frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 08 July 2014
doi: 10.3389/fncom.2014.00068

Learning and stabilization of winner-take-all dynamics through interacting excitatory and inhibitory plasticity

***Jonathan Binas*[1]*, *Ueli Rutishauser*[2,3], *Giacomo Indiveri*[1] and *Michael Pfeiffer*[1]**

[1] *Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland*
[2] *Department of Neurosurgery and Department of Neurology, Cedars-Sinai Medical Center, Los Angeles, CA, USA*
[3] *Computation and Neural Systems Program, Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA*

Edited by:
Cristina Savin, IST Austria, Austria

Reviewed by:
Yanqing Chen, The Neurosciences Institute, USA
Cristina Savin, IST Austria, Austria

****Correspondence:***
Jonathan Binas, Institute of Neuroinformatics, University of Zurich and ETH Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland
e-mail: jbinas@ini.ethz.ch

Winner-Take-All (WTA) networks are recurrently connected populations of excitatory and inhibitory neurons that represent promising candidate microcircuits for implementing cortical computation. WTAs can perform powerful computations, ranging from signal-restoration to state-dependent processing. However, such networks require fine-tuned connectivity parameters to keep the network dynamics within stable operating regimes. In this article, we show how such stability can emerge autonomously through an interaction of biologically plausible plasticity mechanisms that operate simultaneously on all excitatory and inhibitory synapses of the network. A weight-dependent plasticity rule is derived from the triplet spike-timing dependent plasticity model, and its stabilization properties in the mean-field case are analyzed using contraction theory. Our main result provides simple constraints on the plasticity rule parameters, rather than on the weights themselves, which guarantee stable WTA behavior. The plastic network we present is able to adapt to changing input conditions, and to dynamically adjust its gain, therefore exhibiting self-stabilization mechanisms that are crucial for maintaining stable operation in large networks of interconnected subunits. We show how distributed neural assemblies can adjust their parameters for stable WTA function autonomously while respecting anatomical constraints on neural wiring.

Keywords: winner-take-all, competition, plasticity, self-organization, contraction theory, canonical microcircuits, inhibitory plasticity

1. INTRODUCTION

Competition through shared inhibition is a powerful model of neural computation (Maass, 2000; Douglas and Martin, 2007). Competitive networks are typically composed of populations of excitatory neurons driving a common set of inhibitory neurons, which in turn provide global negative feedback to the excitatory neurons (Amari and Arbib, 1977; Douglas and Martin, 1991; Hertz et al., 1991; Coultrip et al., 1992; Douglas et al., 1995; Hahnloser et al., 2000; Maass, 2000; Rabinovich et al., 2000; Yuille and Geiger, 2003; Rutishauser et al., 2011). Winner-take-all (WTA) networks are one instance of this circuit motif, which has been studied extensively. Neurophysiological and anatomical studies have shown that WTA circuits model essential features of cortical networks (Douglas et al., 1989; Mountcastle, 1997; Binzegger et al., 2004; Douglas and Martin, 2004; Carandini and Heeger, 2012). An individual WTA circuit can implement a variety of non-linear operations such as signal restoration, amplification, filtering, or max-like winner selection, e.g., for decision making (Hahnloser et al., 1999; Maass, 2000; Yuille and Geiger, 2003; Douglas and Martin, 2007). The circuit plays an essential role in both early and recent models of unsupervised learning, such as receptive field development (von der Malsburg, 1973; Fukushima, 1980; Ben-Yishai et al., 1995), or map formation (Willshaw and Von Der Malsburg, 1976; Amari, 1980; Kohonen, 1982; Song and Abbott, 2001). Multiple WTA instances can be combined to implement more powerful computations that cannot be achieved with a single instance, such as state dependent processing (Rutishauser and Douglas, 2009; Neftci et al., 2013). This modularity has given rise to the idea of WTA circuits representing *canonical microcircuits*, which are repeated many times throughout cortex and are modified slightly and combined in different ways to implement different functions (Douglas and Martin, 1991, 2004; Rutishauser et al., 2011).

In most models of WTA circuits the network connectivity is given a priori. In turn, little is known about whether and how such connectivity could emerge without precise pre-specification. In this article we derive analytical constraints under which local synaptic plasticity on all connections of the network tunes the weights for WTA-type behavior. This is challenging as high-gain WTA operation on the one hand, and stable network dynamics on the other hand, impose diverging constraints on the connection strengths (Rutishauser et al., 2011), which should not be violated by the plasticity mechanism. Previous models like Jug et al. (2012) or Bauer (2013) have shown empirically that functional WTA-like behavior can arise from an interplay of plasticity on excitatory synapses and homeostatic mechanisms. Here, we provide a mathematical explanation for this phenomenon, using a mean-field based analysis, and derive conditions under which biologically plausible plasticity rules applied to all connections of a network of randomly connected inhibitory and excitatory units produce

a functional WTA network with structured connectivity. Due to plastic inhibitory synapses, convergence of the model does not rely on constant, pre-defined inhibitory weights or other common assumptions for WTA models. We prove that the resulting WTA circuits obey stability conditions imposed by contraction analysis (Lohmiller and Slotine, 1998; Rutishauser et al., 2011). This has important implications for the stability of larger networks composed of multiple interconnected WTA circuits, and thus sheds light onto the mechanisms responsible for the emergence of both local functional cortical microcircuits and larger distributed coupled WTA networks.

This article is structured as follows: We first define the network and plasticity models in sections 2.1 to 2.3. Our main analytical results are given in sections 2.4 and 2.5, and illustrated with simulation results in section 2.6. The results are discussed in section 3, and detailed derivations of the analytical results can be found in section 4.

2. RESULTS

2.1. NETWORK TOPOLOGY

In its simplest abstract form, a WTA circuit (**Figure 1A**) consists of a number of excitatory units that project onto a common inhibitory unit. This unit, in turn, provides recurrent inhibitory feedback to all excitatory units. Given appropriate connection strengths, such inhibition makes the excitatory units compete for activation in the sense that the unit receiving the strongest input signal will suppress the activation of all other units through the inhibitory feedback loop, and "win" the competition.

We design a biologically plausible network by taking into account that inhibitory feedback is local, i.e., it only affects cells within a cortical volume that is small enough such that the relatively short inhibitory axonal arbors can reach their targets. We assume excitatory and inhibitory neurons in this volume to be connected randomly (see **Figure 1B**). Furthermore, we assume that there are a finite number of different input signals, each activating a subset of the excitatory cells in the volume. We construct a mean-field model by grouping the excitatory neurons for each driving input stimulus, summarizing the activity of each group of cells by their average firing rate. This results in a simplified population model of the network which—in the case of two different input signals—consists of two excitatory populations (one for each input), and one inhibitory population (see **Figure 1C**). We assume full recurrent connectivity between all populations. This scheme can easily be extended toward more input groups. In particular, if an excitatory group receives multiple inputs, it can be modeled as a new class.

Since inhibitory axons are (typically) short-range, distant populations can communicate only via excitatory projections. We combine multiple local circuits of the form shown in **Figures 1B,C** by introducing excitatory long-range connections between them, as illustrated in **Figure 1D**. Specifically, we add projections from the excitatory populations of one local group to all excitatory and inhibitory populations of the other group. A similar connectivity scheme for implementing distributed WTA networks has been proposed by Rutishauser et al. (2012). Unlike their model, our network does not require specific wiring, but rather targets any potential cell in the other volume. We will show in section 2.4.4 that this is sufficient to achieve competition between units of spatially distributed WTA circuits.

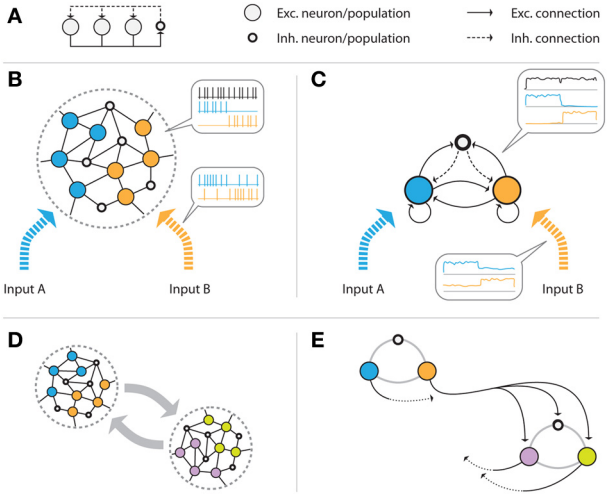

FIGURE 1 | Illustration of the network model. (A) Abstract representation of a WTA circuit, where several excitatory units project onto a common inhibitory unit, and receive global inhibitory feedback from that unit. **(B)** Example volume of the generic cortical structure that is assumed, consisting of (initially randomly connected) excitatory (pyramidal) cells and inhibitory interneurons. The color of the cells indicates the input channel they are connected to: some cells only receive input from the blue, others from the orange source. It is assumed that the volume is sufficiently small, such that all excitatory cells can be reached by the (short-ranged) axons of the inhibitory cells. If the connection strengths are tuned appropriately, the population receiving the stronger input signal will suppress the response of the weaker population via the global inhibitory feedback. **(C)** shows the mean field model of the same network that we construct by grouping excitatory neurons by their input source. The three resulting excitatory and interneuron populations are connected in an all-to-all fashion. **(D,E)** show multiple, distant volumes which are connected via long-range excitatory connections. Projections from one volume to another connect to all cells of the target volume. In **(E)**, the two subgroups are approximated by networks of the type shown in **(C)**, consisting of one inhibitory and several excitatory populations. The black, solid arrows represent exemplary excitatory connections from one population of one group to all populations of the other group. Equivalent connections, indicated by dotted arrows, exist for all of the excitatory populations.

