Biol Cybern (2009) 101:427–444
DOI 10.1007/s00422-009-0346-1

ORIGINAL PAPER

Biological Cybernetics

Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks IV

Structuring synaptic pathways among recurrent connections

Matthieu Gilson · Anthony N. Burkitt · David B. Grayden · Doreen A. Thomas · J. Leo van Hemmen

Received: 23 April 2009 / Accepted: 27 October 2009 / Published online: 24 November 2009

Abstract In neuronal networks, the changes of synaptic strength (or weight) performed by spike-timing-dependent plasticity (STDP) are hypothesized to give rise to functional network structure. This article investigates how this phenomenon occurs for the excitatory recurrent connections of a network with fixed input weights that is stimulated by external spike trains. We develop a theoretical framework based on the Poisson neuron model to analyze the interplay between the neuronal activity (firing rates and the spike-time correlations) and the learning dynamics, when the network is stimulated by correlated pools of homogeneous Poisson spike trains. STDP can lead to both a stabilization of all the neuron firing rates (homeostatic equilibrium) and a robust weight specialization. The pattern of specialization for the recurrent weights is determined by a relationship between the input firing-rate and correlation structures, the network topology, the STDP parameters and the synaptic response properties. We find conditions for feed-forward pathways or areas with strengthened self-feedback to emerge in an initially homogeneous recurrent network.

M. Gilson (✉) · A. N. Burkitt · D. B. Grayden · D. A. Thomas
Department of Electrical and Electronic Engineering,
The University of Melbourne, Melbourne, VIC 3010, Australia
e-mail: mgilson@bionicear.org

M. Gilson · A. N. Burkitt · D. B. Grayden
The Bionic Ear Institute, 384-388 Albert St., East Melbourne,
VIC 3002, Australia

M. Gilson · A. N. Burkitt · D. B. Grayden · D. A. Thomas
NICTA, Victoria Research Lab, University of Melbourne,
Melbourne, VIC 3010, Australia

J. L. van Hemmen
Physik Department (T35) and BCCN–Munich, Technische Universität
München, 85747 Garching bei München, Germany

Keywords Learning · STDP · Recurrent neuronal network · Spike-time correlation

1 Introduction

Synaptic plasticity describes how "connections" between neurons are strengthened (potentiation) or weakened (depression) depending on the neuronal activity. Recent studies have established the importance of the timing of individual spikes in such learning mechanisms (Gerstner et al. 1996). Such spike-timing-dependent plasticity (STDP) has been the subject of both theoretical (Gerstner et al. 1996; Kempter et al. 1999; Gütig et al. 2003; Burkitt et al. 2004; Meffin et al. 2006; Appleby and Elliott 2006; Morrison et al. 2008) and experimental (Markram et al. 1997; Bi and Poo 2001; Sjöström et al. 2001) research, primarily focusing on the single-neuron case for slow learning and the resulting implications for the emergence of structure within feed-forward networks. Since the cortex is dominated by recurrent connections, a crucial step in understanding its behavior involves extending these studies to the case of recurrent networks. This has only begun to be addressed for STDP as well as for synaptic plasticity in general (Karbowski and Ermentrout 2002; Masuda and Kori 2007), mainly using numerical simulation (Song and Abbott 2001; Morrison et al. 2007). A particular effort was made to understand how STDP can modify synchrony properties in neuronal networks (Senn et al. 2002; Câteau et al. 2008; Lubenov and Siapas 2008). In recent articles (Burkitt et al. 2007; Gilson et al. 2009c), we have developed a theoretical framework to model the weight dynamics in a network with any arbitrary architecture and carried out the analysis for the case of a recurrently connected network with no external inputs, both for full and partial connectivity.

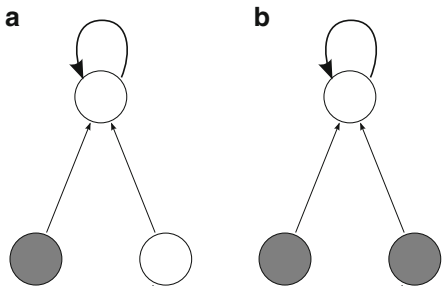


Fig. 1 Schematic illustration of two network configurations studied in this article. The neuronal network (*top circle*) has plastic recurrent connections (*thick arrows*) and is stimulated by two possibly correlated pools of external inputs (*bottom circles, filled* if correlated) with fixed input weights (*thin arrows*)

In this article, we analyze the more biologically interesting case in which the network is excited by external inputs and investigate how STDP can tune its response to the stimulation in an unsupervised fashion. In a similar manner to Gilson et al. (2009a) for learning on the input connections, we extend our previously developed framework (Burkitt et al. 2007) to incorporate the (short-time) effect of post-synaptic responses. We investigate how the weight structure develops within the network when stimulated by input pools with homogeneous firing rates and within-pool (but no between-pool) spike-time correlations, an idea inspired by Kempter et al. (1999). In particular, we focus on the case of two input pools when the input weights are kept fixed, as illustrated in Fig. 1.

After presenting the STDP model (Sect. 2.1) and neuron model (Sect. 2.2) that we use, we derive a dynamical system to describe the evolution of the *plastic* recurrent weights while the input connections are kept *fixed* (Sect. 2.3). We investigate the weight dynamics for a general network configuration (Sect. 3) and then focus on a specific topology, where the external inputs are divided into two homogeneous pools (Sect. 4).

2 Modeling learning and neuronal activity

2.1 Hebbian additive STDP

STDP describes the change in the synaptic weight induced by single spikes and pair of spikes while taking their precise timing into account. In this article, we use the so-called Hebbian additive STDP model (Kempter et al. 1999), keeping in mind its limitations compared to more elaborate STDP versions (van Rossum et al. 2000; Sjöström et al. 2001; Gütig et al. 2003; Meffin et al. 2006; Appleby and Elliott 2006). For two neurons in and out connected by a synapse $in \to out$ with weight J, the weight change δJ induced by a sole pair of pre- and post-synaptic spikes at times t^{in} and t^{out}, respectively, is determined by three additive contributions:

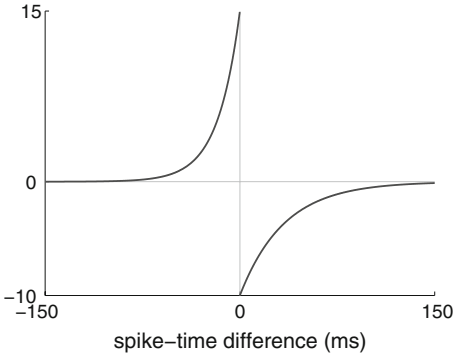


Fig. 2 Example of STDP window function W. It consists of one decaying exponential for potentiation (*left curve*) with time constant 17 ms and one for depression (*right curve*) with 34 ms. See Appendix C for details of the parameters

$$\delta J = \eta \begin{cases} w^{\text{in}} & \text{at time } t^{\text{in}} \\ w^{\text{out}} & \text{at time } t^{\text{out}} \\ W(t^{\text{in}} - t^{\text{out}}) & \text{at time } \max(t^{\text{in}}, t^{\text{out}}). \end{cases} \tag{1}$$

The constant w^{in} (resp. w^{out}) is a rate-based term and accounts for the effect of each pre-synaptic (post-synaptic) spike, which occurs at time t^{in} (t^{out}). The STDP learning window function W determines the last contribution (correlation term) depending on the difference between the spike times $t^{\text{in}} - t^{\text{out}}$ (Gerstner et al. 1996; Kempter et al. 1999). Figure 2 illustrates a typical choice of W, where pre-synaptic spikes that take part in the firing of post-synaptic spikes induce the potentiation (Hebb 1949). These three contributions are scaled by a learning parameter η, typically very small, so that learning occurs very slowly compared to the other neuronal and synaptic mechanisms. We chose η such that the weight change is three orders of magnitude below the corresponding upper bound of the weights. See Gilson et al. (2009a, Sec. 2.1) for details.

2.2 Poisson neuron model

In the Poisson neuron model (Kempter et al. 1999), the spiking mechanism of a given neuron i is approximated by an inhomogeneous Poisson process driven by an intensity function $\rho_i(t)$ in order to generate an output spike-time series $S_i(t)$. The rate function $\rho_i(t)$ is to be related to the soma potential and it evolves over time according to the excitation received from other neurons $j \neq i$ (self-connections are forbidden),

$$\rho_i(t) = \nu_0 + \sum_{j \neq i} \left[J_{ij}(t) \sum_n \epsilon \left(t - t_{j,n} - d_{ij} \right) \right]. \tag{2}$$

The constant ν_0 is the spontaneous firing rate (identical for all the neurons), which accounts for other incoming connections that are not considered in detail. Each pre-synaptic

spike induces a variation of $\rho_i(t)$ taken care of by the post-synaptic potential (PSP), which is determined by the synaptic weights J_{ij}, the post-synaptic response kernel ϵ, and the delays d_{ij}. The kernel function ϵ models the PSP due to the current injected into the post-synaptic neuron as a consequence of one single pre-synaptic spike; $\epsilon(t)$ is normalized to one: $\int \epsilon(t)\,\mathrm{d}t = 1$; and in order to preserve causality, we have $\epsilon(t) = 0$ for $t < 0$. The delays d_{ij} account for the axonal transmission such that the nth spike fired by neuron j at time $t_{j,n}$ reaches the synaptic site $j \to i$ at time $t_{j,n} + d_{ij}$; dendritic delays are not considered in this article. We only consider positive weights here, i.e., excitatory synapses. This neuron model proved to be useful to evaluate the weight dynamics, when learning is very slow compared to the neuronal activation dynamics (Kempter et al. 1999; Gütig et al. 2003). See Gilson et al. (2009a, Sec. 2.2) for more details.

2.3 Dynamical system to predict the weight evolution

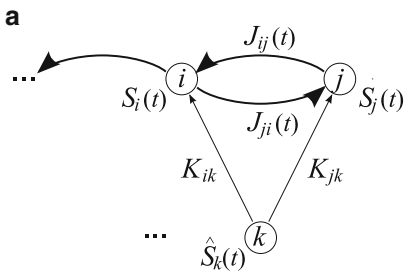
In the following, we present a dynamical system that links the variables of importance to describe the neuronal activity (firing rates and pairwise spike-time correlations) together with the synaptic weights, when STDP modifies the recurrent connections while the input connections are fixed. It is equivalent to the corresponding framework for the converse situation of fixed recurrent connections and plastic input weights developed by Gilson et al. (2009a, Sec. 2.3).

2.3.1 *Description of the network activity*

Let us consider a recurrently connected network of N Poisson neurons stimulated with M Poisson spike trains (a.k.a. external inputs or sources), as shown in Fig. 3a. In addition to receiving synaptic excitation from other neurons via connections that may form feedback loops in the network (but without self-connections), each neuron is also excited by some external inputs. Typically both M and N are large.

Spikes are considered to be instantaneous events compared to the time scale of other neuronal mechanisms (viz. ϵ, delays, etc.). We define $\hat{S}_k(t)$ as the spike-time series of the external input k, $1 \le k \le M$; its value is zero except at the times when a spike is fired and the spike train is described as a sum of Dirac delta-functions (Dirac comb). Likewise the spike-time series of the recurrent neurons are denoted by $S_i(t)$, $1 \le i \le N$.

Figure 3b recapitulates the variables of importance to describe the network activity, which correspond to expectation values for the network seen as a stochastic process. At time t, the firing rate averaged over a given duration T of neuron i is



b

	external inputs	input-to-neuron	network neurons
firing rates	$\hat{\nu}_k(t)$		$\nu_i(t)$
pairwise covariances	$\hat{C}_{kl}(t,u)$	$F_{ik}(t,u)$	$C_{ij}(t,u)$
weights		K_{ik}	$J_{ij}(t)$
delays		$\hat{d}_{ik}$	d_{ij}

Fig. 3 Presentation of the network and the notation. **a** Schematic representation of two of the N neurons (*top circles*, indexed by $1 \le i, j \le N$) and one of the M external inputs (*bottom circle*, $1 \le k \le M$). The recurrent connections have plastic weights $J_{ij}(t)$ (*thick arrows*) while the input connections have fixed weights K_{ik} (*thin arrows*). The spike trains of the input k and neuron i are denoted by $\hat{S}_k(t)$ and $S_i(t)$, respectively. **b** The table shows the variables that describe the neuronal activity: time-averaged firing rates $\hat{\boldsymbol{\nu}}$ and $\boldsymbol{\nu}$; time-averaged covariances $\hat{C}$, F, and C; and the variables related to the synaptic connections: weights K and J; delays $\hat{d}$ and d

$$\nu_i(t) := \frac{1}{T}\int_{t-T}^{t} \langle S_i(t')\rangle \,\mathrm{d}t', \tag{3}$$

where the angular brackets $\langle\cdots\rangle$ denote the ensemble average over the randomness, which is self-averaging (van Hemmen 2001). Likewise the time-averaged firing rate of input k is denoted by $\hat{\nu}_k(t)$. The time-averaged covariances are $C_{ij}(t,u)$ between neurons i and j (u being the time difference between the activities of neurons i and j); $F_{ik}(t,u)$ between neuron i and input k; and $\hat{C}_{kl}(t,u)$ between inputs k and l:

$$\begin{aligned} C_{ij}(t,u) &:= \frac{1}{T}\int_{t-T}^{t} \langle S_i(t')\, S_j(t'+u)\rangle \,\mathrm{d}t' \\ &\quad - \frac{1}{T}\int_{t-T}^{t} \langle S_i(t')\rangle \langle S_j(t'+u)\rangle \,\mathrm{d}t', \\ F_{ik}(t,u) &:= \frac{1}{T}\int_{t-T}^{t} \langle S_i(t')\, \hat{S}_k(t'+u)\rangle \,\mathrm{d}t' \\ &\quad - \frac{1}{T}\int_{t-T}^{t} \langle S_i(t')\rangle \langle \hat{S}_k(t'+u)\rangle \,\mathrm{d}t', \end{aligned} \tag{4}$$

$$\hat{C}_{kl}(t,u) := \frac{1}{T}\int_{t-T}^{t} \left\langle \hat{S}_k(t')\,\hat{S}_l(t'+u)\right\rangle \mathrm{d}t' - \frac{1}{T}\int_{t-T}^{t} \left\langle \hat{S}_k(t')\right\rangle \left\langle \hat{S}_l(t'+u)\right\rangle \mathrm{d}t'.$$

In order to take the delays into account in the learning, we will use the following covariance coefficients (Gilson et al. 2009a, Sect. 2.3.4), adapted from the correlation coefficients used by Burkitt et al. (2007):

$$\begin{aligned} C^{\Psi}_{ij}(t) &:= \int_{-\infty}^{+\infty} \Psi(u)\, C_{ij}(t, u-d_{ik})\,\mathrm{d}u, \\ F^{\Psi}_{ik}(t) &:= \int_{-\infty}^{+\infty} \Psi(u)\, F_{ik}(t, u-\hat{d}_{ik})\,\mathrm{d}u, \\ \hat{C}^{\Psi}_{kl}(t) &:= \int_{-\infty}^{+\infty} \Psi(u)\, \hat{C}_{kl}(t,u)\,\mathrm{d}u, \end{aligned} \tag{5}$$

where Ψ is a given kernel function.

