19}$$

where $m = n/2$.

We differentiate equation (19) twice with respect to x. Differentiating the triangular function $(n + k_2 - |x - y|)$ twice gives a delta function at $x = y$, and we obtain:

$$\lambda \frac{\partial^2}{\partial x^2} w(x) = \int_{-m}^{m} \frac{\partial^2}{\partial x^2}(n + k_2 - |x - y|)w(y)dy = \int_{-m}^{m} -2\delta(x - y)w(y)dy = -2w(x)$$

so

$$w''(x) = -\tfrac{2}{\lambda} w(x)$$

So the antisymmetric eigenfunctions are sin waves and the symmetric eigenfunctions are cos waves with frequency ω related to λ by $\omega^2 = 2/\lambda$. There may also be a cosh solution with negative eigenvalue. The possible frequencies are discretised and depend on k_2. The discretisation condition can be found by explicit integration of the eigenfunction equation.

Figure 15: **Solutions for Frequencies of Eigenfunctions**
The functions $\tan \omega m$ and $\frac{1}{\omega m(k_2/m+1)}$ are shown. At all $m\omega$ where the curves intersect, there is a solution for an eigenfunction $\cos \omega x$. The second curve is shown twice for two different values of k_2, one greater and one smaller than $-m$.

H.2 Discretisation of eigenvalues

Let $w(x) = \cos \omega x$:

$$\begin{aligned} \lambda \cos \omega x \quad = \quad I \quad = & \int_{-m}^{x} \{(n + k_2 - x) + y\} \cos \omega y \, dy \\ & + \int_{x}^{m} \{(n + k_2 + x) - y\} \cos \omega y \, dy \end{aligned}$$

Now using $\int_a^b \cos \omega y \, dy = \frac{1}{\omega} \sin \omega y |_a^b$ and $\int_a^b y \cos \omega y \, dy = \frac{1}{\omega} y \sin \omega y |_a^b + \frac{1}{\omega^2} \cos \omega y |_a^b$:

$$\begin{aligned} I \quad = \quad & \tfrac{1}{\omega}(n + k_2 - x)(\sin \omega x + \sin \omega m) + \tfrac{1}{\omega}(x \sin \omega x - m \sin \omega m) + \tfrac{1}{\omega^2}(\cos \omega x - \cos \omega m) \\ & + \tfrac{1}{\omega}(n + k_2 + x)(\sin \omega m - \sin \omega x) - \tfrac{1}{\omega}(m \sin \omega m - x \sin \omega x) - \tfrac{1}{\omega^2}(\cos \omega m - \cos \omega x) \\ = \quad & \tfrac{2}{\omega}(m + k_2) \sin \omega m - \tfrac{2}{\omega^2} \cos \omega m + \tfrac{2}{\omega^2} \cos \omega x \end{aligned}$$

so we have an eigenfunction $\cos \omega x$ with $\lambda = \frac{2}{\omega^2}$ iff the following discretisation condition holds:

$$\tfrac{2}{\omega}(m + k_2) \sin \omega m - \tfrac{2}{\omega^2} \cos \omega m = 0$$

i.e.

$$\tan \omega m = \frac{1}{\omega m \left(\frac{k_2}{m} + 1 \right)}$$

The solutions to this transcendental discretisation condition are displayed graphically in figure 15.

Similarly for $w(x) = \sin \omega x$, the discretisation condition is found to be:

$$\cos \omega m = 0 \Longrightarrow \omega m = (N + 1/2)\pi$$

Thus the antisymmetric eigenfunctions and eigenvalues are independent of k_2, as stated in section G.1.

Notice that the discretisation condition for the cos solutions has a critical point at $k_2 = -m$, at which the cos and sin eigenfunctions are pairwise degenerate. Beyond the critical point, a single cosh eigenfunction $w(x) = \cosh \omega x$ emerges with $\lambda = -\frac{2}{\omega^2}$ where:

$$-\tanh \omega m = \frac{1}{\omega m \, (k_2/m + 1)}$$

H.3 Limiting properties

By considering figure 15, it can be seen that as $k_2 \to \pm\infty$ the frequencies of the cos solutions all tend to a multiple of π/m. If we use the label a to denote the number of zero–crossings in the function, then we have $\omega_a \to a\pi/2m$. The limiting eigenvalues of the eigenfunctions are $\lambda_a = 2/\omega_a^2 = 2n^2/a^2\pi^2$, except for the case $a = 0$, for which $\omega_0^2 = 1/mk_2$ to leading order, giving $\lambda_0 \to k_2 n$.