2.2. NETWORK DYNAMICS

The activation of a neural populations x_i, which can be excitatory or inhibitory, is described by

$$\tau_i \dot{x}_i(t) = -x_i(t) + \left[\sum_j w_{ij} x_j(t) + I_{\text{ext},i}(t) - T_i \right]_+ , \qquad (1)$$

where τ_i is the time constant of the population, w_{ij} is the weight of the incoming connection from the jth population, $I_{\text{ext},i}(t)$ is an external input given to the population, and T_i is the activation threshold. Furthermore, $[v]_+ := \max(0, v)$ is a half-wave rectification function, preventing the firing rates from taking negative values. Assuming identical time constants for all populations, i.e., $\tau_i = \tau$ for all i, the dynamics of the full system can be written as

$$\tau \dot{\boldsymbol{x}}(t) = -\boldsymbol{x}(t) + [\boldsymbol{W}\boldsymbol{x}(t) + \boldsymbol{I}_{\text{ext}}(t) - \boldsymbol{T}]_+ , \quad (2)$$

where $\boldsymbol{x} = (x_1, \ldots, x_N)$ are the firing rates of the respective populations (excitatory and inhibitory), $\boldsymbol{W}$ is the connectivity matrix (describing local excitatory, local inhibitory, and long-range excitatory connections), $\boldsymbol{I}_{\text{ext}}(t)$ is a vector of external inputs, and $\boldsymbol{T} = (T_1, \ldots, T_N)$ are the activation thresholds of the populations. For the single local microcircuit shown in **Figure 1C**, for example, $\boldsymbol{W}$ would be a 3-by-3 matrix with all entries w_{ij} nonzero except for the inhibitory to inhibitory coupling. For two coupled microcircuits as in **Figure 1E**, the connectivity matrix consists of 4 blocks, with the diagonal blocks describing local connectivity, and the off-diagonal blocks describing long-range projections from excitatory units to the other circuit.

2.3. PLASTICITY MECHANISMS AND WEIGHT DYNAMICS

In our model, we assume that all connections w_{ij} in Equation (2) are plastic, and are subject to the following weight update rule:

$$\dot{w} = \tau_s^2 x_{\text{pre}} x_{\text{post}} \left(x_{\text{post}}(w_{\max} - w) - (\Theta_w + A_w x_{\text{pre}}) w \right). \quad (3)$$

Here, x_{pre} and x_{post} are the pre- and postsynaptic firing rates, respectively, $w_{\max}$ is the maximum possible weight value, and Θ_w, A_w, and τ_s are positive constants, which we set to values that are compatible with experimental findings (see **Table 1**). The learning rate is determined by τ_s, and Θ_w and A_w determine the point at which the rule switches between depression (LTD) and potentiation (LTP). We will show that in a plastic network, global stability and circuit function are determined exclusively by those plasticity parameters. The plasticity rule is derived from the mean-field approximation of the triplet STDP rule by Pfister and Gerstner (2006), which we augment with a weight-dependent term, effectively limiting the weight values to the interval $[0, w_{\max}]$. A more detailed derivation of the learning rule can be found in the Methods (section 4.1). The parameters Θ_w and A_w are set differently for excitatory and inhibitory connections, leading to two types of simultaneously active plasticity mechanisms and weight dynamics, even though the same learning equation is used. We set $\Theta_w = \Theta_{\text{exc}}$ and $A_w = A_{\text{exc}}$ for all excitatory connections, and $\Theta_w = \Theta_{\text{inh}}$ and $A_w = A_{\text{inh}}$ for all inhibitory connections. In particular, we assume A_{inh} to take very low values and set $A_{\text{inh}} = 0$ in our analysis, effectively eliminating any dependence of the fixed point of inhibitory weights on the presynaptic rate. According to fits of the parameters to experimental data (see **Table 1**), this is a plausible assumption. For the sake of simplicity, we also assume the maximum possible weight value $w_{\max}$ to be the same for all excitatory and inhibitory connections. **Figure 2** illustrates the weight change as a function of the pre- and postsynaptic activity.

2.4. STABILITY ANALYSIS

The WTA circuit is assumed to function correctly if it converges to a stable state that represents the outcome of the computation it is supposed to perform. Conditions under which these networks converge to their (single) attractor state exponentially fast were previously derived by Rutishauser et al. (2011). Here, we extend those results to plastic networks and express stability criteria in terms of global learning rule parameters, rather than individual weight values. We first describe criteria for the stabilization of the network and learning rule dynamics, then derive from them conditions on the learning rule parameters. Our analysis leads to very simple sufficient conditions that ensure the desired stable WTA behavior.

The dynamics of the network activation and the weights are given by Equations (2) and (3), respectively. In the following, we will denote them by $\boldsymbol{f}$ and $\boldsymbol{g}$, so the full dynamics can be written as a coupled dynamical system

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w}), \quad (4)$$

$$\dot{\boldsymbol{w}} = \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{w}), \quad (5)$$

where $\boldsymbol{f}$ corresponds to the right hand side of Equation (2), and $\boldsymbol{g}$ combines the update rules for all weights (with different sets of parameters for excitatory and inhibitory connections) in one vector-valued function. We first restrict our analysis to the simplest case of a single winning excitatory population and derive conditions under which the plastic network converges to its fixed point. Later, we extend our analysis to larger systems of multiple coupled excitatory populations.

2.4.1. Analysis of single-node system

Let us first consider a simplified system, in which only one excitatory population is active, e.g., because one population receives much more external input than all others, and the inhibitory feedback suppresses the other populations. As silent populations neither contribute to the network dynamics nor to the weight dynamics, they can be excluded from the analysis. We can therefore reduce the description of the system to a single excitatory population x_{E}, and an inhibitory population x_{I}, together with the connections $w_{\text{E}\to\text{E}}$, $w_{\text{E}\to\text{I}}$, and $w_{\text{I}\to\text{E}}$ between them.

Table 1 | Learning rule parameters $A_2^{\pm}$, and $A_3^{\pm}$ from Pfister and Gerstner (2006).

Model	A_2^+	A_3^+	A_2^-	A_3^-	Θ	A	τ_s^2
All-to-all	5×10^{-10}	6.2×10^{-3}	7×10^{-3}	2.3×10^{-4}	18.19	0.06	1.3×10^{-5}
Nearest spike	8.8×10^{-11}	5.3×10^{-2}	6.6×10^{-3}	3.1×10^{-3}	6.24	2.09	3.6×10^{-5}

The values of Θ, A, and τ_s for the plasticity rule Equation (3) have been computed using Equations (14) to (16). The data corresponds to fits of the triplet Spike-Timing Dependent Plasticity (STDP) model with all-to-all spike interactions (first row) and with nearest spike interactions (second row) to recordings from plasticity experiments in rat visual cortex. Note that the time constants $\tau_{x,y}$ and $\tau_{\pm}$ are not reproduced here. However, they all are of the order of hundreds of milliseconds and can be found in Pfister and Gerstner (2006). In our simulations, we use parameters very similar to the "all-to-all" parameters for inhibitory connections, while for excitatory connections we use ones that are close to the "nearest spike" parameters.

For a given set of (fixed) weights $\boldsymbol{w}_c$, Rutishauser et al. (2011) have shown by means of contraction theory (Lohmiller and Slotine, 1998) that the system of network activations $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w}_c)$ converges to its fixed point $\boldsymbol{x}^*$ exponentially fast if its generalized Jacobian is negative definite. In our case, this condition reduces to

$$\mathrm{Re}\left(w_{\mathrm{E}\to\mathrm{E}} - 2 + \left(w_{\mathrm{E}\to\mathrm{E}}^2 - 4\, w_{\mathrm{I}\to\mathrm{E}}\, w_{\mathrm{E}\to\mathrm{I}}\right)^{1/2}\right) < 0. \quad (6)$$

If condition (6) is met, the system is called contracting and is guaranteed to converge to its attractor state

$$x_{\mathrm{E}}^* = \Lambda I_{\mathrm{ext}}, \quad (7)$$

$$x_{\mathrm{I}}^* = \Lambda w_{\mathrm{E}\to\mathrm{I}} I_{\mathrm{ext}}, \quad (8)$$

exponentially fast for any constant input I_{ext}, where the contraction rate is given by the left hand side of (6), divided by 2τ. Here, $\Lambda = (1 - w_{\mathrm{E}\to\mathrm{E}} + w_{\mathrm{E}\to\mathrm{I}}\, w_{\mathrm{I}\to\mathrm{E}})^{-1}$ corresponds to the network gain. A more detailed derivation of the fixed point can be found in section 4.2. Note that we have set the activation threshold T equal to zero and provide external input I_{ext} to the excitatory population only. This simplifies the analysis but does not affect our results qualitatively.

2.4.2. Decoupling of network and weight dynamics

In the following, we assume that the population dynamics is contracting, i.e., that condition (6) is met, to show that the plasticity dynamics Equation (5) drives the weights $\boldsymbol{w}$ to a state that is consistent with this condition. Essentially, our analysis has to be self-consistent with respect to the contraction of the activation dynamics. If we assume $\boldsymbol{f}$ and $\boldsymbol{g}$ to operate on very different timescales, we can decouple the two systems given by Equations (4) and (5). This is a valid assumption since neural (population) dynamics vary on timescales of tens or hundreds of milliseconds (see **Figure 5** for typical timescales of our system), while synaptic plasticity typically acts on timescales of seconds or minutes. This means that from the point of view of the weight dynamics $\boldsymbol{g}$ the population activation is at its fixed point $\boldsymbol{x}^*$ almost all the time, because it converges to that point exponentially fast. We can thus model the activation dynamics as a quasi-static system, and approximate the learning dynamics as a function of the fixed point of the activation instead of the instantaneous activation.

$$\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{w}) \approx \boldsymbol{g}(\boldsymbol{x}^*, \boldsymbol{w}), \quad (9)$$

The fixed point of this simplified system is found by setting $\boldsymbol{g}(\boldsymbol{x}^*, \boldsymbol{w}) = \boldsymbol{0}$, and according to Equation (3) is given by

$$w^* = \frac{w_{\max} x_{\mathrm{post}}^*}{\Theta_w + A_w x_{\mathrm{pre}}^* + x_{\mathrm{post}}^*}. \quad (10)$$

Combining this expression with Equations (7) and (8) leads to a system of non-linear equations that can be solved for the fixed point weights $w_{\mathrm{E}\to\mathrm{E}}^*$, $w_{\mathrm{E}\to\mathrm{I}}^*$, $w_{\mathrm{I}\to\mathrm{E}}^*$, and activations x_{E}^*, and x_{I}^*. These values solely depend on the learning rule parameters Θ_w, A_w, $w_{\max}$, and the external (training) input I_{ext}.