The plastic recurrent weight from neuron j to neuron i is denoted by $J_{ij}(t)$ and the fixed input weight from input k to neuron i by K_{ik}. The network topology can be arbitrary, n^K being the number of input connections and n^J the number of recurrent connections. Partially connected networks are generated by randomly assigning input-to-neuron and neuron-to-neuron connections. The axonal delays d_{ij} and $\hat{d}_{ik}$ correspond to the definition in Sect. 2.2.

For the sake of simplicity, we use matrix notation in the remainder of the text: vectors $\boldsymbol{\nu}(t)$ and $\hat{\boldsymbol{\nu}}(t)$, and matrices $C^W(t)$, $F^W(t)$, $\hat{C}^W(t)$, $J(t)$, and K.

2.3.2 Slow evolution of the recurrent weights

For slow learning, we can choose the averaging duration T in (3) and (4) to satisfy both $T \ll \eta^{-1}$ and being much larger than the time scale of the synaptic activation mechanisms (separation of time scales). We can then use the ensemble average $\langle\cdots\rangle$ over the randomness, which is self-averaging (van Hemmen 2001), to obtain a differential equation in the expectation value of the recurrent weight $\dot{J}_{ij}(t)$, in a similar way to the derivation by Gilson et al. (2009a, Sec. 2.3.3) for the input weights,

$$\dot{J}_{ij}(t) \simeq \eta\left[w^{\mathrm{in}}\nu_j(t) + w^{\mathrm{out}}\nu_i(t) + \widetilde{W}\nu_j(t)\nu_i(t) + C^W_{ij}(t)\right], \tag{6}$$

where we have used the approximation $\nu_j(t-d_{ij}) \simeq \nu_j(t)$. The constant $\widetilde{W}$ is the integral value of the STDP learning window function W,

$$\widetilde{W} := \int_{-\infty}^{+\infty} W(u)\,\mathrm{d}u. \tag{7}$$

We have used the assumption that each neuron receives inputs from many external inputs and other recurrently connected neurons. This implies weak probabilistic interdependence in the recurrent network required when taking the ensemble average (Burkitt et al. 2007).

The separation of time scales between the "fast" neuronal and synaptic activation mechanisms on the one hand, and the "slow" learning dynamics ($\eta \ll 1$) on the other hand is used in the remainder of this section in order to express $\boldsymbol{\nu}(t)$ and $C^W(t)$ in terms of the input parameters, $\hat{\boldsymbol{\nu}}(t)$ and $\hat{C}(t)$, and of the synaptic weights K and $J(t)$. Such consistency equations for $\boldsymbol{\nu}(t)$ and $C^W(t)$ describe the constraint due to the recurrent connectivity and lead to a dynamical system of the general form

$$\dot{J} = \eta\,\mathbb{F}\left(K, J, \nu_0, \hat{\boldsymbol{\nu}}, \hat{C}\right). \tag{8}$$

The consistency equation for $\boldsymbol{\nu}(t)$ can be found in a companion paper (Gilson et al. 2009a, Eq. 18a) and we focus in the remainder of this section on that for $C^W(t)$.

2.3.3 Remarks on the correlation structure

As explained in Appendix A.1 and B.2, the covariances $C_{ij}(t,u)$ and $\hat{C}_{kl}(t,u)$ by convention do not incorporate the atomic (or point) discontinuity at $u=0$ due to the autocorrelation of the stochastic point processes S_i and $\hat{S}_k$ for $i=j$ and $k=l$, respectively, namely $\langle S_i(t)\rangle\,\delta(u)$ and $\langle \hat{S}_k(t)\rangle\,\delta(u)$, where δ is the Dirac delta function. In other words, we discriminate between the input correlation structure embodied by C and $\hat{C}$ ("spiking information") and the autocorrelation intrinsic to the neuron model.

This implies in particular that, in the learning equation (6) for the weight J_{ii}, the term related to STDP should incorporate the extra term $W(d_{ii})\,\nu_i(t)$ in addition to $C^W_{ii}(t)$, letting aside the learning rate η. However, since we forbid self-connections, we have $J_{ii}=0$ at all times and we can ignore the second term in the learning equation (although it will appear in calculations of Appendix B for the sake of generality).

The matrix coefficient C^V, as defined in (5) for $\Psi = V : u \mapsto W(-u)$, satisfies $C^V = (C^W)^{\mathbf{T}}$ for fixed inputs when the recurrent weights are quasi constant in time (see Appendix A.2); the superscript '$\mathbf{T}$' denotes the transposition. This property is the key in the derivation of the dynamical system below.

The same property applies to the matrix $\hat{C}$ and, for homogeneous pairwise correlation of spikes within an input pool, the elements of the matrix $\hat{C}^W$ corresponding to the input pool are symmetric.

2.3.4 Short durations of the PSP kernel and recurrent delays

The derivation of the consistency equation for C^W is detailed in Appendix B and uses previous results presented in a companion paper (Gilson et al. 2009a). The case of non-identical delays could be rigorously dealt with using Fourier analysis (Hawkes 1971). However, this method does not lead to an easily tractable solution for arbitrary PSP kernel ϵ and distribution of delays. We consider in this article the simplified case of almost identical recurrent delays ($d_{ij} \simeq d$ for all connections $j \to i$) and almost identical input delays ($\hat{d}_{ik} \simeq \hat{d}$ for all connections $k \to i$). The effect of the PSP kernel ϵ and of the recurrent delays d_{ij} can be evaluated when their two distributions in time are narrow in comparison to the width of the learning window W, using the approximation (61) in Appendix B, which gives

$$\begin{aligned} C^W(t) = [\mathbb{1}_N - J(t)]^{-1} \\ \times \left\{ K \left[\hat{C}^{W*\zeta}(t) + [W * \zeta](0)\ \mathrm{diag}\left(\hat{\boldsymbol{\nu}}(t)\right) \right] K^{\mathbf{T}} \right. \\ \left. + W(d)\ \mathrm{diag}\left(\boldsymbol{\nu}(t)\right) \right\} [\mathbb{1}_N - J(t)]^{-1\,\mathbf{T}} \\ - W(d)\ \mathrm{diag}\left(\boldsymbol{\nu}(t)\right). \end{aligned} \tag{9}$$

Equation 9 describes a spatial and temporal filtering on the input covariance $\hat{C}$ to obtain the neuron covariance C. The network connectivity operates through the term $(\mathbb{1}_N - J)^{-1}\ K$ that appears twice in (12b); the same term was found in the consistency equation for the neuron-to-input covariance F by Gilson et al. (2009a, Eq. 16 in Sect. 2.3.5), where it appears once. The function ζ describes the temporal filtering of the PSP kernel function on $\hat{C}$ to obtain C, as does ϵ for F (Gilson et al. 2009a); this effect was ignored in a previous study by Burkitt et al. (2007). The function ζ can be approximated by the self-convolution of ϵ, as illustrated in Fig. 4a,

$$\zeta(r) \simeq \int \epsilon(r + r')\epsilon(r')\mathrm{d}r'; \tag{10}$$

see (64) in Appendix B.3 for details. The following approximation is made in the derivation (Gilson et al. 2009a, Sec. 2.3.5)

$$\int W(u - r)\epsilon(r - d)\,\mathrm{d}r \ \simeq\ W(u). \tag{11}$$

2.3.5 The equations describing the dynamical system

In the limit of large networks ($N \gg 1$ and $M \gg 1$), we can ignore the effects due to autocorrelation, i.e., the terms with 'diag' in (9). The system of equations that describes the dynamics of the firing rates, the covariance coefficients, and the weights reduces to

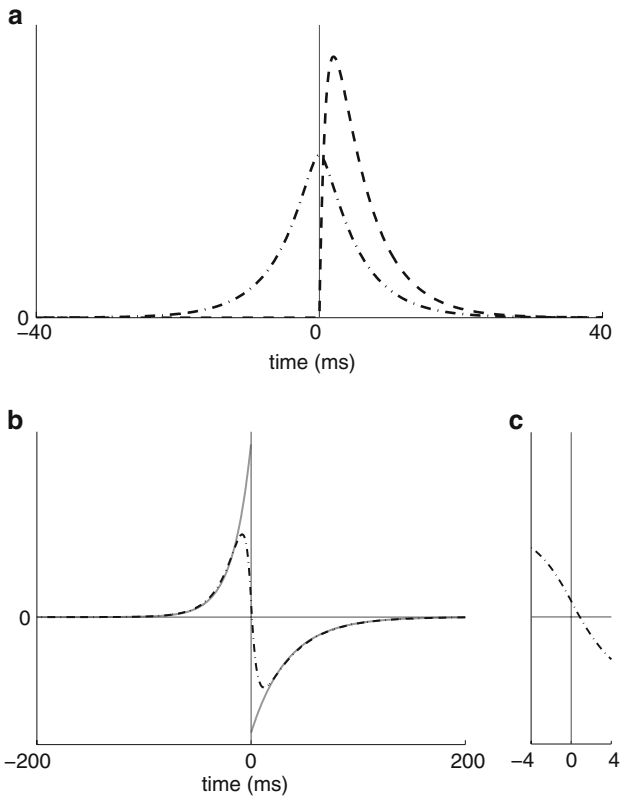


Fig. 4 (**a**) Plots of ϵ (*dashed line*) and ζ (*dashed-dotted line*). (**b**) Plots of W (*gray thick solid line*) and $W * \zeta$ (*black dashed-dotted line*). Globally, the shapes of the two functions are similar, except for small u. We used the parameters listed in Appendix C, for which $W(u) < 0$, but $[W * \zeta](u) > 0$ for very small $u > 0$, see insert (**c**)

$$\boldsymbol{\nu} = (\mathbb{1}_N - J)^{-1} \left(\nu_0\, \mathbf{e} + K\, \hat{\boldsymbol{\nu}}\right), \tag{12a}$$

$$C^W = (\mathbb{1}_N - J)^{-1}\ K\, \hat{C}^{W*\zeta}\, K^{\mathbf{T}}\, (\mathbb{1}_N - J)^{-1\,\mathbf{T}}, \tag{12b}$$

$$\dot{J} = \Phi_J \left(w^{\mathrm{in}}\, \mathbf{e}\, \boldsymbol{\nu}^{\mathbf{T}} + w^{\mathrm{out}}\, \boldsymbol{\nu}\, \mathbf{e}^{\mathbf{T}} + \widetilde{W}\, \boldsymbol{\nu}\, \boldsymbol{\nu}^{\mathbf{T}} + C^W \right). \tag{12c}$$

Time has been rescaled to remove η and the time variable t is omitted from all the vectors and matrices that evolve over time. The N column vector $\mathbf{e}$ has all elements equal to one

$$\mathbf{e} := [1, \ldots, 1]^{\mathbf{T}}. \tag{13}$$

The superscript '$\mathbf{T}$' is the matrix transposition so that, e.g., $\mathbf{e}\,\mathbf{e}^{\mathbf{T}}$ is a $N \times N$ matrix. The projector Φ_J operates on the vector space of $N \times N$ matrices: it nullifies all the matrix components that correspond to missing connections in the network. In particular, Φ_J nullifies all the diagonal terms since we forbid self-connection of the neurons; this assumption is not necessary for the following analysis though.

