Optimal control of transient dynamics in balanced networks supports generation of complex movements

Guillaume Hennequin*[1,2], Tim P. Vogels[1,3], and Wulfram Gerstner[1]

[1] *School of Computer and Communication Sciences and Brain Mind Institute, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne EPFL, Switzerland*

[2] *Department of Engineering, University of Cambridge, Cambridge, UK*

[3] *Centre for Neural Circuits and Behaviour, The University of Oxford, Oxford, UK*

Abstract

Populations of neurons in motor cortex engage in complex transient dynamics of large amplitude during the execution of limb movements. Traditional network models with stochastically assigned synapses cannot reproduce this behavior. Here we introduce a class of cortical architectures with strong and random excitatory recurrence that is stabilized by intricate, fine-tuned inhibition, optimized from a control theory perspective. Such networks transiently amplify specific activity states and can be used to reliably execute multidimensional movement patterns. Similar to the experimental observations, these transients must be preceded by a steady state initialization phase from which the network relaxes back into the background state by way of complex internal dynamics. In our networks, excitation and inhibition are as tightly balanced as recently reported in experiments across several brain areas, suggesting inhibitory control of complex excitatory recurrence as a generic organizational principle in cortex.

Highlights

- Stability-optimized circuits with strong excitatory pathways produce complex neural dynamics.
- They account for single-cell and population recordings in monkey M1 during reaching.
- Their dynamics form a unique basis that can be assembled into complex motor sequences.
- They predict a novel type of excitation/inhibition balance in single neurons.

*Corresponding author, gjeh2@cam.ac.uk

Introduction

The neural basis for movement generation has been the focus of several recent experimental studies (Churchland et al., 2010, 2012). In a typical experiment (Figure 1A), a monkey is trained to prepare a particular arm movement and execute it after the presentation of a go-cue. Concurrent electrophysiological recordings in cortical motor and pre-motor areas show an activity transition from spontaneous firing into a movement-specific preparatory state with firing rates that remain stable until the go-cue is presented (cf. Figure 1B). Following the go-cue, network dynamics begin to display quickly changing, multiphasic firing rate responses that form spatially and temporally complex patterns and eventually relax towards spontaneous activation levels after movement execution (Churchland and Shenoy, 2007).

Recent studies (Afshar et al. (2011); Shenoy et al. (2011)) have suggested a mechanism similar to a spring loaded box, in which motor populations could act as a generic dynamical system that is driven into specific patterns of collective activity by preparatory, external stimuli (Figure 1). When released, intrinsic population dynamics would commandeer the network activity and orchestrate a sequence of motor commands leading to the correct movement. The requirements for a dynamical system of this sort are manifold. It must be highly malleable during the preparatory period, excitable and fast when movement is triggered, and stable enough to return to rest after an activity transient. Moreover, the dynamics must be sufficiently rich to support complex movement patterns (Maass et al., 2002; Sussillo and Abbott, 2009; Laje and Buonomano, 2012).

How the cortical networks at the heart of this black box (Figure 1C) could generate such complex transient amplification through recurrent interactions is still poorly understood. Randomly connected, globally balanced networks of leaky integrate-and-fire neurons exhibit stable background states (van Vreeswijk and Sompolinsky, 1996; Tsodyks et al., 1997; Brunel, 2000; Vogels et al., 2005; Renart et al., 2010) but cannot autonomously produce the substantial yet reliable, spatially patterned departure from background activity observed in the experiments. Networks with strong recurrent pathways can exhibit ongoing, complex rate fluctuations beyond the population mean (Sompolinsky et al., 1988; Rajan et al., 2010; Sussillo and Abbott, 2009; Litwin-Kumar and Doiron, 2012), but do not capture the transient nature of movement-related activity. Moreover, such rate dynamics are chaotic, and sensitivity to noise seems improper in a situation where the initial conditions are thought to dictate the subsequent evolution of the system. Recent studies have shown that chaos can be controlled either through continuous external feedback loops, or modifications of the recurrent connectivity itself (Sussillo and Abbott, 2009; Hoerzer et al., 2012; Laje and Buonomano, 2012). However all of these models violate Dale's principle, according to which neurons can be either excitatory or inhibitory, but not of a mixed type. This in turn makes it difficult to motivate the firing rate (as opposed to spike-based) formulations of cortical dynamics used in these models. In other words, there is currently no biologically plausible network model to implement the spring loaded box of Figure 1c, i.e. a system that well-chosen inputs can prompt to autonomously generate multiphasic transients of large amplitude.

Here we introduce a new class of neuronal networks composed of excitatory and inhibitory neurons which, similarly to chaotic networks, rely on strong and intricate excitatory synaptic pathways. Because traditional homogeneous inhibition is not enough to quench and balance chaotic firing rate fluctuations in these networks, we build a sophisticated inhibitory counter-structure that successfully dampens chaotic behavior but allows strong and fast break-out transients of activity. This inhibitory architecture is constructed with the help of an optimization algorithm that aims to stabilize the activity of each unit by adjusting the strength of existing inhibitory synapses, or by adding or pruning inhibitory connections. The result is a strongly connected, but non-chaotic, balanced network, which otherwise looks random. We refer to such networks as "Stability-Optimized Circuits", or SOCs. We study both a rate-based formulation of SOC dynamics (which lends itself to theoretical analysis) and a more realistic spiking implementation (which allows us to make more focused physiological predictions). We show that external stimuli can force these networks into unique and stable activity states. Interestingly, when input is withdrawn and the activity is left to evolve freely, the subsequent transient dynamics are in good qualitative agreement with the motor cortex data on single-cell and network-wide levels.

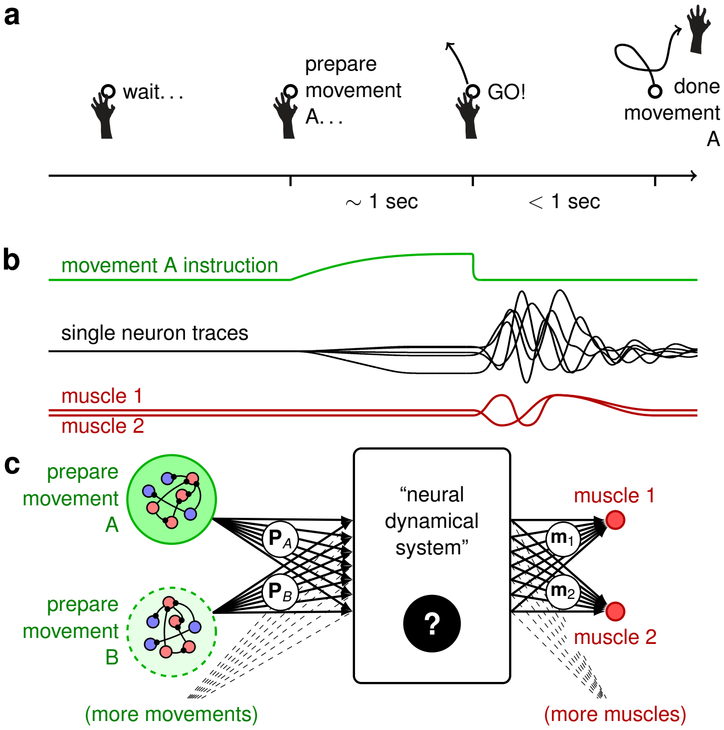


Figure 1: **Dynamical systems view of movement planning and execution.** (**a**) A typical delayed movement generation task starts with the instruction of what movement must be prepared. The arm must then be held still until the go cue is given, upon which the movement is performed. (**b**) During the preparatory period, model neurons receive a ramp input (green). Following the go-cue, that input is withdrawn, leaving the network activity free to evolve from the initial condition set up during the preparatory period. Model neurons then exhibit transient oscillations (black) which drive muscle activity (red). (**c**) Black-box view on movement generation. Muscles (red, right) are thought to be activated by a population of motor cortical neurons ("neural dynamical system", middle). To prepare the movement, this network is initialized in a desired state by the slow activation of a movement-specific pool of neurons (green, left).

We show that SOCs connect hitherto unrelated aspects of balanced cortical dynamics. The mechanism that underlies the generation of large transients here is a more general form of "balanced amplification" (Murphy and Miller, 2009), which was previously discovered in the context of visual cortical dynamics. Additionally, during spontaneous activity in SOCs, a "detailed balance" (Vogels and Abbott, 2009) of excitatory and inhibitory inputs emerges, but much finer than expected from shared population fluctuations (Okun and Lampl, 2008; Cafaro and Rieke, 2010; Renart et al., 2010), beyond also what is possible with recently published inhibitory learning rules (Vogels et al., 2011; Luz and Shamir, 2012) that only alter the weights of inhibitory synapses, but not the structure of the network itself. Prepping such exquisitely balanced systems with an external stimulus will then lead to momentary but dramatic departure from balance, demonstrating how realistically shaped cortical architectures can produce a large library of unique, transient activity patterns that can be decoded into motor commands.

Results

We are interested in studying how neural systems (Figure 1c), can produce the large, autonomous, and stable "spring box dynamics" as described above. We will first investigate how to construct the cortical architectures that display such behavior and show how their activity can be manipulated to produce motor-like activity. We then discuss the implications of this new type of dynamics in terms of how excitation and inhibition interact. While most of our results are obtained within a rate-based formulation of neuronal dynamics, we later confirm them in a more realistic, spiking neural network.

Stability-optimized circuits (SOCs)

We use $N = 200$ interconnected rate units (Dayan and Abbott, 2001; Gerstner and Kistler, 2002; Miller and Fumarola, 2011), of which 100 are excitatory and 100 are inhibitory (later, we will show that our main results also hold in large scale, balanced spiking networks). These "rate units" are best thought of as subgroups of statistically comparable neurons within some area of cortex. We describe the temporal evolution of their “potentials”, gathered in a vector $\mathbf{x}(t)$, according to

$$\tau \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = -\mathbf{x}(t) + \mathbf{I}(t) + \mathbf{W}\,\Delta\mathbf{r}(\mathbf{x}, t) \tag{1}$$

where $\tau = 200$ ms, the combined time constant of membrane and synaptic dynamics, is set to match the dominant timescale in the data of Churchland et al. (2012). $\mathbf{I}(t) = \boldsymbol{\xi}(t) + \mathbf{S}(t)$ denotes all external inputs, i.e. an independent noise term $\xi(t)$ and a specific, patterned, external stimulation $\mathbf{S}(t)$. The vector $\Delta\mathbf{r}(\mathbf{x}, t)$ contains the instantaneous single-unit firing rates, measured relative to a low level of spontaneous activity $r_0 = 5$ Hz. These rates are given by a nonlinear function $\Delta r_i = g(x_i)$ of the potentials (Figure 3e and Experimental Procedures), although we will also consider the linear case $\Delta r_i \propto x_i$ in our analysis. Thus, the final term in Equation 1 accounts for the recurrent dynamics of the system due to its connectivity $\mathbf{W}$.