Figure 3 shows the fixed points of the weight dynamics as a function of Θ_{exc}, and the input strength I_{ext}. Notably, $w_{\mathrm{E}\to\mathrm{E}}^*$ and $w_{\mathrm{E}\to\mathrm{I}}^*$ lie on a fixed line in the $w_{\mathrm{E}\to\mathrm{E}}$-$w_{\mathrm{E}\to\mathrm{I}}$ plane for all parameters Θ_w and A_w. As the weight values are bounded by 0 and $w_{\max}$, the weights converge to a finite value for $I_{\mathrm{ext}} \to \infty$. This is also illustrated in **Figure 4**, which shows the final weight values as a function of $w_{\max}$, both for a finite training input and in the limit $I_{\mathrm{ext}} \to \infty$.

Importantly, the function of a WTA circuit critically depends on the strength of the recurrent connection $w_{\mathrm{E}\to\mathrm{E}}$ (Rutishauser et al., 2011). If $w_{\mathrm{E}\to\mathrm{E}} > 1$, the network operates in "hard" mode, where only one unit can win at a time and the activation of all

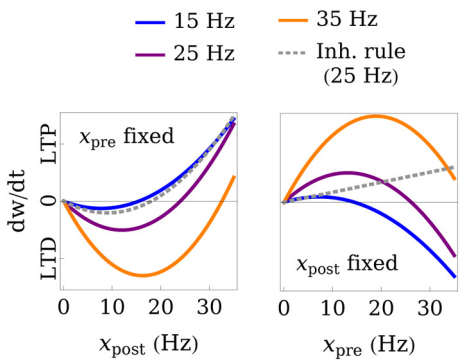

FIGURE 2 | Illustration of the learning rule. The weight change dw/dt is plotted as a function of the post- and presynaptic firing rate for fixed pre- (left) or postsynaptic (right) rates. The gray, dashed line shows the rule that we use for inhibitory connections and whose threshold for LTP, in contrast to excitatory connections, does not depend on the presynaptic rate. The black line marks the transition between LTD and LTP. In this example, the parameters of the learning rules were set to $\Theta_{\mathrm{exc}} = 6$ Hz, $\Theta_{\mathrm{inh}} = 18$ Hz, $A_{\mathrm{exc}} = 2$, $w_{\max} = 4$, and the weight value was fixed at $w_{\max}/3$ for excitatory and $w_{\max}/2$ for inhibitory connections.

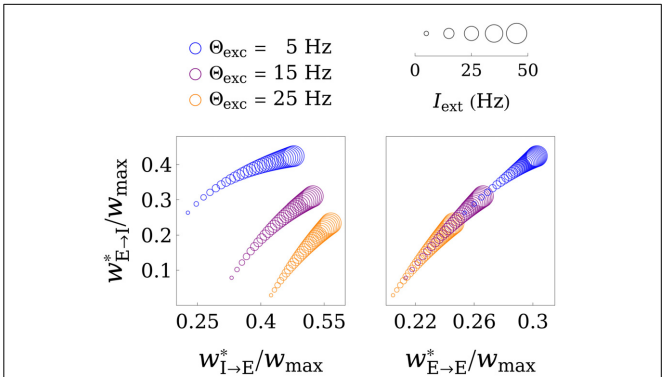

FIGURE 3 | Illustration of the fixed point in weight space. The values of the final weights (in units of $w_{\max}$) are plotted as functions of the parameter Θ_{exc} and the training input strength I_{ext}, where bigger circles correspond to greater I_{ext}. The left panel shows the $w_{\mathrm{E}\to\mathrm{I}}$-$w_{\mathrm{I}\to\mathrm{E}}$ plane, while the right panel shows the $w_{\mathrm{E}\to\mathrm{I}}$-$w_{\mathrm{E}\to\mathrm{E}}$ plane. Interestingly, the fixed point in $w_{\mathrm{E}\to\mathrm{I}}$-$w_{\mathrm{E}\to\mathrm{E}}$ space only gets shifted along a line for different values of Θ_{exc} and I_{ext}. For $I_{\mathrm{ext}} \to \infty$ the weights converge to a limit point, as is illustrated in **Figure 4**. For these plots, the parameters A_{exc} and Θ_{inh} were set to 2 and 18 Hz, respectively, and $w_{\max}$ was set to 4.

other units is zero. On the other hand, if $w_{E\to E}$ is smaller than 1, the network implements "soft" competition, which means that multiple units can be active at the same time. From Equation (27) (Methods) it follows that $w_{E\to E} > 1$ is possible only if $w_{max} > A + 1$. As we will show in the following section, this condition is necessarily satisfied by learning rules that lead to stable WTA circuits.

2.4.3. Parameter regimes for stable network function

We can now use the fixed points found in the previous section to express the condition for contraction given by condition (6) in terms of the learning rule parameters. In general, this new condition does not assume an analytically simple form. However, we can find simple sufficient conditions which still provide a good approximation to the actual value (see Methods section 4.2 for details). Specifically, as a key result of our analysis we derive the following sufficient condition: Convergence to a point in weight space that produces stable network dynamics is guaranteed if

$$A_{exc} + b < w_{max} < 2(1 + A_{exc}), \tag{11}$$

where b is a parameter of the order 1, which is related to the minimum activation x_E (or the minimum non-zero input I_{ext}) during training for which this condition should hold. If the minimum input I_{min} that the network will be trained on is known, then b can be computed from the fixed point $x^*_{E,min} = x^*_E(I_{ext} = I_{min})$, and set to $b = \Theta_{exc}/x^*_{E,min}$. This will guarantee contracting dynamics for the full range of training inputs $I_{ext} \in [I_{min}, \infty)$. In typical scenarios, b can be set to a number of the order 1. This is due to the fact that the network activation is roughly of the same order as the input strength. Setting Θ_{exc} to a value of similar order leads to $b = \Theta_{exc}/x^*_{E,min} \approx 1$.

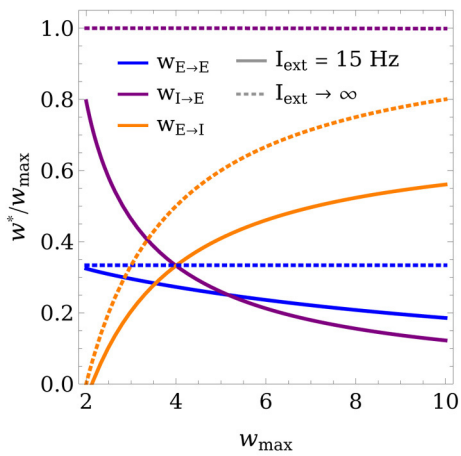

FIGURE 4 | Limit behavior of the fixed point of the weights for weak and strong inputs. The final weight values (in units of w_{max}) are plotted as a function of w_{max}, both for $I_{ext} = 15$ Hz (solid lines) and in the limit of very large inputs $I_{ext} \to \infty$ (dashed lines). In the limit case, $w_{E\to E}$ and $w_{I\to E}$ converge to expressions that are linear in w_{max}, while $w_{E\to I}$ increases superlinearly. The learning rule parameters were set to $\Theta_{exc} = 6$ Hz, $\Theta_{inh} = 18$ Hz, and $A_{exc} = 2$.

Note that condition (11) is independent of Θ_{exc} and Θ_{inh}. This is due to a simplification that is based on the assumption $A_{exc} + b \gg 1$, which can be made without loss of generality. If b and A_{exc} are set to very low values, the full expressions given by 38 and (39) (see Methods section 4.2) apply instead. **Figure 5** shows the the region defined by (11) for different b together with the exact

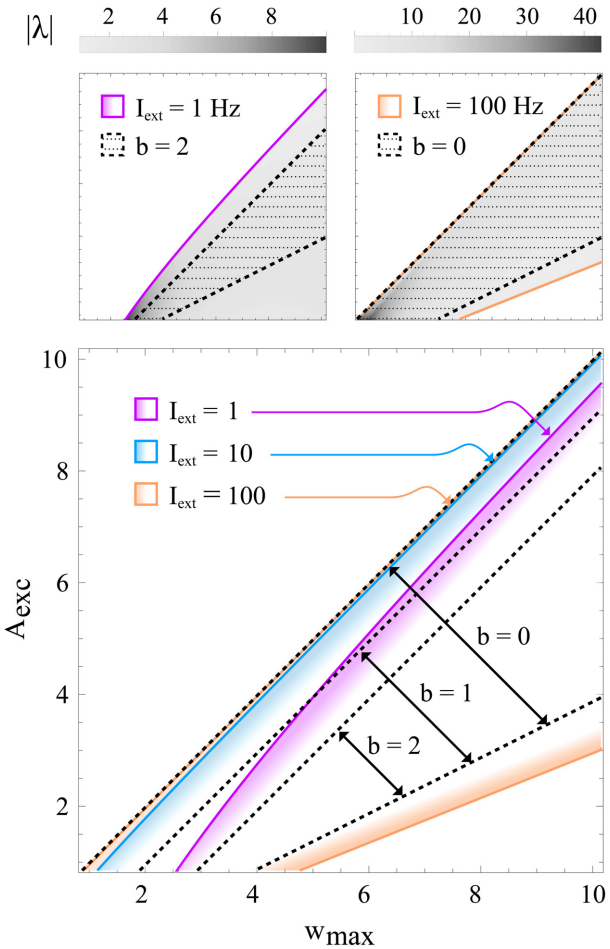

FIGURE 5 | Regions in learning rule parameter space that lead to a stable, contracting network. All panels show the regions of stability in w_{max}-A_{exc} space for different training input strengths. Colored lines correspond to exact solutions, while black, dotted lines correspond to the sufficient condition (11) for different values of b. The top panels illustrate that relatively small values of b (e.g., 2) roughly approximate the exact solution even for very small inputs (e.g., $I_{ext} = 1$ Hz; left), whereas b can be set to lower values (e.g., $b = 0$; right) if the input is larger. The gray-scale value represents the convergence rate $|\lambda|$ (in units of s^{-1}) of the activation dynamics for $\tau = 10$ ms. The bottom panel shows in color the exact regions of contraction for inputs $I_{ext} = 1, 10, 100$ Hz and the approximation given by condition (11) for $b = 0, 1, 2$. Some of the colored regions (and dotted lines) correspond to the ones shown in the upper panels. It can be seen that for higher input strengths the upper bound on A_{exc} (or equivalently, the lower bound on w_{max}) quickly converges to the $b = 0$ diagonal, which represents the asymptotic condition for $I_{exc} \to \infty$. For these plots, the learning rule parameters were set to $\Theta_{exc} = 6$ Hz and $\Theta_{inh} = 18$ Hz.

condition for contraction, indicating that (11) is indeed sufficient and that b can safely be set to a value around 1 in most cases.