2.3.6 Generation of the 'delta-correlated' input spike trains

The inputs that stimulate the network are partitioned into a given number of pools, such that the input spike trains within a pool are correlated but independent from inputs from

different pools. The firing rates of inputs within a pool are all identical. Positive within-pool correlation is generated so that, for any input, a given portion of its spikes occur at the same time as some other spikes within its pool, while the remainder occur at independent times (Gütig et al. 2003; Meffin et al. 2006). In this way, we obtain input spike trains $\hat{S}_k(t)$ with "instantaneous" firing rates $\langle \hat{S}_k(t) \rangle = \hat{\nu}_0$ and pairwise covariances $\hat{C}_{kl}(t,u)$ defined in (4)

$$\hat{C}_{kl}(t,u) \simeq \hat{c}\,\hat{\nu}_0\,\delta(u), \tag{14}$$

where $0 \leq \hat{c} \leq 0.1$ is the correlation strength (chosen to be small) and δ is the Dirac delta-function; hence, the name 'delta-correlated' inputs. See Gilson et al. (2009a, Sect. 2.3.7) for more details.

We thus have for inputs $k \neq l$ within the same correlated pool

$$\hat{C}_{kl}^{W*\zeta} \simeq \hat{c}\,\hat{\nu}_0\,[W * \zeta](0). \tag{15}$$

The sign of the elements of $\hat{C}^{W*\zeta}$ can be either positive or negative depending on the value $[W * \zeta](0)$. For our choice of parameters detailed in Appendix C, we have

$$[W * \zeta](0) = \frac{c_P \tau_P [\tau_A \tau_B + \tau_P(\tau_A + \tau_B)]}{2(\tau_A + \tau_B)(\tau_A + \tau_P)(\tau_B + \tau_P)} - \frac{c_D \tau_D [\tau_A \tau_B + \tau_D(\tau_A + \tau_B)]}{2(\tau_A + \tau_B)(\tau_A + \tau_D)(\tau_B + \tau_D)}, \tag{16}$$

which is of the form $\alpha c_P - \beta c_D$, where the coefficients α and β are positive. Consequently, when taking c_P sufficiently larger than c_D, we obtain $[W * \zeta](0) > 0$. This corresponds to stronger potentiation compared to depression at the origin, as illustrated in Fig. 4.

2.3.7 Analysis of the system dynamics

We aim to study the steady states of the network variables, as well as their stability, through the system of equations (12a–12c) that describes the evolution of the expectation value of these network variables, i.e., the first order of the stochastic process, which we will call *drift*. We will ignore *higher orders* of the stochastic dynamics, for example those related to auto-correlation effects that were studied in previous companion papers (Gilson et al. 2009b,c). The term *mean* (applied to firing rates, weights, etc.) will refer to an average over the neurons, inputs, connections, etc. of the network (topological averaging), whereas *averaged* refers to time averaging, unless otherwise specified. The term *homeostatic equilibrium* will refer to the situation where the mean firing rate and mean weights have reached an equilibrium, although individual firing rates and weights may continue to change. The expression *emergence of weight structure* will refer to the situation where the learning dynamics has imposed a specific weight structure on the network, i.e., further learning may cause individual weights to change but the qualitative character of the distribution (e.g., bimodal) will remain unchanged.

Considerations of the invertibility of the matrix $\mathbb{1}_N - J(t)$ have been discussed previously (Burkitt et al. 2007; Gilson et al. 2009c, Sect. 2.3.2). Because we use the Poisson neuron with a linear activation function, the matrix of recurrent weights J must have eigenvalues in the unit circle; otherwise, the feedback is too strong and the firing rates diverge towards infinity. To enforce this property at all times, we introduce bounds on individual weights in numerical simulation. All the simulation results presented in this paper were run using the neuron and learning parameters listed in Appendix C.

3 General case of learning on the recurrent weights with fixed input weights

In this section we analyze the general solution of (12a–12c), which for fixed inputs (Sect. 2.3.6) and non-learning input weights K reduces to

$$\boldsymbol{\nu} = (\mathbb{1}_N - J)^{-1}\,\widetilde{\boldsymbol{\nu}}, \tag{17a}$$

$$\dot{J} = \Phi_J \Big[w^{\text{in}} \mathbf{e}\boldsymbol{\nu}^{\mathbf{T}} + w^{\text{out}} \boldsymbol{\nu}\mathbf{e}^{\mathbf{T}} + \widetilde{W}\boldsymbol{\nu}\boldsymbol{\nu}^{T} + (\mathbb{1}_N - J)^{-1}\,\widetilde{C}\,(\mathbb{1}_N - J)^{-1\,\mathbf{T}} \Big], \tag{17b}$$

where the following vector and matrix absorb the input parameters

$$\widetilde{\boldsymbol{\nu}} := \nu_0\,\mathbf{e} + K\,\hat{\boldsymbol{\nu}}, \qquad \widetilde{C} := K\hat{C}^{W*\zeta}K^{\mathbf{T}}. \tag{18}$$

We show how STDP applied on the recurrent connections can induce both

- neither saturated nor quiescent stable firing rate for each neuron, which implies stability for the mean incoming weight;
- a specialization of the recurrent weights through splitting of the outgoing weight distribution for each neuron.

Note that the stability of the firing rates here is different from that mentioned in Sect. 2.3.7 and only relates to the learning dynamics.

3.1 Homeostatic equilibrium

We first examine the case of homeostatic equilibrium in the network, namely the situation when the means of the firing rates and of the weights over the whole network have reached an equilibrium. The mean value for a variable averaged over inputs, neurons or connections will be denoted using the subscript 'av'. The following differential equation for J_{av} is derived from (17b)

$$\dot{J}_{\text{av}} \simeq \left(w^{\text{in}} + w^{\text{out}}\right) \nu_{\text{av}} + \left(\widetilde{W} + \frac{\widetilde{C}_{\text{av}}}{\widetilde{\nu}_{\text{av}}^2}\right) \nu_{\text{av}}^2, \tag{19}$$

where we have used the following approximation that comes from the averaging of (17a)

$$\left(1 - n_{\text{av}}^J J_{\text{av}}\right)^{-1} \simeq \frac{\nu_{\text{av}}}{\widetilde{\nu}_{\text{av}}}, \tag{20}$$

$n_{\text{av}}^J := n^J/N$ being the average number of presynaptic recurrent connections for the neurons. This equation has the same form as that in the case with no external inputs and can be analyzed in a similar manner (Burkitt et al. 2007; Gilson et al. 2009c), the change lying in replacement of $\widetilde{W}$ by $\widetilde{W} + \widetilde{C}_{\text{av}}/\widetilde{\nu}_{\text{av}}^2$. The fixed point $(\nu_{\text{av}}^*, J_{\text{av}}^*)$ is

$$\nu_{\text{av}}^* = -\frac{w^{\text{in}} + w^{\text{out}}}{\widetilde{W} + \widetilde{C}_{\text{av}}/\widetilde{\nu}_{\text{av}}^2}, \; J_{\text{av}}^* = \frac{\nu_{\text{av}}^* - \widetilde{\nu}_{\text{av}}}{n_{\text{av}}^J \nu_{\text{av}}^*}. \tag{21}$$

Provided the homeostatic equilibrium is realizable, i.e., the mean firing rate and mean weight have positive equilibrium values, it is stable if and only if

$$\widetilde{W} + \frac{\widetilde{C}_{\text{av}}}{\widetilde{\nu}_{\text{av}}^2} < 0. \tag{22}$$

For weak correlation, this condition reduces to $\widetilde{W} < 0$. In order to ensure that the equilibrium mean firing rate is positive ($\nu_{\text{av}}^* > 0$), we require in addition that $w^{\text{in}} + w^{\text{out}} > 0$. These two conditions are the same as for the case of no external inputs (Burkitt et al. 2007). Note that the weight equilibrium is realizable only if $\nu_{\text{av}}^* > \widetilde{\nu}_{\text{av}}$ (where $\widetilde{\nu}_{\text{av}} \simeq \nu_0 + n_{\text{av}}^K K_{\text{av}} \hat{\nu}_{\text{av}}$), which requires $w^{\text{in}} + w^{\text{out}}$ to be sufficiently large.

3.2 Learning the input correlation structure

We now show that the stabilization corresponding to the homeostatic equilibrium actually holds for all the individual neurons. In addition, STDP can also cause the weights to diverge according to the input correlation structure, thus implementing a robust specialization in a similar way to the case of learning on the input connections (Kempter et al. 1999; Gilson et al. 2009a).

We decompose $\widetilde{C}$ into two components, one proportional to $\widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^{\mathbf{T}}$ and its complement

$$\begin{aligned} &\widetilde{C} = \widetilde{C}_{\parallel} + \widetilde{C}_{\perp}, \\ &\widetilde{C}_{\parallel} := c_{\parallel}\, \widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^{\mathbf{T}} \quad \text{with } c_{\parallel} = \frac{\widetilde{\boldsymbol{\nu}}^{\mathbf{T}} \widetilde{C} \widetilde{\boldsymbol{\nu}}}{(\widetilde{\boldsymbol{\nu}}^{\mathbf{T}} \widetilde{\boldsymbol{\nu}})^2}, \\ &\widetilde{\boldsymbol{\nu}}^{\mathbf{T}} \widetilde{C}_{\perp} \widetilde{\boldsymbol{\nu}} = 0, \end{aligned} \tag{23}$$

where $\widetilde{\boldsymbol{\nu}}$ is defined in (18). Recall that the matrix J belongs to the vector subspace of $\mathbb{R}^{N\times N}$ defined by $\mathbb{M}_J := \{X \in \mathbb{R}^{N\times N}, \Phi_J(X) = X\}$, whose dimension is the number of existing connections n^J.

3.2.1 Sufficient condition for existence of fixed points

We first analyze the special case where $\widetilde{C}_{\perp}$=0. Here, (17a–17b) reduce to the same form as that obtained in the case of no external inputs (Gilson et al. 2009c, Eq. 10), where $\nu_0\mathbf{e}$ and $\widetilde{W}$ are replaced by $\widetilde{\boldsymbol{\nu}}$ and $\widetilde{W}'$, respectively, with

$$\widetilde{W}' := \widetilde{W} + c_{\parallel}. \tag{24}$$

In particular, the fixed points $(\boldsymbol{\nu}^*, J^*)$ of the dynamics correspond to homogeneous neuron firing rates

$$\boldsymbol{\nu}^* = \mu' \mathbf{e}, \tag{25a}$$

$$\left(\mathbb{1}_N - J^*\right) \boldsymbol{\nu}^* = \widetilde{\boldsymbol{\nu}}, \tag{25b}$$

where

$$\mu' := -\frac{w^{\text{in}} + w^{\text{out}}}{\widetilde{W}'}. \tag{26}$$

For weak input correlations, μ' can be approximated by the equilibrium value μ for uncorrelated inputs

$$\mu' \simeq \mu := -\frac{w^{\text{in}} + w^{\text{out}}}{\widetilde{W}}, \tag{27}$$

which means that the firing rates are in this case close to the equilibrium value corresponding to uncorrelated inputs. This characterizes all the fixed points provided each neuron in the network is part of a loop of synaptic connections (Gilson et al. 2009c, Sec. 3.1).

The manifold of all fixed points J^* denoted by $\mathcal{M}^*$ is contained in an affine subspace of $\mathbb{M}_J$ of dimension $n^J - N$, where n^J is the number of recurrent connections, according to (25b). Note that the matrix $\mathbb{1}_N - J^*$ must be invertible to be a valid fixed point (Gilson et al. 2009c, Sec. 2.3.2). When the fixed-point manifold $\mathcal{M}^*$ is attractive, STDP causes J to converge towards $\mathcal{M}^*$, where it evolves due to higher orders of the stochastic process, the drift $\dot{J}$ (cf. Sect. 2.3.7) being zero on $\mathcal{M}^*$. For homogeneous recurrent connections, sufficient conditions such that $\mathcal{M}^*$ is attractive are

$$w^{\text{in}} \gg |w^{\text{out}}| \quad \text{and} \quad \widetilde{W} < 0; \tag{28}$$

refer to Gilson et al. (2009c, Sec. 3.3) for more details of the analysis and on higher stochastic orders than the drift, such as the resulting weight dispersion.

The condition $\widetilde{C}_{\perp} = 0$ occurs in two particular situations: for uncorrelated inputs where $\widetilde{C} = 0$; and in the case of homogeneous input connections since we have then $\widetilde{\boldsymbol{\nu}} \propto \mathbf{e}$ and $\widetilde{C} \propto \mathbf{e}\mathbf{e}^{\mathbf{T}}$. In the first case, the recurrent weights J can compensate for inhomogeneous input firing rates $\hat{\boldsymbol{\nu}}$, causing the neuron firing rates $\boldsymbol{\nu}$ to become homogeneous: the *incoming* recurrent weights of the less stimulated neurons become potentiated and those of the more stimulated neurons become depressed. In both cases, however, the

asymptotic distribution of the *outgoing* recurrent weights is not constrained by STDP and strongly depends on the initial conditions, analogous to what happens in the case of no external inputs. Details are provided by Gilson et al. (2009c, Sec. 4.2).

3.2.2 Weight dynamics for arbitrary matrix $\widetilde{C}_\perp$

The analysis above indicates that a non-zero input correlation structure and inhomogeneous input weights are required so that the recurrent network organizes in a way that represents the input structure, i.e., at least differently from the weight dynamics for the case of no external inputs. In other words, the interesting case in terms of weight specialization corresponds to the absence of a fixed point for the dynamical system.

