J Biol Phys (2007) 33:213–246
DOI 10.1007/s10867-008-9056-5

ORIGINAL PAPER

Going Beyond a Mean-field Model for the Learning Cortex: Second-Order Statistics

M. T. Wilson · Moira L. Steyn-Ross · D. A. Steyn-Ross · J. W. Sleigh

Received: 30 August 2007 / Accepted: 21 January 2008 /
Published online: 18 March 2008

Abstract Mean-field models of the cortex have been used successfully to interpret the origin of features on the electroencephalogram under situations such as sleep, anesthesia, and seizures. In a mean-field scheme, dynamic changes in synaptic weights can be considered through fluctuation-based Hebbian learning rules. However, because such implementations deal with population-averaged properties, they are not well suited to memory and learning applications where individual synaptic weights can be important. We demonstrate that, through an extended system of equations, the mean-field models can be developed further to look at higher-order statistics, in particular, the distribution of synaptic weights within a cortical column. This allows us to make some general conclusions on memory through a mean-field scheme. Specifically, we expect large changes in the standard deviation of the distribution of synaptic weights when fluctuation in the mean soma potentials are large, such as during the transitions between the "up" and "down" states of slow-wave sleep. Moreover, a cortex that has low structure in its neuronal connections is most likely to decrease its standard deviation in the weights of excitatory to excitatory synapses, relative to the square of the mean, whereas a cortex with strongly patterned connections is most likely to increase this measure. This suggests that fluctuations are used to condense the coding of strong (presumably useful) memories into fewer, but dynamic, neuron connections, while at the same time removing weaker (less useful) memories.

Keywords Mean-field · Cortex · Memory · Learning · Modelling

M. T. Wilson (✉) · M. L. Steyn-Ross · D. A. Steyn-Ross
Department of Engineering, University of Waikato, Private Bag 3105,
Hamilton 3240, New Zealand
e-mail: m.wilson@waikato.ac.nz

J. W. Sleigh
Waikato Clinical School, Waikato Hospital, University of Auckland,
Hamilton 3204, New Zealand

1 Introduction

For many years, mean-field models have been used to describe the behavior of populations of neurons, particularly in the cortex and thalamus. Their history includes key contributions from Wilson and Cowan [1], Nunez [2], and Freeman [3], and has been developed more recently, for example, by Wright and Liley [4], Robinson et al. [5], Liley et al. [6], Rennie et al. [7], and Steyn-Ross et al. [8]. The dynamics of such models, including emergent structures and temporal instabilities, have been studied in detail in one dimension by Hutt et al. [9] and Kramer et al. [10]. In addition to the models above, which use the mean soma potential as the state-variable, formulations have been demonstrated based on the refractory period [11]. The motivation for this is the hypothesis that the time elapsed since the last action potential is a better predictor of the probability of firing of a particular neuron than its soma potential.

The major advantage of mean-field models over lower-level (neuron-by-neuron) models, such as Bazhenov et al. [12], Compte et al. [13], and Hill and Tononi [14], is their computational and mathematical simplicity. Additionally, mean-field models are generally more useful for understanding the electroencephalogram (EEG) because the EEG from a single scalp electrode is the result of the behavior of a large population of neurons. These models have proved capable of providing insight on the origins of many of the spectral components of the EEG, e.g., γ activity [7], β, α and spindle oscillations [15, 16], and k complexes [17]. Formulations such as the Bojak and Liley [18] and Wilson et al. [19] models the spatial behavior of the cortex through a two-dimensional grid. In this way, the mean soma potentials at different locations in the cortex are allowed to be instantaneously different. However, assuming spatial homogeneity, the statistics of the time-evolution at different grid-points will be the same.

An understanding of the effects of synaptic plasticity is probably crucial for studies of memory and, particularly, learning. A large number of learning schemes have been studied in cortical models, such as Bienenstock et al. [20], Bienenstock and Lehmann [21], Sandberg et al. [22], and Mongillo et al. [23]. Many are in the spirit of Hebb's theory [24], namely, that synaptic weight tends to grow between neurons whose firing history is strongly correlated. In two recent papers, we have looked at the consequences of Hebb's principle in a mean-field scheme. In the first, Steyn-Ross et al. showed that correlations between excitatory and inhibitory neuron pools increase as the cortical system approaches a saddle-node bifurcation (corresponding to the transition between "up" and "down" states in slow-wave sleep) [25]. This correlation would lead to an increase in excitatory to inhibitory synaptic weights (through Hebb's theory) that the authors associate with a suppression of reverberations in the network. Physically, this could correspond to a weakening of unwanted modes of response to stimuli. In the second paper, Wilson et al. [17] showed that a Hebb-like change in synaptic weights will naturally bring the system towards such a critical point (corresponding to a down-to-up transition) where these correlated fluctuations can take place, and we wish to focus on this situation in this paper. However, use of a Hebbian learning rule within a mean-field scheme is limited because the scheme by definition deals with population averages, and insight into what happens to the synaptic population, and which "memories" are erased, is limited. Our desire to understand better what is happening when population correlations are large leads us to the second-order scheme described in this paper.

Recently, Stetter has modelled pools of neurons within a mean-field scheme to perform a study of cognitive flexibility in the necortex [26]. In this approach, the global symmetry of the mean-field model is broken by allowing different activation currents to enter different pools of neurons, so that "memories" can be represented in the model. We take a different approach in this paper by breaking the symmetry through the synaptic weights—specifically by considering higher-order statistics (i.e., variances of synaptic weights) over the cortex, but still within a mean-field framework. Our intention is to allow some understanding of memory and learning to be developed within a modelling scheme that complements neural-network approaches without the need to lose the advantages of speed and versatility offered by a mean-field approach.

In this paper, we make a step towards this goal by developing a mean-field model in the style of Liley et al. [6] and Steyn-Ross et al. [27] that includes second-order statistics for the synaptic weights. Unfortunately, to close the set of equations, some approximation is required. While we leave the major part of the mathematics to the Appendix, we will outline the methodology in Section 2 below. We then demonstrate through analysis of the equations and through simulations that the behavior of this statistic depends upon the nature of the pre- and postsynaptic neurons (i.e., whether they are excitatory or inhibitory) and the interconnectedness of neurons within the cortex. We compare the mean-field results with predictions from simple neural-network models. Finally, we discuss the implications of our results with particular reference to human sleep. We comment on a recent paper by Tononi and Cirelli in which a purpose of sleep is described in terms of synaptic downscaling [28]. We emphasize that the theory developed here is not complete. In this paper we concentrate on the growth in synaptic weight on the approach to a down-to-up transition in slow-wave sleep, and we intentionally refrain from a detailed explanation of the synaptic downscaling mechanisms or the dynamics of the transition itself. We also do not attempt to model explicitly the slow waves of sleep.

2 Method

We will start with formulating equations at a low-level and then take averages of soma potentials and synaptic variables over spatial regions of the cortex. Whereas in previous work we have taken averages over "first-order" statistics only [17], we now extend the analysis to consider averages over the square of the synaptic weight connections (second-order statistics) and develop a set of self-consistent equations for these.

The basis of mean-field approaches such as Liley et al. [6] and Rennie et al. [7] is the modelling of populations of neurons as opposed to individual neurons. This is a natural approach if the model output is to be compared with EEG because EEG electrodes sample the electric field produced by many thousands of neurons. In this approach, we use variables such as the mean soma (membrane) potentials, synaptic flux rates, etc., averaged over spatial regions of the cortex. We can conveniently use cortical columns for these averages (a cortical column being a collection of neurons arranged in a tubular structure perpendicular to the cortex surface [29]); however, the method does not implicitly rely on a cortex being ordered in this form. We consider a spatially homogeneous cortex—i.e., one where parameters such as the number of connections per neuron, axonal propagation speed, etc., are assumed constant over the cortex. Although not anatomically accurate, it will not affect

the qualitative results and conclusions and greatly simplifies the work. The model we use is presented in the Appendix. Note that the homogeneous cortex does not imply that the variables do not depend on space; for example, mean firing rates vary instantaneously with space, but their long-term time-averages will be constant at all points in space. Although we discuss column-averaged properties, it is important to note that we do not simulate individual columns themselves but consider the averages in a "typical" column in the vicinity of a particular point in space.

We assume that all the neurons in a cortical column are connected to each other neuron within the column with a weight given by w_{jk}^{PQ}. Here, the suffix jk refers to the connection between the j-th presynaptic neuron and the k-th postsynaptic neuron. The superscript PQ tells us that the j-th neuron is of type P [where P is either excitatory (e) or inhibitory (i)] and that the k-th neuron is of type Q (where Q is either e or i). If two neurons are not connected, we can assign $w_{jk}^{\mathrm{PQ}} = 0$. Also, neurons in the cortical column may be connected to others in different columns. In what follows, we will use positive weights for all connections; the sign difference between the effects of the excitatory and inhibitory neurons will be accounted for explicitly through other quantities carrying "+" and "−" signs.

The major state variable of many mean-field models is the "averaged" soma potential of neuron populations. We use this form of mean-field model in this paper, and therefore specifically write V_j^{P} meaning the soma-potential of the j-th neuron (which is of type P). The definition of "average" is important here. In this work, we use a bar [e.g., $\overline{X_j(t)}$] initially to denote the time-average of a quantity (in this case, the arbitrary quantity X) over some time period. Later, we will use it to denote an average over spatially separated regions of the cortex assuming that this measure is the same. The quantity X is, in general, a function of time (t) and neuron (j). An alternative "average" is that over the neurons within a cortical column. We use angle brackets (e.g., $\langle X_j(t) \rangle_j$) to denote the average over neurons j within a column.

To avoid confusion, we will, in most cases, use a subscript on the final angled bracket $\langle \cdots \rangle_j$ to be specific about which neurons we average over.

2.1 Hebbian Learning

There has been considerable development of "learning rules" in low-level modelling studies of cortical networks. Hebb's principle of "fire-together, wire-together" [24] has been modified to account for physical effects such as the balance between long-term potentiation and long-term depression [23]. Nonetheless, Hebb's idea forms the basis of synaptic weight modification in cortical networks and benefits from being both mathematically simple and giving rich behavior. For this reason, we will pursue a Hebbian-based approach in this work, but we acknowledge that modifications are possible and probably appropriate.

The growth of a synaptic weight w_{jk}^{PQ} can be summarized through the correlation between the soma potentials of the presynaptic neuron j and the postsynaptic neuron k [21, 25, 30]:

$$\frac{d}{dt} w_{jk}^{\mathrm{PQ}} = \eta^{\mathrm{PQ}} \left(\overline{V_j^{\mathrm{P}} V_k^{\mathrm{Q}}} - \overline{V_j^{\mathrm{P}}}\, \overline{V_k^{\mathrm{Q}}} \right) \tag{1}$$

where the term η^{PQ} is a constant. Two neurons, j and k, whose soma potentials fluctuate together ("fire-together") will have a high correlation and, therefore, a high rate of change of weight.

2.2 Mean-field Modelling and Ergodicity

Hebbian-learning in the form of (1) relates to individual neurons, not averages over neurons within a cortical column. We can transform to a mean-field form by taking column averages:

$$\left\langle \frac{d}{dt} w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} = \eta^{\mathrm{PQ}} \left\langle \left(\overline{V_j^{\mathrm{P}} V_k^{\mathrm{Q}}} - \overline{V_j^{\mathrm{P}}}\,\overline{V_k^{\mathrm{Q}}} \right) \right\rangle_{jk} \tag{2}$$

where we are averaging over the *pre-* and *post*synaptic neurons independently. For this reason, we can separate the j and k averages to leave:

$$\frac{d}{dt} \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} = \eta^{\mathrm{PQ}} \left(\overline{\left\langle V_j^{\mathrm{P}} \right\rangle_j \left\langle V_k^{\mathrm{Q}} \right\rangle_k} - \overline{\left\langle V_j^{\mathrm{P}} \right\rangle_j}\,\overline{\left\langle V_k^{\mathrm{Q}} \right\rangle_k} \right). \tag{3}$$

Now the quantities are written in terms of column- and time-averages. Equation (3), in fact, contains four equations corresponding to the different P and Q combinations, i.e., P = e, Q = e; P = e, Q = i; P = i, Q = e; and P = i, Q = i. Note that the order of averaging does not matter. In reference [17], we assumed ergodicity to make sense of the time-average in a mean-field model in two spatial dimensions ($\vec{r}$) of the cortex under natural sleep. Under this assumption, the time-average of a quantity is equal to an average over space. Specifically, the time-average denoted by the bars in (3) is assumed to be equal to an average over grid-points in a simulation of the Liley-based cortical equations over two-dimensional space ($\vec{r}$) (i.e., an average over many cortical columns). The cortex must be in a state where it is fluctuating about a stable equilibrium for this argument to apply—the phase-space sampled by a given point ($\vec{r}$) over time t being equivalent to the phase-space sampled at a given moment in time by the ensemble of grid-points ($\vec{r}$). The advantage of making this assumption is that we do not need to keep track of a region's time-history while we perform the simulations. A disadvantage is that the equation set must include "space"; a set of equations with no explicit description of space cannot be used in this way. In what follows, we will implicitly use this hypothesis and use the "bar" to denote an average over a simulation grid. If the grid is large enough, and no spatial inhomogeneities are modelled, we would expect this hypothesis to be reasonable.

2.3 Outline of the Model

We wish to understand how the mean-square of the synaptic weight w_{jk}^{PQ} changes with time. We define the variance of the synaptic weights as:

$$\sigma_{\mathrm{PQ}}^2 = \left\langle w_{jk}^{\mathrm{PQ}} w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} - \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}^2 \tag{4}$$

where the average is taken over all presynaptic neurons j (of type P) and all postsynaptic neurons k (of type Q). We then develop a set of equations for describing $d\sigma_{\mathrm{PQ}}^2/dt$. The full mathematics is presented in the Appendix. In outline, the method follows the steps below.

First, we differentiate (4) with respect to time. This gives terms involving dw_{jk}^{PQ}/dt for which we can substitute from (1). The resultant equation contains averages over a product of three terms, w_{jk}^{PQ}, V_j^{P}, and V_k^{Q}. To reduce to a "second-order" method, we assume that V_j^{P} and V_k^{Q} have small deviations about their *instantaneous* equilibrium values. We are left

with an equation containing the column-averaged pre- and postsynaptic potentials $\left\langle V_j^{\mathrm{P}} \right\rangle_j$ and $\left\langle V_k^{\mathrm{Q}} \right\rangle_k$, and also second order terms, which we denote in this paper by:

$$\xi_{\mathrm{post}}^{\mathrm{PQ}} = \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j V_k^{\mathrm{Q}} \right\rangle_k , \tag{5}$$

$$\xi_{\mathrm{pre}}^{\mathrm{PQ}} = \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_k V_j^{\mathrm{P}} \right\rangle_j . \tag{6}$$

These terms provide information about the correlations between the weights w_{jk}^{PQ} and the neuron's potentials. A pictorial representation is given in Fig. 1. Note that we will usually assume that all connections are reversible ($w_{jk}^{\mathrm{PQ}} = w_{kj}^{\mathrm{QP}}$), so that $\xi_{\mathrm{post}}^{\mathrm{PQ}} = \xi_{\mathrm{pre}}^{\mathrm{QP}}$. The learning rule, as used in (2), is consistent in that it implies all reversible connections $j$to k will remain reversible at future times if $\eta^{\mathrm{PQ}} = \eta^{\mathrm{QP}}$.