2.4.4. Extension to multiple units

So far, we have only studied a small network that can be regarded as a single subunit of a larger, distributed WTA system. However, our results can be generalized to larger systems without much effort. In our model, as illustrated in **Figures 1D,E**, different localized WTA circuits can be coupled via excitatory projections. These projections include excitatory-to-inhibitory connections, as well as reciprocal connections between distant excitatory units. In order to demonstrate the effects of this coupling, we consider two localized subsystems, $\boldsymbol{x} = (x_\mathrm{E},\ x_\mathrm{I})$ and $\boldsymbol{x}' = (x'_\mathrm{E},\ x'_\mathrm{I})$, consisting of one excitatory and one inhibitory unit each. Furthermore, we add projections from x_E to x'_E and x'_I, as required by our model. We denote by $w_{\mathrm{E}\to\mathrm{E}'}$ the strength of the long-range excitatory-to-excitatory connection, while we refer to the long-range excitatory-to-inhibitory connection as $w_{\mathrm{E}\to\mathrm{I}'}$. Note that for the sake of clarity we only consider the unidirectional case $\boldsymbol{x} \to \boldsymbol{x}'$ here, while the symmetric case $\boldsymbol{x} \leftrightarrow \boldsymbol{x}'$ can be dealt with analogously.

We first look at the excitatory-to-inhibitory connections. If only x_E is active and x'_E is silent, then x_I and x'_I are driven by the same presynaptic population (x_E), and $w_{\mathrm{E}\to\mathrm{I}'}$ converges to the same value as $w_{\mathrm{E}\to\mathrm{I}}$. Thus, after convergence, both inhibitory units are perfectly synchronized in their activation when x_E is active, and an equal amount of inhibition can be provided to x_E and x'_E.

Besides synchronization of inhibition, proper WTA functionality also requires the recurrent excitation $w_{\mathrm{E}\to\mathrm{E}'}$ (between the excitatory populations of the different subunits) to converge to sufficiently low values, such that different units compete via the synchronized inhibition rather than exciting each other through the excitatory links. As pointed out by Rutishauser et al. (2012), the network is stable and functions correctly if the recurrent excitation between populations is lower than the recurrent self-excitation, i.e., $w_{\mathrm{E}\to\mathrm{E}'} < w_{\mathrm{E}'\to\mathrm{E}'}$.

We now consider the case where x_E and x'_E receive an external input I_ext. Whenever x'_E alone receives the input, there is no interaction between the two subunits, and the recurrent self-connection $w_{\mathrm{E}'\to\mathrm{E}'}$ converges to the value that was found for the simplified case of a single subunit (section 2.4.2). The same is true for the connection $w_{\mathrm{E}\to\mathrm{E}}$ if x_E alone receives the input. However, in this case x_E and x'_E might also interact via the connection $w_{\mathrm{E}\to\mathrm{E}'}$, which would then be subject to plasticity. As $\boldsymbol{x}$ projects to $\boldsymbol{x}'$, but not vice versa, we require $x_\mathrm{E} > x'_\mathrm{E}$ if both x_E and x'_E receive the same input I_ext, because x_E should suppress x'_E via the long-range competition mechanism. In terms of connection strengths, this means that $w^*_{\mathrm{E}\to\mathrm{I}'}\, w^*_{\mathrm{I}'\to\mathrm{E}'} > w^*_{\mathrm{E}\to\mathrm{E}'}$, i.e., the inhibitory input to x'_E that is due to x_E must be greater than the excitatory input x'_E receives from x_E. In the Methods (section 4.3), we show that a sufficient condition for this to be the case is

$$w_\mathrm{max} > A + b + 1, \tag{12}$$

which alters our results from section 2.4.3 only slightly, effectively shifting the lower bound on w_max by an offset of 1, as can be seen by comparing conditions (11) and (12). On the other hand, making use of the fact that $x'_\mathrm{E} < x_\mathrm{E}$, it can be shown that $w_{\mathrm{E}\to\mathrm{E}'}$ converges to a value smaller than $w_{\mathrm{E}'\to\mathrm{E}'}$ (see Methods section 4.3), as required by the stability condition mentioned above.

2.5. GAIN CONTROL AND NORMALIZATION

In the previous section, we showed how synaptic plasticity can be used to drive the connection strengths toward regimes which guarantee stable network dynamics. Since the actual fixed point values of the weights change with the training input, this mechanism can as well be used to tune certain functional properties of the network. Here we focus on controlling the gain of the network, i.e., the relationship between the strength of the strongest input and the activation of the winning excitatory units within the recurrent circuit, as a function of the training input.

In the case of a single active population, the gain is given by $\Lambda = x_\mathrm{E}/I_\mathrm{ext} = (1 - w^*_{\mathrm{E}\to\mathrm{E}} + w^*_{\mathrm{E}\to\mathrm{I}}\, w^*_{\mathrm{I}\to\mathrm{E}})^{-1}$, as can be inferred from Equation (7). Depending on the gain, the network can either amplify ($\Lambda > 1$) or weaken ($\Lambda < 1$) the input signal.

Figure 6 shows how the gain varies as a function of the learning rule parameters and the training input strength I_ext. Low average input strengths cause the weights to converge to values that lead to an increased gain, while higher training inputs lower the gain. This can be regarded as a homeostatic mechanism, acting to keep the network output within a preferred range. This provides a mechanism for the network to adapt to a wide range of input strengths, while still allowing stable WTA competition.

2.6. SIMULATION RESULTS

As a final step, we verify the analytical results in software simulations of a distributed, plastic WTA network, as illustrated in **Figures 1D,E**. Note that here we consider the case where two subgroups are coupled bidirectionally via excitatory long-range projections, while in section 2.4.4, for the sake of clarity, we focus on the unidirectional case. The desired functionality of

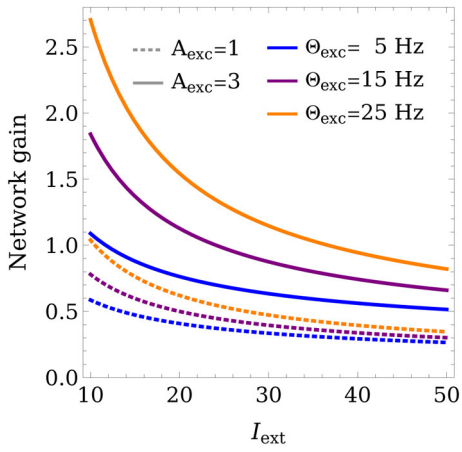

FIGURE 6 | Control of the network gain. The network gain is plotted for different learning rule parameters Θ_exc, A_exc as a function of the training input strength I_ext. Θ_inh was set to 18 Hz and w_max was set to 4.

the resulting network is global competition between the excitatory populations, i.e., the population that receives the strongest input should suppress activation of the other populations, even if the excitatory populations are not directly competing via the same, local inhibitory population. We consider a network with two groups, each consisting of two excitatory populations and one inhibitory population (see **Figure 1E**). While the excitatory populations are connected in an all-to-all manner, inhibitory populations can only target the excitatory populations within their local groups, but do not form long-range projection. Initially, all connection weights (excitatory and inhibitory) are set to random values between 0.3 and 1.8. Note that those values could potentially violate the conditions for contraction defined in (6), but we will show empirically that the plasticity mechanism can still drive the weights toward stable regimes. As training input, we present 1000 constant patterns for 2 s each. In every step, four input values in the ranges 5 ± 2 Hz, 10 ± 2 Hz, 15 ± 2 Hz, and 20 ± 2 Hz are drawn from uniform distributions and applied to the four excitatory units. The different input signals are randomly assigned to the populations in every step, such that a randomly chosen population receives the strongest input. Thereby, each population only receives one of the four inputs.

Figure 7 shows the activation of the different populations before and after learning. Before learning (left), the network does not necessarily implement stable competition between the different excitatory populations. Instead, it may end up in an oscillating state or amplify the wrong winning unit. However, after training (**Figure 7**, right), the network always converges to a stable state representing the winner of the competition. Furthermore, it can be seen that the inhibitory populations perfectly synchronize, as described in section 2.4.4.

The change of weights is illustrated in **Figure 8**: Initially (top), all weights were set to random values in the range [0.3, 1.8]. Since all populations receive the same average input, the weight matrices should converge to symmetric states. For the specific set of learning rule parameters we chose in this example, and the specific input rates described above, $w_{\mathrm{E}\to\mathrm{E}}$ converges to a value around 1, which means that the network is at the edge of the transition between hard and soft WTA behavior. The weights $w_{\mathrm{E}\to\mathrm{I}}$, connecting excitatory to inhibitory units, converge to values around 2. Furthermore, the weights $w_{\mathrm{I}\to\mathrm{E}}$, which connect inhibitory to excitatory units all converge to very similar values (around 1.1), such that inhibition is synchronized across the whole network. Note that not all connections between excitatory populations have converged to the same value. This is because as soon as the network is close to the hard WTA regime, some connections cannot change anymore as only one excitatory unit is active at a time, and the weight change is zero if either the pre- or the post-synaptic unit is inactive.

3. DISCUSSION

We have shown how neural circuits of excitatory and inhibitory neurons can self-organize to implement stable WTA competition. This is achieved through an interplay of excitatory and inhibitory plasticity mechanisms operating on all synapses of the network. As a key result, we provide analytical constraints on the learning rule parameters, which guarantee emergence of the desired network function.