When $\widetilde{C}_\perp \neq 0$, the equation $\dot{J} = 0$ in (17b) may have no solution, but the evolution of the recurrent weights J can still be related to the manifold $\mathcal{M}^*$. Equation 17b can be rewritten

$$\dot{J} = \Phi_J \Big[-w^{\text{out}} \mathbf{e}\mathbf{u}^{\mathbf{T}} - w^{\text{in}} \mathbf{u}\mathbf{e}^{\mathbf{T}} + \widetilde{W}' \mathbf{u}\mathbf{u}^{\mathbf{T}} + (\mathbb{1}_N - J)^{-1} \widetilde{C}_\perp (\mathbb{1}_N - J)^{-1\,\mathbf{T}} \Big], \tag{29}$$

where the vector $\mathbf{u} := \boldsymbol{\nu} - \mu' \mathbf{e}$ evaluates for each neuron the difference between its firing rate and the common equilibrium value. For weakly correlated inputs, $\widetilde{C}_\perp$ is small and, since $(\mathbb{1}_N - J)^{-1}$ is kept invertible with bounded norm, the last term in (29) can be considered to be a small perturbation on the evolution of $\mathbf{u}$. When the conditions (28) such that the manifold $\mathcal{M}^*$ is attractive are met, there exist initial conditions on the recurrent weights such that $\mathbf{u}$ will converge towards zero and remains in a neighborhood of zero for bounded perturbations $\widetilde{C}_\perp$. This means that individual neuron firing rates are then all quasi-stable and close to the equilibrium value $\mu' \simeq \mu$, and that J converges towards the vicinity of $\mathcal{M}^*$, provided the input correlations are sufficiently small. A rigorous analysis would consider the domain of attraction for $\mathbf{u}$, namely the set of initial recurrent weights for which J will be driven towards $\mathcal{M}^*$. We assume that this domain is sufficiently large based on our study of the homeostatic equilibrium in Sect. 3.1, which suggests that homogeneous initial recurrent weights will converge towards $\mathcal{M}^*$.

While remaining in the neighborhood of $\mathcal{M}^*$, $\widetilde{C}_\perp \neq 0$ may cause ongoing structural evolution of the weights J when it becomes the leading order while the sum of the terms involving $\mathbf{u}$ in (29) converge to a quasi-equilibrium around zero. However, the analysis is difficult in the general case and we will only consider a simple but biologically relevant special case.

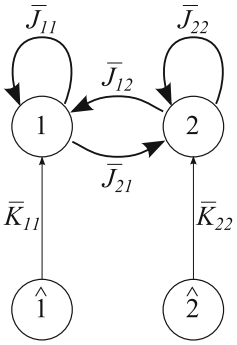


Fig. 5 The recurrently connected neurons are divided into two groups (*top circles*), each being stimulated by one pool of external inputs (*bottom circles*). The pools and groups have homogeneous characteristics within each. The *overline* and the *subscripts* '1' and '2' correspond to the mean variables ($\hat{\boldsymbol{\nu}}$, $\boldsymbol{\nu}$, K, J, ...) over each pool of external inputs, group of neurons, etc.

4 Network with two distinct input pathways

We now consider a network with two homogeneous input pools that each excite half of the recurrently connected neurons, as illustrated in Fig. 5. The input pools have homogeneous characteristics (viz. firing rates and correlations) within each of them and no correlation between them; the neuron groups also have homogeneous characteristics. The connectivity is homogeneous from pools to groups and between groups. This network topology can be obtained after symmetry breaking by applying STDP on the input connections with balanced correlation strength between the two input pools (Gütig et al. 2003; Gilson et al. 2009b). We show in this section how the recurrent weights can then organize in an unsupervised way according to the input correlation structure, leading to the emergence of functional organization.

Similar to the case of learning input weights K with fixed recurrent weights J (Gilson et al. 2009a, Appendix B.1), symmetries in the network allow us to reduce the dimensionality of the vector space $\mathbb{M}_J$ where the matrix J evolves, in order to study the drift of J. For example, we define the variable $\bar{J}_{12}$ as the sum of the incoming weight from group 2 averaged over the neurons of group 1, which gives, for full recurrent connectivity,

$$\bar{J}_{12} \simeq \frac{2}{N} \sum_{1 \leq i \leq N/2} \sum_{N/2+1 \leq j \leq N} J_{ij}. \tag{30}$$

Likewise, $\bar{\boldsymbol{\nu}}_1$ is the mean firing rate of group 1. The variables in the reduced space correspond to equivalence classes defined modulo redistributions of incoming weights that do not modify the drift $\dot{J}$: in other words, two weight matrixes J and J' are in the same class $\bar{J}$ if they induce the same drift $\dot{J} = \dot{J}'$ expressed in (17b). For the topology described in Fig. 5, these variables are

$$\bar{K} = \begin{pmatrix} \bar{K}_{11} & 0 \\ 0 & \bar{K}_{22} \end{pmatrix},$$
$$\overline{\widetilde{C}} = [W * \zeta](0) \begin{pmatrix} \bar{K}_{11}^2 \hat{c}_1 \bar{\hat{\boldsymbol{\nu}}}_1 & 0 \\ 0 & \bar{K}_{22}^2 \hat{c}_2 \bar{\hat{\boldsymbol{\nu}}}_2 \end{pmatrix}, \tag{31}$$
$$\bar{J} = \begin{pmatrix} \bar{J}_{11} & \bar{J}_{12} \\ \bar{J}_{21} & \bar{J}_{22} \end{pmatrix},$$

$$\bar{\boldsymbol{\nu}} = \begin{pmatrix} \bar{\nu}_1 \\ \bar{\nu}_2 \end{pmatrix}.$$

The expression for $\overline{\overline{C}}$ with the correlation strengths $\hat{c}_1$ and $\hat{c}_2$ comes from (15) and is an approximation for large N.

The inverse of $\mathbb{1}_2 - \bar{J}$ is

$$\left(\mathbb{1}_2 - \bar{J}\right)^{-1} = \frac{1}{\Theta_{\bar{J}}} \begin{pmatrix} 1 - \bar{J}_{22} & \bar{J}_{12} \\ \bar{J}_{21} & 1 - \bar{J}_{11} \end{pmatrix}, \tag{32}$$

where $\Theta_{\bar{J}}$ is the determinant of $\mathbb{1}_2 - \bar{J}$:

$$\Theta_{\bar{J}} := \left(1 - \bar{J}_{11}\right)\left(1 - \bar{J}_{22}\right) - \bar{J}_{12}\bar{J}_{21}. \tag{33}$$

Substituting (32) in (17a), we obtain

$$\begin{pmatrix} \bar{\nu}_1 \\ \bar{\nu}_2 \end{pmatrix} = \frac{1}{\Theta_{\bar{J}}} \begin{pmatrix} (1 - \bar{J}_{22})\tilde{\bar{\nu}}_1 + \bar{J}_{12}\tilde{\bar{\nu}}_2 \\ \bar{J}_{21}\tilde{\bar{\nu}}_1 + (1 - \bar{J}_{11})\tilde{\bar{\nu}}_2 \end{pmatrix}. \tag{34}$$

The learning equation (17b) becomes

$$\begin{aligned} \dot{\bar{J}} &= w^{\text{in}} \begin{pmatrix} \bar{\nu}_1 & \bar{\nu}_2 \\ \bar{\nu}_1 & \bar{\nu}_2 \end{pmatrix} + w^{\text{out}} \begin{pmatrix} \bar{\nu}_1 & \bar{\nu}_1 \\ \bar{\nu}_2 & \bar{\nu}_2 \end{pmatrix} \\ &\quad + \widetilde{W} \begin{pmatrix} \bar{\nu}_1^2 & \bar{\nu}_1\bar{\nu}_2 \\ \bar{\nu}_1\bar{\nu}_2 & \bar{\nu}_2^2 \end{pmatrix} + \Omega, \\ \Omega &:= \frac{[W * \zeta](0)}{\Theta_{\bar{J}}^2} \begin{pmatrix} 1 - \bar{J}_{22} & \bar{J}_{12} \\ \bar{J}_{21} & 1 - \bar{J}_{11} \end{pmatrix} \\ &\quad \begin{pmatrix} \bar{K}_{11}^2 \hat{c}_1 \tilde{\bar{\nu}}_1 & 0 \\ 0 & \bar{K}_{22}^2 \hat{c}_2 \tilde{\bar{\nu}}_2 \end{pmatrix} \begin{pmatrix} 1 - \bar{J}_{22} & \bar{J}_{21} \\ \bar{J}_{12} & 1 - \bar{J}_{11} \end{pmatrix}. \end{aligned} \tag{35}$$

Since the weights J are all positive and the spectrum of the matrix J is in the unit circle (Burkitt et al. 2007; Gilson et al. 2009c, Appendix A), we have $0 \leq \sum_j \bar{J}_{ij} < 1$ for all i.

When the correlations strengths $\hat{c}_1$ and $\hat{c}_2$ are non-zero, the expression for $\overline{\overline{C}}$ in (31) generally corresponds to $\overline{\overline{C}}_\perp \neq 0$; hence, the equation $\dot{\bar{J}} = 0$ has no solution except for specific choices of learning parameters. In the remainder of this section, we assume that the stability conditions in (28) are met and the correlation strengths $\hat{c}_1$ and $\hat{c}_2$ are small. As explained in Sect. 3.2.2, (35) causes the firing rates $\bar{\nu}_1$ and $\bar{\nu}_2$ to stabilize over time close to $\mu' \simeq \mu$, as illustrated in Fig. 6; the weight matrix J thus remains in the neighborhood of $\mathcal{M}^*$. This implies in particular that $\bar{J}_{11}\bar{\nu}_1 + \bar{J}_{12}\bar{\nu}_2 = \bar{\nu}_1 \simeq \text{const.}$, and consequently $\bar{J}_{11}$ and $\bar{J}_{12}$ will evolve in opposite directions. Likewise, $\bar{J}_{21}\bar{\nu}_1 + \bar{J}_{22}\bar{\nu}_2 = \bar{\nu}_2 = \text{const.}$ and $\bar{J}_{21}$ and $\bar{J}_{22}$ will diverge from each other. These two diverging behaviors are determined by the matrix $\overline{\overline{C}}_\perp$ when it is non-zero, which induces a specialization of the recurrent connections. In the following, we analyze Ω directly in order to study the specialization scheme, instead of $\overline{\overline{C}}_\parallel$ and $\overline{\overline{C}}_\perp$ separately.

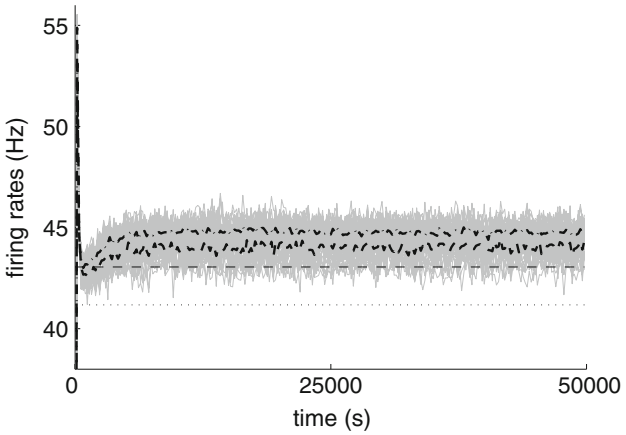


Fig. 6 Evolution of the neuron firing rates. The network consisted of $N = 60$ neurons such that each half (#1–30 and #31–60) was stimulated by just one pool of 30 inputs, as illustrated by Fig. 5. The input firing rates were $\hat{\nu}_1 = 35\,\text{Hz}$ and $\hat{\nu}_2 = 30\,\text{Hz}$; only input pool $\hat{1}$ had correlation ($\hat{c}_1 = 0.1$). After a transient from quiescence at time $t = 0$ up to 55 Hz, the individual firing rates (*gray bundle*) quickly converged towards the predicted equilibrium value μ' (calculated using (21), *black thin dashed line*), close to the value μ corresponding to uncorrelated inputs (*black thin dotted line*), and then remained quasi-homogeneous in the neighborhood of μ'. The *thick dashed line* (*bottom*) represents the mean firing rate for group 1 and the *dashed-dotted line* (*top*) the mean for group 2

4.1 One input pool with spike-time correlation and one uncorrelated pool

We first consider the case where only one pool has homogeneous non-zero spike-time correlation ($\hat{c}_1 > 0$) while the other has none ($\hat{c}_2 = 0$). The term of the learning equation (35) related to $\overline{\overline{C}}$ is

$$\Omega = \frac{[W * \zeta](0)\bar{K}_{11}^2 \hat{c}_1 \tilde{\bar{\nu}}_1}{\Theta_{\bar{J}}^2} \begin{pmatrix} (1-\bar{J}_{22})^2 & (1-\bar{J}_{22})\bar{J}_{21} \\ (1-\bar{J}_{22})\bar{J}_{21} & \bar{J}_{21}^2 \end{pmatrix}. \tag{36}$$

Subtracting the second term from the first term in the first line of (36) to evaluate the evolution of $\bar{J}_{11} - \bar{J}_{12}$, we obtain

$$\Omega_{11} - \Omega_{12} = \frac{[W * \zeta](0)\bar{K}_{11}^2 \hat{c}_1 \tilde{\bar{\nu}}_1}{\Theta_{\bar{J}}^2}(1-\bar{J}_{22})(1-\bar{J}_{22}-\bar{J}_{21}). \tag{37}$$

The invertibility of $\mathbb{1}_N - J$ to ensure that the firing rates $\bar{\boldsymbol{\nu}}$ do not diverge implies that $\bar{J}_{22} + \bar{J}_{21} < 1$; it follows that $\bar{J}_{22} < 1$ also holds. Consequently, the evolution of $\bar{J}_{11}$ and $\bar{J}_{12}$ is determined by the sign of $[W * \zeta](0)$: there is potentiation of $\bar{J}_{11}$ at the expense of $\bar{J}_{12}$ for $[W * \zeta](0) > 0$, as illustrated in Fig. 7.