The time-variation of the column-average potential $\left\langle V_k^{\mathrm{Q}} \right\rangle_k$ has been the subject of many papers, as referenced in the introduction. In our work, we draw on the equations developed by Liley et al. [6] and used by Steyn-Ross et al. [27] and Wilson et al. [19] for the sleeping cortex, although we remark that other self-consistent descriptions are valid here. The exact equations are described in the Appendix; they are a set of coupled differential equations for various variables, such as average excitatory and inhibitory soma potential and synaptic flux rates, and we shall refer to these variables as the Liley-style variables.

To close the set of equations (i.e., to have a set of first-order differential equations in time that describe how $\left\langle V_k^{\mathrm{Q}} \right\rangle_k$, σ_{PQ}^2 etc., vary in time), we need to know how $\xi_{\mathrm{post}}^{\mathrm{PQ}}$ changes with time. Differentiating (5), we can substitute for dw_{jk}^{PQ}/dt from (1) and V_k^{Q} (from the Liley-style equations). We are left with a term of the form:

$$F^{\mathrm{PQR}} = \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j \left\langle w_{lk}^{\mathrm{RQ}} \right\rangle_l \right\rangle_k = \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j \left\langle w_{kl}^{\mathrm{QR}} \right\rangle_l \right\rangle_k = \left\langle w_{jk}^{\mathrm{PQ}} w_{kl}^{\mathrm{QR}} \right\rangle_{jkl} . \tag{7}$$

This is represented pictorially in Fig. 1c. Unfortunately, there are three means in this expression, i.e., the size of the variable set has grown. To give a manageable (bounded) equation set, we need to truncate this set, and do this by estimating F^{PQR} through terms involving just two means.

When P $\neq$ R, we assume the PQ and RQ weights to be uncorrelated, so F^{PQR} splits into $\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} \left\langle w_{lk}^{\mathrm{RQ}} \right\rangle_{lk}$, which is simply a product of column-averaged weights and is in the form we require.

2.4 Network Structure

When P = R, F^{PQR} has the form $\left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j^2 \right\rangle_k$. To close the equation set, we wish to relate this to the variance σ_{PQ}^2, as defined in (4). To do this, we need a relationship between the term $\left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j^2 \right\rangle_k$ and $\left\langle w_{jk}^{\mathrm{PQ}} w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}$. It is clear that, in the case where weights are completely independent (e.g., Fig. 2a), we can use the central limit theorem to relate these two quantities; i.e., the variance in $\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j$ over postsynaptic neurons k is equal

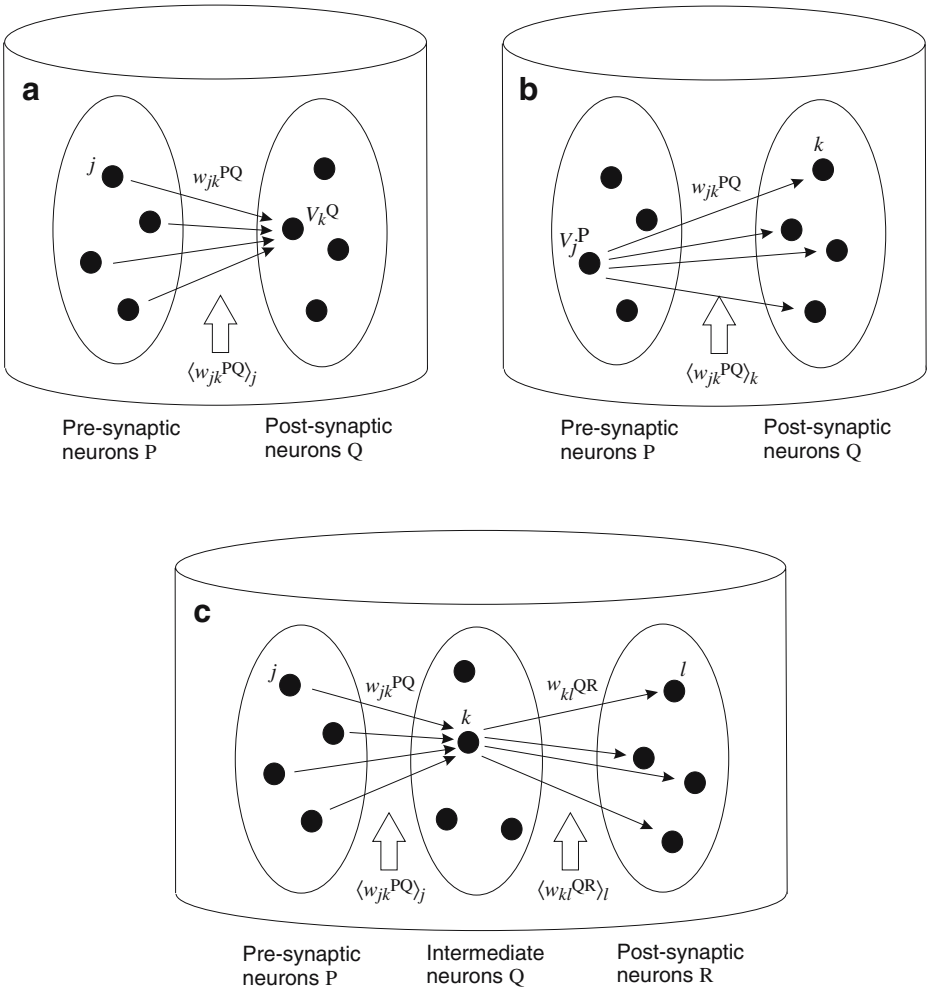

Fig. 1 A pictorial representation of the terms $\xi_{\text{post}}^{\text{PQ}}$ (**a**), $\xi_{\text{pre}}^{\text{PQ}}$ (**b**), and F^{PQR} (**c**). In each case, the *ellipses* denote pools of neurons (pre- or postsynaptic; in the case of **c**, there is also an intermediate pool). The averages denoted by the *block arrows* are carried out first, followed by averages over the remaining neuron pool

to the variance in w_{jk}^{PQ} over all neurons j and k, divided by the number of type-P to type-Q connections. Generally, however, the weights are not independent. In this case, we speculatively introduce a parameter s_{PQ} to describe the *effective* number of *independent* type-P to type-Q connections and use the central limit theorem. We define s_{PQ} through the equation:

$$s_{\text{PQ}} = \frac{N_{\text{PQtotal}} - N_{\text{PQ}}}{N_{\text{PQtotal}}} \tag{8}$$

where N_{PQtotal} is the total number of type-P to type-Q connections each postsynaptic neuron has within a cortical column and N_{PQ} is the total number of *independent* connections. This

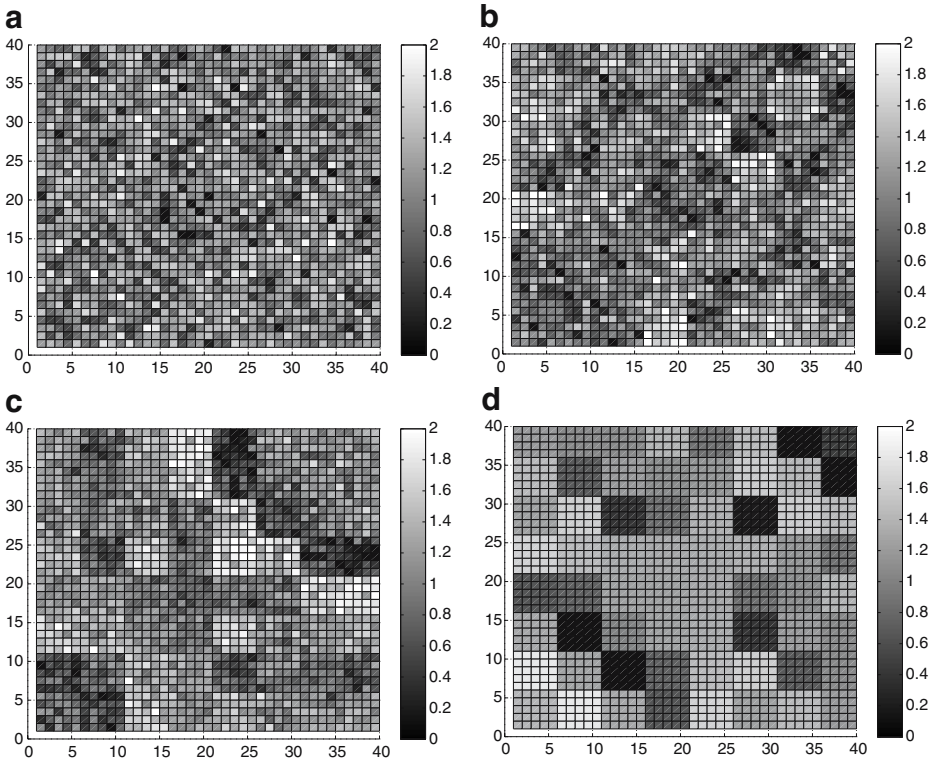

Fig. 2 A representation of the role of the parameter s for describing the structure of the weights network. All four parts (**a**–**d**) show example weight matrices for a neural-network consisting of 40 neurons. The x axis denotes the presynaptic neuron, the y axis the postsynaptic neuron. The strength of the weight between the two is denoted by a grayscale value from 0 to 2. All four parts have the same *mean* weight (namely, 1.0) and the same variance in the weights (namely, 0.4). However, whereas **a** has no "structure' (knowledge of one weight tells us nothing about any other weight), **d** has considerable structure. The s-parameter describes this trend; $s = 0.0$, 0.26, 0.57, and 0.75 for **a**, **b**, **c**, and **d**, respectively

latter term can be defined more formally in (58). Physically, s_{PQ} represents the degree of "structure" that is present in the synaptic weights. For a hypothetical cortical structure where synaptic connections are mostly independent of each other (presumably one that contains no useful memories), $s_{\mathrm{PQ}} \approx 0$. For a cortex whose synapses are strongly correlated with each other, $s_{\mathrm{PQ}} \approx 1$. For a physiologically realistic cortex, we would expect s_{PQ} to lie between these extremes. Figure 2 explains this in a diagrammatic form. However, note that, in the simulations of the mean-field model, we do not model *individual* synaptic weights. Instead, in our simulations, we *specify* a value of s (from 0 to 1) and find that this qualitatively influences the results.

Interconnections of systems have been the subject of recent discussion. The biological reviews of Douglas and Martin [31] and Thomson and Bannister [32] describe patterns of cortical connectivity. Moreover, Tononi and Sporns use the idea of a "complexity" to describe elements of a system that can integrate information among themselves [33], and Albert and Barabási present a comprehensive review of network descriptions [34]. We expect the parameter s_{PQ} to be related to such measures but we do not discuss this further.

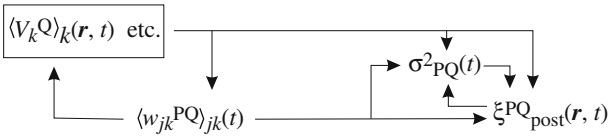

Fig. 3 The interdependence of the state variables in the differential equations of the model. Note that the *box* around the $\left\langle V_k^{\mathrm{Q}} \right\rangle_k$ implies that there is the complete set of Liley-style equations to include here. A quantity that is modelled over space through a series of grid-points contains the symbol $\boldsymbol{r}$; the symbol t denotes that all these quantities depend on time. The left-hand pair of quantities is not influenced by the right-hand pair

In summary, we have a set of self-consistent first-order differential equations for column-averaged quantities. Our state variables describing the system are (1) the Liley-style variables (see Appendix), which describe the time-evolution of the column-averaged membrane potentials, of which $\left\langle V_k^{\mathrm{Q}} \right\rangle_k$ is key; (2) the synaptic weights $\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}$ (four of these for P, Q = e and i); (3) the variances σ_{PQ}^2 (four of these); and (4) the second-order weight-potential combination $\xi_{\mathrm{post}}^{\mathrm{PQ}}$ (four of these).

The equations of the model are described by (1) the Liley-style equations given in the Appendix, (24–31); (2) for $d\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}/dt$, (3); (3) for $d\,\sigma_{\mathrm{PQ}}^2/dt$, (47) of the Appendix; and (4) for $\mathrm{d}\xi_{\mathrm{post}}^{\mathrm{PQ}}/\mathrm{dt}$, (55) of the Appendix. The dependencies of the equation set are shown in Fig. 3.

2.5 Equilibrium Conditions

It is easy to show (see Appendix) that an equilibrium for σ_{PQ}^2 exists for the case of:

$$\sigma_{\mathrm{PQ}}^2 = 0, \tag{9}$$

$$\xi_{\mathrm{post}}^{\mathrm{PQ}} = \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} \left\langle V_k^{\mathrm{Q}} \right\rangle_k . \tag{10}$$

Physically, this equilibrium condition corresponds to a cortex where all synaptic weights are identical (no variance in the weights, so that each weight is equal to its column average.) This is encouraging because it implies that previous modelling (e.g., Steyn-Ross et al. [27] and Wilson et al. [19]), which have not included variances in weights, can be considered as being performed at an equilibrium state for the variance. However, the stability of this equilibrium is hard to assess analytically. In the Appendix, we present an argument as to why we would expect the equilibrium to be *unstable*.

3 Simulations

We now present results from simulations of this extended mean-field model, namely, the Liley-style equations of the Appendix and (3), (47), and (55). The various parameters are presented in Table 1. For convenience, we simulate the cortex as a 50 × 50-cm, two-dimensional sheet, with the length scale chosen so that its area broadly matches that known anatomically (about 2,600 cm^2 [29]). We choose a 16 × 16 spatial grid to model the spatial

Table 1 The standard parameters used throughout this paper, except where stated otherwise

Parameter	Description	Standard value
$\tau_{e,i}$	Membrane time constants	0.04, 0.04 s^{-1}
$Q_{\max}^{e,i}$	Maximum firing rates	30, 60 s^{-1}
$\theta_{e,i}$	Sigmoid thresholds	−58.5, −58.5 mV
$\sigma_{e,i}$	Standard deviation for threshold	4.0, 6.0 mV
$\rho_{e,i}$	Gain per synapse at resting voltage	0.001, −0.00105 mV · s
$V_{\text{rev}}^{e,i}$	Reversal potentials at synapse	0, −70 mV
$V_{\text{rest}}^{e,i}$	Cell resting potential	−64, −64 mV
N_{ea}^{α}	Long-range e to e or i connectivity	3710
N_{ea}^{β}	Short-range e to e or i connectivity	410
N_{ia}^{β}	Short-range i to e or i connectivity	800
$\langle\phi_{ea}^{\text{sc}}\rangle$	Mean e to e or i subcortical flux	750 s^{-1}
$\langle\phi_{ia}^{\text{sc}}\rangle$	Mean i to e or i subcortical flux	1500 s^{-1}
γ_{ea}	Excitatory synaptic rate constant	300 s^{-1}
γ_{ia}	Inhibitory synaptic rate constant	65 s^{-1}
$L_{x,y}$	Spatial length of cortex in model	500 mm
a_{mc}	Area of cortical column	1 mm^2
Λ_{ea}	Characteristic inverse length-scale for connections	0.2 mm^{-1}
v	Mean axonal conduction speed	1400 mm s^{-1}

In this table, the superscript or suffix a can take on the labels e and i. The values are taken mostly from the paper of Rennie et al. [7]. Although there is considerable uncertainty in these parameters, they form a plausible set that is sufficient for the purposes of elucidating much of the physics of the cortical model. It is quite possible that further physical effects can be produced by varying these parameters sufficiently

variation of the variables as a balance between fidelity and simulation speed. Periodic boundary conditions are used because it is not clear how to "terminate" the cortex in such a model; however, a large length (50 cm) reduces the impact of this uncertainty.