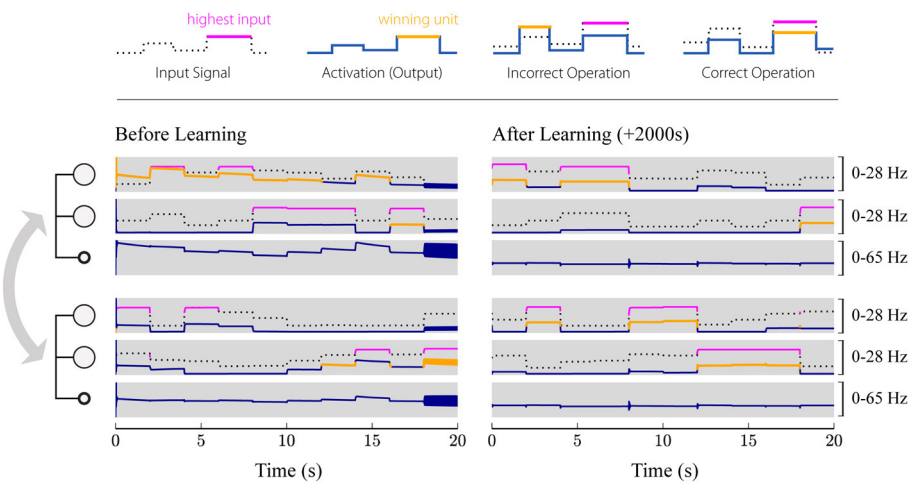

FIGURE 7 | Simulated evolution of the plastic network. 1000 input patterns were applied for 2 s each. Two local subsystems, consisting of two excitatory units and one inhibitory unit each, are coupled via all-to-all excitatory connections, while inhibitory feedback is provided only locally. The first three rows show the populations of the first group, where the first two rows correspond to the two excitatory populations, and the third row shows the activation of the inhibitory population. The last three rows show the activations of the second group. The left panel shows the first 20 s after initialization (before learning), while the right panel shows the network activity during the last 20 s (after learning). Solid blue and orange lines correspond to the firing rate of the respective population, whereby the highlighted segments (orange) mark the winner among the four excitatory populations. The input given to the individual units is plotted as a dotted black or solid magenta line, where the highlighted segments (magenta) correspond to the strongest signal among the four. Thus, the network operates correctly if magenta and orange lines are aligned (the one population that receives the strongest input wins the competition), while misaligned lines (the population receiving the strongest input does not win the competition) indicate incorrect operation. Initially, the network frequently selects the wrong winning unit and even starts oscillating for some input patterns (around 18 s). After training (right), the network converges to a stable state with only the winning unit active for different input patterns.

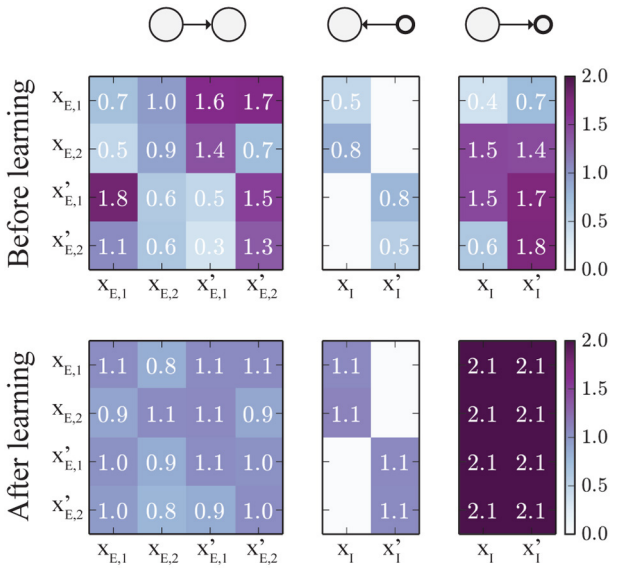

FIGURE 8 | Evolution of the connectivity matrices. The top panels show the (initially random) weight matrices, and the lower panels show the final weight matrices after 2000 s of simulated training time. The left panel shows the connections between excitatory populations. The panel in the middle represents the connections from inhibitory to excitatory populations. Note that some inhibitory connections are set to zero as inhibitory units can only project to targets within their own local group, according to our model. The right panel shows the connections from excitatory to inhibitory units. Here, connections from the same source converge to the same value, leading to perfect synchronization of the two inhibitory units. The (rounded) weight values are displayed on top of the image.

Although constraints on the weights for stable competition in recurrent excitatory-inhibitory circuits have been derived before (Xie et al., 2002; Hahnloser et al., 2003; Rutishauser and Douglas, 2009; Rutishauser et al., 2012), it has remained unclear how a network can self-tune its synaptic weights to comply with these conditions. The presented model achieves this and provides important insights regarding the mechanisms responsible for this self-tuning. Our results predict a relationship between the maximum synaptic weight $w_{\max}$ in a circuit and the learning rule parameter A_{exc}, which controls the contribution of the presynaptic rate to the shifting of the threshold between potentiation and depression. Furthermore, our model predicts a relationship between the network gain and the amount of excitatory input into the circuit during development or training (see **Figure 6**), indicating that high gain (amplification) should be expected for weak inputs, and low gain for strong inputs, which is in accordance with common assumptions about homeostasic mechanisms (Turrigiano, 2011).

From a developmental perspective, the self-configuration of functional WTA circuits through plasticity has the advantage of requiring a smaller number of parameters to be encoded genetically to obtain stable and functional network structures (Zubler et al., 2013). With self-tuning mechanisms like the ones suggested here, only the parameters for the two different types of plasticity in excitatory and inhibitory synapses, rather than the strengths of all synaptic connections, need to be specified, and the network can adapt to the statistics of inputs it receives from its environment and from other brain regions.

Besides guaranteeing stability, it is also desirable to control functional properties of the circuit, such as its gain. Experimental data suggests that cortical recurrent circuits often operate in a high gain regime and with strong (larger than unity) recurrent excitatory feedback (Douglas et al., 1995). The strength of this feedback determines whether the WTA is "soft" (multiple excitatory units can be active at the same time) or "hard" (only one unit can be active at a time, i.e., the network operates in a nonlinear regime) (Rutishauser et al., 2011). Many interesting computations that can be realized with these types of networks rely on the non-linearities introduced by such strong recurrent excitation (e.g., Vapnik, 2000), therefore it is important that similar conditions can be achieved with our model. In addition, various forms of learning rely on balanced WTA competition (Masquelier et al., 2009; Habenschuss et al., 2012; Nessler et al., 2013), which requires an adaptation of the gain as the excitatory connections into the circuit undergo plasticity. In our network, the resulting network gain is a function of both the learning rule parameters and the strength of the training input signals. As a consequence, our system can switch between high and low gain, and hard or soft WTA behavior simply by receiving input stimuli of different (average) strengths. Thus, different parts of the network might develop into different functional modules, depending on the inputs they receive.

Our model does not specifically address the question of how the network structure, which leads to our results (essentially random all-to-all connectivity) might develop in the first place. For instance, if certain long-range connections between multiple subcircuits do not exist initially, they will never be established by our model, and the units of the different subcircuits can never compete. On the one hand, this might be a desired effect, e.g., to construct hierarchies or asymmetric structures for competition, in which some parts of the network are able to suppress other parts, but not vice-versa. On the other hand, structural plasticity could account for the creation of missing synaptic connections, or the removal of ineffective connections if the desired stable function cannot be achieved with the anatomical substrate. There is increasing evidence for activity dependent synapse formation and elimination in both juvenile and adult brains (Butz et al., 2009), in particular a coordinated restructuring of inhibitory and excitatory synapses for functional reorganization (Chen and Nedivi, 2013). Another approach, recently investigated in simulations by Bauer (2013), is to set up the right network topology by developmental self-construction processes in a first step, and the tune the network using synaptic plasticity in a second step.

Our model is based on a weight-dependent variation of the learning rule proposed by Pfister and Gerstner (2006), but this is by no means the only learning rule capable of the self-calibration effect we describe in this article. By changing its parametrization, the rule can subsume a wide variety of commonly used Hebbian, STDP-like, and homeostatic plasticity mechanisms. Indeed, further experiments, which are not presented in this manuscript, indicate that a whole class of learning rules with depression at

low and potentiation at high postsynaptic firing rates would lead to similar results. We chose the triplet rule to demonstrate our findings as its parameters have been mapped to experiments, and also because it can be written in an analytically tractable form. We have assumed here a specific type of inhibitory plasticity, which analytically is of the same form as the simultaneous excitatory plasticity, but uses different parameters. With the parameters we chose for the inhibitory plasticity rule, we obtain a form that is very similar to the one proposed by Vogels et al. (2011). By introducing inhibitory plasticity it is no longer necessary to make common but biologically unrealistic assumptions, like pre-specified constant and uniform inhibitory connection strengths (Oster et al., 2009), or more abstract forms of summing up the excitatory activity in the circuit (Jug et al., 2012; Nessler et al., 2013), because inhibitory weights will automatically converge toward stable regions. Inhibitory plasticity has received more attention recently with the introduction of new measurement techniques, and has revealed a great diversity of plasticity mechanisms, in line with the diversity of inhibitory cell types (Kullmann and Lamsa, 2011; Kullmann et al., 2012). Our model involves only a single inhibitory population per local sub-circuit, which interacts with all local excitatory units. Not only is this a common assumption in most previous models, and greatly simplifies the analysis, but also is in accordance with anatomical and electrophysiological results of relatively unspecific inhibitory activity in sensory cortical areas (Kerlin et al., 2010; Bock et al., 2011). However, recent studies have shown more complex interactions of different inhibitory cell types (Pfeffer et al., 2013), making models based on diverse cell types with different properties an intriguing target for future studies. The assumption of a common inhibitory pool that connects to all excitatory units is justified for local circuits, but violates anatomical constraints on the length of inhibitory axons if interacting populations are far apart (Binzegger et al., 2005). Our results easily generalize to the case of distributed inhibition, by adapting the model of Rutishauser et al. (2012) (see **Figure 1E**). Our contribution is to provide the first learning theory for these types of circuits.