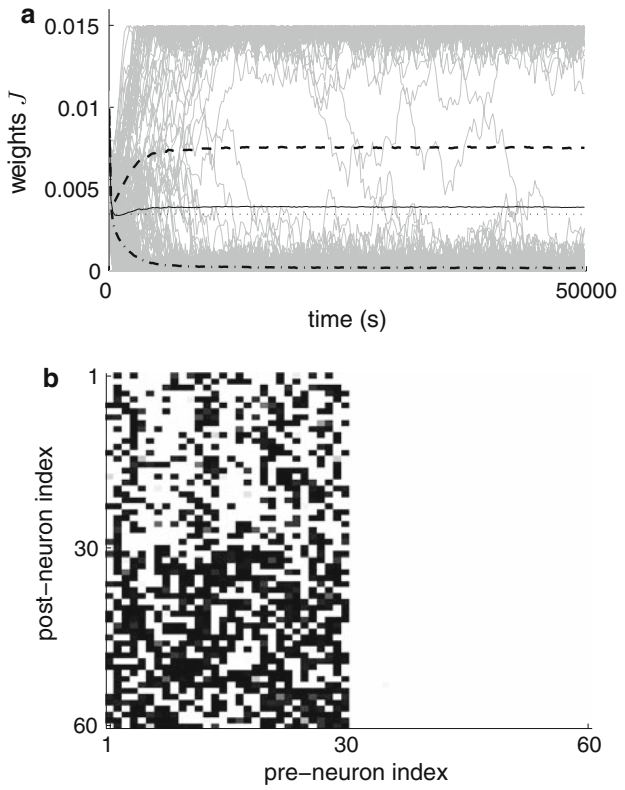


Fig. 7 Strengthening of the outgoing weights of the neuron group that receives correlated input. The network has the same parameters as that in Fig. 6. **a** Evolution of the recurrent weights J. The individual weights (*gray bundle*, only a representative portion is plotted) diverged towards the bounds, while the homeostatic equilibrium (J_{av} in *black thin solid line*) was satisfied close to the predicted equilibrium value (*black thin dotted line*) calculated using (21). The recurrent weights coming from the first half ($\bar{J}_{11}+\bar{J}_{21}$ in *thick dashed line*) increased at the expense of those from the second half ($\bar{J}_{12}+\bar{J}_{22}$ in *thick dashed-dotted line*), i.e., the weights coming out of the first group that received correlated inputs were potentiated. **b** Weight matrix J after the emergence of the structure. Darker pixels indicate potentiated weights. Weights on the left side (corresponding to the mean in *dashed line*) were more potentiated than those on the right side

Likewise for $\bar{J}_{21}-\bar{J}_{22}$, subtracting the second term from the first term in the second line of (36) leads to

$$\Omega_{21}-\Omega_{22}=\frac{[W*\zeta](0)\bar{K}_{11}^2\hat{c}_1\bar{\hat{\nu}}_1}{\Theta_{\bar{J}}^2}\,\bar{J}_{21}(1-\bar{J}_{22}-\bar{J}_{21}), \tag{38}$$

and when $[W*\zeta](0)>0$, $\bar{J}_{21}$ will be potentiated and $\bar{J}_{22}$ depressed.

Putting it all together, neuron group 1, which receives correlated input, has its outgoing weights $\bar{J}_{11}$ and $\bar{J}_{21}$ potentiated when $[W*\zeta](0)>0$. The converse situation occurs when $[W*\zeta](0)<0$. Note that the particular value $W(0)$ does not play any role in this analysis.

In any case, the weights will diverge due to the drifts in (37) and (38) until reaching the bounds. The distribution of the neuron firing rates may become bimodal but the discrepancies between $\bar{\nu}_1$ and $\bar{\nu}_2$ remain small for inputs with small delta-correlation, as illustrated in Fig. 6. Recall that the discrepancies between the individual firing rates and $\mu'\simeq\mu$ (weak correlation) relate to the absence of a fixed point for the dynamical system. The same applies for the homeostatic equilibrium of the weights in Fig. 7a.

In summary, the input correlation structure determines the asymptotic distribution of the recurrent weights by strengthening (or weakening) the outgoing weights of the group that receives correlated inputs when $[W*\zeta](0)$ is positive (negative). The divergence of the weights ($\bar{J}_{11}$ and $\bar{J}_{21}$ vs. $\bar{J}_{12}$ and $\bar{J}_{22}$) induces a robust specialization, cf. Fig. 7b. These results are similar to those described by Gilson et al. (2009a, Sect. 4) when STDP is applied on the input connections with fixed recurrent weights: stability of some components of the weight matrix (in that case, the mean incoming weight for each neuron) in parallel to a diverging behavior corresponding to a splitting between the different weight sets according to the input correlations.

This specialization described above contrasts with the "redistribution" of the incoming weights in the case of uncorrelated inputs, which results in homogeneous equilibrium firing rates for the neurons, as explained in Sect. 3.2.1. In particular, when $\hat{\nu}_1>\hat{\nu}_2$, we have

$$\left(1-\bar{J}_{22}-\bar{J}_{21}\right)\bar{\hat{\nu}}_1=\left(1-\bar{J}_{11}-\bar{J}_{12}\right)\bar{\hat{\nu}}_2, \tag{39}$$

which means that the incoming weights to group 2, $\bar{J}_{22}+\bar{J}_{21}$, are potentiated at the expense of the weights to group 1, $\bar{J}_{22}+\bar{J}_{21}$. This weight compensation is a consequence of an equilibrium, hence it is weaker than the potentiation resulting from the diverging behavior due to input correlation.

4.2 Two input pools with balanced spike-time correlations

For $\hat{c}_1>0$ and $\hat{c}_2>0$, we use an equivalent equation to (36) for $\hat{c}_2$ by permuting the indices in order to obtain an equation similar to (37) that gives the direction of the evolution of $\bar{J}_{11}-\bar{J}_{12}$,

$$\begin{aligned}\Omega_{11}-\Omega_{12}=&\frac{[W*\zeta](0)\bar{K}_{11}^2\hat{c}_1\bar{\hat{\nu}}_1}{\Theta_{\bar{J}}^2}(1-\bar{J}_{22})(1-\bar{J}_{22}-\bar{J}_{21})\\&-\frac{[W*\zeta](0)\bar{K}_{22}^2\hat{c}_2\bar{\hat{\nu}}_2}{\Theta_{\bar{J}}^2}\,\bar{J}_{12}(1-\bar{J}_{11}-\bar{J}_{12}).\end{aligned} \tag{40}$$

Consider balanced input firing rates equal to $\hat{\nu}_{\text{av}}$, balanced correlation strengths equal to $\hat{c}_{\text{av}}$, and balanced input weights $\bar{K}_{11}=\bar{K}_{22}$. For homogeneous initial recurrent weights equal to J_{av}, (40) reduces to

$$\Omega_{11} - \Omega_{12} = \frac{[W * \zeta](0)\bar{K}_{11}^2 \hat{c}_{\mathrm{av}} \hat{\nu}_{\mathrm{av}}}{\Theta_{\bar{J}}^2} (1 - n_{\mathrm{av}}^J J_{\mathrm{av}})^2. \quad (41)$$

Consequently, for $[W * \zeta](0) > 0$, $\bar{J}_{11}$ will be initially potentiated at the expense of $\bar{J}_{21}$. Likewise, $\bar{J}_{22}$ will be initially potentiated at the expense of $\bar{J}_{12}$. When starting from homogeneous recurrent weights, the weight evolution satisfies $\bar{J}_{11} \simeq \bar{J}_{22}$ and $\bar{J}_{12} \simeq \bar{J}_{21}$ over time. This leads to

$$\Omega_{11} - \Omega_{12} = \frac{[W * \zeta](0)\bar{K}_{11}^2 \hat{c}_{\mathrm{av}} \hat{\nu}_{\mathrm{av}}}{\Theta_{\bar{J}}^2} (1 - \bar{J}_{11} - \bar{J}_{12})^2. \quad (42)$$

As a result, $\bar{J}_{11}$ becomes increasingly potentiated and $\bar{J}_{12}$ depressed. The condition $[W * \zeta](0) > 0$ implies the strengthening of the within-group connections due to the input correlation when starting from homogeneous initial weights J, as illustrated in Fig. 8. This conclusion still holds when there are small inhomogeneities between the input firing rates, correlations, or input weights. In the converse situation, when $[W * \zeta](0) < 0$, the within-group connections were weakened and the between-group connections were strengthened.

Recall that the analysis of Sect. 2.3 assumed very short recurrent delays d. Simulations run using $d = 0.4 \pm 0.2$ ms showed the expected outcomes described above for both $[W * \zeta](0)$ positive and negative. However, for longer delays $d \simeq 3$–50 ms, results similar to that in Sect. 4.1, or even the opposite of the expected behavior for $[W * \zeta](0) > 0$ were observed, i.e., depression instead of potentiation of the within-group weights. The specialization observed in numerical simulation was weaker in this case than that described in Sect. 4.1. The desired strengthening of within-group connections was obtained for larger recurrent delays ($d = 3 \pm 1$ ms) using a different learning window function W shifted such that $W(u) = 0$ for $u = t^{\mathrm{in}} - t^{\mathrm{out}} = 1$ ms, which corresponds to potentiation around the origin, as illustrated in Fig. 9. The potentiation observed in simulations is weaker for larger recurrent delays; shifting the curve of W more to the right allows the use of longer delays. This indicates the importance of the shape of W around the origin when interacting with the narrowly correlated neuronal activity within the network, due to the narrow within-pool input correlations. This is illustrated in Fig. 10, which shows that the correlation extends over the domain from -20 to $+20$ ms. The discrepancies between the theoretical prediction and the simulation result relates to the approximation made to derive (9), which only considered the first order of recurrence for the feedback connections. Involving higher order of recurrence implies iterated convolutions of ζ with the delayed PSP kernel ϵ (and its mirrored function), which leads to a more spread predicted curve; this explains why the actual distribution is less peaked and broader (the integrals of both curves are comparable).

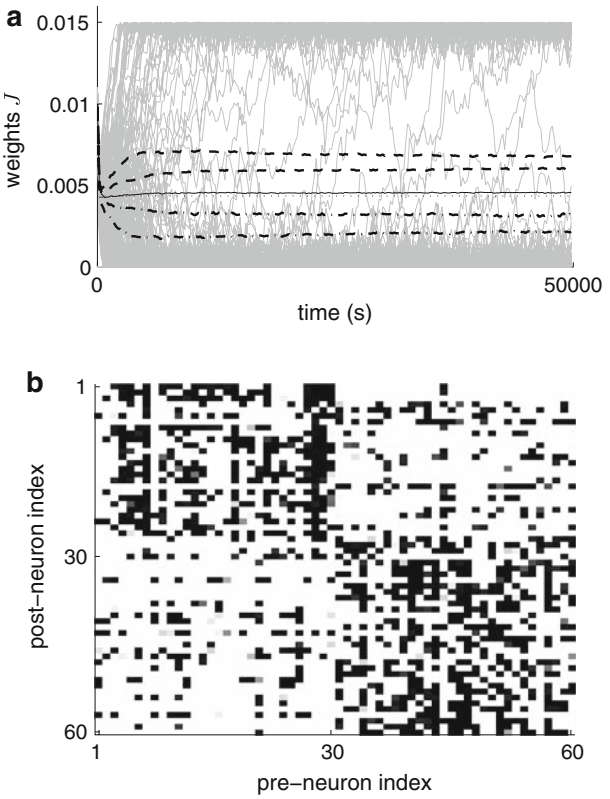


Fig. 8 Within-group strengthening of the recurrent connections due to stimulation by correlated inputs. The network and figure are similar to Fig. 7, except that the two input pools have the same firing rate $\hat{\nu}_1 = \hat{\nu}_2 = 30$ Hz and balanced within-pool correlation ($\hat{c}_1 = \hat{c}_2 = 0.1$). **a** Evolution of the recurrent weights J. The individual weights (*gray bundle*, only a representative portion is plotted) diverged towards the bounds, while the homeostatic equilibrium was satisfied (mean J_{av} in *black thin solid line*, prediction in *black thin dotted line*). The two means of the within-group connections ($\bar{J}_{11}$ and $\bar{J}_{22}$ in *dashed lines*) were potentiated while those of the between-group connections ($\bar{J}_{12}$ and $\bar{J}_{21}$ in *dashed-dotted lines*) were depressed. **b** Matrix J after the emergence of the weight structure. Darker pixels indicate potentiated weights. The weights in the top-left and bottom-right quarters (corresponding to the two means in *dashed line*) were more potentiated than those in the top-right and bottom-left quarters

5 Discussion

In this article, we have presented a mathematical framework to analyze the weight dynamics in neuronal networks where the recurrent connections are subject to STDP. The main novelty compared to previous work (Burkitt et al. 2007; Gilson et al. 2009c) lies in incorporating the effect of the PSP time course and delays, which allows us to predict the behavior of networks excited by external pulse trains. The derivation of the neuron covariance self-consistency equation (9) is a cornerstone of this analysis, which was made tractable by using the Poisson neuron model Kempter et al. (1999). This equation is crucial to evaluate spike-driven effects of STDP in recurrent networks, which cannot be captured by rate-based