A second-order predictor-corrector method [35] is used for the stochastic time integration with a time step of 0.2 ms. The rate of the learning parameter, η^{PQ}, has been set to 1×10^2 mV^{-2} s^{-1} for all PQ. Because correlations between neighboring cortical columns are small (except close to a critical point), this is a fairly small learning rate. This parameter describes the dynamics of the change in the weights—it does not influence, for example, the slow waves, which are not explicitly modelled in this paper. The system is driven by white noise, as described in the Appendix. Although white-noise driving is clearly not experienced by a human cortex, even during sleep, we use it because it is relatively easy to implement numerically and it is adequate for bringing out the qualitative behavior in our model.

In our implementation, it is important to realize that, in this model, we do not consider individual synaptic connections explicitly, but describe their nature using the state-variables' mean weight $\left\langle w_{jk}^{\text{PQ}} \right\rangle_{jk}$ (four of these for the four P, Q = e, i combinations), variance in the weight σ_{PQ}^2 (four of these), and structure in the weight s_{PQ} (in principle, four of these).

We focus our simulations away from the equilibrium point. This is where the behavior is most interesting and relevant because a real cortex clearly does not have a zero variance in its synaptic weights. Also, at the equilibrium point, simulations become difficult. Numerically, negative variances are possible in this model, but physically, it is clear that this should not happen—as the variance approaches zero from above, the change in variance approaches zero. However, if the variance is set in the model to be negative, the equations allow it to fall further, exaggerating an unphysical situation.

We illustrate the case where we start with $\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} = 1$ and $\sigma_{\mathrm{PQ}}^2 = 0.5$ for all PQ combinations. To bring out the importance of the structure in the cortex, we contrast the case of a strongly linked cortex (low N_{PQ}, $s_{\mathrm{PQ}} \approx 1$) with the obviously unrealistic and hypothetical case of a nonlinked cortex (high $N_{\mathrm{PQ}}, s_{\mathrm{PQ}} \approx 0$). Results for the weights $\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}$ and variances σ_{PQ}^2 for both cases are shown in Fig. 4.

In part a, we present the mean excitatory soma potential, $\left\langle V_j^e \right\rangle_j$, as a function of time. The potential exhibits small fluctuations about an equilibrium that is increasing with time. The size of the fluctuation is dependent on the strength of the noise terms in the underlying equations. The dynamic nature of the equilibrium point is a direct result of changes in

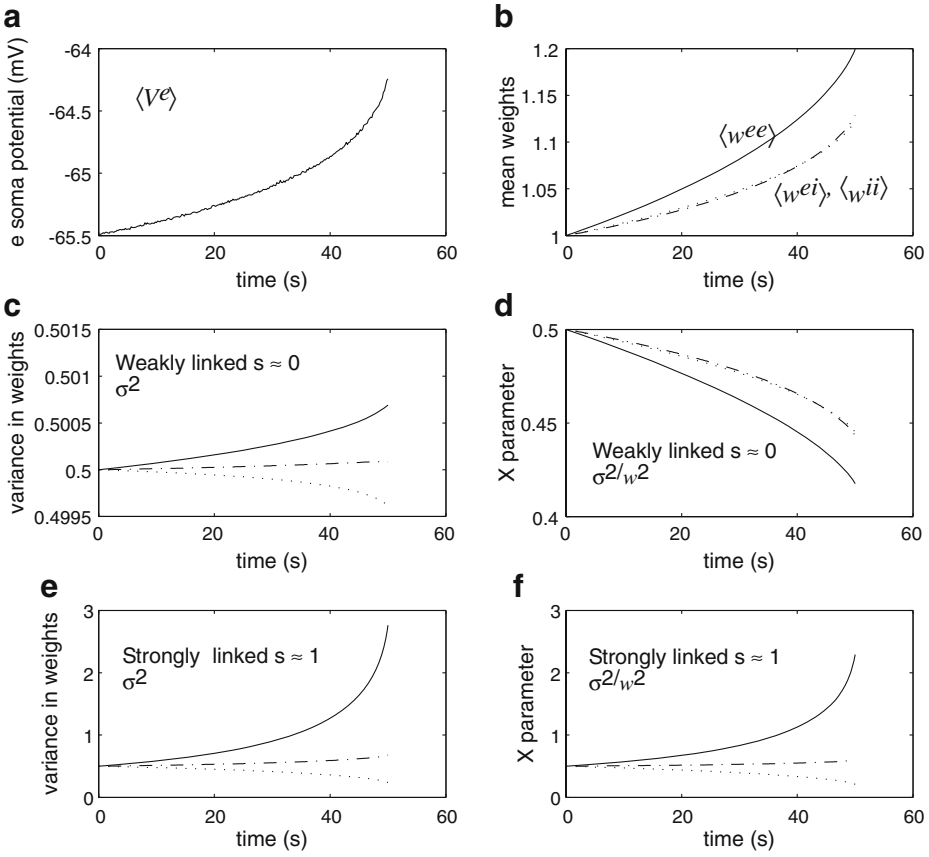

Fig. 4 Development of weights for a nonequilibrium situation. **a** The excitatory soma potential as a function of time. It exhibits small fluctuations (of order 0.01 mV) about a dynamically increasing equilibrium point. The change in equilibrium is due to the movement in the mean weights, as described in reference [17]. **b** The mean weights $\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}$ against time. **c** The variance in the weights σ_{PQ}^2 as a function of time, for a weakly linked cortex ($N_{\mathrm{PQ}} = 1000$, $s_{\mathrm{PQ}} \approx 0$). **d** X_{PQ}, i.e., the variance divided by the mean weight squared, against time, for $N_{\mathrm{PQ}} = 1000$, $s_{\mathrm{PQ}} \approx 0$. **e** and **f** Similar to **c** and **d** but for a highly linked cortex ($N_{\mathrm{PQ}} = 1$, $s_{\mathrm{PQ}} \approx 1$). Key for all plots: *ee*, *solid*; *ie* = *ei*, *dot-dash*; *ii*, *dotted*

the weights $\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}$, due to (3), which are plotted in part b. In reference [17], the authors describe how the mean weights change in such a manner as to bring the system towards a critical point. This is why the rate of change of the plotted values increases as we approach about 50 s in time; we comment further on this below.

In parts c and e, we see that variance in the *ee* terms grows, whereas the variance in the *ii* term falls, with the variance in the $ei = ie$ term remaining approximately constant. The rapid increase in gradient as the time approaches 50 s is due to the system organizing itself towards a critical point, as described in Wilson et al. [17]. It is clear that the changes occur much faster for the strongly linked cortex (e) compared to the weakly linked cortex (c). Indeed, for the weakly linked cortex, the variances have hardly changed.

What is perhaps more interesting than the change in σ_{PQ}^2 is the change in σ_{PQ}^2 in *relation* to the mean weights, because the mean weights also change. A relevant dimensionless measure would be $X_{\mathrm{PQ}} = \sigma_{\mathrm{PQ}}^2 / \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}^2$, which describes the *relative width* of the distribution of the weights in terms of the mean of the distribution. In effect, X is a "normalized" variance. A network with high X would have a high variation in "relative" weights. This is important because the straight learning rule (1) has no mechanism for preventing the weights from growing to infinity. A simple correction, consistent with the second-order scheme, is discussed below.

Plots of X_{PQ} against time are also shown in Fig. 4d and f for the different numbers of independent connections. In this case, there is a clear qualitative difference between the weakly linked (d) and strongly linked (f) cortex. In the case of the weakly linked cortex, the relative width of the weights distribution falls (that is, all the normalized weights become more similar). However, in the strongly linked cortex, the *ee* variance grows quickly enough so that the relative width of the distribution is increased.

3.1 The Approach to Bifurcation

Why do the quantities $\left\langle V_k^e \right\rangle_k$, $\left\langle w_{jk}^{ee} \right\rangle_{jk}$, etc., appear to diverge with time? In a previous paper, we have described how a Liley-style mean-field model with Hebb's principle self-organizes to a saddle-node bifurcation, where a tiny increase in excitation (or, more likely, a drop in inhibition [36]) would cause a discontinuous change transition to a higher-firing state [17]. In Fig. 4, just beyond the 50-s time frame of the plot, the system carries out such a discontinuous change (not shown because the numerics of the simulation break down at this point), and the quantities rapidly change on the approach to this jump. Moreover, fluctuations in quantities will increase on the approach to the transition, as the strength of the stable attractor diminishes. Therefore, in the direct vicinity of the jump, we would not expect our analysis to be valid because we have explicitly assumed small fluctuations in soma potential from its instantaneous equilibrium value (through Eq. 43) to reduce our equation set to second order.

The self-organization can be stabilized through, for example, breaking the symmetry between the $e \to Q$ and $i \to Q$ synaptic weights. As an example, Fig. 5 shows $\left\langle V_k^e \right\rangle_k$ for the same case as Fig. 4a but with $\eta^{i\mathrm{Q}}$ three times higher than $\eta^{e\mathrm{Q}}$. We see that the pattern of growth in the mean weights causes $\left\langle V_k^e \right\rangle_k$ to eventually stabilize. Note that, as $\left\langle V_k^e \right\rangle_k$ reaches its limit, its fluctuations are large, corresponding to the approach to the bifurcation. The stability is understood by realizing that, very close to the jump, the fluctuations in σ_{ei}^2 increase very rapidly as the excitatory and inhibitory neuron populations become

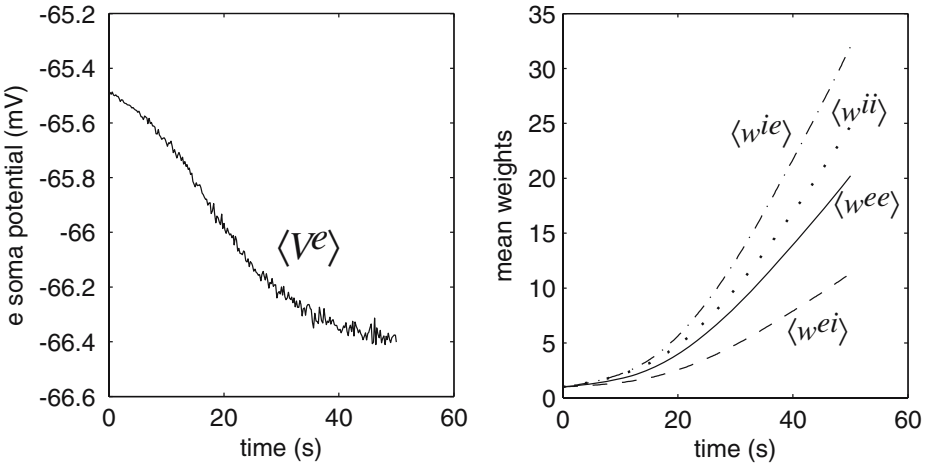

Fig. 5 *Left*: A plot of the mean excitatory soma potential $\langle V_k^e \rangle_k$ against time, at one spatial grid-point, for the case where $\eta^{iQ} = 3\eta^{eQ}$. The fluctuations in $\langle V_k^e \rangle_k$ increase as time increases, corresponding to the approach to a saddle-node bifurcation. However, the mean value stabilizes. *Right*: The mean weights as a function of time. Although the mean soma potential reaches an equilibrium in this scheme, the mean weights do not (see reference [17])

correlated [25]. This growth provides a rapid increase in the $\left\langle w_{jk}^{ei} \right\rangle_{jk}$ and $\left\langle w_{jk}^{ie} \right\rangle_{jk}$ terms, which stabilize the system.

3.2 Including Limitation on Weights

An alternative approach to stabilizing the equations would be to limit the growth of the synaptic weights. There are many approaches given in the literature (e.g., Blumenfeld et al. [37]), but they are not mostly tractable to our second-order analysis. One workable approach, however, is to include an exponential decay-like term on (2), in the manner of Bienenstock and Lehmann [21]. This has the additional advantage of ensuring that noncorrelated or anticorrelated neurons have a reduced synaptic weight between them. We remark, however, that the primary thrust of this paper is to look at the changes in synaptic weights on the approach to the down-to-up transitions and, therefore, we choose not to emphasize the precise neurophysiology behind the control of synaptic weights. An exponential term is simply a convenient way of ensuring that weights do not diverge.

We now write:

$$\frac{dw_{jk}^{\mathrm{PQ}}}{dt} = \eta^{\mathrm{PQ}} \left(\overline{V_j^{\mathrm{P}} V_k^{\mathrm{Q}}} - \overline{V_j^{\mathrm{P}}} \, \overline{V_k^{\mathrm{Q}}} \right) + \mu \left(k^{\mathrm{PQ}} - w_{jk}^{\mathrm{PQ}} \right) \tag{11}$$

where μ is a decay constant and k^{PQ} is a "resting" weight value, i.e., the equilibrium value in the absence of Hebbian learning. Following the previous steps of the method, we obtain the change in the mean weight over a cortical column as:

$$\left\langle \frac{d}{dt} w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} = \eta^{\mathrm{PQ}} \left(\overline{\left\langle V_j^{\mathrm{P}} \right\rangle_j \left\langle V_k^{\mathrm{Q}} \right\rangle_k} - \overline{\left\langle V_j^{\mathrm{P}} \right\rangle_j} \, \overline{\left\langle V_k^{\mathrm{Q}} \right\rangle_k} \right) + \mu \left(k^{\mathrm{PQ}} - w_{jk}^{\mathrm{PQ}} \right). \tag{12}$$

The variance follows from the equation:

$$\frac{d}{dt}\sigma^2_{\mathrm{PQ}} = \frac{d}{dt}\left\langle w^{\mathrm{PQ}}_{jk} w^{\mathrm{PQ}}_{jk}\right\rangle_{jk} - \frac{d}{dt}\left(\left\langle w^{\mathrm{PQ}}_{jk}\right\rangle^2_{jk}\right)$$
$$= 2\left\langle w^{\mathrm{PQ}}_{jk}\frac{d}{dt}w^{\mathrm{PQ}}_{jk}\right\rangle_{jk} - 2\left\langle w^{\mathrm{PQ}}_{jk}\right\rangle_{jk}\frac{d}{dt}\left\langle w^{\mathrm{PQ}}_{jk}\right\rangle_{jk}, \tag{13}$$