To study the effect of internal network dynamics we consider two different network structures $\mathbf{W}$. First, we build a reference network in which excitatory and inhibitory connections are drawn randomly with probability $p = 0.1$. The connection sign depends on the nature of the presynaptic neuron (excitatory or inhibitory), and all connections of the same type have equal strengths. Strong excitatory connections result in an unstable activity regime in which small perturbations in firing rates propagate chaotically across the network (Sompolinsky et al., 1988; Rajan et al., 2010). The connectivity matrix $\mathbf{W}$ is characterized by a circular distribution of eigenvalues (Rajan and Abbott, 2006). The unstable, chaotic dynamics manifests itself in the presence of several eigenvalues with real parts much larger than unity (Figure 2b) and can be observed in the large switch-like fluctuations (Figure 2e) in the neurons’ firing rates in the absence of stimulation ($\mathbf{S}(t) = 0$).

In a second, stability-optimized circuit (SOC, Figure 2a) the excitatory connections are identical to those in the reference network (Figure 2c), but the inhibitory connections are no longer drawn randomly. Instead, they have been precisely matched against the excitatory connectivity. This “matching” is achieved by an algorithmic optimization procedure that modifies the inhibitory weights and wiring patterns of the reference network, aiming to pull the unstable eigenvalues of $\mathbf{W}$ towards stability (Experimental Procedures, and Supplementary Movie 1). We end up with an eigenvalue spectrum that is compressed horizontally, with all eigenvalues in the stable regime (Figure 2b, black dots). The total number of inhibitory connections in the inhibition-stabilized network is increased and the distribution of their strengths is wider, but the mean inhibitory weight is kept the same (Figure 2d). The resulting background activity is stable and non-chaotic (Figure 2e), with small noisy fluctuations around the mean caused by $\boldsymbol{\xi}(t)$. Shuffling the optimal inhibitory connectivity results in chaotic dynamics similar to the reference network (not shown), indicating that it is not the broad, sparse distribution of inhibitory weights but the precise inhibitory wiring pattern that stabilizes the dynamics.

SOCs exhibit complex transient amplification

To test whether SOCs can produce the type of complex transient behavior seen in experiments (Churchland and Shenoy (2007), Churchland et al. (2012), cf. Figure 1), we momentarily clamp each unit to a specific firing rate, and observe the network as it relaxes to the background state (later we will model the preparatory period explicitly). Depending on the spatial pattern of initial stimulation, the network activity exhibits a variety of transient behaviors. Some initial conditions result in fast monotonous decay towards rest, while others drive large transient deviations from baseline rate in most neurons.

To quantify this amplifying behavior of the network in response to a stimulus, we introduce the notion of “evoked energy” $\mathcal{E}(\mathbf{a})$, that measures both the amplitude and duration of the collective transient evoked by initial condition $\mathbf{a}$ for a given SOC compared to an unconnected network (Experimental Procedures). Of all initial conditions $\Delta\mathbf{r}$ with constant power $\sigma^2 = \sum_i \Delta r_i^2 / N$, we find the one that maximizes this energy and call it $\mathbf{a}_1$. We repeat this procedure among all patterns orthogonal to $\mathbf{a}_1$ to obtain the second best pattern $\mathbf{a}_2$, and iterate until we have filled a full basis of $N = 200$ orthogonal initial conditions

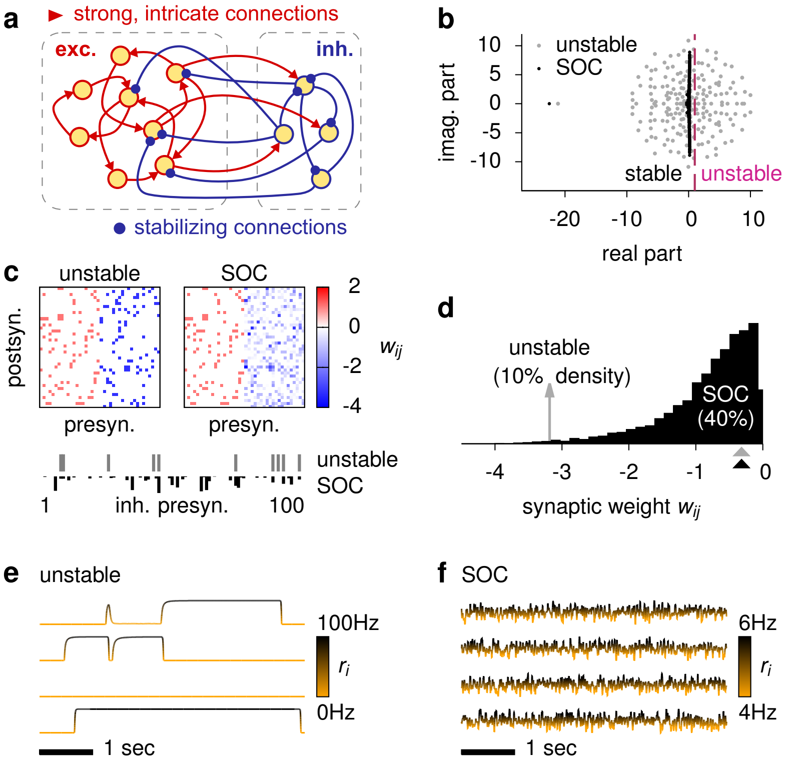


Figure 2: **Random inhibition-stabilized networks (SOCs).** (**a**) Schematic of a SOC. A population of rate units is recurrently connected, with strong and intricate excitatory pathways (red) that would normally produce unstable activity. Stabilization is achieved through fine-tuned inhibitory feedback (blue). (**b**) Eigenvalue spectrum of the connectivity matrix of a SOC (black), and that of the unstable random network from which it is derived (gray). Stability requires all eigenvalues to lie to the left of the dashed vertical line. Note the large negative real eigenvalue, which corresponds to the spatially uniform activity pattern. (**c**) Matrices of synaptic connectivity before (unstable) and after (SOC) stability optimization through inhibitory tuning. By design, the excitatory weights are the same in both matrices. Matrices were thinned out to 40 × 40 (instead of 200 × 200) for visualization purposes. The bottom row shows the strengths of all the inhibitory input synapses to a single sample neuron, in the unstable network (gray) and in the corresponding SOC (black). (**d**) Distribution of inhibitory weights in the unstable network (10% connection density, gray peak at $w_{ij} \simeq -3.18$) and in the stabilized version (40% connection density, black). The mean inhibitory weight is the same before and after optimization ($\simeq -0.318$, gray and black arrowheads). (**e,f**) Spontaneous activity in the unstable network (e) and in the SOC (f), for four example units. Note the difference in firing rate scales.

$\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N\}$ (an analytical solution exists for the linear case, $\Delta r_i \propto x_i$; see Experimental Procedures). A large set of these orthogonal initial conditions are transiently amplified by the connectivity of the network, with the strongest states evoking energies ∼25 times greater than expected from the exponential decay of activity in unconnected neurons (Figure 3a). Interestingly, for these strongly amplifying states, the population-averaged firing rate remains roughly constant during the transient (red line in Figure 3b, middle), but the average *absolute* deviation from baseline firing rate per unit can grow dramatically (Figure 3b, top), because some units become more active and others become less active than baseline. Amplifying behavior progressively attenuates but subsists for roughly the first half of the basis $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{\sim 100})$. Eventually, amplification disappears, and even turns into active dampening of the initial condition (Figure 3a, green dots). For $\mathbf{a}_{200}$, the least amplifying initial condition, return to rest occurs three times faster than it would in unconnected neurons (Figure 3b). Here, the least-preferred state $\mathbf{a}_{200}$ corresponds to the uniform spatial mode of activity $(1, 1, \ldots, 1)$, the trivial case in which all neurons are initialized slightly above (or below) their baseline rate. Notice also that only a fraction of the 200 possible independent initial states evoke an energy that is substantially larger than $\mathcal{E}_0$ – which is the energy that random initial conditions are expected to evoke (cf. black arrowhead Figure 3a) – and thus produce responses that are discernible from a noisy background state.

Finally, if we increase the firing rate standard deviation σ in the initial condition, such that a substantial number of (excitatory *and* inhibitory) neurons will reach lower saturation and stop firing during the transient, the duration of the response increases (Figure 3c,d). For $\sigma > 3$ Hz the network response begins to self-sustain in the chaotic regime (data not shown). This behavior is beyond the scope of our study, and in the following we set $\sigma = 1.5$ Hz, which results in transients of ~ 1 second duration. Note