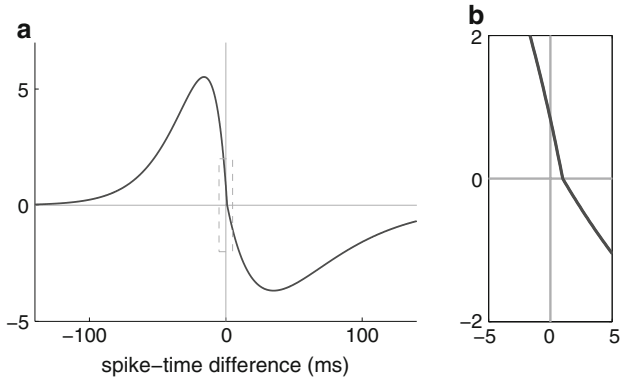


Fig. 9 **a** Example of a different STDP window function W. Each branch is an alpha function with the same time constants 17 ms (potentiation) and 34 ms (depression) as for Fig. 2. **b** Enlarged portion of (**a**) around the origin (**a**: *dashed box*) illustrating that the curve has been shifted to the right such that its zero corresponds to $u = 1$ ms (see insert on the right)

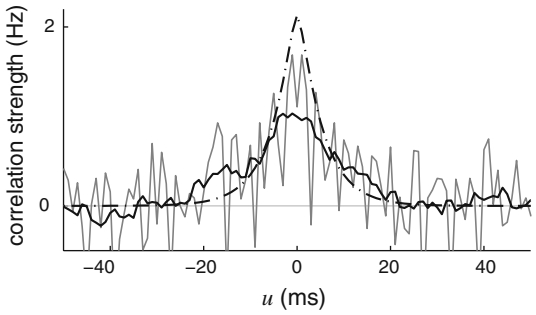


Fig. 10 Cross-correlogram (*black thin line*) for two neurons from the same group in the network of Fig. 8 with $d = 0.1$ ms simulated for 100 s (without learning). The time bin of the histogram is 1 ms and the graph has been rescaled accordingly. The *black thick line* represents the curve smoothed over 10 ms. The *dashed-dotted line* indicates the theoretical prediction at the first order of recurrence using (9)

learning. We then used a fixed-point analysis to predict the evolution of the neuron firing rates and the emergence of an asymptotic weight structure in a network stimulated by fixed external inputs with non-learning input weights and a correlation structure.

Stability of the neuron firing rates, which also implies that of the mean incoming weight for each neuron, can be obtained for a wide range of learning parameters. In this way, the asymptotic weights are neither all saturated nor all quiescent, which is necessary to generate an effective weight specialization. For weak input correlations, sufficient conditions are $w^{\text{in}} > |w^{\text{out}}| > 0$ and $\widetilde{W} < 0$. They correspond, respectively, to an increase of the weight due to each pre-synaptic spike, by a greater amount than the effect of each single post-synaptic spike (either potentiation or depression), and to more depression than potentiation induced by the STDP window function W for uncorrelated inputs (negative integral value). This is in agreement with the theoretical analysis of the learning dynamics in a recurrently connected network with no external inputs (Burkitt et al. 2007; Gilson et al. 2009c) and earlier studies of numerical simulations of recurrently connected integrate-and-fire neurons (Song and Abbott 2001; Morrison et al. 2007).

The rate-based learning constants w^{in} and w^{out} were used in order to obtain homeostatic equilibrium for the weights, so that a structure could emerge depending on the input correlation. Unlike the STDP learning window function W, we did not choose them relying on experimental results; note that they do not impair the local character of the learning rule (no global information is required). We expect the stability conclusions to hold in most cases for similar stabilizing mechanisms, such as weight normalization (van Rossum et al. 2000) or a suitable weight-dependency for W (van Rossum et al. 2000; Gütig et al. 2003), so long as the resulting weight dynamics leads to an effective homeostatic equilibrium.

In order to obtain a non-trivial specialization of the recurrent weights during the firing rate equilibrium, a network topology is necessary where different neuron groups receive distinct inputs with correlation. Otherwise, the weight dynamics is equivalent to that in a network with no external inputs: the weight drift is driven by a rate-based rule and no significant weight structure emerges Gilson et al. (2009c). When conditions are met such as those described in Sects. 4.1 and 4.2, the individual weights exhibit strong competition that can result in the emergence of a feed-forward synaptic pathway or the strengthening of within-group connections for learning on J, as illustrated in Fig. 11 for two input pools. Very short recurrent delays were required to obtain within-group strengthening of recurrent connections (Sect. 4.2), which corresponds to the assumption in Sect. 2.3.4 that was made in order to derive the dynamical system (12a–12c). No reliable trend was found in numerical simulation when using larger delays; further study is required to understand the weight dynamics for longer post-synaptic response and recurrent delays, which were shown to be of interest to optimize information processing by neurons (Pfister et al. 2006; Toyoizumi et al. 2007).

This robust specialization scheme, which results from both the (partial) homeostatic equilibrium and the diverging behavior related to spike-time correlations, is determined by $\hat{C}^{W*\zeta}$ in (12b). Rate-based learning cannot generate such mixed weights dynamics. This matrix embodies the interplay between the correlation structure of the external inputs, the STDP window function W, and the PSP response kernel ϵ. The asymptotic weight distribution will be determined by the input correlations, when they are sufficiently large, rather than the input firing rates. For the network configurations in Fig. 11, the durations of ϵ and of the delays played a crucial role when STDP modifies the recurrent connections. This is in contrast to the case of plastic input connections (Gilson et al. 2009a,b) where the weights specialize in the

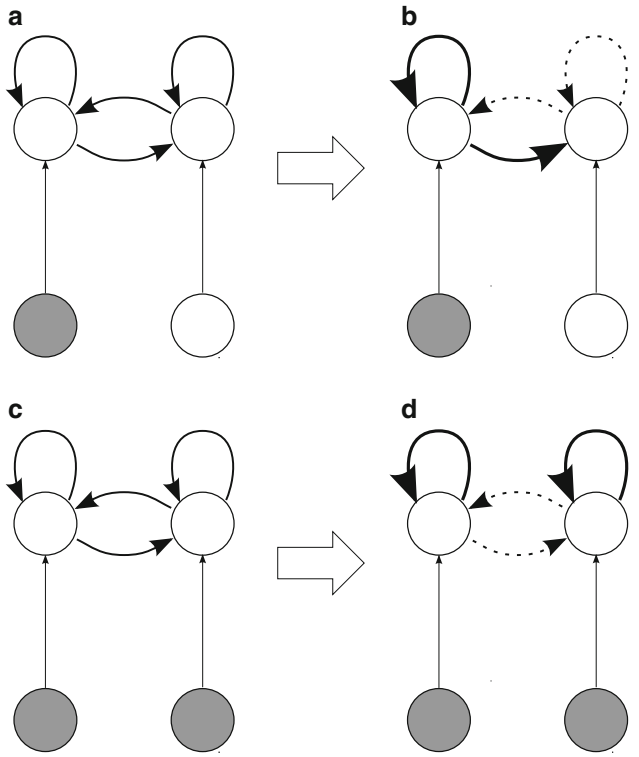


Fig. 11 Schematic representation of the recurrent weight specialization (**a** & **c**) before and (**b** & **d**) after the learning epoch for two input configurations (top vs. bottom). In each case, among the two sets of incoming connections to each neuron group (*top circle*) in the network, one becomes potentiated at the expense of the other. (**a** ⇒ **b**) When one input pool dominates in terms of spike-time correlations (*filled bottom circle*) compared to the other (*open bottom circle*), the neuron group that receives more correlations takes over in the recurrent network through the potentiation of its outgoing recurrent weights (*very thick arrow*), while those of the other group are depressed (*dashed thick arrow*). (**c** ⇒ **d**) For two input pools with balanced spike-time correlations (*filled bottom circles*), the within-group recurrent connections are potentiated (*very thick arrow*) while the between-group connections are depressed (*dashed thick arrow*)

same fashion irrespective of the delays and shapes of ϵ and W for Hebbian STDP. The different schemes of potentiation vs. depression that were observed depending upon $\hat{C}^{W*\zeta}$ may explain the contradictory behaviors observed in numerical simulations, which generated debate about whether STDP induces more (Izhikevich et al. 2004) or less (Iglesias et al. 2005) synchronization in recurrent networks. For example, in a network where nearby neurons have more chance to be connected than distant neurons, stronger synchrony can be obtained by potentiating local connections, which corresponds to the within-group connections in our network (nearby neurons receiving similar external inputs).

We also expect the following behavior when considering the generalization to the case where the network receives inputs from more than two pools with spike-time within-pool correlations, but no between-pool correlation:

- Differences in the input firing-rate excitation received by the neuron groups (given by $K\hat{\boldsymbol{\nu}}$) lead to a redistribution of the *incoming* recurrent weights; the incoming weights of more stimulated groups will be depressed compared to those of other groups.
- Neuron groups that receive strong (positive) spike-time correlation will experience a potentiation of their *outgoing* weights and then dominate the recurrent weight structure in the network, provided the conditions on $W * \zeta$ and on the recurrent delays are met. The second trend usually overrides the first one, since it relies on a diverging behavior.

Future directions. The present study deliberately made minimal assumptions on the network topology and on the input firing-rate and correlation structures (coincident spiking times, i.e., narrow correlation distribution between inputs) in order to reproduce self-organization schemes. The introduction of other mechanisms, some of which are mentioned below, is likely to introduce further complexity into the dynamics, where STDP would be the driving force underlying the specialization of the synaptic weights. This may, in particular, help to lift the constraint of very short recurrent delays required to obtain within-group strengthening among recurrent connections.

Non-linear neuronal activation mechanisms may also play a significant role in determining the neuron covariance structure. For sigmoidal Poisson neurons, a loop expansion technique can be used to evaluate the firing rates and the correlations around a fixed point of the dynamics (Roberts 2004), which corresponds to the expansion of $[\mathbb{1}_N - J]^{-1}$ for our version of the Poisson neuron model, cf. Appendix B.3. A challenge is to apply this framework to more complex neuron models, such as a Poisson neuron with non-linear activation function and the integrate-and-fire neuron (Burkitt 2006; Moreno-Bote et al. 2008).

In this article as well as in the companion papers of this series, we constrained our study to using only narrow distributions of axonal delays and did not investigate the evolution of synchronization between neurons. Previous work showed that STDP can induce non-trivial synchrony structure between neurons (Câteau et al. 2008), and that dendritic delays play an important role (Senn et al. 2002; Lubenov and Siapas 2008). The present framework can be adapted to incorporate these aspects.

Weight-dependence of STDP modifies the dynamics of the synaptic weight compared to additive STDP, in particular concerning their stabilization: a stable unimodal weight distribution can be obtained (van Rossum et al. 2000; Gütig et al. 2003; Morrison et al. 2007). This can be desirable in some cases, e.g., initial unimodal weight distribution and uncorrelated inputs. An interesting question is whether a single STDP rule can combine such a non-specialization for uncorrelated

inputs together with the qualitative properties of the weight splitting presented in this article for correlated inputs. This will be the subject of a subsequent study.

Acknowledgements The authors are greatly indebted to Chris Trengove, Sean Byrnes, Hamish Meffin, Michael Eager, and Paul Friedel for their constructive comments. They are also grateful to Iven Mareels, Konstantin Borovkov, Dragan Nesic, and Barry Hughes for helpful discussions. MG is funded by scholarships from the University of Melbourne and NICTA. MG also benefited from an enjoyable stay at the Physik Department (T35) of the Technische Universität München. LvH gratefully acknowledges a most enjoyable stay at the Department of Electrical and Electronic Engineering at the University of Melbourne. LvH is partially supported by the BCCN–Munich. Funding is acknowledged from the Australian Research Council (ARC Discovery Project #DP0771815). The Bionic Ear Institute acknowledges the support it receives from the Victorian Government through its Operational Infrastructure Support Program.

Appendix A: Remarks on the input covariance structure

A.1 Definition of the external input covariance

In (4) the following definition for the covariance between two external inputs k and l is used

$$\begin{aligned}&\mathrm{Cov}[\hat{S}_k(t), \hat{S}_l(t+u)]\\ &\quad := \big\langle \hat{S}_k(t)\hat{S}_k(t+u)\big\rangle - \big\langle \hat{S}_k(t)\big\rangle\big\langle \hat{S}_l(t+u)\big\rangle. \end{aligned} \tag{43}$$

The inputs $\hat{S}_k$ are second-order stationary processes, which means that these functions $\mathrm{Cov}[\hat{S}_k(t), \hat{S}_l(t+u)]$ are constant in t. Similar to Hawkes (1971), we take the convention that all the $\mathrm{Cov}[\hat{S}_k(t), \hat{S}_l(t+u)]$ are continuous at $u = 0$, which means that they do not include the atomic discontinuity for $u = 0$ and $k = l$ due to the autocorrelation of the stochastic point-processes $\hat{S}_k$. We refer to 'complete covariance' for the second moment that includes the extra contribution $\langle \hat{S}_k(t)\rangle\delta(u)$ for each pair $k = l$, where δ is the Dirac delta function and $\langle \hat{S}_k(t)\rangle$ the constant firing rate.

This convention aims to discriminate between the intrinsic covariance resulting from autocorrelation (always present even for uncorrelated inputs) and the correlation structure that encodes spike synchronization. For uncorrelated inputs, the matrix $\hat{C}(t,u)$ defined in (4) satisfies $\hat{C}(t,u) = 0$ for all $u \in \mathbb{R}$. Therefore, it can be related to the spike-timing information conveyed by the external inputs and encoded in their covariance. However, in the derivation of the self-consistency covariance equations (9), we incorporate terms related to the autocorrelation of the external inputs in order to assess their impact on the learning dynamics.