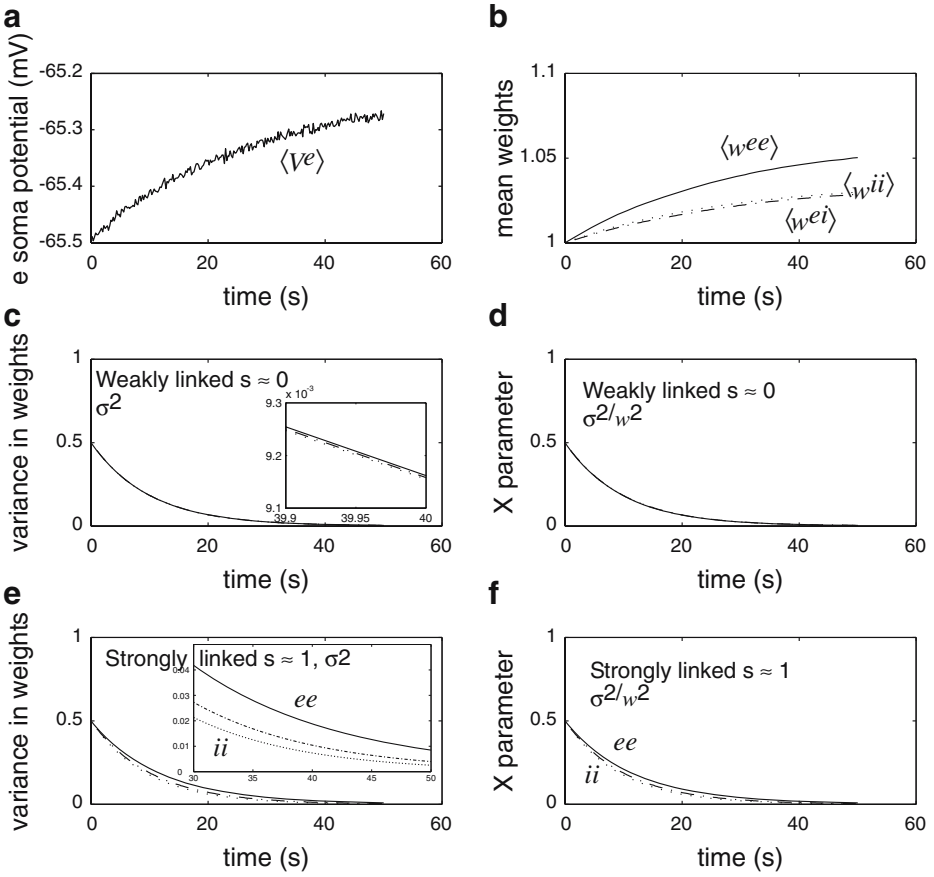

Fig. 6 Development of statistics when an exponential decay factor is included. **a** A plot of the excitatory soma potential $\left\langle V^e_j\right\rangle_j$ as a function of time. It follows an upward trend as the weights grow, but stabilizes as the weights settle. Note that it exhibits small fluctuations about the *instantaneous* equilibrium values. **b** A plot against time of the mean weights $\left\langle w^{\mathrm{PQ}}_{jk}\right\rangle_{jk}$. These settle at large times to an equilibrium value. **c** The variance in the weights σ^2_{PQ}as a function of time, for a weakly linked cortex ($N_{\mathrm{PQ}} = 1000$). The three lines *ee*, $ie = ei$, and *ii* are indistinguishable, except on the *insert* shown for large times. **d** X_{PQ}, i.e., the variance divided by the mean weight squared, against time, for $N_{\mathrm{PQ}} = 1000$. **e** and **f** Similar to **c** and **d** but for a highly linked cortex ($N_{\mathrm{PQ}} = 1$). The *insert* in **e** expands the traces for large times. Here, the lines *ee*, $ei = ie$, and *ii* can be distinguished, with the *ee* trace always having the largest variance; although, in this scheme, all variances eventually reach zero, the *ee* variance takes the longest to fall. Key for all plots: *ee*, *solid*; $ie = ei$, *dot-dash*; *ii*, *dotted*

which gives from (11) and (12)

$$\frac{d}{dt}\sigma^2_{\mathrm{PQ}} = 2\eta^{\mathrm{PQ}}\left[\left\langle w^{\mathrm{PQ}}_{jk}\left(\overline{V^{\mathrm{P}}_j V^{\mathrm{Q}}_k} - \overline{V^{\mathrm{P}}_j}\,\overline{V^{\mathrm{Q}}_k}\right)\right\rangle_{jk} - \left\langle w^{\mathrm{PQ}}_{jk}\right\rangle_{jk}\left\langle \overline{V^{\mathrm{P}}_j V^{\mathrm{Q}}_k}\right\rangle_{jk} + \left\langle w^{\mathrm{PQ}}_{jk}\right\rangle_{jk}\left\langle \overline{V^{\mathrm{P}}_j}\right\rangle_{j}\left\langle \overline{V^{\mathrm{Q}}_k}\right\rangle_{k}\right] - 2\mu\sigma^2_{\mathrm{PQ}}. \tag{14}$$

This is the same as (40), except with an additional exponential decay term for σ^2_{PQ}.

To illustrate this effect, we rerun the simulations of Fig. 4 for the case of $\mu = 0.05\ \mathrm{s}^{-1}$ and $k^{\mathrm{PQ}} = 1$ for the weakly linked (low s) and strongly linked (high s) cases. Results are shown in Fig. 6. We see in parts a and b that stabilization has occurred in soma potential and weights, respectively. However, a clear drawback with this approach to stabilization is that all variances eventually reach zero (parts c and e), indicating that all weights within a given PQ population approach identical values. There are some consistencies with Fig 4; note that this drop in variance is slowest for the case of *ee* correlations (i.e., at any stage, the variation is greatest in the $e \to e$ weights), and differences between the PQ sets are most pronounced in the strongly linked case.

4 Low-level Simulations

To provide some confirmation for the results of the second-order mean field model, we have carried out simulations of small networks of simple neurons. Note that this approach is not intended to be physiologically accurate but to provide some verification of our mean-field results using a non-mean-field scheme. By doing this, our intention is not to make a detailed analysis of the synaptic learning characteristics of these networks but rather to give some validity to the results from the second-order analysis of the mean-field model.

4.1 Neurons with Somas

We first use a simple model that reduces to a similar form as the Liley-style equations under a mean-field approximation. In this scheme, we model individual e and i neurons, assuming a slow soma response but a fast synaptic response. A presynaptic neuron j (of type P) is connected to a postsynaptic neuron k (of type Q) with a weight w^{PQ}_{jk}. Neurons of type Q (e or i) have a soma potential V^{Q}_k that evolves according to the equation:

$$\tau\frac{dV^{\mathrm{Q}}_k}{dt} = \left(V_{\mathrm{rest}} - V^{\mathrm{Q}}_k\right) + \frac{\chi_e}{M^e}\sum_{\mathrm{excits}\,j=1}^{M^e} w^{e\mathrm{Q}}_{jk}\psi^{e\mathrm{Q}}_k q(V^e_j) + \frac{\chi_i}{M_i}\sum_{\mathrm{inhibs}\,j=1}^{M^i} w^{i\mathrm{Q}}_{jk}\psi^{i\mathrm{Q}}_k q(V^i_j), \tag{15}$$

where τ is a time-constant (40 ms); V_{rest} is a resting potential (-64 mV); M^e and M^i are the numbers of individual excitatory and inhibitory neurons used in the simulation, respectively (20 of each); χ_e is an effective excitatory neuron impact (1.0 mV s); χ_i is an effective inhibitory neuron impact (-1.0 mV s); and the function q of the soma potential is a firing rate given by:

$$q(V) = K[1 + \tanh a(V - V_{\mathrm{rest}})] + \upsilon^{noise}. \tag{16}$$

The firing-rate function (ignoring the noise term) clearly approaches zero as V is large and negative (i.e., well below the resting potential of −64 mV), crosses K when $V = V_{\text{rest}}$, and approaches $2K$ when V is large and positive (i.e., well above rest). The width of this transition is governed by a. In our simulations, we choose $K = 15\ \text{s}^{-1}$ and $a = 0.1\ \text{mV}^{-1}$. The ψ_k^{PQ} terms bring in the reversal potentials $V_{\text{rev}}^{\text{P}}$, where:

$$\psi_k^{\text{PQ}} = \frac{V_{\text{rev}}^{\text{P}} - V_k^{\text{Q}}}{V_{\text{rev}}^{\text{P}} - V_{\text{rest}}}. \tag{17}$$

Physically, the reversal potentials constrain V_k^{Q} to the range $V_{\text{rev}}^i < V_k^{\text{Q}} < V_{\text{rev}}^e$. In accordance with the mean field simulations, we choose $V_{\text{rev}}^e = 0$ mV and $V_{\text{rev}}^i = -70$ mV.

The υ^{noise} term introduces white noise into the system; its statistics are:

$$\overline{\upsilon^{\text{noise}}(t)} = 0 \tag{18}$$

$$\overline{\upsilon^{\text{noise}}(t)\upsilon^{\text{noise}}(t')} = C^2\delta(t - t') \tag{19}$$

with $C^2 = 0.001\ \text{s}^{-1}$. To keep this model simple, we have not considered the the details of the shapes of the excitatory and inhibitory postsynaptic potentials, but we assume they occur quickly. This assumption is akin to the adiabatic approximation made by Steyn-Ross et al. [8, 27]. Numerical values are chosen to be roughly consistent with those of the Liley-style model of the Appendix.

The synaptic weights grow according to a Hebb-style rule:

$$\begin{aligned} \frac{dw_{jk}^{\text{PQ}}}{dt} &= \eta\left(q(V_j^{\text{P}}) - K\right)\left(q(V_k^{\text{Q}}) - K\right)/K^2 \qquad V_j^{\text{P}}, V_k^{\text{Q}} > V_{\text{rest}} \\ &= 0 \qquad \text{otherwise.} \end{aligned} \tag{20}$$

In effect, when a presynaptic and a postsynaptic neuron are both high-firing, we increment the weight, with the greatest increment being when they are both at their highest firing rates (i.e., both have soma potentials greater than zero). In our simulations, we choose $\eta = 0.25\ \text{s}^{-1}$.

To apply the model, we need to define a starting matrix for w_{jk}, denoting the initial strengths of connections between the presynaptic neuron j and the postsynaptic neuron k. To model an "unlinked" cortex, we can generate each w_{jk} independently from a Gaussian distribution (but we ensure symmetry between *ei* and *ie* connections). This would correspond to $s = 0$ (or high N_{PQ}) in our model; see Fig. 2a. To model a "linked" cortex, we need to remove some of the independency between weights. A simple way to do this is to put the weights matrix w_{jk} into a "block" form and assign weights to synapses in different blocks from Gaussian distributions with different means. Figure 7a illustrates this. In this case, $s = 0.50$, corresponding to a low value of N_{PQ}. (See also Fig. 2c.)

For each run of the model, we can then extract, at each time step, the value for the mean weights for the *ee*, *ei*, *ie* and *ii* connections. Also, we can find the variance in these weights, equivalent to σ_{PQ}^2 of the second-order mean-field model.

We run this model 250 times with different starting values of the weights matrix w_{jk}. This allows us to estimate a statistical uncertainty in the mean and variance values. One example of the weights matrix after a time evolution is given in Fig. 7b. Statistical results are shown in Fig. 8. Again, we look at the "unlinked" case (independent weights, plots a and b, with $s \approx 0$) and the "linked case" (weights are correlated in a "block" form, plots c and d, with, in this case, $s \approx 0.5$). We see much of the same behavior as in Fig. 4. Looking

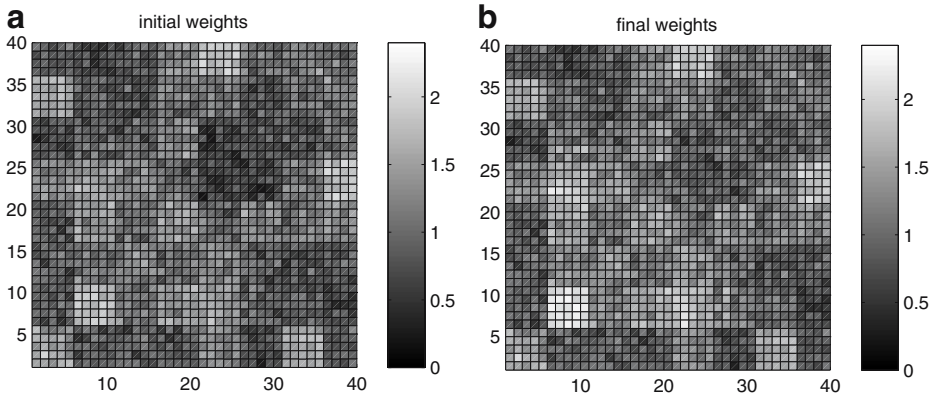

Fig. 7 An example of a weights matrix w_{jk} in which weights are assigned from a Gaussian distribution, but with correlation between weights in particular blocks. **a** before propagation, **b** after propagation. Note how some connections have become much stronger. *White*, strong connections; *black*, weak connections

at the variances in Fig. 8a for the weakly linked case, we see that, initially, the variance in the *ee* term increases, the variance in the *ii* term decreases, and the variance in the $ei = ie$ term stays roughly constant. For the equivalent strongly linked case in part c, we see the changes are much stronger, in agreement with the second-order mean-field results. However, in this simulation, all variances eventually increase, with the *ii* variance turning direction (at about 30 s in Fig. 8c), whereas in the mean-field simulations of Fig. 4c and e, the *ii* variance continues to decrease. The plots of Fig. 8b and d show the variances with respect to the mean weights squared; in these cases, it is clear that the changes are small, showing that the squared weights change in a similar manner to the variances. However, part d shows an increase in this measure for the excitatory to excitatory connections, in agreement with Fig. 4f, whereas the weakly linked case in part b shows a decay. An increase is also obtained for the excitatory to inhibitory connections. Overall, we are encouraged by the correspondence in many features between this simple neuron-by-neuron model and the mean-field scheme.

We can introduce limitations on the synaptic weights in a similar way as for the mean-field approach, using an exponential decay function. Equation (20) becomes:

$$\begin{aligned} \frac{dw_{jk}^{\mathrm{PQ}}}{dt} &= \eta \left(q(V_j^{\mathrm{P}}) - K\right)\left(q(V_k^{\mathrm{Q}}) - K\right)/K^2 + \mu(\kappa - w_{jk}^{\mathrm{PQ}}) \qquad V_j^{\mathrm{P}}, V_k^{\mathrm{Q}} > V_{\mathrm{rest}} \\ &= \mu(\kappa - w_{jk}^{\mathrm{PQ}}) \qquad \text{otherwise.} \end{aligned} \tag{21}$$

We illustrate this with the case of $\eta = 0.25\ \mathrm{s}^{-1}$, $\mu = 0.03\ \mathrm{s}^{-1}$, and $\kappa = 1$, in Fig. 9, for weakly and strongly linked starting weight matrices. In the case of the weakly linked weights, Fig. 9a shows the decay in variance to zero is relatively fast and follows the same trend as for the mean-field analysis of Fig. 6c—the *ee*, $ei = ie$, and *ie* traces are indistinguishable. For the strongly linked case in Fig. 9c, the behavior is consistent with Fig. 6e—the variances fall, with the *ee* falling slightly more slowly and *ii* falling slightly more quickly. The variances measured in relation to the mean weight squared, as shown in parts b and d for the weakly and strongly linked cases, respectively, all fall rapidly, in agreement with the mean-field case shown in Figs. 4d and f.

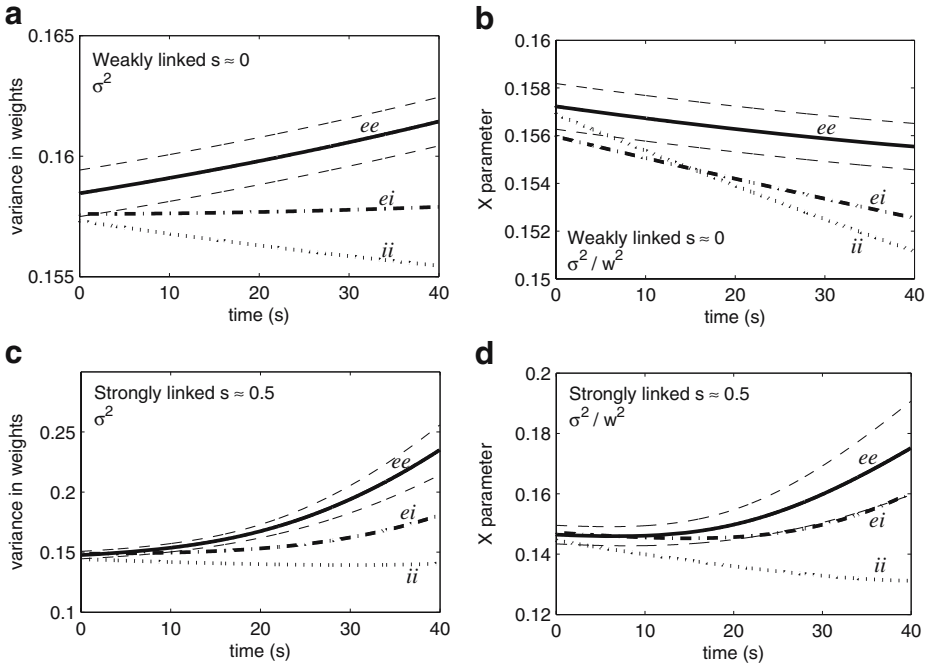

Fig. 8 Results from the soma neuron network model. **a** A plot of the variance in weights against time for uncorrelated initial weights. **b** A plot of variance in the weights divided by the mean weight squared, i.e., the *X*-parameter, for the uncorrelated initial weights. **c** The variance in weights against time for correlated initial weights, of the form of Fig. 2c. **d** A plot of variance in the weights divided by mean weight squared, for the correlated initial weights. Key: *ee*, *solid*; *ie* = *ei*, *dot-dash*; *ii*, *dotted*. The *dashed lines* on either side of the *ee* line indicate the standard uncertainty in the mean, for a total of 250 trials

4.2 Binary Neurons

We have also performed simulations with a simpler, binary model, where each neuron is either "firing" or "quiet." This is similar to the model in (15) above, but with the width of the hyperbolic tangent function in (16) being zero (i.e., $a \to \infty$) and the removal of the exponential decay from the rest term.