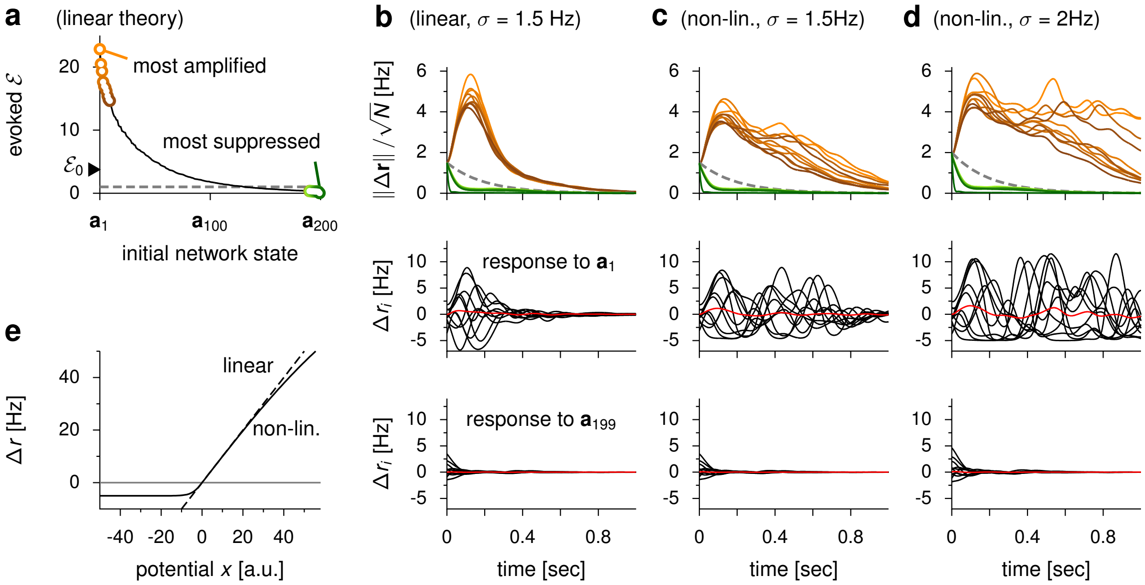


Figure 3: **Transient amplification in SOCs** – (**a**) The energy $\mathcal{E}$ evoked by $N = 200$ orthogonal initial conditions $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N\}$ as the network evolves linearly ($\Delta r_i = x_i$) with no further input according to Equation 1. The energy (Equation 3) is normalized such that it equals 1 for an unconnected network ($\mathbf{W} = 0$) irrespective of the initial condition (dashed horizontal line). Each successive initial condition $\mathbf{a}_i$ is defined as the one that evokes maximum energy, within the subspace orthogonal to all previous input patterns $\mathbf{a}_{j<i}$ (Experimental Procedures). The black arrowhead indicates the mean, or the expected evoked energy $\mathcal{E}_0$ when the neurons are initialized in random activity states. (**b**) Dynamics of the SOC in the linear regime ($\Delta r_i = x_i$). Top: time-evolution of $\|\Delta\mathbf{r}(t)\| / \sqrt{N}$ – which measures the momentary spread of firing rates in the network above or below baseline – as the dynamics unfold from any of the 10 best or 10 worst initial states (same color code as in panel (a)). Initial states have a standard deviation of $\sigma = 1.5$ Hz across the population. The dashed gray line shows $\sigma \exp(-t/\tau)$, i.e. the behavior of an unconnected pool of neurons. Bottom: sample firing rate responses of 10 randomly chosen neurons following initialization in state $\mathbf{a}_1$ or $\mathbf{a}_{199}$. The red line indicates the momentary population-averaged firing rate. (**c,d**) Same as in (b), now with the nonlinear gain function shown in (e). Unlike in the linear case, the dynamics now depend on the spread σ of the initial firing rates across the network (1.5 Hz in (c) as in (b), 2 Hz in (d)). The larger this spread, the longer the duration of the population transient. When $\sigma > 3$ Hz, the network initiates self-sustained (chaotic) activity (not shown). (**e**) Single-unit input-output nonlinearity (solid line, $\Delta r_i = g(x_i)$ given by Equation 2), and its linearization (dashed line, $\Delta r_i = x_i$).

also that we did not observe this return to chaotic behavior in the full spiking network (see below), even though firing rates in the initial conditions deviated more dramatically from baseline.

SOC dynamics are consistent with experimental data

In Churchland et al. (2012), monkeys were trained to perform 27 different cued and delayed arm movements (Figure 1a). The activity of the neurons recorded during this task (Figure 4a) displayed transient activity following the go cue that was similar to the responses of SOCs to appropriate initialization (cf. Figure 3c). To model this behavior, we assume that each of the 27 instructed movements is associated with a pool of prefrontal cortical neurons (Figure 1c), each of which feeds the motor network through a set of properly tuned input weights (Experimental Procedures). For a given movement, the corresponding command pool becomes progressively more active during the one-second-long delay period (Fuster and Alexander, 1971; Amit and Brunel, 1997; Wang, 1999). Remarkably, this simple input drives the SOC into a stable steady state (Figure 4b). By adjusting the movement-specific input weights (Experimental Procedures), we can manipulate this steady state, making it possible to force the network into a specific spatial arrangement of activity. This is not possible in generic chaotic networks in which external inputs are overwhelmed by a strong and uncontrolled recurrent activity. Here, we chose the input weights such

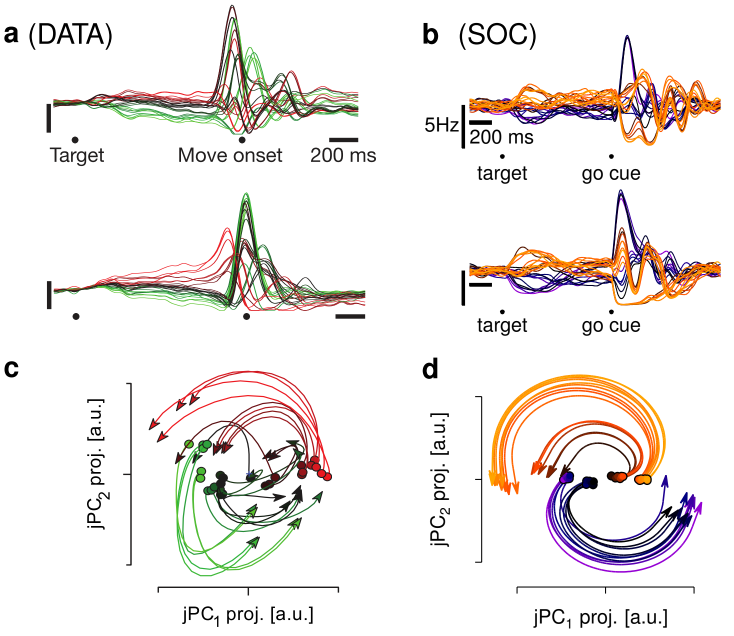


Figure 4: **SOCs agree with experimental data.** (**a**) Experimental data, adapted with permission from Churchland et al. (2012). Each trace denotes the trial-averaged firing rate of a single cell (two sample cells are shown here) during a delayed reaching task. Each trace corresponds to one of 27 different movements. Vertical scale bars denote 20 spikes/sec. The go cue is not explicitly marked here, but occurs about 200 ms before movement onset. (**b**) Time-varying firing rates of two neurons in the SOC, for 27 "conditions", each characterized by a different collective steady state of preparatory activity (see text). (**c**) Experimental data adapted from Churchland et al. (2012), showing the first 200 ms of movement-related population activity projected onto the top jPC plane. Each trajectory corresponds to one of the 27 conditions mentioned in (a). (**d**) Same analysis as in (c), for the SOC.

that, by the end of the delay period, the network arrives at a state that is one of 27 different linear combinations of $\mathbf{a}_1$ and $\mathbf{a}_2$, i.e. the two orthogonal activity states that evoke the strongest collective responses. The go cue quickly silences the command pool, leaving the network free to depart from its preparatory state and to engage in transient amplification as shown above. The recurrent dynamics produce strong, multiphasic, and movement-specific responses in single units (Figure 4b), qualitatively similar to the data.

Churchland and colleagues reported another important aspect of the transient collective dynamics following the go cue: the complexity of the single-neuron multiphasic responses was in fact hiding orderly rotational dynamics on the population level. A plane of projection could be found in which the vector of population firing activity (corresponding to vector $\Delta\mathbf{r}(t)$ in our model) would start rotating after the go cue, and consistently rotate in the same direction for all movements (Figure 4c). Surprisingly, our model analyzed with the same dynamical variant of principal component analysis (jPCA, Churchland et al. (2012), Experimental Procedures), displays the same phenomenon (Figure 4d, see also Discussion).

SOCs can generate complex movements

The complicated, multiphasic nature of the firing rate transients in SOCs suggests the possibility of reading out equally complex patterns of muscle activity. We illustrate this idea in a task in which the joint activation of two muscles must produce one of two target movements ("snake" or "butterfly" in Figure 5), within 500 ms following the go cue. Similarly to Figure 4, the preparatory input for the "snake" movement is chosen such that, by the arrival of the go cue, the network activity matches the network's preferred initial condition $\mathbf{a}_1$. Planning the "butterfly" movement sets the network in its second most preferred state $\mathbf{a}_2$. Two readout units ("muscles") compute a weighted sum of all neuronal activities in the network, that we take to directly reflect the horizontal and vertical coordinates of the movement. The weights are trained through least-squares regression (Experimental Procedures). One hundred noisy trials for each movement are enough to successfully train the system to map each command onto the correct trajectory (compare the five test trials in Figure 5a).

We conclude that the SOC's single neuron responses form a set of basis functions that is rich enough to allow read-out of non-trivial movements (Figure 5a). This is not possible in untuned, chaotic balanced

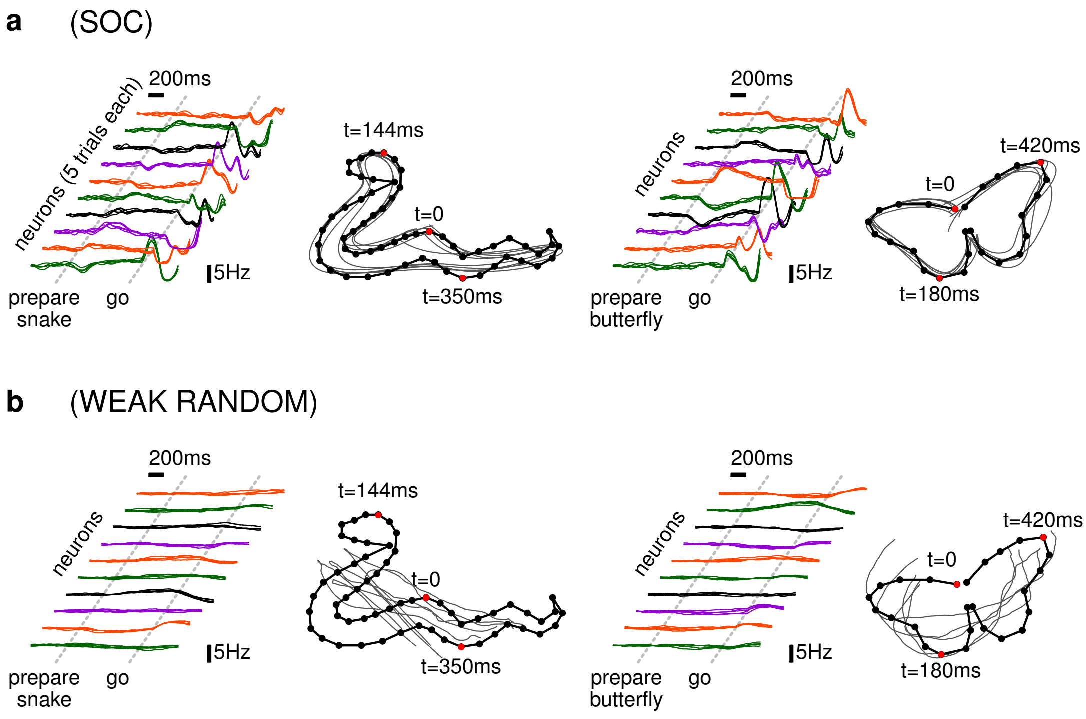


Figure 5: **Generation of complex movements through SOC dynamics**. (**a**) Firing rates versus time for 10 neurons of the SOC, as the system prepares and executes either of the two target movements (snake, left or butterfly, right). Five test trials are shown for each neuron. The corresponding muscle trajectories following the go cue are shown for the same five test trials (thin traces) and compared to the target movement (black trace and dots). (**b**) Same as in (a), for a weakly connected (untuned) random balanced network (Experimental Procedures).

networks without exquisite feedback loops or supervised learning of lateral connections (Sussillo and Abbott, 2009; Hoerzer et al., 2012; Laje and Buonomano, 2012) because of the high sensitivity to noise. Further, in balanced networks with connections that are weak enough to prevent chaotic dynamics, activity transients resemble simple exponential decays (Figure 5b), and the resulting basis functions are so redundant that the network cannot robustly learn the snake and butterfly trajectories (Figure 5b).