Since our model is purely rate-based, a logical next step is to investigate how it translates into the spiking neural network domain. Establishing similar constraints on spike-based learning rules that enable stable WTA competition remains an open problem for future research, although Chen et al. (2013) have shown empirically that WTA behavior in a circuit with topologically ordered input is possible under certain restrictions on initial synapse strengths, and in the presence of STDP and short-term plasticity. Spiking WTA circuits can potentially utilize the richer temporal dynamics of spike trains in the sense that the order of spikes and spike-spike correlations have an effect on the connectivity.

Potential practical applications of our model, and future spiking extensions, lie in neuromorphic VLSI circuits, which have to deal with the problem of device mismatch (Indiveri et al., 2011), and can thus not be precisely configured a priori. Our model could provide a means for the circuits to self-tune and autonomously adapt to the peculiarities of the hardware.

4. MATERIALS AND METHODS

4.1. DERIVATION OF THE PLASTICITY MECHANISM

The learning rule given by Equation (3) is based on the triplet STDP rule by Pfister and Gerstner (2006). Since we are interested in the rate dynamics, we use the mean-field approximation of this rule, which is provided by the authors and leads to an expected weight change of

$$\dot{w} = x_{\text{pre}} x_{\text{post}} \left(A_2^+ \tau_+ - A_2^- \tau_- + A_3^+ \tau_+ \tau_y x_{\text{post}} - A_3^- \tau_- \tau_x x_{\text{pre}} \right), \tag{13}$$

where x_{pre}, x_{post} are the pre- and postsynaptic activations and $A_2^{\pm}$, $A_3^{\pm}$, $\tau_{\pm}$, $\tau_{x,y}$ are parameters that determine the amplitude of weight changes in the triplet STDP model. All of the parameters are assumed to be positive. Through a substitution of constants given by

$$\tau_s^2 := A_3^+ \tau_+ \tau_y, \tag{14}$$

$$\Theta_w := \left(A_2^- \tau_- - A_2^+ \tau_+ \right) / \tau_s^2, \tag{15}$$

$$A_w := A_3^- \tau_- \tau_x / \tau_s^2, \tag{16}$$

the rule in Equation (13) can be written in the simpler form

$$\dot{w} = \tau_s^2 x_{\text{pre}} x_{\text{post}} \left(x_{\text{post}} - (\Theta_w + A_w x_{\text{pre}}) \right), \tag{17}$$

where Θ_w is in units of a firing rate and A_w is a unitless constant. The terms in parentheses on the right of Equation (17) can be divided into a positive (LTP) part that depends on x_{post}, and a negative (LTD) part that depends on x_{pre}. In order to constrain the range of weights, we add weight-dependent terms $m_+(w)$ and $m_-(w)$ to the two parts of the rule, which yields

$$\dot{w} = \tau_s^2 x_{\text{pre}} x_{\text{post}} \left(x_{\text{post}} m_+(w) - (\Theta_w + A_w x_{\text{pre}}) m_-(w) \right). \tag{18}$$

Throughout this manuscript, we use a simple, linear weight dependence $m_+ = w_{\max} - w$ and $m_- = w$, which effectively limits the possible values of weights to the interval $[0, w_{\max}]$. We chose this form, which is described by a single parameter, for reasons of analytical tractability and because it is consistent with experimental findings (Gütig et al., 2003). In Pfister and Gerstner (2006), values for the parameters $\tau_{x,y}$, $\tau_{\pm}$, and $A_{2,3}^{\pm}$ of the rule Equation (13) were determined from fits to experimental measurements in pyramidal cells in visual cortex (see **Table 1**) and hippocampal cultures (Bi and Poo, 1998, 2001; Sjöström et al., 2001; Wang et al., 2005). We used these values to calculate plausible values for Θ_w, A_w, and τ_s using Equations (14) to (16). In our simulations, we use parameters very similar to the experimentally derived values in **Table 1**. Specifically, for inhibitory connections we use parameters very similar to the ones found from fits of experimental data to the triplet STDP model with all-to-all spike interactions. On the other hand, we choose parameters for the excitatory plasticity rules which are close to fits of the triplet STDP rule with nearest-neighbor spike interactions. The parameters that were used in software simulations and to obtain most of the numeric results are listed in **Table 2**. Note that for the

Table 2 | Model parameters used in software simulation.

Parameter	Value	Description
Θ_{exc}	6 Hz	Learning rule parameter
Θ_{inh}	18 Hz	Learning rule parameter
A_{exc}	2	Learning rule parameter
w_{max}	4	Maximum weight value
$\tau^2_{s,exc}$	3.6 ms^2	Exc. connection learning rate parameter
$\tau^2_{s,inh}$	1.3 ms^2	Inh. connection learning rate parameter
τ_{exc}	5 ms	Exc. population time constant
τ_{inh}	1 ms	Inh. population time constant

weight-dependent rule in Equation (18) we have assumed that the parameter Θ_w influences only the LTD part. According to the definition in Equation (15), this is the case if $A_2^- \gg A_2^+$, or $\Theta_w \approx A_2^- \tau_-/\tau_s$, respectively. Otherwise Θ_w contains both a potentiating (A_2^+) and a depressing (A_2^-) component, and Equation (18) should be replaced with a more complex expression of the form of Equation (13).

4.2. DERIVATION OF THE STABILITY CRITERIA

In section 2.4, we outlined how the fixed points and stability criteria for the WTA system can be found. In this section, we provide the detailed derivations that led to these results.

As described in section 2.4, we first consider a simplified system of one excitatory and one inhibitory population, x_E and x_I, which yield an activation vector $\boldsymbol{x} = (x_E, x_I)^T$. They are coupled recurrently through a weight matrix $\boldsymbol{W} = \begin{bmatrix} w_{E\to E} & w_{I\to E} \\ w_{E\to I} & 0 \end{bmatrix}$, receive external inputs $I_{ext}(t)$ with weights μ_E and μ_I respectively, and have thresholds T_E, T_I. Assuming that both units are active, i.e., their total synaptic input is larger than their thresholds, their dynamics are described by

$$\tau_{exc}\dot{x}_E = -x_E + w_{E\to E}x_E - w_{I\to E}x_I + \mu_E I_{ext} - T_E, \quad (19)$$

$$\tau_{inh}\dot{x}_I = -x_I + w_{E\to I}x_E + \mu_I I_{ext} - T_I, \quad (20)$$

where τ_{exc}, τ_{inh} are the population time constants. The fixed points of the activations can be found by setting $\dot{x}_E = \dot{x}_I = 0$. If we assume, for simplicity, that $T_E = T_I = 0$ this yields the fixed points

$$x_E^* = \Lambda I_{ext}\left(\mu_E - w_{I\to E}\mu_I\right), \quad (21)$$

$$x_I^* = \Lambda I_{ext}\left(w_{E\to I}\mu_E - (w_{E\to E} - 1)\mu_I\right). \quad (22)$$

where

$$\Lambda = \left(1 - w_{E\to E} + w_{E\to I}w_{I\to E}\right)^{-1} \quad (23)$$

is the network gain. Furthermore, we can make the assumption that $\mu_I = 0$ and $\mu_E = 1$, effectively disabling the external input to the inhibitory population. This reduces Equations (21) and (22) to

$$x_E^* = \Lambda I_{ext}, \quad (24)$$

$$x_I^* = \Lambda w_{E\to I} I_{ext}. \quad (25)$$

These simplifications do not change the results of our analysis qualitatively and can be made without loss of generality.

Approximating x_{pre} and x_{post} by their fixed point activities (as described in section 2.4), and setting $\dot{w} = 0$ in the learning rule Equation (18), the fixed point of the weight dynamics (with $w > 0$) takes the form

$$w^* = \frac{w_{max}x_{post}^*}{\Theta_w + A_w x_{pre}^* + x_{post}^*}. \quad (26)$$

Note that this fixed point in weight space always exists for any given x_{pre} and x_{post}, and is stable for the weight dependence $m_+(w) = w_{max} - w$; $m_-(w) = w$ that we chose in Equation (18). In fact, this is true for all choices of the weight dependence satisfying $\partial m_+/\partial w < 0$ and $\partial m_-/\partial w > 0$, as can be shown by means of a linear stability analysis.

We now derive the fixed points for the weights $w_{E\to E}$, $w_{E\to I}$, and $w_{I\to E}$ of the simplified system. For $w_{E\to E}$, Equation (26) can be simplified by noting that $x_{pre}^* = x_{post}^* = x_E^*$, leading to an expression that depends on the activation of the excitatory population x_E^*:

$$w_{E\to E}^* = \frac{w_{max}}{\Theta_{exc}/x_E^* + A_{exc} + 1}. \quad (27)$$

Similarly, we can compute the fixed point of $w_{E\to I}$ as a function of x_E^*, noting that $x_{post}^* = x_I^* = w_{E\to I}x_E^*$ [see Equations (24) and (25)]:

$$w_{E\to I}^* = w_{max} - \Theta_{exc}/x^* - A_{exc}. \quad (28)$$

Finally, using the relationship $x_I^* = w_{E\to I}x_E^*$ from Equations (24) and (25), and the previously computed value of $w_{E\to I}$ from Equation (28) with the fixed point equation for $w_{I\to E}$, we obtain

$$w_{I\to E}^* = \frac{w_{max}}{\Theta_{inh}/x_E^* - A_{inh}\left(\Theta_{exc}/x_E^* + (A_{exc} - w_{max})\right) + 1}. \quad (29)$$

In the following, we set $A_{inh} = 0$, as described in section 2.3. An exact solution for the activation x_E^* at the fixed point of the system is obtained by inserting $w_{E\to E}^*$, $w_{E\to I}^*$, and $w_{I\to E}^*$ into Equation (24), and solving the resulting fixed-point problem $x_E^* = f(x_E^*)$. This corresponds to finding the roots of the third order polynomial

$$P(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 = 0 \quad (30)$$

with coefficients

$$a_0 = \Theta_{\text{exc}}\Theta_{\text{inh}} I_{\text{ext}}, \tag{31}$$

$$a_1 = -\Theta_{\text{exc}}\Theta_{\text{inh}} + \Theta_{\text{exc}} I_{\text{ext}} + \Theta_{\text{inh}} I_{\text{ext}} + \Theta_{\text{inh}} A_{\text{exc}} I_{\text{ext}} + \Theta_{\text{exc}}^2 w_{\max}, \tag{32}$$

$$a_2 = -\Theta_{\text{exc}} - \Theta_{\text{inh}} - \Theta_{\text{inh}} A_{\text{exc}} + I_{\text{ext}} + A_{\text{exc}} I_{\text{ext}} + \Theta_{\text{exc}} w_{\max} + \Theta_{\text{inh}} w_{\max} + 2\Theta_{\text{exc}} A_{\text{exc}} w_{\max} - \Theta_{\text{exc}} w_{\max}^2, \tag{33}$$

$$a_3 = -1 - A_{\text{exc}} + w_{\max} + A_{\text{exc}} w_{\max} + A_{\text{exc}}^2 w_{\max} - w_{\max}^2 - A_{\text{exc}} w_{\max}^2. \tag{34}$$

The activation of the excitatory population x_{E} at the fixed point is then given by the positive, real root of Equation (30).