The state of a neuron at the next time step is then determined simply by whether the input from the excitatory connections to the neuron, weighted by the w_{jk}^{PQ}, outweighs the input from the inhibitory connections. That is, if we denote the state of neuron k at discrete time t by $S_k(t)$, where $S_k = 1$ denotes firing and $S_k = 0$ denotes quiet, the state of a neuron at discrete time $t + 1$ is given by:

$$S_k(t+1) = \operatorname{sign}\left(\sum_j p_j w_{jk}(t) S_j(t) + \nu_k\right) + 1 \tag{22}$$

where p_j represents the sign of the presynaptic neuron—i.e., if it is an e neuron $p_j = 1$ and if it is an i neuron $p_j = -1$. The term ν_k represents the "noise" input and can be taken, for example, from a Gaussian distribution of mean 0 and some adjustable standard deviation. Alongside this equation, a Hebbian-style learning rule is used—namely, that if two neurons

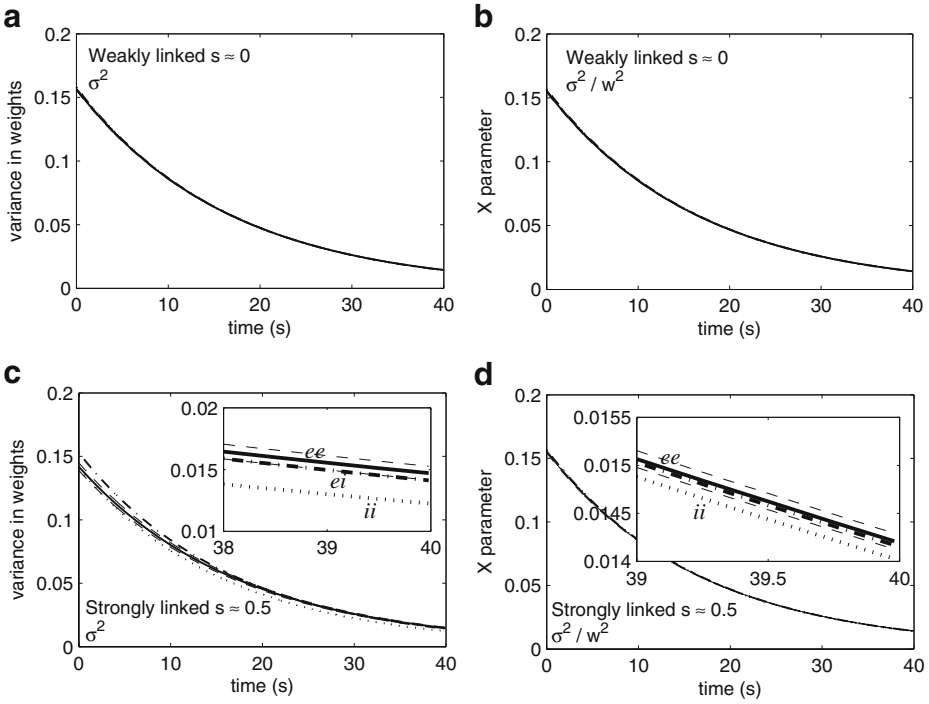

Fig. 9 Results from the soma neuron network model when weights are limited by an exponential decay term. **a** A plot of the variance in weights against time for uncorrelated initial weights; the three curves are virtually identical. **b** A plot of variance in the weights divided by the mean weight squared, i.e., the X parameter, for the uncorrelated initial weights; again, the three curves are virtually identical. **c** The variance in weights against time for correlated initial weights, of the form of Fig. 2c. An *insert* shows the difference in the three traces at larger times. **d** A plot of variance in the weights divided by mean weight squared, for the correlated initial weights. Key: *ee*, *solid*; $ie = ei$, *dot-dash*; *ii*, *dotted*. The *dashed lines* on either side of the *ee* line indicate the standard uncertainty in the mean, for a total of 250 trials

j and k both fire together at time t, the weight between them increments by a predetermined amount η, i.e.:

$$S_j(t) = 1 \quad \text{and} \quad S_k(t) = 1 \quad \Rightarrow \quad w_{jk}(t+1) = w_{jk}(t) + \eta. \tag{23}$$

Depending upon the synaptic weights, the model can exhibit stable firing patterns and limit cycles, which can be loosely attributed to "memories" following Hopfield [38]. Although biophysically lacking in many regards, the model has the advantage of being mathematically simple while maintaining physically dynamic behavior [39]. We run the simulation with equal numbers of e and i neurons so that, on average, there will be equal numbers of firing and quiet neurons.

Results are shown in Fig. 10 for a total of one thousand trials. Again, we see similar behavior as in previous simulations. Looking at parts a and c, we see that, initially, the variance in the *ee* term increases, the variance in the *ii* term decreases, and the variance in the $ie = ei$ term stays roughly constant. Again, the rate of these changes is largest for the strongly linked cases (c), although the difference is clearly not as marked as for the mean-field scheme. Eventually, the *ii* variance increases, as it does in Fig. 8. The changes in the

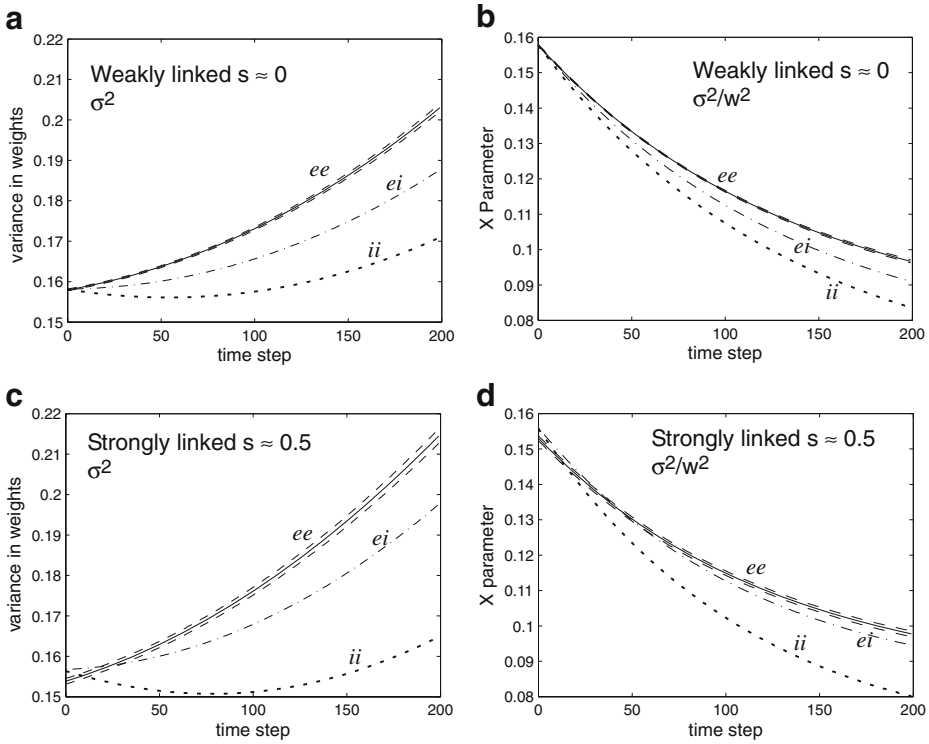

Fig. 10 Results from the binary neuron network model. **a** A plot of the variance in weights against time for uncorrelated initial weights. **b** A plot of variance in the weights divided by the mean weight squared, i.e., the *X* parameter, for the uncorrelated initial weights. **c** The variance in weights against time for correlated initial weights, of the form of Fig. 2c. **d** A plot of variance in the weights divided by mean weight squared, for the correlated initial weights. Key: *ee*, *solid*; $ie = ei$, *dot-dash*; *ii*, *dotted*. The *dashed lines* on either side of the *ee* line indicate the standard uncertainty in the mean, for a total of 1,000 trials. Note that, initially, the *ee* variance rises, *ii* variance falls, and the $ei = ie$ variance stays constant, but all the variances rise for later times

variance terms in relation to the mean weights squared (parts b and d) are mostly in line with the mean-field results of Fig. 4, although the *ee* term does not show an increase with time for the highly correlated case.

It is fair to say that there are some differences between the low-level simulations and the predictions of the mean-field model, as discussed above. In particular, there is the increase in variance of the *ei* and *ii* terms at large times in Figs. 8c and 10a and c. This we attribute to the fact that the mean-field analysis has assumed that fluctuations in soma potential from the local equilibrium values are small; in the neuron models we have analyzed, this is not necessarily the case, except at small times when, as expected, the predictions of the models are more consistent. However, we draw encouragement from the fact that the variance in the *ee* weights, in particular, shows an increase with time when weights are not bounded, with that increase most significant when the weights are correlated.

5 Discussion

What implications do these results have? Broadly, the results indicate the following trends:

- A growth in the variance of the $e \to e$ synaptic weights, but a fall in the variance of the $i \to i$ synaptic weights, for unstabilized connections.
- An accelerated growth of synaptic weights in a highly linked cortex. If s_{PQ} is large enough, the growth in the variance may be greater than that in the mean square weight.
- An unstable equilibrium at $\sigma^2_{PQ} = 0$.

The growth in the variance of the $e \to e$ connections suggests that the strong connections are growing faster than the weak connections. In other words, the network is emphasizing particular $e \to e$ pathways above others. Physiologically, this suggests that the cortex is coding "memories" in terms of a few key $e \to e$ connections, allowing other connections to become available for "new" memories. Abraham and Robins [40] discuss the possibility of a network requiring dynamically changing weights to be able to allow encoding of new memories while retaining the ability to recall old ones efficiently, and Horn et al. [41] and Pantic et al. [42] consider the benefits of synaptic depression to recall of memories in neural networks. Conversely, given the drop in the variance of the $i \to i$ connections, the implication is that these are not so important for memory, but rather to keep the network stable. The number of i neurons in the cortex is substantially lower than the number of e neurons, consistent with this interpretation. Just as interesting is the difference in behavior between a strongly linked and a weakly linked cortex. In the case of no physical limitations on weights, the strongly linked cortex, with relatively few independent connections, shows a much greater increase in σ^2_{ee} than a weakly linked cortex. This suggests that the *ee* weights in more highly linked regions of a cortex will be spread more than the weakly linked regions. When physical limitations on weights are introduced, we could expect that the highly linked key pathways will become cemented at the expense of weakly linked pathways, whose weights will equalize, resulting in the removal of weak memories and more efficient coding of strong memories. The result is a cortex that maintains the strongest memories but is in a state where it can efficiently learn new memories [40]. The precise coding pathways would be expected to be constantly changing.

We remark that we expect the growth in variance in excitatory to excitatory weights to be a robust result when no physical limit is put on the weights. This is because (48) in the Appendix, which describes the growth in variance in terms of weights and soma potentials, comes from the expression of Hebb's rule in (3) and is independent of the exact form of the equations for soma potential. The requirements are: (1) the soma potentials do not deviate greatly from their local equilibriums, and (2) the physically reasonable assumption that a depolarizing of soma potential results in an increased firing rate. We would expect assumption 1 to be reasonable, for example, in the down state of slow-wave sleep, where the soma potentials are relatively constant [43], though not actually at the transition to the up state itself, when the potentials rise quickly.

Before we consider possible neurophysiological implications, we remark on the validity of the learning rule for slow-wave sleep. In a down state, neurons have very low firing rates, and so it is reasonable to ask whether any synaptic weight changes are possible. The changes predicted by the model are a direct consequence of the well-used covariance form of the learning rule in (1) and (3), which implies that correlated fluctuations in membrane

potential (as we would expect to occur at a transition [25]) will induce a growth in synaptic weight. There are at least three possible explanations as to why this effect might occur in practice: (1) it is the very sparse (but correlated) action potentials that produce synaptic weight changes (indeed, Crochet et al. [44] have shown that postsynaptic response is very sensitive to presynaptic stimulation when the neurons are in a down state, and Massimini et al. [45] have demonstrated a similar effect just prior to the transition to the up state); (2) that the correlated subthreshold fluctuations themselves directly change synaptic weights through intracellular dendritic mechanisms; or (3) that the subthreshold fluctuations act to "prime" the cortex so that the initial burst of action potentials, which occur as the neurons transition synchronously to the up state, are particularly efficacious in synaptic weight modification.

5.1 Implications

Tononi and Cirelli have recently put forward an argument that a homeostatic purpose of slow-wave sleep is to adjust synaptic gain [28]. They describe a thought-experiment in which there are synapses of three different weights. After a period of slow-wave sleep, they speculated that, although all synaptic weights are reduced, there could be a *relative* strengthening of the strongest synapse and loss of the weakest one. Assuming that synaptic weight changes can occur in the low-firing down state, our results could provide a quantitative basis for their postulate—that a period of nonspecific episodic white noise stimulation will cause a relative increase in the variance of the gain in excitatory to excitatory synaptic weights in *structured* neuronal assemblies (see Fig. 6) when transitions between down and up states occur (such as in slow-wave sleep). That is, the strong have gotten stronger, the weak have gotten weaker. In less structured assemblies, which presumably do not carry useful memories, the variance is reduced (i.e., the weights are equalized). Further experimental research is required to test whether the neuromodulatory environment of slow-wave sleep is conducive to the plasticity changes predicted by our model.

According to (1), these synaptic changes will take place quickest when the correlations between neurons are greatest. This is close to the points of transition between low- and high-firing states [17, 25] (although we expect our analysis to break down at the transition itself). This naturally suggests a role for the slow oscillation of slow-wave sleep [43, 46], where the cortex makes multiple jumps between low- and high-firing states. Experimentally, Battaglia et al. have reported that electroencephalographic sharp waves from the hippocampus are correlated with the down–up transition of slow-wave sleep, suggestive of a role for slow waves in long-term memory [47]. Additionally, Marshall et al. have demonstrated that the transcranial application of slow-wave potentials improves declarative memory recall in human subjects [48].