Interaction between excitation and inhibition in SOCs

To understand the mechanism by which SOCs amplify their preferred inputs, we dissociated the excitatory (c_E) and inhibitory (c_I) synaptic inputs each unit received from other units in the network in the absence of specific external stimulation ($\mathbf{S}(t) = 0$). We quantified the E/I balance by $r_{\mathrm{EI}}(t)$, the momentary Pearson correlation coefficient between c_E and c_I across the network population. We observed that the preferred initial states of the SOC momentarily produce substantially negative E/I input correlations (Figure 6a), indicating an average mismatch between excitatory and inhibitory inputs. Balance is then quickly restored by internal network dynamics, with $r_{\mathrm{EI}}(t)$ reaching ~ 0.8 at the peak of the transient triggered by initial condition $\mathbf{a}_1$. The effect subsists, although progressively attenuated, for roughly the first 100 preferred initial states ($\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{100}$), which are also the initial states that trigger amplified responses.

Notably, the patterns of neuronal activity after 100 ms of recurrent processing have a larger amplitude than – but bear little spatial resemblance to – the initial condition. This is reflected by a rapid decay (within 100 ms) of the correlation coefficient between the momentary network activity and the initial state (Figure 6b, black). However, considering the excitatory and inhibitory populations separately shows that

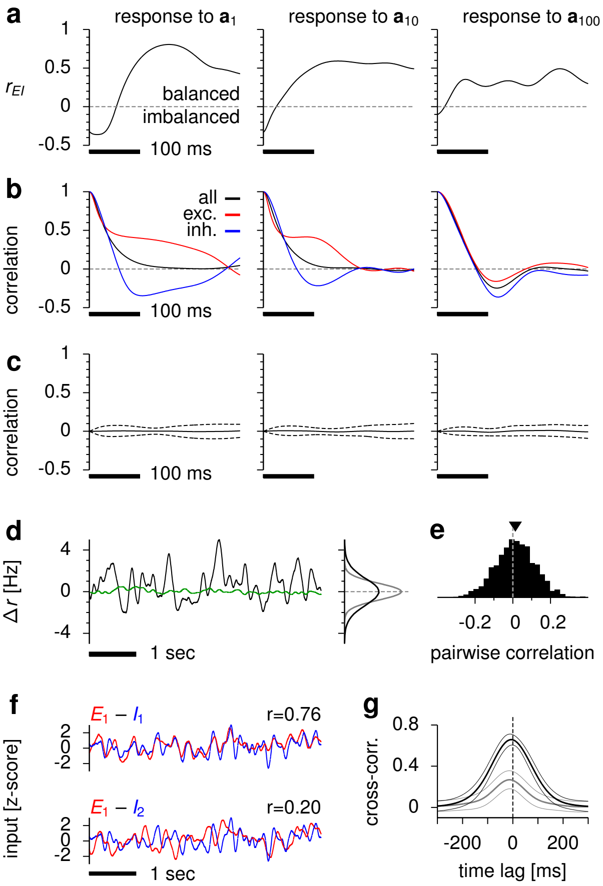


Figure 6: **Precise balance of excitation and inhibition in SOCs.** The network is initialized in state $\mathbf{a}_1$ (left), $\mathbf{a}_{10}$ (middle) or $\mathbf{a}_{100}$ (right), and runs freely thereafter. The amplitude of the initial condition is chosen weak enough for the dynamics of amplification to remain linear (cf. Figure 3). (**a**) Temporal evolution of the Pearson correlation coefficient r_{EI} between the momentary excitatory and inhibitory recurrent inputs across the population. (**b**) Corresponding time course of the correlation coefficients between the network activity and the initial state, calculated from the activity of the entire population (black), of the excitatory subpopulation (red), and of the inhibitory one (blue). (**c**) Temporal evolution of the correlation coefficient between the network activity when initialized in state $\mathbf{a}_i$ ($i = 1$ (left), 10 (middle) or 100 (right)) and when initialized in a different state $\mathbf{a}_j$ ($j \neq i$, $j < 100$). Solid lines denote the average across j, and the dashed flanking lines indicate one standard deviation. Small values indicate that the responses to the various initial conditions $\mathbf{a}_i$ are roughly decorrelated. (**d**) Black: spontaneous fluctuations around baseline rate of a sample unit in the network. The corresponding rate distribution is shown on the right (black), and compared to the distribution obtained if the unit were not connected to the rest of the network (gray). The green line denotes the momentary population average rate, which fluctuates much less. (**e**) Histogram of pairwise correlations between neuronal firing rates estimated from 100 seconds of spontaneous activity. The black triangular mark indicates the mean (~ 0.014). (**f**) Excitatory (red) and inhibitory (blue) inputs taken in the same sample unit (top) or in a pair of different units (bottom), and normalized to z-scores. The corresponding Pearson correlation coefficients are indicated above each combination, and computed from 100 seconds of spontaneous dynamics. (**g**) Black: lagged cross-correlogram of excitatory and inhibitory inputs to single units, each normalized to z-score (cf. (f), top row). The solid line is an average across all neurons; flanking lines denote ± 1 standard deviation. Inhibition lags behind excitation by a few milliseconds. Cross-correlating the E input into one unit with the I input into another unit (cf. (f), bottom row) yields the gray curve, which is an average over 1,000 randomly chosen such pairs in the SOC.

the excitatory subpopulation remains largely in the same spatial activity mode throughout the transient, i.e. neurons that were initially active (resp. inactive) tend to remain active (resp. inactive) throughout the relaxation (Figure 6b, red). In contrast, the inhibitory subpopulation becomes negatively correlated with its initial pattern after only 60 ms (Figure 6b, blue). In other words, it is the swift reversal of the pattern of inhibitory activity that quenches a growing excitatory transient and pulls the system back to rest.

The amplifying dynamics of excitation and inhibition seen on the level of transient responses to some initial conditions also shape the spontaneous background activity in SOCs (Figure 2f and Figure 6d). In the absence of additional stimuli, neurons are driven by private noise $\boldsymbol{\xi}(t)$ (Experimental Procedures), such that firing rate fluctuations can be observed even in the unconnected case ($\mathbf{W} = 0$, Figure 6d, gray histogram). The recurrent SOC connectivity amplifies these unstructured fluctuations by one third (Figure 6d, black histogram), because the noise stimulates each of the $\mathbf{a}_i$ modes evenly, and although some modes are amplified by the recurrent dynamics and others are suppressed (Figure 3a), the net result is a mild amplification (Figure 3a, black arrowhead). Furthermore, since only a few activity modes experience very strong amplification (cf Figure 3a), the resulting distribution of pairwise correlations among neurons is wide with a small positive mean (Figure 6e).

SOCs exhibit an exquisite temporal match between excitatory and inhibitory inputs to single neurons during spontaneous activity (Figure 6f). The correlation between these two input streams averages to ~ 0.66 across units, because any substantial mismatch between recurrent E and I inputs is instantly converted into a pattern of activity in which those inputs match again (cf. Figure 6a). The amplitude of such reactions is larger than the typical response to noise, so the network is constantly in a state of detailed E/I balance (Vogels and Abbott, 2009). Furthermore, we have seen that it is mostly the spatial pattern of inhibitory activity that reverses during the course of amplification to restore the balance, while the excitatory activity is much less affected (Figure 6b). Thus, during spontaneous activity, inhibitory inputs are expected to lag behind excitatory inputs by a few milliseconds, which can indeed be seen in their average cross-correlogram (Figure 6g) and has also been observed experimentally (Okun and Lampl, 2008; Cafaro and Rieke, 2010).

The small temporal co-fluctuations in the firing rates of the excitatory and inhibitory populations are known to translate into correlated excitatory and inhibitory inputs to single neurons, in densely connected circuits (Renart et al., 2010). Here, interestingly, excitatory and inhibitory inputs are correlated more strongly than expected from the magnitude of such shared population fluctuations. This can be seen by correlating the excitatory input stream taken in one unit with the inhibitory input stream taken in another unit (Figure 6f, bottom row). Such correlations average to ~ 0.26 only (to be compared with 0.66 above – see Figure 6g).

Spiking implementation of a SOC

So far we have described neuronal activity on the level of firing rates. An important question is whether the dynamical features of rate-based SOCs are borne out in more realistic models of interconnected spiking neurons. To address this issue, we built a large-scale model of a SOC composed of 15'000 leaky integrate-and-fire model neurons, separated into 12'000 excitatory and 3'000 inhibitory neurons. The network was structured such that each neuron belonged to one of 200 excitatory or 200 inhibitory small neuron subgroups (of size 60 and 15 respectively), whose average momentary activities can be interpreted as the "rate variables" discussed until here.

In order to keep the network in the asynchronous and irregular firing regime, the whole network was – in part – randomly and sparsely connected, in this respect similar to traditional models (van Vreeswijk and Sompolinsky, 1996; Brunel, 2000; Vogels et al., 2005; Renart et al., 2010). In addition to those random synapses, others were added on the basis of which subgroups the pre- and post-synaptic neurons belonged to, and following connectivity patterns between subgroups given by a 400×400 SOC matrix obtained similarly to $\mathbf{W}$ in Figure 2 (c.f. Experimental Procedures). Overall, the average connection probability between spiking neurons was 0.2. Note also that neurons were not given any source of randomness, twork, thus any apparent stochasticity was internally generated.