H.4 Components of fixed point

In the limits $k_2 \to \pm\infty$, the DC component n_a of the ath normalised eigenfunction $w^{(a)}(x)$ is (to leading order):

$$
\begin{aligned}
n_a &= \int_{-m}^{m} w^{(a)}(y) \cdot \frac{1}{\sqrt{n}} dy \\
&= \int_{-m}^{m} \frac{\cos \omega_a y}{\sqrt{m}\sqrt{n}} dy \quad \text{for } a > 0 \text{ and even} \\
&= \frac{\sqrt{2}}{m\omega_a} \sin m\omega_a
\end{aligned}
$$

The argument $m\omega_a \to \pi a/2$, since

$$\tan \omega m = \frac{1}{\omega m\,(k_2/m + 1)} \to 0$$

So we can use $\sin m\omega_a \simeq (-1)^{a/2} \tan m\omega_a \simeq \frac{(-1)^{a/2}}{\omega_a k_2}$ in the above expression for n_a:

$$
\begin{aligned}
n_a &= \frac{\sqrt{2}}{m\omega_a} \frac{(-1)^{a/2}}{\omega_a k_2} \\
&= \frac{(-1)^{a/2} \lambda_a}{\sqrt{2} m k_2} \quad \text{for } a > 0 \text{ and even}
\end{aligned}
$$

where we have substituted $\lambda_a = 2/\omega_a^2$.

We can obtain the components of the fixed point in the eigenvector basis from this:

$$
\begin{aligned}
w_a^{FP} &= -\frac{k_1\sqrt{n}}{\lambda_a} n_a \\
&= -\frac{k_1}{k_2} \frac{(-1)^{a/2}}{\sqrt{m}} \quad \text{for } a > 0 \text{ and even, and } k_2 \to \pm\infty
\end{aligned}
$$

So all these components have equal magnitude, to leading order, in the limit $k_2 \to \pm\infty$. w_0^{FP} differs by a factor of $\sqrt{2}$: $w_0^{FP} = -\frac{k_1}{k_2}\frac{1}{\sqrt{2m}}$. Carrying the integrals and normalisation to $O(k_2^{-2})$, we obtain:

$$w_0^{FP} = -k_1/\sqrt{2m}(k_2 + \tfrac{2}{3}m)$$

And the location of the constraint surface is, for $k_2 \to -\infty$:

$$\sum_i w_i = \sqrt{n} w_0^{\mathrm{FP}} = k_1/|k_2 + \tfrac{2}{3}m|$$

Acknowledgements

D.J.C.M. is supported by a Caltech Fellowship and a Studentship from SERC, UK.

K.D.M. thanks M.P. Stryker for encouragement and financial support while this work was undertaken. K.D.M. was supported by an N.E.I. Fellowship and and by a Human Frontiers Science Program Grant to M.P. Stryker (T. Tsumoto, Coordinator).

This collaboration would have been impossible without the internet/NSFnet, long may their daemons flourish.

References

[1] F.V. Atkinson (1964). *Discrete and continuous boundary problems,* Academic Press.

[2] R.N. Bracewell (1965). *The Fourier transform and its applications,* McGraw–Hill.

[3] R. Linsker (1986). From basic network principles to neural architecture (series), *Proc. Nat. Acad. Sci. USA* **83**, 7508–7512, 8390–8394, 8779–8783.

[4] R. Linsker (1988). Self–organization in a perceptual network, *Computer* **21**(3), 105–117.

[5] R. Linsker (1988). Development of feature-analyzing cells and their columnar organization in a layered self-adaptive network. In *Computer Simulation in Brain Science*, R.M.J. Cotterill, Ed., Cambridge University Press, 416–431.

[6] D.J.C. MacKay and K.D. Miller (1990). Analysis of Linsker's simulations of Hebbian rules, *Neural Computation* **2**, 169–182.

[7] D.J.C. MacKay and K.D. Miller (1990). Analysis of Linsker's simulations of Hebbian rules, in *Advances in Neural Information Processing Systems II*, D. Touretzky, Ed. (Morgan Kaufman, San Mateo CA), 694–701.

[8] E. Merzbacher (1970). *Quantum mechanics,* 2nd ed., Wiley.

[9] K.D. Miller (1989). Orientation–selective cells can emerge from a Hebbian mechanism through interactions between ON– and OFF–center inputs, *Soc. Neurosc. Abst.* **15,** 794.

[10] K.D. Miller (1990). Correlation-based mechanisms of neural development, in *Neuroscience and Connectionist Theory*, M.A. Gluck and D.E. Rumelhart, Eds., Lawrence Erlbaum Associates, Hillsboro NJ, 267–353.

[11] K.D. Miller (1990). Derivation of Linear Hebbian Equations from a Nonlinear Hebbian Model of Synaptic Plasticity. To appear in *Neural Computation* **2**(3).

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

[13] F. Oberhettinger (1972). *Tables of Bessel transforms,* Springer.

[14] E. Oja. A Simplified Neuron Model as a Principal Component Analyzer, *J. Math Biology*, **15**, 1982, pp. 267–273.

[15] E. Seneta (1973). *Non–negative matrices,* Wiley.

[16] D.S. Tang (1989). Information–theoretic solutions to early visual information processing: Analytic results. *Phys. Rev. A,* **40**(11), 6626–6635.