6 Conclusions

In this paper, we have presented an adaptation to a mean-field model that allows the study of the distribution of synaptic weights as opposed to simply the mean synaptic weights. The goal has been to produce a versatile model that has the advantages of both mean-field and neural network approaches, and although the mathematics is only approximate, our model is a step towards this. Specifically, the changes in this distribution under a Hebbian learning rule have been followed; however, other schemes can be incorporated

into the model. In the absence of physical limits on the size of the weights, the standard deviation of the distribution of excitatory to excitatory synaptic weights grows with time, with the rate of growth dependent upon the degree of independence between neuron connections. The model requires that consideration be given to the degree of structure in the connectivity of the network—i.e., "what does the weight w_{ij} imply about the weight w_{jk}?" The standard deviation of the excitatory to excitatory synaptic weights of a realistic cortical network with highly linked connections grows more quickly than that of a hypothetical network of independent connections. When limits are placed on the growth of weights (albeit somewhat artificially), networks of independently connected neurons decrease their standard deviation of the weights, implying that the weights begin to equalize. However, sufficiently linked networks will grow their standard deviations for excitatory to excitatory connections. This suggests that strong memories within the network, under driving by noise, are gradually encoded with fewer neurons (higher standard deviation), whereas weak memories are removed. These processes take place most quickly where fluctuations in soma potential are greatest, suggesting that slow-wave sleep is important. The standard deviation of inhibitory to inhibitory neurons tends to decrease, suggesting that these connections are not so important for coding memories. Further work would be required to establish the robustness of these results.

7 Appendix

7.1 Equations for the Mean-field Model

First, we describe the complete set of equations for the Liley-style mean-field cortical model as used and presented in Wilson et al. [19], drawing from Liley et al. [6] and Rennie et al. [7].

$$\tau_e \frac{d\langle V^e\rangle}{dt} = V^e_{\text{rest}} - \langle V^e\rangle + \langle w^{ee}\rangle \rho_e \psi^{ee} \Phi^{ee} + \langle w^{ie}\rangle \rho_i \psi^{ie} \Phi^{ie}; \tag{24}$$

$$\tau_i \frac{d\langle V^i\rangle}{dt} = V^i_{\text{rest}} - \langle V^i\rangle + \langle w^{ei}\rangle \rho_e \psi^{ei} \Phi^{ei} + \langle w^{ii}\rangle \rho_i \psi^{ii} \Phi^{ii}; \tag{25}$$

$$\left(\frac{d^2}{dt^2} + 2\gamma_{ee}\frac{d}{dt} + \gamma_{ee}^2\right)\Phi^{ee} = \gamma_{ee}^2\left(N_{ee}^{\alpha}\phi^{ee} + N_{ee}^{\beta}Q^e + \phi_{ee}^{\text{sc}}\right); \tag{26}$$

$$\left(\frac{d^2}{dt^2} + 2\gamma_{ei}\frac{d}{dt} + \gamma_{ei}^2\right)\Phi^{ei} = \gamma_{ei}^2\left(N_{ei}^{\alpha}\phi^{ei} + N_{ei}^{\beta}Q^e + \phi_{ei}^{\text{sc}}\right); \tag{27}$$

$$\left(\frac{d^2}{dt^2} + 2\gamma_{ie}\frac{d}{dt} + \gamma_{ie}^2\right)\Phi^{ie} = \gamma_{ie}^2\left(N_{ie}^{\beta}Q^i + \phi_{ie}^{\text{sc}}\right); \tag{28}$$

$$\left(\frac{d^2}{dt^2} + 2\gamma_{ii}\frac{d}{dt} + \gamma_{ii}^2\right)\Phi^{ii} = \gamma_{ii}^2\left(N_{ii}^{\beta}Q^i + \phi_{ii}^{\text{sc}}\right); \tag{29}$$

$$\left(\frac{\partial^2}{\partial t^2} + 2v\Lambda_{ee}\frac{\partial}{\partial t} + v^2\Lambda_{ee}^2 - v^2\nabla^2\right)\phi^{ee} = v^2\Lambda_{ee}^2 Q^e; \tag{30}$$

$$\left(\frac{\partial^2}{\partial t^2} + 2v\Lambda_{ei}\frac{\partial}{\partial t} + v^2\Lambda_{ei}^2 - v^2\nabla^2\right)\phi^{ei} = v^2\Lambda_{ei}^2 Q^e. \tag{31}$$

The column-averaged soma potentials excitatory e and inhibitory i soma potentials are denoted by $\langle V^e \rangle$ and $\langle V^i \rangle$, respectively. Time is denoted by t.

In these equations, V^e_{rest} and V^i_{rest} are the excitatory and inhibitory neurons' resting potentials and ρ_e and ρ_i are the strengths of the excitatory postsynaptic potential (EPSP) and inhibitory postsynaptic potential (IPSP) response functions (i.e., the *area* of the plot of postsynaptic potential response function against time). Note that the inhibitory effect is modelled with a *negative* ρ_i. The terms $\langle w^{ee} \rangle$, $\langle w^{ie} \rangle$, $\langle w^{ei} \rangle$, and $\langle w^{ii} \rangle$ represent column-averaged synaptic weights—namely, $\langle w^{\text{PQ}} \rangle = \left\langle w^{\text{PQ}}_{jk} \right\rangle_{jk}$, etc., where P and Q can take on the labels e and i. In the modelling of Steyn-Ross et al. [27], each of the four weights is implicitly assumed to be unity. The variables ψ^{ee}, ψ^{ei}, ψ^{ie}, and ψ^{ii} are weighting functions dependent upon the soma potentials. They are given by:

$$\psi^{ab} = \frac{V^a_{\text{rev}} - \langle V^b \rangle}{V^a_{\text{rev}} - V^b_{\text{rest}}}. \tag{32}$$

Here, V^a_{rev} is the reversal potential of the type a synapse, due to the concentrations of the neurotransmitters AMPA and GABA. The superscripts a and b can take on the labels e and i. The terms τ_e and τ_i describe the time-constants for the e and i neurons.

The terms Φ^{ab} describe the synaptic flux rate for the connections *from* type a *to* type b. The γ_{ab} terms are synaptic rate-constants; their reciprocals give the time-scales over which the EPSPs and IPSPs occur. Long-range synaptic flux-rate is represented through the terms ϕ_{ea} (note that there are no long-range inhibitory connections). The N^{β}_{ab} represent numbers of *local* intracolumn connections from type a neurons to type b (again, a and b can take on the labels e and i) and the N^{α}_{ea} the number of *long-range* connections from type e neurons to type a. The mean axonal velocity for long-range interactions is given by v, and the characteristic length for long-range interactions is given by $1/\Lambda_{ea}$. Short-range interactions are not modelled with axonal propagation but are assumed to be instantaneous; the N^{β}_{jk} terms couple directly with the population firing rates Q^k in (26–29).

The sigmoidal functions Q^e and Q^i, describing the population firing-rate of neurons, are given by:

$$Q^e\big(\langle V^e \rangle\big) = \frac{Q^e_{\max}}{1 + \exp\big[-\pi(\langle V^e \rangle - \theta_e)/\sqrt{3}\sigma_e\big]}; \tag{33}$$

$$Q^i\big(\langle V^i \rangle\big) = \frac{Q^i_{\max}}{1 + \exp\big[-\pi(\langle V^i \rangle - \theta_i)/\sqrt{3}\sigma_i\big]}. \tag{34}$$

Here, we have introduced further variables $Q^e_{\max}$ and $Q^i_{\max}$, the maximum firing rates for the excitatory and inhibitory neurons, respectively; θ_e and θ_i, the inflexion point voltage; and σ_e and σ_i, the standard deviation of the threshold potential.

The ϕ^{sc}_{ab} terms provide the subcortical white-noise driving of the model. The time variation of the subcortical noise is assumed to obey the equation:

$$\phi^{\text{sc}}_{ab}(\vec{r}, t) = \langle \phi^{\text{sc}}_{ab} \rangle + \upsilon_{ab}(\vec{r}, t) \left\langle \overline{\phi^{\text{sc}}_{ab}} \right\rangle^{\frac{1}{2}}. \tag{35}$$

The $\upsilon_{ab}(\vec{r}, t)$ describe white noise; their statistics obey:

$$\langle \upsilon_{ab}(\vec{r}, t)\rangle = 0, \tag{36}$$

$$\left\langle \upsilon_{ab}(\vec{r}, t)\upsilon_{cd}(\vec{r'}, t')\right\rangle = \delta_{ac}\delta_{bd}\delta(t-t')\delta(\vec{r}-\vec{r'}). \tag{37}$$

Here, $\langle\cdots\rangle$ denotes an average over space and time.

The list of standard parameters used is given in Table 1. In the anesthesia modelling of Steyn-Ross et al. [8], and the sleep modelling of Wilson et al. [19], these equations have been slightly modified by various scaling parameters to describe the effects of drugs and neuromodulators.

7.2 Developing the Equations to Include Variances

We wish to understand how the mean-square of the synaptic weight w_{jk}^{PQ} changes with time. We define the variance of the synaptic weights across a cortical column as:

$$\sigma_{\mathrm{PQ}}^2 = \left\langle w_{jk}^{\mathrm{PQ}} w_{jk}^{\mathrm{PQ}}\right\rangle_{jk} - \left\langle w_{jk}^{\mathrm{PQ}}\right\rangle_{jk}^2 \tag{38}$$

where the average is taken over all presynaptic neurons j (of type P) and all postsynaptic neurons k (of type Q). Taking the time-derivative of the variance, we get:

$$\frac{d}{dt}\sigma_{\mathrm{PQ}}^2 = 2\left\langle w_{jk}^{\mathrm{PQ}}\frac{d}{dt}w_{jk}^{\mathrm{PQ}}\right\rangle_{jk} - 2\left\langle w_{jk}^{\mathrm{PQ}}\right\rangle_{jk}\frac{d}{dt}\left\langle w_{jk}^{\mathrm{PQ}}\right\rangle_{jk}. \tag{39}$$

Substituting from (1) and (3), we obtain:

$$\begin{aligned}\frac{d}{dt}\sigma_{\mathrm{PQ}}^2 = 2\eta^{\mathrm{PQ}}\Bigg[&\left\langle w_{jk}^{\mathrm{PQ}}\left(\overline{V_j^{\mathrm{P}}V_k^{\mathrm{Q}}} - \overline{V_j^{\mathrm{P}}}\;\overline{V_k^{\mathrm{Q}}}\right)\right\rangle_{jk}\\ &- \left\langle w_{jk}^{\mathrm{PQ}}\right\rangle_{jk}\left\langle\overline{V_j^{\mathrm{P}}V_k^{\mathrm{Q}}}\right\rangle_{jk} + \left\langle w_{jk}^{\mathrm{PQ}}\right\rangle_{jk}\left\langle\overline{V_j^{\mathrm{P}}}\right\rangle_{j}\left\langle\overline{V_k^{\mathrm{Q}}}\right\rangle_{k}\Bigg].\end{aligned} \tag{40}$$

In (40), we are bringing in higher-order correlation terms. In particular, we will now need to consider terms of the form $\left\langle w_{jk}^{\mathrm{PQ}}\overline{V_j^{\mathrm{P}}V_k^{\mathrm{Q}}}\right\rangle$. The terms are of third-order in the sense that they are several averages over a product of three different quantities. As a first step away from considering purely the means of quantities, we wish to reduce this to second order. Therefore, we take the usual approach of considering only small variations in V_j^{P} and V_k^{Q} from their *instantaneous* column-averaged values. We write, without loss of generality:

$$V_j^{\mathrm{P}} = \left\langle V_j^{\mathrm{P}}\right\rangle_j + \Delta V_j^{\mathrm{P}} \tag{41}$$

$$V_k^{\mathrm{Q}} = \left\langle V_k^{\mathrm{Q}}\right\rangle_k + \Delta V_k^{\mathrm{Q}}. \tag{42}$$

We therefore have, ignoring terms of order ΔV^2:

$$\left\langle w_{jk}^{\mathrm{PQ}} V_j^{\mathrm{P}} V_k^{\mathrm{Q}} \right\rangle_{jk} \approx -\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} \left\langle V_j^{\mathrm{P}} \right\rangle_j \left\langle V_k^{\mathrm{Q}} \right\rangle_k + \left\langle V_j^{\mathrm{P}} \right\rangle_j \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j V_k^{\mathrm{Q}} \right\rangle_k + \left\langle V_k^{\mathrm{Q}} \right\rangle_k \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_k V_j^{\mathrm{P}} \right\rangle_j . \tag{43}$$

These quantities are implicitly time-varying (e.g., $V_j^{\mathrm{P}} = V_j^{\mathrm{P}}(t)$); the "$t$" has been dropped for clarity. Note carefully the order of averaging. We now have terms that are second-order in the sense of averages over a product of two quantities. However, there is a new kind of entity, involving a combined average over the synaptic weight connections w_{jk}^{PQ}, and the soma potentials of the pre- and postsynaptic neurons (V_j^{P} and V_k^{Q} respectively). We define two new variables for these averages:

$$\xi_{\mathrm{post}}^{\mathrm{PQ}} = \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j V_k^{\mathrm{Q}} \right\rangle_k , \tag{44}$$

$$\xi_{\mathrm{pre}}^{\mathrm{PQ}} = \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_k V_j^{\mathrm{P}} \right\rangle_j . \tag{45}$$

Here, $\xi_{\mathrm{post}}^{\mathrm{PQ}}$ is an average over all postsynaptic neurons k of the soma potential multiplied by the average synaptic weight *onto* that neuron. Conversely, $\xi_{\mathrm{pre}}^{\mathrm{PQ}}$ is an average over all presynaptic neurons j of the soma potential multiplied by the average synaptic weight *from* that neuron. If there is symmetry in the weight connections, i.e., $w_{jk}^{\mathrm{PQ}} = w_{kj}^{\mathrm{QP}}$, then $\xi_{\mathrm{post}}^{\mathrm{PQ}} = \xi_{\mathrm{pre}}^{\mathrm{QP}}$ (note order of superscripts). Substituting (43–45) into (40), we obtain:

$$\frac{d}{dt}\sigma_{\mathrm{PQ}}^2 = 2\eta^{\mathrm{PQ}} \left(-2 \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} \overline{\left\langle V_j^{\mathrm{P}} \right\rangle_j \left\langle V_k^{\mathrm{Q}} \right\rangle_k} + 2 \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} \overline{\left\langle V_j^{\mathrm{P}} \right\rangle_j} \, \overline{\left\langle V_k^{\mathrm{Q}} \right\rangle_k} + \overline{\left\langle V_j^{\mathrm{P}} \right\rangle_j \xi_{\mathrm{post}}^{\mathrm{PQ}}} - \overline{\left\langle V_j^{\mathrm{P}} \right\rangle_j} \, \overline{\xi_{\mathrm{post}}^{\mathrm{PQ}}} + \overline{\left\langle V_k^{\mathrm{Q}} \right\rangle_k \xi_{\mathrm{pre}}^{\mathrm{PQ}}} - \overline{\left\langle V_k^{\mathrm{Q}} \right\rangle_k} \, \overline{\xi_{\mathrm{pre}}^{\mathrm{PQ}}} \right) \tag{46}$$

Using the definition of covariance as $\mathrm{cov}(A, B) = \overline{AB} - \overline{A}\,\overline{B}$, we can write this as:

$$\frac{d}{dt}\sigma_{\mathrm{PQ}}^2 = 2\eta^{\mathrm{PQ}} \left[-2 \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} \mathrm{cov} \left(\left\langle V_j^{\mathrm{P}} \right\rangle_j , \left\langle V_k^{\mathrm{Q}} \right\rangle_k \right) + \mathrm{cov} \left(\xi_{\mathrm{post}}^{\mathrm{PQ}} , \left\langle V_j^{\mathrm{P}} \right\rangle_j \right) + \mathrm{cov} \left(\xi_{\mathrm{pre}}^{\mathrm{PQ}} , \left\langle V_k^{\mathrm{Q}} \right\rangle_k \right) \right] , \tag{47}$$

where the covariances can be calculated in a simulation by considering the variation of these terms over a two-dimensional grid (i.e., over the cortex). Using further the identity $\mathrm{cov}(\lambda A - \mu B, C) = \lambda \mathrm{cov}(A, C) - \mu \mathrm{cov}(B, C)$, and (45), we obtain the expression:

$$\frac{d}{dt}\sigma_{\mathrm{PQ}}^2 = 2\eta^{\mathrm{PQ}} \left\{ \mathrm{cov}^* \left[\mathrm{cov}_k \left(\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j , V_k^{\mathrm{Q}} \right) , \left\langle V_j^{\mathrm{P}} \right\rangle_j \right] + \mathrm{cov}^* \left[\mathrm{cov}_j \left(\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_k , V_j^{\mathrm{P}} \right) , \left\langle V_k^{\mathrm{Q}} \right\rangle_k \right] \right\} , \tag{48}$$

where cov_k denotes the covariance over all postsynaptic neurons within the cortical column, cov_j denotes the covariance over all presynaptic neurons in the cortical column, and cov^* denotes the covariance over all columns (i.e., over the cortex).