The random connections were assumed to trigger fast PSPs (10 ms decay time, e.g. as if they were located proximal to the soma), while the structured synapses were modeled as slow ones (100 ms decay time, as if they were more distal down the dendritic tree – see Branco and Häusser (2011) and Experimental Procedures). This separation of timescales endows the spiking network with all of the above-described dynamical properties of rate-based SOCs (Figure 7 and Figure 8).

The spiking SOC operated in a balanced regime, with large sub-threshold membrane potential fluctuations and occasional action potential firing (Figure 7a) with realistic rate and interspike interval statistics (Figure 7c). Spiking events were fully desynchronized on the level of the entire population, whose momentary activity was approximately constant at ~ 6 Hz (a more detailed description of spontaneous activity in SOCs is given below in Figure 8).

Similar to our rate-based SOCs, the spiking network could be initialized in any desired activity state through the injection of specific ramping input currents into each neuron (Figure 7a). The go cue triggered sudden input withdrawal, resulting in large and complex transients in the trial-averaged spiking activities of single cells (Figure 7a, middle). These transients lasted for about 500 ms. Note that the

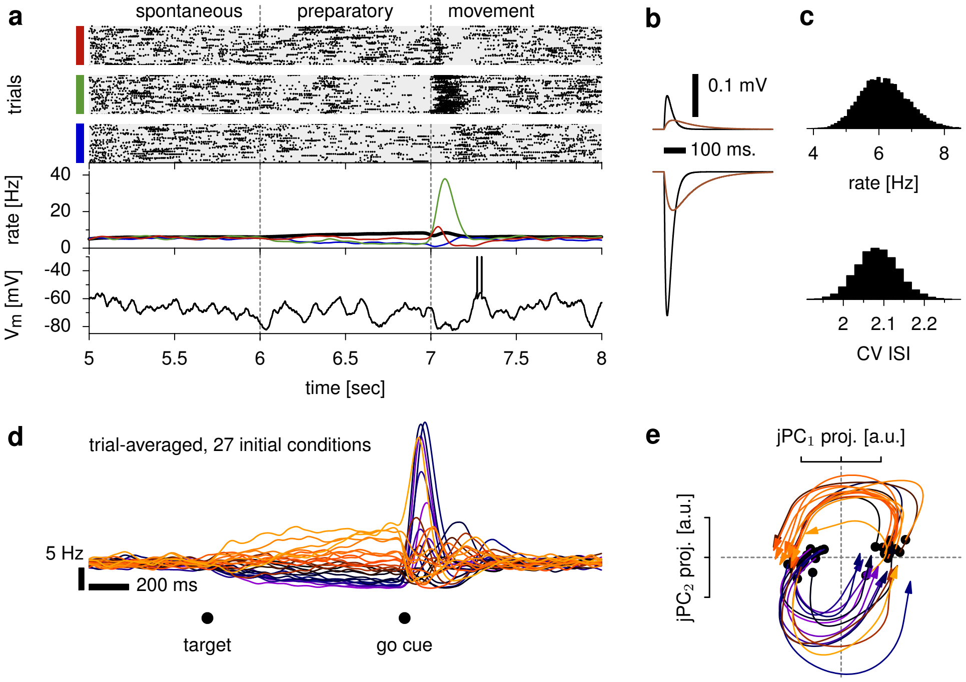


Figure 7: **Transient dynamics in a spiking SOC. (a)** The network is initialized in a mixture of its top two preferred initial states during the preparatory period. Top: raster plot of spiking activity over 200 trials for three cells (red, green, blue). Middle: temporal evolution of the trial-averaged activity of those cells (same color code), and that of the overall population activity (black). Rate traces were computed over 1'000 trials and smoothed with a Gaussian kernel (width 20 ms), to reproduce the analysis of Churchland et al. (2012). Bottom: sample voltage trace of a randomly chosen neuron. **(b)** Fast (black) and slow (brown) synaptic PSPs, corresponding to random and structured connections in the spiking circuit, respectively. **(c)** Distribution of average firing rates (top) and ISI CVs (bottom) during spontaneous activity. **(d)** Trial-averaged firing rate traces for a single sample cell, when the preparatory input drives the SOC into one of 27 random mixtures of its first and second preferred initial conditions. Averages were computed over 1'000 trials, and smoothed as described in (a). **(e)** First 200 ms of movement-related population activity, projected onto the top jPC plane. Each trajectory corresponds to a different initial condition in (d), using the same color code.

desired preparatory state was reached only on average (as it is in the data of Churchland et al.), i.e. there was substantial trial-to-trial variability in the spiking activity, even on the level of subgroup activities. It is therefore interesting that, despite such inaccuracies in the network preparation, large multiphasic transients can still be observed in the average evoked responses.

The trial-averaged firing rate responses to 27 different initial conditions, chosen in the same way as in Figure 4, as well as the diversity of single cell responses, were qualitatively similar to the data in Churchland et al. (2012) (Figure 7d and Supplementary Figure 1). When projected onto the top jPC plane, the population activity showed orderly rotations, as it did in our rate SOC (Figure 7e) and again similar to the experimental data.

Spontaneous activity in spiking SOCs is markedly different from that in traditional balanced circuit models. Subgroups of neurons display large, slow and graded activity fluctuations (Figure 8a), and pairwise correlations between subgroup activities are accurately predicted by a linear stochastic model similar to Equation 1 (Figure 8b). These correlations are much weaker in magnitude in a control, traditional random network (Experimental Procedures; see also Supplementary Figure 2) with equivalent synaptic input statistics (compare the black and brown histograms in Figure 8b).

The distribution of spike correlations in the SOC (Figure 8c) is wide with a very small positive mean $\bar{r} \approx 0.0027$), indicating that cells fire asynchronously. The same is true in the control random network ($\bar{r} \approx 0.0005$; see Renart et al. (2010)). However, within SOC subgroups, spiking was substantially

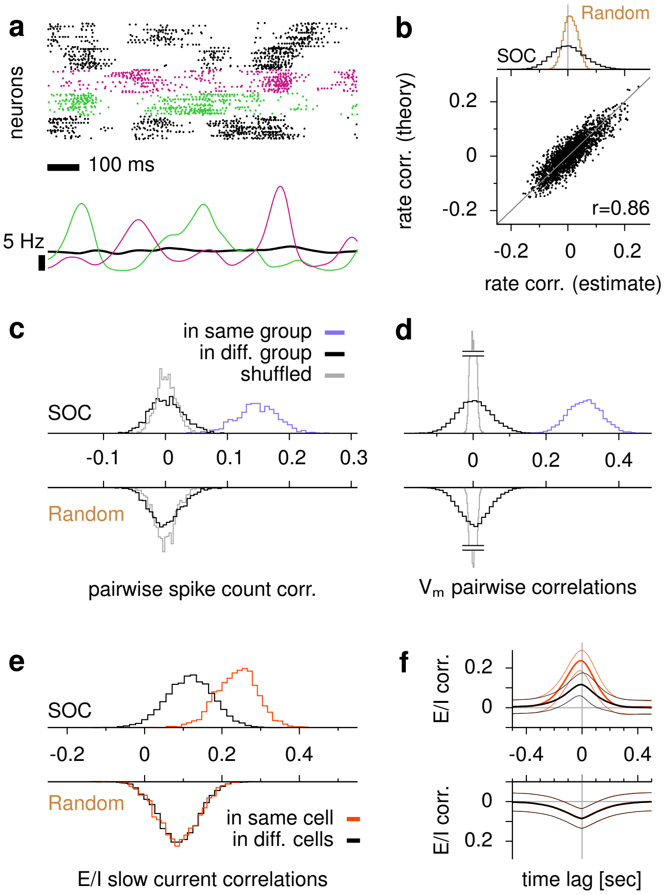


Figure 8: **Spontaneous activity in spiking SOCs.** (**a**) Top: raster plot of spontaneous spiking activity in the SOC. Only the neurons in the first 5 subgroups (300 neurons) are shown. Bottom: momentary activity of the whole population (black), and of the second (green) and third (magenta) subgroups. Traces were smoothed using a Gaussian kernel of 20 ms width. (**b**) Pairwise correlations between instantaneous subgroup firing rates in the SOC, as empirically measured from a 1'000 second-long simulation (x-axis) versus theoretically predicted from a linear stochastic model (y-axis). Rate traces were first smoothed using a Gaussian kernel (20 ms width) as in (a). Distributions of pairwise correlations are shown at the top, for the SOC (black) and for a control random network with equivalent synaptic input statistics (brown; see Experimental Procedures). (**c**) Distributions of pairwise spike correlations in the SOC (top) and in the control random network the random network (bottom), between pairs of neurons belonging to the same subgroup (blue), or to different subgroups (black). Spike trains were first convolved with a Gaussian kernel of width 100 ms. Gray curves were obtained by shuffling the ISIs, thus destroying correlations while preserving the ISI distribution. (**d**) Distributions of subthreshold membrane potential correlations. Colors have the same meaning as in (c). Voltage traces were cut off at the spike threshold. Gray curves correspond to temporally shuffled traces. (**e**) Distributions of pairwise correlations between the slow excitatory and inhibitory currents, taken in the same cell (red) or in two different cells (black). Data was binned for 10'000 such E/I current pairs for each histogram. (**f**) Full lagged cross-correlograms between the slow excitatory and inhibitory currents, taken in the same cell (red) or in two different cells. Thick lines denote averages over 10'000 such E/I current pairs, and thin flanking lines indicate standard deviations. The peak at negative time lag corresponds to E currents leading I currents.

correlated (Figure 8c, blue; $\bar{r} \approx 0.17$), and particularly so on the 100 ms timescale. Not surprisingly, membrane potentials followed a similar pattern of correlations (Figure 8d).

Importantly, the detailed balance prediction made above for the rate-based scenario (Figure 6f,g) remains true on the level of single cells in the spiking network. Slow E and I inputs (corresponding to the structured SOC recurrent synapses) to single neurons are substantially more correlated ($\bar{r} \approx 0.24$) than pairs of E and I currents taken from different neurons ($\bar{r} \approx 0.12$; compare red and black in Figure 8e,f). This is not true in the control random network, in which the balance of excitation is merely a reflection of the synchronized fluctuations of the excitatory and inhibitory populations as a whole.