The fixed point of the activation x_{E}^*, and thus the fixed points of the weights, are monotonic functions of the training input strength I_{ext} (see **Figure 3**, for example). In the following, we investigate the behavior of the fixed point weight values for very large and very small external inputs during training, respectively. This helps us to find conditions on the learning rule parameters that lead to stable dynamics (of the network activation) for any training input strength. We define a positive constant $b := \Theta_{\text{exc}}/x_{\text{E}}^*$, and plug it into Equations (27)–(29). This yields

$$w_{\text{E}\to\text{E}}^* = \frac{w_{\max}}{A_{\text{exc}} + b + 1}, \tag{35}$$

$$w_{\text{E}\to\text{I}}^* = w_{\max} - A_{\text{exc}} - b, \tag{36}$$

$$w_{\text{I}\to\text{E}}^* = \frac{w_{\max}\Theta_{\text{exc}}}{b\,\Theta_{\text{inh}} + \Theta_{\text{exc}}}. \tag{37}$$

Inserting Equations (35)–(37) into the condition for contraction of the activation dynamics given by (6), we can describe the condition in terms of the learning rule parameters, and a new constant $\tilde{\Theta} := \Theta_{\text{exc}}/(\Theta_{\text{exc}} + b\,\Theta_{\text{inh}})$:

$$\frac{1}{(1 + A_{\text{exc}} + b)} < \tilde{\Theta} < 1, \tag{38}$$

$$(A_{\text{exc}} + b)\left(1 + \frac{1}{\tilde{\Theta}(1 + A_{\text{exc}} + b)^2 - 1}\right) < w_{\max} < 2(1 + A_{\text{exc}} + b), \tag{39}$$

Assuming $A_{\text{exc}} + b \gg 1$ (note that we can always set A_{exc} to a sufficiently large value), the conditions reduce to

$$0 < \tilde{\Theta} < 1, \tag{40}$$

$$A_{\text{exc}} + b < w_{\max} < 2(1 + A_{\text{exc}} + b), \tag{41}$$

whereby the first condition can be dropped, since $\tilde{\Theta} \in [0, 1]$ always holds. The second condition still depends on b, and therefore on x_{E}^*. We will illustrate how to eliminate this dependence under very weak assumptions. First, in the limit of very large inputs x_{E}^* also takes very large values, leading to $b \to 0$ for $I_{\text{ext}} \to \infty$. In that case, condition (41) becomes independent of b and can be written as

$$A_{\text{exc}} < w_{\max} < 2(1 + A_{\text{exc}}). \tag{42}$$

On the other hand, in the case of very small inputs we have to include the effects of b, as b can in principle take very large values. In typical scenarios the output of the network can be assumed to be roughly of the order of its input. If Θ_{exc} is chosen to be of the same order, then $b \approx 1$. For any finite b, we can express the stability condition that is valid for all inputs as the intersection of the conditions for large inputs, condition (42), with the one for arbitrarily small inputs, condition (41), leading to

$$A_{\text{exc}} + b < w_{\max} < 2(1 + A_{\text{exc}}). \tag{43}$$

Note that this condition can be met for any finite b by choosing sufficiently large A_{exc} and $w_{\max}$. However, as discussed above, choices of the parameter b of the order 1 should be sufficient for typical scenarios, whereas higher values would guarantee stable dynamics for very low input strengths (e.g., $I_{\text{ext}} \ll \Theta_{\text{exc}}$). This is illustrated in **Figure 5**, where the exact regions of stability as a function of $w_{\max}$ and A_{exc} are shown for different training input strengths, together with the sufficient conditions given by (43). In practice, a good starting point for picking a value b for which the stability conditions should hold is to determine the minimum non-zero input $I_{\min}$ encountered during training for which this condition should hold, and setting $b = \Theta_{\text{exc}}/x_{\text{E,min}}^*$, where $x_{\text{E,min}}^*$ is the fixed point activation for $I_{\text{ext}} = I_{\min}$.

4.3. EXTENSION TO MULTIPLE UNITS

In this section, we illustrate how multiple subunits, as analyzed in the previous section, can be combined to larger WTA networks with distributed inhibition. For the sake of simplicity, we only consider the unidirectional case, where a subunit $\boldsymbol{x} = (x_{\text{E}}, x_{\text{I}})$ projects onto another subunit $\boldsymbol{x}' = (x'_{\text{E}}, x'_{\text{I}})$ via excitatory connections $w_{\text{E}\to\text{E}'}$ and $w_{\text{E}\to\text{I}'}$. The bidirectional case $\boldsymbol{x} \leftrightarrow \boldsymbol{x}'$ can be analyzed analogously. If x_{E} and x'_{E} receive the same input, the response of x'_{E} should be weaker, such that activation of x_{E} causes suppression of x'_{E} rather than excitation. This means that

$$w_{\text{E}\to\text{E}'}^* < w_{\text{E}\to\text{I}'}^* w_{\text{I}'\to\text{E}'}^* \tag{44}$$

must hold. We assume that both subsystems have been trained on inputs of the same average strength, such that their local connections have converged to the same weights, i.e., $w_{\text{E}'\to\text{I}'}^* = w_{\text{E}\to\text{I}}^*$ and $w_{\text{I}'\to\text{E}'}^* = w_{\text{I}\to\text{E}}^*$. Furthermore, we assume that condition (44) is true initially. This can be guaranteed by setting the initial value of $w_{\text{E}\to\text{E}'}$ to a sufficiently small number. Our task then is to show that condition (44) remains true for all time. The values of $w_{\text{E}'\to\text{I}'}^*$ and $w_{\text{I}'\to\text{E}'}^*$, or $w_{\text{E}\to\text{I}}^*$ and $w_{\text{I}\to\text{E}}^*$ respectively, are described by Equations (36) and (37). On the other hand, according to Equation (26), the value of $w_{\text{E}\to\text{E}'}^*$ is given by

$$w_{\text{E}\to\text{E}'}^* = \frac{w_{\max} x_{\text{E}}'^*}{\Theta_{\text{exc}} + A_{\text{exc}} x_{\text{E}}^* + x_{\text{E}}'^*}. \tag{45}$$

Plugging all this into condition (44) and simplifying the expression, leads to the condition

$$w_{\max} > A_{\text{exc}} + b + \frac{x'_{\text{E}}}{A_{\text{exc}} x_{\text{E}} + x'_{\text{E}} + \Theta}, \tag{46}$$

which can be replaced by the sufficient condition

$$w_{\max} > A_{\text{exc}} + b + 1, \tag{47}$$

that guarantees $x^*_{\text{E}'} < x^*_{\text{E}}$ if both excitatory populations receive the same input. On the other hand, this result implies $w^*_{\text{E}\to\text{E}'} < w^*_{\text{E}'\to\text{E}'}$, which is required for stable network dynamics (Rutishauser et al., 2012), and can be verified by comparing the respective fixed point equations

$$w^*_{\text{E}\to\text{E}'} = x'_{\text{E}} / \left(\Theta_{\text{exc}} + A_{\text{exc}} x_{\text{E}} + x'_{\text{E}}\right), \tag{48}$$

$$w^*_{\text{E}'\to\text{E}'} = x'_{\text{E}} / \left(\Theta_{\text{exc}} + A_{\text{exc}} x'_{\text{E}} + x'_{\text{E}}\right). \tag{49}$$

4.4. SOFTWARE SIMULATION

Software simulations of our model were implemented using custom Python code based on the "NumPy" and "Dana" packages, and run on a Linux workstation. Numerical integration of the system dynamics was carried out using the forward Euler method with a 1 ms timestep.

AUTHOR CONTRIBUTIONS

Jonathan Binas, Ueli Rutishauser, Giacomo Indiveri, Michael Pfeiffer conceived and designed the experiments. Jonathan Binas performed the experiments and analysis. Jonathan Binas, Ueli Rutishauser, Giacomo Indiveri, Michael Pfeiffer wrote the paper.

FUNDING

The research was supported by the Swiss National Science Foundation Grant 200021_146608, and the European Union ERC Grant "neuroP" (257219).

ACKNOWLEDGMENT

We thank Rodney Douglas, Peter Diehl, Roman Bauer, and our colleagues at the Institute of Neuroinformatics for fruitful discussion.

REFERENCES

Amari, S., and Arbib, M. (1977). "Competition and cooperation in neural nets," in *Systems Neuroscience*, ed J. Metzler (San Diego, CA: Academic Press), 119–165.

Amari, S.-I. (1980). Topographic organization of nerve fields. *Bull. Math. Biol.* 42, 339–364. doi: 10.1007/BF02460791

Bauer, R. (2013). *Self-Construction and -Configuration of Functional Neuronal Networks*. PhD Thesis, ETH Zürich.

Ben-Yishai, R., Bar-Or, R. L., and Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 92, 3844–3848. doi: 10.1073/pnas.92.9.3844

Bi, G., and Poo, M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* 18, 10464–10472.