This equation is independent of that used to describe how V_k^{Q} changes with time, but is a direct consequence of the mean-field form of Hebb's rule, (3).

To complete our equation set, we need to find out how the terms $\xi_{\mathrm{post}}^{\mathrm{PQ}}$ and $\xi_{\mathrm{pre}}^{\mathrm{PQ}}$ change with time, and restrict them to second order. We can write:

$$\frac{d}{dt}\xi_{\mathrm{post}}^{\mathrm{PQ}} = \frac{d}{dt}\left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j V_k^{\mathrm{Q}} \right\rangle_k = \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j \frac{dV_k^{\mathrm{Q}}}{dt} \right\rangle_k \tag{49}$$

where the final part assumes that the synaptic weights change much more slowly than the mean soma potentials. There is a similar expression for $\xi_{\mathrm{pre}}^{\mathrm{PQ}}$.

We now require a description of the variation of the soma potential of a particular neuron, V_k^{Q}, as a function of time, and the variation of the column-averaged soma potential, $\left\langle V_k^{\mathrm{Q}} \right\rangle_k$, as a function of time.

The latter of these quantities is straightforward; it is the state variable that is modelled in a large number of mean-field schemes. Any scheme can be used here; for illustration purposes, we will use that of Liley et al. [6] and used by Steyn-Ross et al. [27] and Wilson et al. [19], already presented. The mean soma potential $\left\langle V_k^{\mathrm{Q}} \right\rangle_k$ of type Q neurons follows the equation:

$$\tau_{\mathrm{Q}} \frac{d\left\langle V_k^{\mathrm{Q}} \right\rangle_k}{dt} = \left(V_{\mathrm{rest}}^{\mathrm{Q}} - \left\langle V_k^{\mathrm{Q}} \right\rangle_k \right) + \rho_e \left\langle w_{lk}^{\mathrm{eQ}} \right\rangle_{lk} \Phi^{e\mathrm{Q}} \psi^{e\mathrm{Q}} + \rho_i \left\langle w_{mk}^{\mathrm{iQ}} \right\rangle_{mk} \Phi^{i\mathrm{Q}} \psi^{i\mathrm{Q}} \tag{50}$$

where τ_{Q} is the average somatic time-constant for neurons of type Q (Q = e or i); $V_{\mathrm{rest}}^{\mathrm{Q}}$ is the mean resting potential; ρ_e is the average time-integrated area of the EPSP; ρ_i is the average time-integrated area of the IPSP; $\Phi^{e\mathrm{Q}}$ is the average synaptic flux from excitatory neurons onto neurons of type Q; $\Phi^{i\mathrm{Q}}$ is the average synaptic flux from inhibitory neurons onto neurons of type Q; and $\psi^{e\mathrm{Q}}$ and $\psi^{i\mathrm{Q}}$ are weighting functions, depending on the reversal potentials, describing how the susceptibility of a neuron to synaptic input changes with its soma potential: $\psi^{e\mathrm{Q}} = \left(V_{\mathrm{rev}}^{e} - \left\langle V_k^{\mathrm{Q}} \right\rangle_k \right) / (V_{\mathrm{rev}}^{e} - V_{\mathrm{rest}}^{\mathrm{Q}})$ where V_{rev}^{e} is the reversal potential at excitatory synapses. A similar expression exists for $\psi^{i\mathrm{Q}}$. Note that ρ_i is negative—this accounts for the hyperpolarizing effect of the inhibitory synapses.

The equation for the soma potential of a particular neuron, V_k^{Q}, requires further discussion. What is of importance is that we recover the mean-field equation (50) when we take the mean over the neurons within a cortical column. It is not critical to model the action-potentials explicitly because our intention is to look at the correlations in behavior of neurons to determine the growth of the weight of the connection between them. This correlation can be modelled more simply and without loss of applicability by ignoring the spikes. With this in mind, we deconstruct (50) to obtain an *approximate* equation for the potential change for a single neuron that is fit for the purpose of this model. That is:

$$\tau_{\mathrm{Q}} \frac{dV_k^{\mathrm{Q}}}{dt} = \left(V_{\mathrm{rest}}^{\mathrm{Q}} - V_k^{\mathrm{Q}} \right) + \rho_e \sum_l w_{lk}^{\mathrm{eQ}} \tilde{\Phi}_{lk}^{e\mathrm{Q}} \tilde{\psi}_k^{e\mathrm{Q}} + \rho_i \sum_m w_{mk}^{\mathrm{iQ}} \tilde{\Phi}_{mk}^{i\mathrm{Q}} \tilde{\psi}_k^{i\mathrm{Q}} \tag{51}$$

where the first sum is over excitatory neurons k only and the second sum is over inhibitory neurons m only. The terms $\tilde{\Phi}_{lk}^{e\mathrm{Q}}$ is the synaptic flux rate from, specifically, the l-th

presynaptic (of type e) to the k-th postsynaptic neuron (of type Q). The term $\tilde{\psi}_k^{e\mathrm{Q}}$ describes the effect of the reversal potential at excitatory synapses on the susceptibility of neuron k through $\tilde{\psi}_k^{e\mathrm{Q}} = (V_{\mathrm{rev}}^e - V_k^{\mathrm{Q}})/(V_{\mathrm{rev}}^e - V_{\mathrm{rest}}^{\mathrm{Q}})$. A similar expression exists for $\tilde{\psi}_k^{i\mathrm{Q}}$.

We can now substitute (51) into (49) to give:

$$\tau_{\mathrm{Q}} \frac{d\xi_{\mathrm{post}}^{\mathrm{PQ}}}{dt} = -\left(\left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j V_k^{\mathrm{Q}} \right\rangle_k - \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} V_{\mathrm{rest}}^{\mathrm{Q}}\right) + \rho_e \sum_l \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j w_{lk}^{e\mathrm{Q}} \tilde{\Phi}_{lk}^{e\mathrm{Q}} \tilde{\psi}_k^{e\mathrm{Q}} \right\rangle_k + \rho_i \sum_m \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j w_{mk}^{i\mathrm{Q}} \tilde{\Phi}_{mk}^{i\mathrm{Q}} \tilde{\psi}_k^{i\mathrm{Q}} \right\rangle_k . \tag{52}$$

In order for Eq. (51) to reduce to Eq. (50) on averaging over neurons k, we recognize that:

$$\left\langle \sum_l \tilde{\Phi}_{lk}^{e\mathrm{Q}} \tilde{\psi}_k^{e\mathrm{Q}} \right\rangle_k = \Phi^{e\mathrm{Q}} \psi^{e\mathrm{Q}}, \tag{53}$$

$$\left\langle \sum_m \tilde{\Phi}_{mk}^{i\mathrm{Q}} \tilde{\psi}_k^{i\mathrm{Q}} \right\rangle_k = \Phi^{i\mathrm{Q}} \psi^{i\mathrm{Q}}, \tag{54}$$

and so, Eq. (52) becomes:

$$\tau_{\mathrm{Q}} \frac{d\xi_{\mathrm{post}}^{\mathrm{PQ}}}{dt} = -\left(\xi_{\mathrm{post}}^{\mathrm{PQ}} - \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} V_{\mathrm{rest}}^{\mathrm{Q}}\right) + \rho_e \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j \left\langle w_{lk}^{e\mathrm{Q}} \right\rangle_l \right\rangle_k \Phi^{e\mathrm{Q}} \psi^{e\mathrm{Q}} + \rho_i \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j \left\langle w_{mk}^{i\mathrm{Q}} \right\rangle_m \right\rangle_k \Phi^{i\mathrm{Q}} \psi^{i\mathrm{Q}} \tag{55}$$

where we also assume no correlation between the synaptic flux rates $\tilde{\Phi}_{lk}^{e\mathrm{Q}}$ onto neuron k from neuron l and the weight $w_{lk}^{e\mathrm{Q}}$ (similarly for $\tilde{\Phi}_{mk}^{i\mathrm{Q}}$ and $w_{mk}^{i\mathrm{Q}}$).

This equation contains second-order terms in w. To close the set of equations completely, we relate them to $\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}$ and σ_{PQ}^2.

The second order-terms in w in (55) have the form:

$$F^{\mathrm{PQR}} = \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j \left\langle w_{lk}^{\mathrm{RQ}} \right\rangle_l \right\rangle_k = \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j \left\langle w_{kl}^{\mathrm{QR}} \right\rangle_l \right\rangle_k = \left\langle w_{jk}^{\mathrm{PQ}} w_{kl}^{\mathrm{QR}} \right\rangle_{jkl} \tag{56}$$

where the second step follows if we assume symmetry in the weights: $w_{jk}^{\mathrm{PQ}} = w_{kj}^{\mathrm{QP}}$. Because the Hebbian rule (1) is symmetric, weights that start symmetric will stay symmetric at all times. Here, P, Q, and R can each mark the populations e or i. We see that F^{PQR} is, in effect, the mean of the matrix elements resulting from the product of the w^{PQ} and w^{QR} matrices, and it represents a second-order effect of population P influencing population R via population Q (or R influencing P via population Q.) The term F^{PQR} unfortunately involves three means, and this is not conducive to physical modelling because it generates further terms that must be considered as time-varying quantities. To make the modelling manageable, we seek to write F^{PQR} approximately with terms involving just two means.

Let us consider the case when P = R (=*e* or *i*). This will always be the case for *one* of the two final terms on the right-hand-side of (55). The populations w^{PQ} and w^{RQ} are certainly correlated. Then, we have:

$$F^{\mathrm{PQR}} = \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j \left\langle w_{lk}^{\mathrm{PQ}} \right\rangle_l \right\rangle_k = \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j^2 \right\rangle_k . \tag{57}$$

We wish to relate this to the variance $\sigma_{\mathrm{PQ}}^2 = \left\langle w_{jk}^{\mathrm{PQ2}} \right\rangle_{jk} - \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}^2$. The simplest way of doing this is to assume, through the central limit theorem, that the variance (over neurons k) in $\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j$ is equal to the variance in w_{jk}^{PQ} (over all j and k) divided by the number N_{PQ} of *independent* P to Q connections in the cortical column for each postsynaptic neuron of type Q. This allows us to write:

$$\begin{aligned} \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j^2 \right\rangle_k - \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}^2 &= \sigma_{\mathrm{PQ}}^2 / N_{PQ} \\ \Rightarrow \left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j^2 \right\rangle_k &= \sigma_{\mathrm{PQ}}^2 / N_{PQ} + \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}^2 . \end{aligned} \tag{58}$$

Equation (58) can be substituted for *one* of the terms in (55), depending on whether P = R = *e* or P = R = *i*. This leaves a second term, namely, when P ≠ R. In this case, we will assume that the populations w^{PQ} and w^{RQ} are uncorrelated. Therefore, we will write:

$$\left\langle \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_j \left\langle w_{lk}^{\mathrm{RQ}} \right\rangle_l \right\rangle_k = \left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk} \left\langle w_{lk}^{\mathrm{RQ}} \right\rangle_{lk} , \tag{59}$$

which is now in terms of means over weights (to the power 1). Therefore, we have written (55) in terms of the set of variables $\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}$ and σ_{PQ}^2, where P and Q can take the values *e* and *i*. This completes the set of equations, namely, (3), (47), and (55), with assumptions (58) and (59) and symmetry in the weights matrix. Our state variables are $\left\langle V_j^{\mathrm{P}} \right\rangle_j$, $\left\langle w_{jk}^{\mathrm{PQ}} \right\rangle_{jk}$, σ_{PQ}^2 and $\xi_{\mathrm{post}}^{\mathrm{PQ}}$, and these equations describe how these quantities change with time. (We also require the other Liley-style equations for Φ^{PQ}, etc., for the model of Steyn-Ross et al. [27]).

For the modelling to proceed, we need to specify N_{PQ} as a parameter. To provide for a more physical description, we transform this to a parameter s given by:

$$s_{\mathrm{PQ}} = \frac{N_{\mathrm{PQtotal}} - N_{\mathrm{PQ}}}{N_{\mathrm{PQtotal}}} \tag{60}$$

where N_{PQtotal} is the averaged total number of connections for each post-synaptic neuron of type Q. This is different for *e* and *i* neurons, but we use a value of 1000 as a physically reasonable approximation (see Table 1). This means that $s_{\mathrm{PQ}} \approx 0$ represents a cortex that is "weakly-linked" in that its synaptic weights are mostly independent of each other, but $s_{\mathrm{PQ}} \approx 1$ represents a cortex that is "strongly-linked," in that its synaptic weights have strong dependencies. Figure 2 explains this pictorially.

7.3 Analysis of the Equilibrium

It is easy to show that an equilibrium for σ^2_{PQ} exists for the case of:

$$\sigma^2_{\mathrm{PQ}} = 0, \tag{61}$$

$$\xi^{\mathrm{PQ}}_{\mathrm{post}} = \left\langle w^{\mathrm{PQ}}_{jk} \right\rangle_{jk} \left\langle V^{\mathrm{Q}}_{k} \right\rangle_{k}. \tag{62}$$

Physically, this corresponds to a cortex where all weights are the same. To see this equilibrium, first of all, put (61) and (62) into (58) and then into (55). This allows us to take the mean weight out as a common factor to give:

$$\frac{d\xi^{\mathrm{PQ}}_{\mathrm{post}}}{dt} = \left\langle w^{\mathrm{PQ}}_{jk} \right\rangle_{jk} \frac{d\left\langle V^{\mathrm{Q}}_{k} \right\rangle_{k}}{dt}, \tag{63}$$

where we have used (50). Integrating gives:

$$\xi^{\mathrm{PQ}}_{\mathrm{post}} = \left\langle w^{\mathrm{PQ}}_{jk} \right\rangle_{jk} \left\langle V^{\mathrm{Q}}_{k} \right\rangle_{k} + \mathrm{Const} \tag{64}$$

where the constant can be set to zero by the initial condition (62). Therefore, (62) remains valid at later times. Likewise, putting (61) and (62) into (40), using the assumption that $\eta^{\mathrm{PQ}} = \eta^{\mathrm{QP}}$ so $\xi^{\mathrm{QP}}_{\mathrm{pre}} = \xi^{\mathrm{PQ}}_{\mathrm{post}}$, gives us:

$$\frac{d}{dt}\sigma^2_{\mathrm{PQ}} = 0, \tag{65}$$

which completes the analysis. Note that $\xi^{\mathrm{PQ}}_{\mathrm{post}}$ itself is not constant, but it always follows (62). In effect, we can define a (spatially dependent) variable

$$\Delta^{\mathrm{PQ}} = \xi^{\mathrm{PQ}}_{\mathrm{post}} - \left\langle w^{\mathrm{PQ}}_{jk} \right\rangle_{jk} \left\langle V^{\mathrm{Q}}_{k} \right\rangle_{k} = \mathrm{cov}_j\left(\left\langle w^{\mathrm{PQ}}_{jk} \right\rangle_{k}, V^{\mathrm{P}}_{j} \right) \tag{66}$$

which is constant along with σ^2_{PQ} under these equilibrium conditions. Note that Δ^{PQ} is the inner covariance of (48).