Discussion

The motor cortex data of Churchland et al. (2012) showcase two seemingly conflicting characteristics. On the one hand, motor cortical areas appear to be precisely controllable during movement preparation, and dynamically stable with firing rates evolving well below saturation during movement execution. In most network models, such stability arises from weak recurrent interactions. On the other hand the data shows rich transient amplification of specific initial conditions, a phenomenon that requires strong recurrent excitation. To reconcile these opposing aspects, we introduced and studied the concept of "stability-optimized circuits" (SOCs), broadly defined as precisely balanced networks with strong and complex recurrent excitatory pathways. In SOCs, strong excitation mediates fast activity break-outs following appropriate input, while inhibition keeps track of the activity and acts as a retracting spring force. In the presence of intricate excitatory recurrence, inhibition cannot instantaneously quench such activity growth, leading to transient oscillations as excitation and inhibition waltz their way back to a stable background state. This results in spatially and temporally rich firing rate responses, qualitatively similar to those recorded by Churchland and colleagues (Churchland et al., 2012).

To build SOCs, we used progressive optimal refinement of the inhibitory synaptic connectivity within a normative, control-theoretic framework. Our method makes use of recent techniques for stability optimization (Vanbiervliet et al., 2009), and can in principle produce stability-optimized circuits from any given excitatory connectivity. In simple terms, we iteratively refined both the absence/presence and the strengths of the inhibitory connections to pull all the unstable eigenvalues of the network's connectivity matrix back into the stable regime (Figure 2b). Even though we constrained the procedure to yield plausible network connectivity, notably one that respects Dale's law (Dayan and Abbott (2001), chap. 7), it does not constitute – and is not meant to be – a synaptic plasticity rule. However, the phenomenology achieved by recent models of inhibitory synaptic plasticity (Vogels et al., 2011; Luz and Shamir, 2012; Kullmann et al., 2012) is similar to, though more crude than, that of our SOCs. It raises the possibility that nature solves the problem of network stabilization through a form of inhibitory plasticity, potentially aided by appropriate pre- and re-wiring during development (Terauchi and Umemori, 2012). The shortcut of minimizing the "smoothed spectral abscissa" (Experimental Procedures) that we used here to stabilize the network may thus be difficult to observe experimentally in its current control-theoretic variant, but may implemented through several sequential network mechanisms. It is also likely to be of great use in future studies of feedback-mediated control in neuronal networks (see also Sussillo and Abbott (2009); Hoerzer et al. (2012); Laje and Buonomano (2013)).

In a protocol qualitatively similar to the experimental design of Churchland et al. (2012) (Figure 1), we could generate complex activity transients by forcing the SOC into one of a few specific preparatory states through the delivery of appropriate inputs, which were then withdrawn to release the network into free dynamics (Figure 4). Those "engine dynamics" (Shenoy et al., 2011; Churchland et al., 2012) could easily be converted into actual muscle trajectories. Simple linear readouts, with weights optimized through least-squares regression, were sufficient to produce fast and elaborate 2-dimensional movements (Figure 5). Three aspects of the SOC dynamics make this possible. First, the firing rates strongly deviate from baseline during the movement period, effectively increasing the signal-to-noise ratio in the network response. Second, the transients are multiphasic (Figure 4b), as opposed to simple rise-and-decay, allowing the readouts not to overfit on multi-curved movements. Third, the preferred initial conditions of the SOC are converted into activity modes that are largely non-overlapping (Figure 6c). Thus, not only is the system highly excitable from a large set of states, but those states produce responses that are distinguishable from one another, ensuring that different motor commands can be mapped onto distinct muscle trajectories (Figure 5).

The network activity during the transients displayed collective rotational behavior that was also present in the multi-neuron data of Churchland et al. (2012). In strictly technical terms, this is due to a peculiarity of the eigenvalues of random matrices. In the unstable network we started with, the rightmost eigenvalue is a pair of complex conjugate eigenvalues and corresponds to activity modes of fastest expansion, or largest "energy" growth (such a pair with large real and small imaginary parts is bound to exist, although it may not be the rightmost one). These activity modes are to a large extent determined by the excitatory connectivity, which is left untouched during the process of inhibitory stabilization. Thus, the large responses of the SOC to one of its most preferred initial conditions $(\mathbf{a}_1, \mathbf{a}_2)$ is likely related to the dynamics of the original network in its principal eigenplane. In this eigenplane, the population activity vector rotates slowly, at a speed governed by the imaginary part of the eigenvalue pair, and expands exponentially at a rate given by the real part. In the SOC, since inhibition tends to quench neuronal activity, the remaining observable effect is a transiently expanding rotation, as seen in Churchland et al. (2012).

To explore the properties of transient dynamics in SOCs, we were able to calculate and rank the maximally (or minimally) amplified initial states analytically in the linear regime. In the more realistic, nonlinear model in which neurons can saturate at zero and maximum firing rates, the numerical ranking of initial states remained nearly identical to the linear case, showing that it is predominantly the connectivity of the network that determines the response to momentary disturbances. This is not surprising, as the onset of amplification after a weak perturbation relies on the connectivity matrix of SOCs being mathematically "nonnormal", which is a linear property (Ganguli et al. (2008); Murphy and Miller (2009); Goldman (2009); Hennequin et al. (2012); see also our discussion of balanced amplification in SOCs below).

Relation to balanced amplification and relevance to sensory circuits Transient amplification in SOCs is in principle an extended, more intricate form of "balanced amplification", first described by Murphy and Miller (2009) in a model of V1 synaptic organization. There, the authors argued that, in networks with strong excitation balanced by equally strong (or stronger) inhibition, small patterns of spatial imbalance between excitation and inhibition ("difference modes") can drive large activity transients in which neighboring excitatory and inhibitory neurons fire in unison ("sum modes"). Due to the absence of a topology in SOCs, it is impossible to tell which neuron is a neighbor to which, making sum and difference modes difficult to define. Nevertheless, sum and difference modes can be understood more broadly as patterns of average balance and imbalance in the excitatory and inhibitory synaptic *inputs* to single cells. With this definition, we showed here (Figure 6a) that the phenomenology of amplification in SOCs is similar to balanced amplification, i.e. small stimulations of difference modes drive large activations of sum modes. This accounts for the large transient firing rate deflections of individual neurons that follow appropriate initialization. A key difference between SOCs and Murphy and Miller's model of V1 is the apparent complexity in the lateral excitatory connections, which here gave rise to dynamics that are richer than monophasic rise-and-decay responses (Figure 3 and Figure 4). This made it possible to learn complex muscle trajectories in which the same direction of motion must be encoded at several different times during the movement (Figure 5). Furthermore, although the "spring-box" analogy may not apply directly to sensory cortices, SOCs (as inhibition-stabilized networks) could still provide an appropriate conceptual framework for such cortical areas, as suggested by Ozeki et al. (2009). Likewise, the method we have used here to build such circuits could prove useful in finding conditions for inhibitory stabilization of known excitatory connectivities that are not easily reducible to analytically tractable models (see e.g. Ahmadian et al. (2013)).

Relation to detailed excitation/inhibition balance SOCs make a strong prediction regarding how excitation and inhibition interact in cortical networks: E and I synaptic inputs in single neurons should be temporally correlated in a way that cannot be explained by the activity co-fluctuations that occur on the level of the entire population.

During spontaneous activity in SOCs, balanced amplification of external noise (or intrinsically generated chaos, as in our spiking SOC) results in strongly correlated E/I inputs in single units. This phenomenon is a recurrent equivalent to what has been referred to as "detailed balance" in feedforward network models (Vogels and Abbott, 2009; Vogels et al., 2011; Luz and Shamir, 2012), and cannot be attributed to mere co-fluctuations of the overall activity of E and I neurons. Such co-variations can be substantial in balanced networks (Vogels and Abbott, 2005; Kriener et al., 2008; Murphy and Miller, 2009), but have been quenched here by requiring inhibitory synaptic connections to be three times stronger than excitatory ones on average (Renart et al., 2010; Hennequin et al., 2012). The residual shared population fluctuations accounted for only one third ($r_{\mathsf{shared}} \simeq 0.26$) of the total E/I input correlation ($r \simeq 0.66$, Figure 6f,g). The excess correlation can only be explained by the comparatively large fluctuations of balanced, zero-mean activity modes (the responses to the preferred initial conditions of the SOC; Figure 6a).

A certain degree of such E/I balance has been observed in several brain areas, and on levels as different as trial-averaged E and I synaptic input conductances in response to sensory stimuli (Wehr and Zador

(2003); Mariño et al. (2005); Froemke et al. (2007); Dorrn et al. (2010); but see Haider et al. (2013)) single-trial synaptic responses in which the trial-average has been removed ("residuals", Cafaro and Rieke (2010)), and spontaneous activity (Okun and Lampl, 2008; Cafaro and Rieke, 2010). However, experimental evidence for the relative contribution of shared population fluctuations to the E/I balance is still inconclusive. Early evidence for precise excitatory-inhibitory input tracking during spontaneous activity was collected in paired recordings of neighboring cells (Okun and Lampl, 2008) in the mouse barrel cortex. Thus, by experimental design, only E/I correlations originating from population-level fluctuations were measured. Conversely, a recent study in the mouse retina measured E and I input conductances (near)-simultaneously in single cells, but did not perform cross-measurements in pairs of different cells (Cafaro and Rieke, 2010). Hence, both experiments lacked the respective control measure to assess the contribution of detailed vs. global balance of excitation and inhibition. If the former dominates the network dynamics, SOCs are a plausible candidate structure.

Spiking models of SOCs The simplicity and analytical tractability of rate models make them appealing to theoretical studies such as ours. One may worry, however, that some fundamental aspects of collective dynamics are being overseen when spiking events are reduced to their probabilities of occurence, i.e. to rate variables. To verify our results, we embedded a SOC in a standard balanced spiking network, in which millions of randomly assigned synapses connect two populations of excitatory and inhibitory neurons. The SOC structure was embodied by additional connections between subgroups of these neurons, each containing on the order of tens of spiking cells. The resulting network displayed simultaneous firing rate and spiking variability (Churchland and Abbott, 2012), thus phenomenologically similar to the network of Litwin-Kumar and Doiron (2012), although the slow rate fluctuations here are more graded and emerge through a different mechanism. The sea of random synapses in our network induces strong excitatory and inhibitory inputs to single cells that cancel each other on average, leaving large subthreshold fluctuations in membrane potential and therefore irregular spiking whose variability is mostly "private" to each neuron. This feature is common to all traditional balanced network models (van Vreeswijk and Sompolinsky, 1996; Brunel, 2000; Vogels et al., 2005; Renart et al., 2010). On the level of subgroups of neurons, this source of variability is not entirely lost to averaging: although all the cells in a given subgroup n fire at the same rate r_n at any given time, receiver neurons in another subgroup m will only "sense" a noisy sample estimate $\hat{r}_n$ of this rate, because n connects onto m through a finite number of synapses. Now, because the connectivity between subgroups is strong, but stabilized, this intrinsic source of noise (the "residual" $\xi_n = \hat{r}_n - r_n$) is continuously amplified into large, structured firing rate fluctuations on the level of subgroups. The underlying mechanism is the same as for the rate model, i.e. balanced amplification of noise (Murphy and Miller, 2009), with the notable difference that the noise in the spiking network is intrinsically generated (the external excitatory drive that each neuron receives was chosen constant here to make this point).