Bi, G., and Poo, M. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Ann. Rev. Neurosci.* 24, 139–166. doi: 10.1146/annurev.neuro.24.1.139

Binzegger, T., Douglas, R. J., and Martin, K. (2004). A quantitative map of the circuit of cat primary visual cortex. *J. Neurosci.* 24, 8441–8453. doi: 10.1523/JNEUROSCI.1400-04.2004

Binzegger, T., Douglas, R. J., and Martin, K. A. (2005). Axons in cat visual cortex are topologically self-similar. *Cereb. Cortex* 15, 152–165. doi: 10.1093/cercor/bhh118

Bock, D. D., Lee, W.-C. A., Kerlin, A. M., Andermann, M. L., Hood, G., Wetzel, A. W., et al. (2011). Network anatomy and *in vivo* physiology of visual cortical neurons. *Nature* 471, 177–182. doi: 10.1038/nature09802

Butz, M., Wörgötter, F., and van Ooyen, A. (2009). Activity-dependent structural plasticity. *Brain Res. Rev.* 60, 287–305. doi: 10.1016/j.brainresrev.2008.12.023

Carandini, M., and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* 13, 51–62. doi: 10.1038/nrn3136

Chen, J. L., and Nedivi, E. (2013). Highly specific structural plasticity of inhibitory circuits in the adult neocortex. *Neuroscientist* 19, 384–393. doi: 10.1177/1073858413479824

Chen, Y., McKinstry, J. L., and Edelman, G. M. (2013). Versatile networks of simulated spiking neurons displaying winner-take-all behavior. *Front. Comput. Neurosci.* 7:16. doi: 10.3389/fncom.2013.00016

Coultrip, R., Granger, R., and Lynch, G. (1992). A cortical model of winner-take-all competition via lateral inhibition. *Neural Netw.* 5, 47–54. doi: 10.1016/S0893-6080(05)80006-1

Douglas, R., Koch, C., Mahowald, M., Martin, K., and Suarez, H. (1995). Recurrent excitation in neocortical circuits. *Science* 269, 981–985. doi: 10.1126/science.7638624

Douglas, R. J., and Martin, K. A. (1991). Opening the grey box. *Trends Neurosci.* 14, 286–293. doi: 10.1016/0166-2236(91)90139-L

Douglas, R. J., and Martin, K. A. (2004). Neuronal circuits of the neocortex. *Ann. Rev. Neurosci.* 27, 419–451. doi: 10.1146/annurev.neuro.27.070203.144152

Douglas, R. J., and Martin, K. A. (2007). Recurrent neuronal circuits in the neocortex. *Curr. Biol.* 17, R496–R500. doi: 10.1016/j.cub.2007.04.024

Douglas, R. J., Martin, K. A., and Whitteridge, D. (1989). A canonical microcircuit for neocortex. *Neural Comput.* 1, 480–488. doi: 10.1162/neco.1989.1.4.480

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202. doi: 10.1007/BF00344251

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

Habenschuss, S., Bill, J., and Nessler, B. (2012). "Homeostatic plasticity in Bayesian spiking networks as Expectation Maximization with posterior constraints," in *Proceedings of Neural Information Processing Systems (NIPS)*, 782–790.

Hahnloser, R., Douglas, R., Mahowald, M., and Hepp, K. (1999). Feedback interactions between neuronal pointers and maps for attentional processing. *Nat. Neurosci.* 2, 746–752. doi: 10.1038/11219

Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., and Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405, 947–951. doi: 10.1038/35016072

Hahnloser, R. H., Seung, H. S., and Slotine, J.-J. (2003). Permitted and forbidden sets in symmetric threshold-linear networks. *Neural Comput.* 15, 621–638. doi: 10.1162/089976603321192103

Hertz, J., Krogh, A., and Palmer, R. (1991). *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison-Wesley.

Indiveri, G., Linares-Barranco, B., Hamilton, T. J., Van Schaik, A., Etienne-Cummings, R., Delbruck, T., et al. (2011). Neuromorphic silicon neuron circuits. *Front. Neurosci.* 5:73. doi: 10.3389/fnins.2011.00073

Jug, F., Cook, M., and Steger, A. (2012). "Recurrent competitive networks can learn locally excitatory topologies," in *International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Kerlin, A. M., Andermann, M. L., Berezovskii, V. K., and Reid, R. C. (2010). Broadly tuned response properties of diverse inhibitory neuron subtypes in mouse visual cortex. *Neuron* 67, 858–871. doi: 10.1016/j.neuron.2010.08.002

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybern.* 43, 59–69. doi: 10.1007/BF00337288

Kullmann, D. M., and Lamsa, K. P. (2011). LTP and LTD in cortical GABAergic interneurons: emerging rules and roles. *Neuropharmacology* 60, 712–719. doi: 10.1016/j.neuropharm.2010.12.020

Kullmann, D. M., Moreau, A. W., Bakiri, Y., and Nicholson, E. (2012). Plasticity of inhibition. *Neuron* 75, 951–962. doi: 10.1016/j.neuron.2012.07.030

Lohmiller, W., and Slotine, J.-J. E. (1998). On contraction analysis for non-linear systems. *Automatica* 34, 683–696. doi: 10.1016/S0005-1098(98)00019-3

Maass, W. (2000). On the computational power of winner-take-all. *Neural Comput.* 12, 2519–2536. doi: 10.1162/089976600300014827

Masquelier, T., Guyonneau, R., and Thorpe, S. (2009). Competitive STDP-based spike pattern learning. *Neural Comput.* 21, 1259–1276. doi: 10.1162/neco.2008.06-08-804

Mountcastle, V. B. (1997). The columnar organization of the neocortex. *Brain* 120, 701–722. doi: 10.1093/brain/120.4.701

Neftci, E., Binas, J., Rutishauser, U., Chicca, E., Indiveri, G., and Douglas, R. J. (2013). Synthesizing cognition in neuromorphic electronic systems. *Proc. Natl. Acad. Sci. U.S.A.* 110, E3468–E3476. doi: 10.1073/pnas.1212083110

Nessler, B., Pfeiffer, M., Buesing, L., and Maass, W. (2013). Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity. *PLoS Comput. Biol.* 9:e1003037. doi: 10.1371/journal.pcbi.1003037

Oster, M., Douglas, R., and Liu, S.-C. (2009). Computation with spikes in a winner-take-all network. *Neural Comput.* 21, 2437–2465. doi: 10.1162/neco.2009.07-08-829

Pfeffer, C. K., Xue, M., He, M., Huang, Z. J., and Scanziani, M. (2013). Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nat. Neurosci.* 16, 1068–1076. doi: 10.1038/nn.3446

Pfister, J.-P., and Gerstner, W. (2006). Triplets of spikes in a model of spike timing-dependent plasticity. *J. Neurosci.* 26, 9673–9682. doi: 10.1523/JNEUROSCI.1425-06.2006

Rabinovich, M. I., Huerta, R., Volkovskii, A., Abarbanel, H. D. I., Stopfer, M., and Laurent, G. (2000). Dynamical coding of sensory information with competitive networks. *J. Physiol. (Paris)* 94, 465–471. doi: 10.1016/S0928-4257(00)01092-5

Rutishauser, U., and Douglas, R. J. (2009). State-dependent computation using coupled recurrent networks. *Neural Comput.* 21, 478–509. doi: 10.1162/neco.2008.03-08-734

Rutishauser, U., Douglas, R. J., and Slotine, J.-J. (2011). Collective stability of networks of winner-take-all circuits. *Neural Comput.* 23, 735–773. doi: 10.1162/NECO-a-00091

Rutishauser, U., Slotine, J.-J., and Douglas, R. J. (2012). Competition through selective inhibitory synchrony. *Neural Comput.* 24, 2033–2052. doi: 10.1162/NECO-a-00304

Sjöström, P. J., Turrigiano, G. G., and Nelson, S. B. (2001). Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron* 32, 1149–1164. doi: 10.1016/S0896-6273(01)00542-6

Song, S., and Abbott, L. F. (2001). Cortical development and remapping through spike timing-dependent plasticity. *Neuron* 32, 339–350. doi: 10.1016/S0896-6273(01)00451-2

Turrigiano, G. (2011). Too many cooks? Intrinsic and synaptic homeostatic mechanisms in cortical circuit refinement. *Ann. Rev. Neurosci.* 34, 89–103. doi: 10.1146/annurev-neuro-060909-153238

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. New York, NY: Information Science and Statistics, Springer.

Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C., and Gerstner, W. (2011). Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* 334, 1569–1573. doi: 10.1126/science.1211095

von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* 14, 85–100. doi: 10.1007/BF00288907

Wang, H.-X., Gerkin, R. C., Nauen, D. W., and Bi, G.-Q. (2005). Coactivation and timing-dependent integration of synaptic potentiation and depression. *Nat. Neurosci.* 8, 187–193. doi: 10.1038/nn1387

Willshaw, D. J., and Von Der Malsburg, C. (1976). How patterned neural connections can be set up by self-organization. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 194, 431–445. doi: 10.1098/rspb.1976.0087

Xie, X., Hahnloser, R. H., and Seung, H. S. (2002). Selectively grouping neurons in recurrent networks of lateral inhibition. *Neural Comput.* 14, 2627–2646. doi: 10.1162/089976602760408008

Yuille, A., and Geiger, D. (2003). "Winner-take-all networks," in *The Handbook of Brain Theory and Neural Networks*, ed M. Arbib (Cambridge, MA: MIT Press), 1228–1231.

Zubler, F., Hauri, A., Pfister, S., Bauer, R., Anderson, J. C., Whatley, A. M., et al. (2013). Simulating cortical development as a self constructing process: a novel multi-scale approach combining molecular and physical aspects. *PLoS Comput. Biol.* 9:e1003173. doi: 10.1371/journal.pcbi.1003173

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 April 2014; accepted: 16 June 2014; published online: 08 July 2014.
Citation: Binas J, Rutishauser U, Indiveri G and Pfeiffer M (2014) Learning and stabilization of winner-take-all dynamics through interacting excitatory and inhibitory plasticity. Front. Comput. Neurosci. ***8****:68. doi: 10.3389/fncom.2014.00068*
This article was submitted to the journal Frontiers in Computational Neuroscience.