A.2 Properties of the matrix $\hat{C}^W$

Since the stochastic processes $\hat{S}_k(t)$ have time-invariant first and second stochastic moments, we have $\langle \hat{S}_k(t)\rangle = \text{const.}$ and

$$\big\langle \hat{S}_k(t)\,\hat{S}_l(t+u)\big\rangle = \big\langle \hat{S}_l(t)\,\hat{S}_k(t-u)\big\rangle \tag{44}$$

for all indices k and l, which implies $\hat{C}(t,u) = \hat{C}^{\mathbf{T}}(t,-u)$ because the input firing rates $\langle \hat{S}_k(t)\rangle$ are constant for all k. When convoluting with a given kernel $\Psi(t)$, we obtain

$$\begin{aligned}\int_{-\infty}^{+\infty} \Psi(-u)\,\hat{C}(t,u)\,\mathrm{d}u &= -\int_{+\infty}^{-\infty} \Psi(u)\,\hat{C}(t,-u)\,\mathrm{d}u\\ &= \int_{-\infty}^{+\infty} \Psi(u)\,\hat{C}^{\mathbf{T}}(t,u)\,\mathrm{d}u, \end{aligned} \tag{45}$$

where we have used a change of variable $u \to -u$. In particular, this implies that $\hat{C}^V = (\hat{C}^W)^{\mathbf{T}}$ for the time-reverse of the STDP window function $V(u) = W(-u)$.

Moreover, for homogeneous pairwise correlation of inputs, we have

$$\big\langle \hat{S}_k(t)\,\hat{S}_l(t+u)\big\rangle = \big\langle \hat{S}_l(t)\,\hat{S}_k(t+u)\big\rangle, \tag{46}$$

which implies that the function $u \mapsto \langle \hat{S}_k(t)\hat{S}_l(t+u)\rangle$ is symmetric in u and thus the matrix $\hat{C}^W$ is symmetric in this case.

Appendix B: Neuron covariance consistency equations

In this Appendix, we derive the self-consistency equations for the covariance coefficient C^W presented in Sect. 2.3.4, which leads to (9). This analysis includes the spike-triggering effects induced by the autocorrelation of the external inputs and of the neurons (Kempter et al. 1999), which were neglected in Burkitt et al. (2007). In addition, we incorporate the fine-timing effects such as delays and the time course of the PSP response.

The instantaneous neuron covariance is given by

$$\begin{aligned}&\mathrm{Cov}[S_i(t), S_j(t+u)]\\ &\quad := \big\langle S_i(t')S_j(t'+u)\big\rangle - \big\langle S_i(t')\big\rangle\big\langle S_j(t'+u)\big\rangle. \end{aligned} \tag{47}$$

As in Appendix 43 for the input covariances, we consider that the function of u in (47) and thus $C_{ij}(t,u)$ in (4) are continuous in $u = 0$.

B.1 Taking the autocorrelation into account

Similar to Gilson et al. (2009a, Appendix A.2.1), we use the definition of $\rho_i(t)$,

$$\big\langle S_i(t)\,S_j(t+u)\big\rangle = \big\langle \rho_i(t)\,S_j(t+u)\big\rangle, \tag{48}$$

where the rhs incorporates terms due to autocorrelation. However, the derivation of the consistency equation for the neuron-to-neuron covariance is a bit more complex than for

the input-to-neuron covariance, due to the recurrent connections and the probabilistic interdependence with the past spiking history of the network that they imply. Here, we adapt the original derivation in Hawkes (1971, (12) and (24)).

We freeze t and consider the Fourier transform of $C(t,u)$ by integrating u, using the adiabatic assumption that the weights J are quasi-constant. The key-point of this derivation is that C satisfies

$$C(t,-u) = C^{\mathbf{T}}(t,u), \tag{49}$$

under the assumption of slow learning for the weights (quasi time-invariant first and second stochastic moments), as explained in Appendix A.2. To calculate $C(t,u)$, we use (48)

$$\begin{aligned}
&\langle S_i(t)\, S_j(t+u)\rangle \\
&\quad = \nu_0 \langle S_j(t+u))\rangle \\
&\qquad + \sum_{i'} J_{ii'} \langle (\epsilon * S_{i'})\,(t-d_{ii'})\; S_j(t+u)\rangle \\
&\qquad + \sum_{k} K_{ik} \langle \left(\epsilon * \hat{S}_k\right)(t-\hat{d}_{ik})\; S_j(t+u)\rangle \\
&\qquad + J_{ij}\,\epsilon(-u-d_{ij})\,\langle S_j(t+u)\rangle.
\end{aligned} \tag{50}$$

The last term is a spike-triggering effect due to a recurrent connection, corresponding to $t-r-d_{ij} = t+u$. It is important to note that this equality is only valid for the half-plane $u<0$.

Following (2), we use the following expression (Gilson et al. 2009a, Eq. 13)

$$\begin{aligned}
\langle S_i(t)\rangle &= \langle \rho_i(t)\rangle \\
&= \nu_0 + \sum_{j\neq i} J_{ij} \langle \epsilon * S_j(t-d_{ij})\rangle \\
&\quad + \sum_{k} K_{ik}(t) \langle \epsilon * \hat{S}_k(t-\hat{d}_{ik})\rangle
\end{aligned} \tag{51}$$

and proceed to the time averaging over $[t-T,t]$ in order to obtain

$$\begin{aligned}
C_{ij}(t,u) &= \sum_{i'} J_{ii'} \int \epsilon(r-d_{ii'}) C_{i'j}(t,u+r)\,\mathrm{d}r \\
&\quad + \sum_{k} K_{ik} \int \epsilon(r-\hat{d}_{ik}) F_{jk}(t,-u-r)\,\mathrm{d}r \\
&\quad + J_{ij}\,\epsilon(-u-d_{ij})\,\nu_j(t).
\end{aligned} \tag{52}$$

The following changes of variable were made: $t \to t+r+d_{ii'}$ and $r \to r-d_{ii'}$ for terms in the first sum in i' on the rhs, and $t \to t-u+d_{ij}$ and $r \to r-\hat{d}_{ik}$ for the second sum in k. Note the negative sign for u in F_{jk}. Recall that this is only valid for $u<0$, so information from the past is used to evaluate the impact of the autocorrelation on the covariance structure and we evaluate the total effect using the symmetry in u of C, cf. (49).

We now assume, as in Gilson et al. (2009a, Appendix A.2.5), that the delays $d_{ij} = d$ and $\hat{d}_{ik} = \hat{d}$ are sharply distributed. We would like to take the Fourier transform of (52) in order to obtain an equivalent of the consistency equation for the input-to-neuron covariance (Gilson et al. 2009a, Eq. 63). However, (52) is only valid for $u<0$. As in Hawkes (1971, Eq. 22), we introduce the matrix $\breve{C}$ defined by the Fourier transform on the rhs of (52) less the Fourier transform of C, namely $\mathcal{F}C(\omega)$

$$\begin{aligned}
\breve{C}(\omega) &:= -\mathcal{F}C(\omega) + \underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\,\mathcal{F}C(\omega) \\
&\quad + \underline{K}(-\omega)\,\mathcal{F}\epsilon(\omega)\,\mathcal{F}F^{\mathbf{T}}(-\omega) \\
&\quad + \underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\,\mathrm{diag}\,(\boldsymbol{\nu}).
\end{aligned} \tag{53}$$

We have defined $\underline{K}_{ik}(\omega) := K_{ik}\exp(\mathbf{i}\hat{d}_{ik}\omega)$ and $\underline{J}_{ij}(\omega) := J_{ij}\exp(\mathbf{i}d_{ij}\omega)$. The matrix $\breve{C}(\omega)$ thus defined incorporates the effects induced by autocorrelation for the "future" ($u>0$ in (50)). An argument on the regularity of $\omega \mapsto \breve{C}(\omega)$ (i.e., $\breve{C}$ is holomorphic) is used to evaluate it; we reproduce in the remainder of this section the calculations presented by Hawkes (1971), where further details are provided.

Expressing $\mathcal{F}C$ in terms of $\breve{C}$ from (53)

$$\begin{aligned}
&\mathcal{F}C(\omega) \\
&\quad = \left[\mathbb{1}_N - \underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\right]^{-1} \\
&\qquad \times \Big[-\breve{C}(\omega) + \underline{K}(-\omega)\,\mathcal{F}\epsilon(\omega)\,\mathcal{F}F^{\mathbf{T}}(-\omega) \\
&\qquad\quad + \underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\,\mathrm{diag}\,(\boldsymbol{\nu})\Big] \\
&\quad = \left[\mathbb{1}_N - \underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\right]^{-1} \\
&\qquad \times \left[\underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\,\mathrm{diag}\,(\boldsymbol{\nu}) - \breve{C}(\omega)\right] \\
&\qquad + \left[\mathbb{1}_N - \underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\right]^{-1} \underline{K}(\omega)\,\mathcal{F}\epsilon(-\omega) \\
&\qquad \times \left[\mathcal{F}\hat{C}(\omega) + \mathrm{diag}\left(\hat{\boldsymbol{\nu}}\right)\right] \\
&\qquad \times \underline{K}^{\mathbf{T}}(-\omega)\,\mathcal{F}\epsilon(\omega)\left[\mathbb{1}_N - \underline{J}(-\omega)\,\mathcal{F}\epsilon(\omega)\right]^{-1\,\mathbf{T}},
\end{aligned} \tag{54}$$

using the expression of $\mathcal{F}F$ (Gilson et al. 2009a, Appendix A.2.5):

$$\begin{aligned}
\mathcal{F}F(\omega) &= \left[\mathbb{1}_N - \underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\right]^{-1} \\
&\quad \times \underline{K}(\omega)\,\mathcal{F}\epsilon(-\omega)\left[\mathcal{F}\hat{C}(\omega) + \mathrm{diag}\left(\hat{\boldsymbol{\nu}}\right)\right],
\end{aligned} \tag{55}$$

and also $\mathcal{F}\hat{C}(-\omega) = \mathcal{F}\hat{C}^{\mathbf{T}}(\omega)$, cf. Appendix A.2.

Now using (49) with the expression of $\mathcal{F}C$ in (54), we obtain

$$\begin{aligned}
&\left[\mathbb{1}_N - \underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\right]^{-1} \\
&\quad \times \left[\underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\,\mathrm{diag}\,(\boldsymbol{\nu}) - \breve{C}(\omega)\right] \\
&\quad = \left[\underline{J}(-\omega)\,\mathcal{F}\epsilon(\omega)\,\mathrm{diag}\,(\boldsymbol{\nu}) - \breve{C}(-\omega)\right]^{\mathbf{T}} \\
&\quad \times \left[\mathbb{1}_N - \underline{J}(-\omega)\,\mathcal{F}\epsilon(\omega)\right]^{-1\,\mathbf{T}},
\end{aligned} \tag{56}$$

since the last term involving $\hat{C}$ and $\mathrm{diag}(\hat{\boldsymbol{\nu}})$ in the rhs of (54) satisfies an equality similar to (49).

Equation 56 can be reorganized

$$\begin{aligned}&\underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\,\mathrm{diag}\,(\boldsymbol{\nu})\\&\quad+\left[\mathbb{1}_N-\underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\right]\check{C}^{\mathbf{T}}(-\omega)\\&=\mathrm{diag}\,(\boldsymbol{\nu})\;\mathcal{F}\epsilon(\omega)\,\underline{J}^{\mathbf{T}}(-\omega)\\&\quad+\check{C}(\omega)\,\left[\mathbb{1}_N-\underline{J}(-\omega)\,\mathcal{F}\epsilon(\omega)\right]^{\mathbf{T}}.\end{aligned}\tag{57}$$

The equality (57) allows us to define a function that is regular on the whole plane ω, each side being regular for the half of the plane related to the sign of the imaginary part of ω. This requires assumptions on exponentially fast decay of $\epsilon(u)$ for $u\to+\infty$ and of elements of $\check{C}$. The so-defined holomorphic function vanishes when $|\omega|\to\infty$, which implies that it is actually zero on the whole plane. Consequently, we have the expression of $\check{C}$ in terms of the weight matrices K and J (modified to incorporate the effect of the PSP kernel ϵ and delays), of the covariance between the neurons and the inputs (through F) and of the autocorrelation of the processes ($\mathrm{diag}(\boldsymbol{\nu})$)

$$\begin{aligned}\check{C}(\omega)=&-\mathrm{diag}\,(\boldsymbol{\nu})\;\mathcal{F}\epsilon(\omega)\,\underline{J}^{\mathbf{T}}(-\omega)\\&\times\left[\mathbb{1}_N-\underline{J}(-\omega)\,\mathcal{F}\epsilon(\omega)\right]^{-1\,\mathbf{T}}.\end{aligned}\tag{58}$$