In order to match the timescale of the rate transients in our spiking SOC to those in the data of Churchland et al. (2012), we assumed that the structured SOC synapses had slower time constants than the random ones. Importantly, the nature of the phenomenon was not affected by the changes in time constant. It just became slower. The existence of separate populations of fast/slow synapses in the cortex could be motivated by recent experiments in which the distance from soma along the dendritic arbor was shown to predict the magnitude of the NMDA component in the corresponding somatic PSPs (Branco and Häusser, 2011). Thus, distal synapses tend to evoke slower PSPs than proximal ones. It is in fact an interesting and testable prediction of our model that distal synapses are actively recruited in motor cortex during movement preparation and generation.

Our current spiking implementation of a SOC implies that the "rate units" in the rate-based model of Equation 1 are best thought of as describing the momentary activities of small groups of statistically comparable spiking neurons. Similarly, our "random connectivity" on the level of rate variables (the random E $\leftrightarrow$ E connections in Fig 2a) is to be seen as a way of modelling "complex" interactions between neuronal subgroups, and is conceptually different from the "random connectivity" that characterizes traditional balanced spiking networks (and part of ours). However, we used subgroups of neurons to encode rate variables only because we were not able to run our SOC stabilization algorithm ($\mathcal{O}(N^3)$ complexity) on a matrix larger than $> 5'000 \times 5'000$, which would be required for a spiking implementation. This structural assumption could be relaxed, as long as the network connectivity remains "low-rank", thus embedding strong effective interactions between only a handful of activity modes, (e.g. in the order of 100 in a network of size $10'000$), which need not formally be "subgroups". Without fine-tuned dynamic inhibitory feedback, these interactions would drive the network unstable. Weaker connectivity may exist among the thousands of remaining dimensions and generate the apparent chaos that characterizes single spikes on fast timescales. This is a research question of its own (Stern et al., 2013; DePasquale et al., 2013), and is beyond the scope of our work.

Summary In summary, we have shown that specific, recurrent inhibition is a powerful means of stabilizing otherwise unstable, complex circuits. The resulting networks are collectively excitable, and display rich transient responses to appropriate stimuli that resemble the activity recorded in the (pre-)motor cortex (Churchland et al., 2012) on both single-neuron and populations levels. We found that SOCs can be used as "spring loaded motor engines" to generate complicated and reliable movements. The intriguing parallels to the detailed balance of excitatory and inhibitory inputs in cortical neurons, as well as to recent theories that apply specifically to the visual cortex (Ozeki et al., 2009; Murphy and Miller, 2009), suggest cortical-wide relevance for this new class of cortical architectures.

Experimental Procedures

Network setup and dynamics

Single-neuron dynamics followed Equation 1, which we integrated using a standard fourth-order Runge-Kutta method. Following Rajan et al. (2010), we used the gain function

$$g(x) = \begin{cases} r_0 \tanh\left[x/r_0\right] & \text{if} \quad x < 0 \\ (r_{max} - r_0) \tanh\left[x/(r_{max} - r_0)\right] & \text{if} \quad x \geq 0 \end{cases} \tag{2}$$

with baseline firing rate $r_0 = 5$ Hz and maximum rate $r_{max} = 100$ Hz (Figure 3e). Unless indicated otherwise, the input $I(t) = \boldsymbol{\xi}(t) + \mathbf{S}(t)$ included a noise term $\boldsymbol{\xi}(t)$, which we modelled as an independent Ornstein-Uhlenbeck process for each neuron, with time constant $\tau_\xi = 50$ ms. We set the variance of these processes to $\sigma_0^2(\tau + \tau_\xi)/\tau_\xi$, such that, in the limit of very weak synaptic connectivity, the firing rate of each cell in the network fluctuated around baseline with a standard deviation $\sigma_0 = 0.2$ Hz.

In order to set the initial network state $\mathbf{x}$ into a specific pattern $\mathbf{a}_k$, we used two different schemes. For the mathematical analysis, we set $\mathbf{S}(t)$ to $\mathbf{a}_k \delta(t)$ directly (Figures 3 and 6). For the comparison to experimental data, we modelled the preparatory period explicitly by delivering a slowly ramping input to each cell during ongoing activity (Figures 4 and 5). This input was delivered as vector $\mathbf{S}(t) = r(t)\mathbf{P}_k$, in which $r(t)$ denotes the ramp activation of the input pool k and $\mathbf{P}_k$ are the projection weights from pool k onto the motor network (Figure 1b,c). The ramp $r(t)$ had a slow exponential rise with time constant 400 ms beginning with the target cue at $t = -1$sec, followed by a fast exponential decay with time constant 2 ms after the go cue. The projection weights were set to $\mathbf{P}_k = \mathbf{a}_k - \mathbf{W}g(\mathbf{a}_k)$ in order to guarantee $\mathbf{x}(t = 0) \simeq \mathbf{a}_k$.

In Figure 4b, the 27 arm reaching movements in Churchland et al. (2012) were modeled as 27 different initial conditions $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{27}$ for the SOC. We chose each $\mathbf{b}_k$ as a random linear combination of the SOC's first and second preferred initial conditions $\mathbf{a}_1$ and $\mathbf{a}_2$ (see next section). More precisely, $\mathbf{b}_k = \sum_{\ell=1,2} s_{k\ell} z_{k\ell} \mathbf{a}_\ell$ where $s_{k\ell}$ is a random sign, and the $z_{k\ell}$ were drawn uniformly between 0.5 and 1.

Preferred initial states

To find the preferred initial conditions of the SOC, we restricted ourselves to the linear regime in which $\Delta r_i \simeq x_i$. To quantify the response evoked by some unit-norm initial condition $\Delta\mathbf{r}(t=0) \equiv \mathbf{a}$ we defined the "energy" $\mathcal{E}(\mathbf{a})$ of the response as

$$\mathcal{E}(\mathbf{a}) = \frac{2}{\tau} \int_0^\infty \|\Delta\mathbf{r}(t)\|^2 \, dt. \tag{3}$$

We assumed that the network dynamics run freely without noise ($\xi(t) = 0$) from $t = 0$. Here $2/\tau$ is a normalizing factor such that $\mathcal{E} = 1$ for an unconnected network ($\mathbf{W} = 0$), irrespective of the (unit-norm) initial condition $\mathbf{a}$ (in which case $\|\Delta\mathbf{r}(t)\|^2 = \exp(-2t/\tau)$). Since the SOC is linearly stable, $\mathcal{E}$ is finite, in the sense that any initial condition is bound to decay (exponentially) after sufficiently long periods of time.

The "best" input direction is then defined as the initial condition $\mathbf{a}_1$ that maximizes $\mathcal{E}(\mathbf{a})$. In the linear regime, this maximization can be performed analytically. Equation 3 can be rewritten as

$$\mathcal{E}(\mathbf{a}) = \mathbf{a}^T \left[\frac{2}{\tau} \int_0^\infty \mathrm{d}t \; e^{(\mathbf{W}^T - \mathbb{1})t/\tau} \; e^{(\mathbf{W} - \mathbb{1})t/\tau} \right] \mathbf{a} \overset{\mathsf{def}}{=} \mathbf{a}^T \, \mathbf{Q} \, \mathbf{a} \tag{4}$$

where $(\cdot)^T$ denotes the matrix transpose. The last equality defines $\mathbf{Q}$ as the matrix integral inside square brackets. $\mathbf{Q}$ is a symmetric, positive-definite matrix, so its principal eigenvector is precisely the initial condition $\mathbf{a}_1$ that maximizes the evoked energy, which is then given by the corresponding principal eigenvalue of $\mathbf{Q}$. In fact, the full eigenbasis of $\mathbf{Q}$, ranked in decreasing order of the associated eigenvalues, defines a collection $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N)$ of N orthogonal input states that each maximize the evoked energy within the subspace orthogonal to all previous best input directions. Again, the eigenvalues are the corresponding evoked energies. We use this energy formalism again below to explain the optimal inhibitory stabilization algorithm. Note that in the linear regime, $\mathbf{a}_k$ and $-\mathbf{a}_k$ evoke the exact same energy. In the non-linear network, this need not be the case, and we resolved this sign ambiguity by picking the condition that evoked most energy. Note also that $\mathbf{Q}$ is the solution to the Lyapunov equation

$$(\mathbf{W} - \mathbb{1})^T \, \mathbf{Q} + \mathbf{Q} \, (\mathbf{W} - \mathbb{1}) = -2 \cdot \mathbb{1} \tag{5}$$

which is easily solved numerically (Bartels and Stewart, 1972), e.g. using the Matlab function `lyap`.

Construction of the SOC architecture

Random connectivity matrices of size $N = 2M$, with M positive (excitatory) columns and M negative (inhibitory) columns, were generated as in Hennequin et al. (2012) with connectivity density $p = 0.1$. Non-zero excitatory (resp. inhibitory) weights were set to $w_0/\sqrt{N}$ (resp. $-\gamma w_0/\sqrt{N}$), where $w_0 = R/\sqrt{p(1-p)(1+\gamma^2)/2}$ and R is the desired spectral radius (before stability optimization, Rajan and Abbott (2006)).