We now consider the stability of the equilibrium conditions (61) and (62). We look at the equation for the variation in Δ^{PQ} (66). We have, from (50) and (52):

$$\begin{aligned}
\tau_{\mathrm{Q}} \frac{d\Delta^{\mathrm{PQ}}}{dt} &= \tau_{\mathrm{Q}} \frac{d}{dt}\left(\xi^{\mathrm{PQ}}_{\mathrm{post}} - \left\langle w^{\mathrm{PQ}}_{jk} \right\rangle_{jk} \left\langle V^{\mathrm{Q}}_{k} \right\rangle_{k} \right) \\
&= -\left(\xi^{\mathrm{PQ}}_{\mathrm{post}} - \left\langle w^{\mathrm{PQ}}_{jk} \right\rangle_{jk} \left\langle V^{\mathrm{Q}}_{k} \right\rangle_{k} \right) \\
&\quad + \rho_e \left[\left\langle \left\langle w^{\mathrm{PQ}}_{jk} \right\rangle_{j} \left\langle w^{e\mathrm{Q}}_{lk} \right\rangle_{l} \right\rangle_{k} - \left\langle w^{\mathrm{PQ}}_{jk} \right\rangle_{jk} \left\langle w^{e\mathrm{Q}}_{lk} \right\rangle_{lk} \right] \Phi^{e\mathrm{Q}} \psi^{e\mathrm{Q}} \\
&\quad + \rho_i \left[\left\langle \left\langle w^{\mathrm{PQ}}_{jk} \right\rangle_{j} \left\langle w^{i\mathrm{Q}}_{mk} \right\rangle_{m} \right\rangle_{k} - \left\langle w^{\mathrm{PQ}}_{jk} \right\rangle_{jk} \left\langle w^{i\mathrm{Q}}_{mk} \right\rangle_{mk} \right] \Phi^{i\mathrm{Q}} \psi^{i\mathrm{Q}}.
\end{aligned} \tag{67}$$

Looking at the two terms in $[\cdots]$ brackets, we can see that one of them will be zero, depending upon whether P = *e* or *i*. The other becomes the variance $\sigma^2_{\mathrm{PQ}}/N_{\mathrm{PQ}}$, from (58). This leaves us:

$$\tau_{\mathrm{Q}}\frac{d\Delta^{\mathrm{PQ}}}{dt} = -\Delta^{\mathrm{PQ}} + \frac{\rho_P \Phi^{\mathrm{PQ}} \psi^{\mathrm{PQ}} \sigma^2_{\mathrm{PQ}}}{N_{\mathrm{PQ}}}. \tag{68}$$

The negative sign in front of the Δ^{PQ} on the right-hand-side of this equation ensures that, for all σ^2_{PQ} (assuming for the moment that σ^2_{PQ} does not change), Δ^{PQ} will approach its equilibrium value.

We now need to look at the change in σ^2_{PQ}. However, linearizing in σ^2_{PQ} and Δ^{PQ} is not straightforward because of the covariances in (48), so we must look at the signs of the time-derivatives of the quantities using simple arguments. Equation (48) describes how σ^2_{PQ} changes with time. The covariance terms describe the correlation between quantities, and we can again determine the signs of these from inspection of the relevant equations.

7.3.1 Case of P = Q = e

Here, the two terms on the right-hand-side of (48) are identical. Let us look at the innermost covariance term, $\mathrm{cov}_k\left(\left\langle w^{ee}_{jk}\right\rangle_j, V^e_k\right)$, where the covariance is taken over postsynaptic neurons *k*. If this covariance is positive, it means physically that postsynaptic neurons with higher V_e values will have the higher average pre–post weight. If this is the case, it follows that the presynaptic neurons with these higher weights must also have higher V_e values, so that the increased excitatory effect from the higher-weight neurons drives up the soma potential of the postsynaptic neurons. Conversely, if the inner covariance is negative, this means that the postsynaptic neurons with the lowest V_e values would have the higher-than-average pre–post weight (i.e., low-firing neurons have the strongest weight). Therefore, the excitatory input from the presynaptic neurons to such postsynaptic neurons would be low. Therefore, we expect the outer covariance cov*, that is the covariance between the inner covariance and the postsynaptic soma-potential, also to be positive. This means that the rate of increase of variance σ^2_{ee} will be positive. This result requires only the assumption of higher soma-potentials giving higher firing rates, and therefore, we would expect it to be robust to changes in the models for evolution of the soma-potential. Moreover, it is reasonable to assume that a more highly structured cortex would have greater covariances and, therefore, more rapid changes in variance of weight.

7.3.2 Case of P = Q = i

Again, let us look at the innermost covariance term. Again, assume that this covariance is positive, so that the postsynaptic neurons with the highest soma potentials are also the ones with the highest average pre–post weights. However, because the presynaptic neurons are *inhibitory*, this must mean that the presynaptic neurons where the weights are maximum have a lower-than-average firing-rate and soma potential; otherwise, they would cause the postsynaptic soma potential to be *lower* than average. Therefore, we would expect the outer covariance cov* to be *negative*.

7.3.3 Case of P = e, Q = i

The two terms of (48) are no longer equal. However, by the same argument of the case P=Q= *e*, we would expect both terms to be positive—a positive correlation between the

weights and the postsynaptic soma potential must be as a result of a high firing (high soma potential) presynaptic neurons, and vice-versa [first term on right-hand side of (48)]. Similarly, a positive correlation between the weights and the presynaptic soma potential must imply high postsynaptic soma potential [second term on right-hand side of (48)]. Therefore, the outer covariance terms will be positive.

7.3.4 Case of P = i, Q = e

By the same argument as the case of P = Q = *i*, we would expect both terms to be negative. We note that, because we are using a symmetric learning rule, (1), we would expect $\sigma_{ei}^2 = \sigma_{ie}^2$. However, the arguments above would suggest that the former grows with time and the latter would fall with time. This contradiction is reconciled when we realize that the correlations between populations of *e* and *i* neurons would normally be significantly lower than the correlations between populations of *e* and *e* neurons, or *i* and *i* neurons [25]. In other words, we would expect the growth of these terms to be approximately zero in the low noise limit assumed by (43). Note also that we would not physically expect σ_{ei}^2, σ_{ie}^2, or σ_{ii}^2 to fall below zero. Overall, therefore, we might expect the equilibrium point to be unstable, given the growth in the term σ_{ee}^2.

References

1. Wilson, H.R., Cowan, J.D.: Excitatory and inhibitory interactions in localized populations of model neurons. Biophys. J. **12**, 1–24 (1972)
2. Nunez, P.L.: The brain wave function: a model for the EEG. Math. Biosci. **21**, 279–297 (1974)
3. Freeman, W.J.: Predictions on neocortical dynamics derived from studies in paleocortex. In: Basar, E., Bullock, T.H. (eds.) Induced Rhythms of the Brain, chap. 9, pp. 183–199. Birkhaeuser, Boston (1992)
4. Wright, J.J., Liley, D.T.J.: Dynamics of the brain at global and microscopic scales: neural networks and the EEG. Behav. Brain Sci. **19**, 285–316 (1996)
5. Robinson, P.A., Rennie, C.J., Wright, J.J.: Propagation and stability of waves of electrical activity in the cerebral cortex. Phys. Rev. E **56**, 826–840 (1997)
6. Liley, D.T.J., Cadusch, P.J., Wright, J.J.: A continuum theory of electro-cortical activity. Neurocomputers **26–27**, 795–800 (1999)
7. Rennie, C.J., Wright, J.J., Robinson, P.A.: Mechanisms for cortical electrical activity and emergence of gamma rhythm. J. Theor. Biol. **205**, 17–35 (2000)
8. Steyn-Ross, M.L., Steyn-Ross, D.A., Sleigh, J.W.: Modelling general anaesthesia as a first-order phase transition in the cortex. Prog. Biophys. Mol. Biol. **85**, 369–385 (2004)
9. Hutt, A., Bestehorn, M., Wennekers, T.: Pattern formation in intracortical neuronal fields. Network **14**, 351–368 (2003)
10. Kramer, M.A., Kirsch, H.E., Szeri, A.J.: Pathological pattern formation and epileptic seizures. J. R. Soc. Lond. Interface **2**, 113 (2005)
11. Chizhov, A.V., Graham, L.J., Turbin, A.A.: Simulation of neural population dynamics with a refractory density approach and a conductance-based threshold neuron model. Neurocomputing **70**(1–3), 252–262 (2006)
12. Bazhenov, M., Timofeev, I., Steriade, M., Sejnowski, T.J.: Model of thalamocortical slow-wave sleep oscillations and transitions to activated states. J. Neurosci. **22**, 8691–8704 (2002)
13. Compte, A., Sanchez-Vives, M.V., McCormick, D.A., Wang, X.J.: Cellular and network mechanisms of slow oscillatory activity (<1 Hz) and wave propagations in a cortical network model. J. Neurophysiol. **89**, 2707–2725 (2003)
14. Hill, S., Tononi, G.: Modeling sleep and wakefulness in the thalamocortical system. J. Neurophysiol. **93**, 1671–1698 (2005)
15. Robinson, P.A., Rennie, C.J., Rowe, D.L., O'Connor, S.C., Wright, J.J., Gordon, E., Whitehouse, R.W.: Neurophysical modeling of brain dynamics. Neuropsychopharmacology **28**, S74–S79 (2003)

16. Robinson, P.A., Rennie, C.J., Wright, J.J., Bahramali, H., Gordon, E., Rowe, D.L.: Prediction of electroencephalographic spectra from neurophysiology. Phys. Rev. E **63**, 021,903 (2001)
17. Wilson, M.T., Steyn-Ross, D.A., Sleigh, J.W., Steyn-Ross, M.L., Wilcocks, L.C., Gillies, I.P.: The k-complex and slow oscillation in terms of a mean-field cortical model. J. Comput. Neurosci. **21**, 243–257 (2006)
18. Bojak, I., Liley, D.T.J.: Modelling the effects of anaesthesia on the electroencephalogram. Phys. Rev. E **71**, 41902 (2005)
19. Wilson, M.T., Steyn-Ross, M.L., Steyn-Ross, D.A., Sleigh, J.W.: Predictions and simulations of cortical dynamics during natural sleep using a continuum approach. Phys. Rev. E **72**, 051910 1–14 (2005)
20. Bienenstock, E.L., Cooper, L.N., Munro, P.W.: Theory for the development of neuron selectivity: orientation specificity and binocular interation in visual cortex. J. Neurosci. **2**, 32–48 (1982)
21. Bienenstock, E., Lehmann, D.: Regulated criticality in the brain? Adv. Complex Systems **1**, 361–384 (1998)
22. Sandberg, A., Tegnér, J., Lansner, A.: A working memory model based on fast Hebbian learning. Netw. Comput. Neural Syst. **14**, 789–802 (2003)
23. Mongillo, G., Amit, D.J., Brunel, N.: Retrospective and prospective persistent activity induced by Hebbian learning in a recurrent cortical network. Eur. J. Neurosci. **18**, 2011–2024 (2003)
24. Hebb, D.O.: The Organization of Behaviour. Wiley, New York (1949)
25. Steyn-Ross, M.L., Steyn-Ross, D.A., Sleigh, J.W., Wilson, M.T., Wilcocks, L.C.: A mechanism for learning and memory erasure in a white-noise driven sleeping cortex. Phys. Rev. E **72**, 061,910 (2005)
26. Stetter, M.: Dynamic functional tuning of nonlinear cortical networks. Phys. Rev. E **73**, 031903 (2006)
27. Steyn-Ross, D.A., Steyn-Ross, M.L., Sleigh, J.W., Wilson, M.T., Gillies, I.P., Wright, J.J.: The sleep cycle modelled as a cortical phase transition. J. Biophys. **31**, 547–569 (2005)
28. Tononi, G., Cirelli, C.: Sleep function and synaptic homeostatis. Sleep Med. Rev. **10**, 49–62 (2006)
29. Mountcastle, V.B.: The columnar organization of the neocortex. Brain **120**, 701–722 (1997)
30. Sejnowski, T.J.: Storing covariance with nonlinearly interacting neurons. J. Math. Biol. **4**, 303–321 (1977)
31. Douglas, R.J., Martin, K.A.: Recurrent neuronal circuits in the neocortex. Curr. Biol. **17**(13), R496 (2007)
32. Thomson, A.M., Bannister, A.P.: Interlaminar connections in the neocortex. Cerebral Cortex **13**, 5–14 (2003)
33. Tononi, G., Sporns, O.: Measuring information integration. BMC Neurosci. **4**, 31 (2003)
34. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**(1), 47–97 (2002)
35. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastc Differential Equations. Springer, Berlin (1992)
36. Rudolph, M., Pospischil, M., Timofeev, I., Destexhe, A.: Inhibition determines membrane potential dynamics and controls action potential generation in awake and sleeping cat cortex. J. Neurosci. **27**(20), 5280–5290 (2007)
37. Blumenfeld, B., Preminger, S., Sagi, D.: Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity. Neuron **52**, 383–394 (2006)
38. Hopfield, J.J.: Neural networks and physical systems with emergent computational abilities. Proc. Natl. Acad. Sci. U. S. A. **78**, 2554–2558 (1982)
39. Hopfield, J.J.: Neurons with graded response have collective computational properties like those of two state neurons. Proc. Natl. Acad. Sci. U. S. A. **81**, 3088–3092 (1984)
40. Abraham, W.C., Robins, A.: Memory retention—the synaptic stability versus plasticity dilemma. Trends Neurosci. **28**(2), 73–78 (2005)
41. Horn, D., Levy, N., Ruppin, E.: Memory maintenance via neuronal regulation. Neural Comput. **10**, 1–18 (1998)
42. Pantic, L., Torres, J.J., Kappen, H.J., Gielen, S.C.A.M.: Associate memory with dynamic synapses. Neural Comput. **14**, 2903–2923 (2002)
43. Steriade, M., Núnez, A., Amzica, F.: A novel slow ($<$1 Hz) oscillation of neocortical neurons *in vivo*: depolarizing and hyperpolarizing components. J. Neurosci. **13**, 3252–3265 (1993)
44. Crochet, S., Chauvette, S., Boucetta, S., Timofeev, I.: Modulation of synaptic transmission in neocortex by network activities. Eur. J. Neurosci. **21**, 1030–1044 (2005)
45. Massimini, M., Rosanova, M., Mariotti, M.: EEG slow ($\sim$1 Hz) waves are associated with nonstationarity of thalamo-cortical sensory processing in the sleeping human. J. Neurophysiol. **89**, 1205–1213 (2003)

46. Steriade, M., Timofeev, I., Grenier, F.: Natural waking and sleep states: a view from inside neocortical neurons. J. Neurophysiol. **85**, 1969–1985 (2001)
47. Battaglia, F.P., Sutherland, G.R., McNaughton, B.L.: Hippocampal sharp wave bursts conincide with neocortical "up-state" transitions. Learn. Mem. **11**, 697–704 (2004)
48. Marshall, L., Helgadóttir, H., Mölle, M., Born, J.: Boosting slow oscillations during sleep potentiates memory. Nature **444**, 610–613 (2006)