Finally, we obtain the expression for $\mathcal{FC}$ by using (58) in (54), similar to the expression of $\mathcal{FF}$ in (55)

$$\begin{aligned}&\mathcal{FC}(\omega)\\&=\left[\mathbb{1}_N-\underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\right]^{-1}\\&\quad\times\Big\{\underline{K}(\omega)\,\mathcal{F}\epsilon(-\omega)\,\left[\mathcal{F}\hat{C}(\omega)+\mathrm{diag}\,(\hat{\boldsymbol{\nu}})\right]\\&\qquad\times\underline{K}^{\mathbf{T}}(-\omega)\,\mathcal{F}\epsilon(\omega)\\&\qquad+\underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\,\mathrm{diag}\,(\boldsymbol{\nu})+\mathrm{diag}\,(\boldsymbol{\nu})\;\mathcal{F}\epsilon(\omega)\,\underline{J}^{\mathbf{T}}(-\omega)\\&\qquad-\underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\,\mathrm{diag}\,(\boldsymbol{\nu})\;\mathcal{F}\epsilon(\omega)\,\underline{J}^{\mathbf{T}}(-\omega)\Big\}\\&\qquad\times\left[\mathbb{1}_N-\underline{J}(-\omega)\,\mathcal{F}\epsilon(\omega)\right]^{-1\,\mathbf{T}}\\&=\left[\mathbb{1}_N-\underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\right]^{-1}\;\underline{K}(\omega)\,\mathcal{F}\epsilon(-\omega)\\&\qquad\times\left[\mathcal{F}\hat{C}(\omega)+\mathrm{diag}\,(\hat{\boldsymbol{\nu}})\right]\\&\qquad\times\underline{K}^{\mathbf{T}}(-\omega)\,\mathcal{F}\epsilon(\omega)\,\left[\mathbb{1}_N-\underline{J}(-\omega)\,\mathcal{F}\epsilon(\omega)\right]^{-1\,\mathbf{T}}\\&\quad+\left[\mathbb{1}_N-\underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\right]^{-1}\;\mathrm{diag}\,(\boldsymbol{\nu})\\&\qquad\times\left[\mathbb{1}_N-\underline{J}(-\omega)\,\mathcal{F}\epsilon(\omega)\right]^{-1\,\mathbf{T}}\\&\quad-\mathrm{diag}\,(\boldsymbol{\nu}).\end{aligned}\tag{59}$$

B.2 Remark on the autocorrelation structure due to the recurrent connections

The autocorrelation effects are the three terms in the first expression of $\mathcal{FC}(\omega)$ in (59) involving 'diag'. Note that the complete covariance impacts upon STDP, i.e., including the first-order autocorrelation that corresponds to $u=0$, which is added to $C(t,u)$ in the convolution with W in the learning equation (cf. Sect. 2.3.3). This actually corresponds to the last term $\mathrm{diag}\,(\boldsymbol{\nu})$ in the second expression of $\mathcal{FC}(\omega)$ in (59). Refer to Appendix A.1 for the distinction between covariance and complete covariance (Hawkes 1971), and its relationship to the encoding of neuronal information.

By naively taking only the spike-triggering effect as an extra contribution in (48), one obtains $\underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\,\mathrm{diag}\,(\boldsymbol{\nu})$. The consideration of the double expansion

$$\left\langle S_i(t)\,S_j(t+u)\right\rangle=\left\langle\rho_i(t)\,\rho_j(t+u)\right\rangle\tag{60}$$

using the expression for $\rho_i(t)$ in (2) may lead to the expression of (59), since it preserves the symmetry between the neurons i and j. However, care must be taken to ensure that the terms involving $\mathrm{diag}\,(\boldsymbol{\nu})$ are correctly considered.

B.3 Short recurrent delays

The function $\mathcal{FC}^W$ can be obtained by multiplying the lhs of (59) by $\exp(-\mathbf{i}d\omega)\,\mathcal{F}W(-\omega)$, to incorporate the impact of W and the delays d_{ij}. The inverse of $\mathbb{1}_N-\underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)$ could be developed in a power series in order to obtain a rigorous expression of $C^W(t)$. This actually leads to a double series because of the two occurrences of the inverse matrix. For delta-correlated inputs and an STDP window function W with compact support, only a finite number of terms remain in the double series when ϵ is different from the Dirac delta function. Unlike the input-to-neuron covariance, a larger value for the delay d does not lead to a single term and the expression is more difficult to handle.

To simplify the result, we proceed to the same approximations assuming the short durations of ϵ and d similar to Gilson et al. (2009a, Appendix A.2.7) in the expression of $\mathcal{FC}$ in (59), i.e.,

$$\underline{J}(\omega)\,\mathcal{F}\epsilon(-\omega)\;\simeq\;\underline{J}(0)\,\mathcal{F}\epsilon(0)\;=\;J\,,\tag{61}$$

in order to deal with the inverses and express C in the time domain u using the inverse Fourier transform. This leads to the following expression of the term $C^W(t)+W(d)\,\mathrm{diag}(\boldsymbol{\nu}(t))$ due to STDP in the rhs of the learning matrix equation of J (cf. Sect. 2.3.3)

$$\begin{aligned}&C^W(t)+W(d)\,\mathrm{diag}\,(\boldsymbol{\nu}(t))\\&\quad=[\mathbb{1}_N-J]^{-1}\;K\,\left[\hat{C}^{W*\zeta}(t)+[W*\zeta](0)\,\mathrm{diag}\,(\hat{\boldsymbol{\nu}})\right]\\&\qquad\times K^{\mathbf{T}}\,[\mathbb{1}_N-J]^{-1\,\mathbf{T}}\\&\qquad+W(d)\,[\mathbb{1}_N-J]^{-1}\,\mathrm{diag}\,(\boldsymbol{\nu})\,[\mathbb{1}_N-J]^{-1\,\mathbf{T}}.\end{aligned}\tag{62}$$

The function ζ describes the impact of the PSP kernel on the input covariance structure. It corresponds to the inverse Fourier transform of (cf. (59)

$$\exp(\mathbf{i}\,\hat{d}\,\omega)\,\mathcal{F}\epsilon(-\omega)\;\exp(-\mathbf{i}\,\hat{d}\,\omega)\,\mathcal{F}\epsilon(\omega)\;\exp(-\mathbf{i}\,d\,\omega),\tag{63}$$

Table 1 Table of simulation parameters

Time step	10^{-4} s
Simulation duration	10^5 s
Input Poisson spike trains	
Firing rates	$\hat{\nu}_{\text{av}} = 30$ Hz
Correlation strength	$\hat{c}_{\text{av}} = 0 - 0.1$
Poisson neurons	
Instantaneous firing rate	$\nu_0 = 5$ Hz
Synapses	
Rise time constant	$\tau_A = 1$ ms
Decay time constant	$\tau_B = 5$ ms
Mean of recurrent delays	$d = 0.4$ ms
Spread of recurrent delays	± 0.2 ms
Mean of input delays	$\hat{d} = 7$ ms
Spread of input delays	± 1 ms
STDP	
Learning parameter	$\eta = 5 \times 10^{-7}$
Pre-synaptic rate-based coeff.	$w^{\text{in}} = 4$
Post-synaptic rate-based coeff.	$w^{\text{out}} = -0.5$
Potentiation time constant	$\tau_P = 17$ ms
Potentiation scaling coefficient	$c_P = 15$
Depression time constant	$\tau_D = 34$ ms
Depression scaling coefficient	$c_D = 10$

but reversed in time since the convolution with W corresponds to $\mathcal{F}W(-\omega)$, i.e.,

$$\begin{aligned}\zeta(r) &:= \int \epsilon(r + r' + d)\epsilon(r')\,\mathrm{d}r' \\ &\simeq \int \epsilon(r + r')\epsilon(r')\,\mathrm{d}r'. \end{aligned} \tag{64}$$

The expression in (62) differs from its equivalent in Burkitt et al. (2007) by the autocorrelation terms involving 'diag' and the convolution $W * \zeta$. The recurrent connections induce intrinsic correlation structure within the network, which are at the first order of the recurrence described by ζ. This may partly explain the small discrepancies between theoretical predictions and simulation results in Burkitt et al. (2007).

Appendix C: Simulation parameters

The results in this article were obtained using Poisson neurons simulated in discrete-time with the parameters listed in Table 1, unless stated otherwise. The STDP window function W is given by

$$W(u) = \begin{cases} c_P \exp(u/\tau_P) & \text{for } u < 0 \\ -c_D \exp(-u/\tau_D) & \text{for } u > 0. \end{cases} \tag{65}$$

The PSP kernel ϵ is defined by

$$\epsilon(t) = \begin{cases} \frac{\exp(t/\tau_B) - \exp(t/\tau_A)}{\tau_B - \tau_A} & \text{for } t \geq 0 \\ 0 & \text{for } t < 0. \end{cases} \tag{66}$$

These parameters are in the same range as those used in previous studies (Kempter et al. 1999; Burkitt et al. 2007).

References

Appleby PA, Elliott T (2006) Stable competitive dynamics emerge from multispike interactions in a stochastic model of spike-timing-dependent plasticity. Neural Comput 18(10):2414–2464

Bi GQ, Poo MM (2001) Synaptic modification by correlated activity: Hebb's postulate revisited. Annu Rev Neurosci 24:139–166

Burkitt AN (2006) A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input. Biol Cybern 95(1):1–19

Burkitt AN, Meffin H, Grayden DB (2004) Spike-timing-dependent plasticity: the relationship to rate-based learning for models with weight dynamics determined by a stable fixed point. Neural Comput 16(5):885–940

Burkitt AN, Gilson M, van Hemmen JL (2007) Spike-timing-dependent plasticity for neurons with recurrent connections. Biol Cybern 96(5):533–546

Câteau H, Kitano K, Fukai T (2008) Interplay between a phase response curve and spike-timing-dependent plasticity leading to wireless clustering. Phys Rev E 77(5):051909

Gerstner W, Kempter R, van Hemmen JL, Wagner H (1996) A neuronal learning rule for sub-millisecond temporal coding. Nature 383(6595):76–78

Gilson M, Burkitt AN, Grayden DB, Thomas DA, van Hemmen JL (2009a) Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks I: input selectivity–strengthening correlated input pathways. Biol Cybern 101(2):81–102

Gilson M, Burkitt AN, Grayden DB, Thomas DA, van Hemmen JL (2009b) Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks II: input selectivity–symmetry breaking. Biol Cybern 101(2):103–114

Gilson M, Burkitt AN, Grayden DB, Thomas DA, van Hemmen JL (2009c) Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks III: partially connected neurons driven by spontaneous activity. Biol Cybern doi:10.1007/s00422-009-0343-4

Gütig R, Aharonov R, Rotter S, Sompolinsky H (2003) Learning input correlations through nonlinear temporally asymmetric Hebbian plasticity. J Neurosci 23(9):3697–3714

Hawkes AG (1971) Point spectra of some mutually exciting point processes. J Roy Stat Soc Ser B 33(3):438–443

Hebb DO (1949) The organization of behavior: a neuropsychological theory. Wiley, New York

Iglesias J, Eriksson J, Grize F, Tomassini M, Villa A (2005) Dynamics of pruning in simulated large-scale spiking neural networks. Biosystems 79:11–20

Izhikevich EM, Gally JA, Edelman GM (2004) Spike-timing dynamics of neuronal groups. Cereb Cortex 14:933–944

Karbowski J, Ermentrout GB (2002) Synchrony arising from a balanced synaptic plasticity in a network of heterogeneous neural oscillators. Phys Rev E 65(3):031902

Kempter R, Gerstner W, van Hemmen JL (1999) Hebbian learning and spiking neurons. Phys Rev E 59(4):4498–4514

Lubenov EV, Siapas AG (2008) Decoupling through synchrony in neuronal circuits with propagation delays. Neuron 58(1):118–131

Markram H, Lübke J, Frotscher M, Roth A, Sakmann B (1997) Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex. J Physiol (Lond) 500(2):409–440

Masuda N, Kori H (2007) Formation of feedforward networks and frequency synchrony by spike-timing-dependent plasticity. J Comput Neurosci 22(3):327–345

Meffin H, Besson J, Burkitt AN, Grayden DB (2006) Learning the structure of correlated synaptic subgroups using stable and competitive spike-timing-dependent plasticity. Phys Rev E 73(4):041911

Moreno-Bote R, Renart A, Parga N (2008) Theory of input spike auto- and cross-correlations and their effect on the response of spiking neurons. Neural Comput 20(7):1651–1705

Morrison A, Aertsen A, Diesmann M (2007) Spike-timing-dependent plasticity in balanced random networks. Neural Comput 19(6):1437–1467

Morrison A, Diesmann M, Gerstner W (2008) Phenomenological models of synaptic plasticity based on spike timing. Biol Cybern 98(6):459–478

Pfister JP, Toyoizumi T, Barber D, Gerstner W (2006) Optimal spike-timing-dependent plasticity for precise action potential firing in supervised learning. Neural Comput 18(6):1318–1348

Roberts PD (2004) Recurrent biological neuronal networks: the weak and noisy limit. Phys Rev E 69(3):031910

Senn W, Schneider M, Ruf B (2002) Activity-dependent development of axonal and dendritic delays, or, why synaptic transmission should be unreliable. Neural Comput 14(3):583–619

Sjöström PJ, Turrigiano GG, Nelson SB (2001) Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. Neuron 32(6):1149–1164

Song S, Abbott LF (2001) Cortical development and remapping through spike timing-dependent plasticity. Neuron 32(2):339–350

Toyoizumi T, Pfister JP, Aihara K, Gerstner W (2007) Optimality model of unsupervised spike-timing-dependent plasticity: synaptic memory and weight distribution. Neural Comput 19(3):639–671

van Hemmen JL (2001) Theory of synaptic plasticity. In: Moss F, Gielen S (eds) Handbook of biological physics, vol 4: Neuro-informatics and neural modelling. Elsevier, Amsterdam, pp 771–823

van Rossum MCW, Bi GQ, Turrigiano GG (2000) Stable Hebbian learning from spike timing-dependent plasticity. J Neurosci 20(23):8812–8821