To generate a SOC, we generated such a random connectivity matrix with $R = 10$, producing unstable, chaotic network behavior. After the creation of the initial $\mathbf{W}$, all excitatory connections remained fixed. To achieve robust linear stability of the dynamics, we refined the inhibitory synapses to minimize the "smoothed spectral abscissa" $\tilde{\alpha}_\epsilon(\mathbf{W})$, a relaxation of the spectral abscissa (the largest real part in the eigenvalues of $\mathbf{W}$) that – among other advantages – leads to tractable optimization (Vanbiervliet et al., 2009). We describe it in more details below, but in short, the inhibitory weights followed a gradient descent on $\tilde{\alpha}_\epsilon(\mathbf{W})$, subject to three constraints. First, we constrained the inhibitory weights to remain inhibitory, i.e. negative. Second, we enforced a constant ratio between the average magnitude of the inhibitory weights, and its excitatory counterpart ($\gamma = 3$, see Discussion). Third, the density of inhibitory connections was restricted to less than 40%. This constrained gradient descent usually converged within a few hundred iterations.

In more technical terms, the smoothed spectral abscissa can be introduced by way of considering, as above, the energy evoked through linear network dynamics by some initial condition $\mathbf{a}$:

$$\mathcal{E}(\mathbf{W}, \mathbf{a}) = \mathbf{a}^T \mathbf{Q}(1) \mathbf{a} \tag{6}$$

where $\mathbf{Q}(s)$ is defined more generally as

$$\mathbf{Q}(s) = \frac{2}{\tau} \int_0^\infty e^{(\mathbf{W}^T - s\mathbb{1})t/\tau} \, e^{(\mathbf{W} - s\mathbb{1})t/\tau} \, dt \quad . \tag{7}$$

For the network to be stable, the energy must remain finite for any initial condition $\mathbf{a}$. It is easy to show that for any $\mathbf{a}$, $\mathcal{E}(\mathbf{W}, \mathbf{a})$ is upper-bounded by the trace of $\mathbf{Q}(1)$. Thus, if $\mathsf{tr}\,[\mathbf{Q}(1)] < \epsilon^{-1}$ for some given $\epsilon > 0$, then the energy evoked by any $\mathbf{a}$ is less than ϵ^{-1}, so the network dynamics of Equation 1 are guaranteed to be stable. In a network that is not (yet) linearly stable, we ask: how far must the system be "shifted", $\mathbf{W} \mapsto \mathbf{W} - s\mathbb{1}$ (cf. Equation 7), for $\mathsf{tr}\,[\mathbf{Q}(s)]$ to become smaller than ϵ^{-1}? The ϵ-smoothed spectral abscissa answers this question. Mathematically, $\tilde{\alpha}_\epsilon(\mathbf{W})$ is the unique root of $s \mapsto \mathsf{tr}\,[\mathbf{Q}(s)] - \epsilon^{-1}$, which is a monotonically decreasing function of s. If the shift s is smaller than the spectral abscissa $\alpha(\mathbf{W})$, some of the eigenvalues of $\mathbf{W} - s\mathbb{1}$ will have positive real parts, causing $\mathsf{tr}\,[\mathbf{Q}(s)]$ to diverge. The smoothed spectral abscissa $\tilde{\alpha}_\epsilon(\mathbf{W})$ is therefore necessarily greater than $\alpha(\mathbf{W})$, which means that stability can be achieved by minimizing $\tilde{\alpha}_\epsilon(\mathbf{W})$ instead of $\alpha(\mathbf{W})$. This is an advantageous replacement as, unlike α, $\tilde{\alpha}_\epsilon$ is a smooth function of the synaptic weights.

The tractability of the approach stems from the computability of $\tilde{\alpha}_\epsilon(\mathbf{W})$ and its derivatives w.r.t $\mathbf{W}$. For any $s > \alpha(\mathbf{W})$, the matrix $\mathbf{Q}(s)$ defined in Equation 7 is known to be the solution to the following Lyapunov equation

$$(\mathbf{W} - s\mathbb{1})^T \, \mathbf{Q}(s) + \mathbf{Q}(s) \, (\mathbf{W} - s\mathbb{1}) = -2 \cdot \mathbb{1} \tag{8}$$

Solving this equation numerically can be done efficiently (Bartels and Stewart, 1972). Knowing that $\mathsf{tr}\,[\mathbf{Q}(s)] - \epsilon^{-1}$ is a decreasing function of s, one can apply standard root-finding methods to identify $\tilde{\alpha}_\epsilon(\mathbf{W})$. Finally, Vanbiervliet et al. (2009) also provide its derivatives w.r.t the synaptic weights, needed to perform our gradient-based minimization of $\tilde{\alpha}_\epsilon$:

$$\frac{\partial \, \tilde{\alpha}_\epsilon(\mathbf{W})}{\partial \mathbf{W}} = \frac{\mathbf{Q}(\tilde{\alpha}_\epsilon)\mathbf{P}(\tilde{\alpha}_\epsilon)}{\mathsf{tr}\,[\mathbf{Q}(\tilde{\alpha}_\epsilon)\mathbf{P}(\tilde{\alpha}_\epsilon)]} \tag{9}$$

where

$$\mathbf{P}(s) = 2 \int_0^\infty e^{t(\mathbf{W} - s\mathbb{1})} \, e^{t(\mathbf{W} - s\mathbb{1})^T} \, dt \tag{10}$$

solves a Lyapunov equation analogous to Equation 8:

$$(\mathbf{W} - s\mathbb{1}) \, \mathbf{P}(s) + \mathbf{P}(s) \, (\mathbf{W} - s\mathbb{1})^T = -2 \cdot \mathbb{1} \tag{11}$$

The iterative constrained gradient descent on $\tilde{\alpha}_\epsilon(\mathbf{W})$ entails the following steps:

1. Compute the current value of the smoothed spectral abscissa $\tilde{\alpha}_\epsilon(\mathbf{W})$. This implies multiple iterations of a numerical root-finding method (e.g. bisection) on $s \mapsto \mathsf{tr}[\mathbf{Q}(s)] - \epsilon^{-1}$. Each iteration requires solving Equation 8 for $\mathbf{Q}(s)$, but we show below that this step can in fact be bypassed.

2. Solve Equation 8 and Equation 11 with $s = \tilde{\alpha}_\epsilon$ found in step 1. This gives matrices $\mathbf{Q}(\tilde{\alpha}_\epsilon)$ and $\mathbf{P}(\tilde{\alpha}_\epsilon)$, which must be multiplied to form the desired gradient (Equation 9).

3. Move the inhibitory weights by a small amount in the direction of the negative gradient. That is, for every existing inhibitory synapse w_{ij} (only 40% of all possible inhibitory connections exist at any given time, see step 6), set $w_{ij} \leftarrow w_{ij} - \eta \, [\partial \tilde{\alpha}_\epsilon / \partial \mathbf{W}]_{ij}$. Here $\eta = 10$ is a learning rate.

4. Enforce negativity constraint, by clipping all positive inhibitory weights to zero

5. Enforce the constraint that inhibition must be $\gamma = 3$ times stronger than excitation on average. This is done by writing the connectivity matrix $\mathbf{W}$ block-wise as

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}^{E\to E} & \mathbf{W}^{I\to E} \\ \mathbf{W}^{E\to I} & \mathbf{W}^{I\to I} \end{pmatrix}, \tag{12}$$

and multiplicatively rescaling both inhibitory blocks by $-\gamma\overline{\mathbf{W}}^{E\to E}/\overline{\mathbf{W}}^{I\to E}$ and $-\gamma\overline{\mathbf{W}}^{E\to I}/\overline{\mathbf{W}}^{I\to I}$ respectively, where $\overline{\mathbf{W}}^{X\to Y}$ denotes the average over all matrix elements. This step is not necessary for stability optimization, but is essential to make sure that the high correlation of excitatory and inhibitory input currents that emerges from optimization is not overwhelmed by the baseline correlation contributed by shared population fluctuations (see Discussion, and Renart et al. (2010); Hennequin et al. (2012)).

6. Enforce the maximum density of inhibitory connections, by removing any existing w_{ij} that step 4 may have set to zero, and replacing it by a new connection w_{ik} (which does not exist yet) where inhibitory neuron k is chosen randomly. Set the strength of these new connections to zero initially. Again, this constraint is not strictly required, but adds to the biological plausibility of the resulting connectivity.

Steps 1 through 6 are then repeated until convergence of the spectral abscissa.

In this procedure, ϵ modulates the distance between the spectral abscissa α and its upper bound $\tilde{\alpha}_\epsilon$: if ϵ is decreased, $\tilde{\alpha}_\epsilon$ becomes a tighter upper bound to the spectral abscissa. In pilot studies, we realized that stability could be reached much faster if ϵ was set to decrease progressively during the course of the gradient descent. Empirically, it seemed a good idea to keep the ratio $\tilde{\alpha}_\epsilon/\alpha$ constant, and to adjust ϵ in every iteration to meet this need. Mathematically, this means that the cost function ($\mathbf{W} \mapsto \tilde{\alpha}_\epsilon(\mathbf{W})$) keeps moving, but it becomes a progressively tighter upper bound on α, and crucially, we no longer need to compute $\tilde{\alpha}_\epsilon$ in step 1! We thus capitalized on this observation and set $\tilde{\alpha}_\epsilon(\mathbf{W}) = C\alpha(\mathbf{W})$ in every iteration, with $C = 1.5$, empirically found to be a good choice. Note that this automatically constrains ϵ to a value of $1/\mathrm{tr}[\mathbf{Q}(C\alpha)]$, where $\mathbf{Q}(\cdot)$ is defined in Equation 7. Steps 2 to 6 are then performed as prescribed above. Note that computationally, the cost is still of order $K \cdot N^3$, but the large constant K implied by the iterative root-finding method of Step 1 is eliminated. Note also that Step 2 requires solving Equation 8 and Equation 11, for which we need to compute a Schur decomposition of $\mathbf{W}$ only once (Bartels and Stewart, 1972). As a byproduct, the Schur decomposition also returns the spectral abscissa at no further cost, so α needs not be computed separately. The above simplified procedure is very effective, except for one small detail. As $\tilde{\alpha}_\epsilon$ gets closer to α (which is bound to happen as learning progresses, since $\tilde{\alpha}_\epsilon = C\alpha$ and $\tilde{\alpha}_\epsilon$ decreases), it also becomes increasingly non-smooth as a function of $\mathbf{W}$. For its gradient to remain well-behaved, we therefore kept $\tilde{\alpha}_\epsilon$ some safe margin away from α, by setting $\tilde{\alpha}_\epsilon$ to $C\alpha$ or $\alpha + B$ in every iteration, whichever was the greatest. We used $B = 0.2$.

Analysis of rotational dynamics

The plane of projection of Figure 4D was found with jPCA, a dynamical variant of principal component analysis (Churchland et al., 2012) used to extract low-dimensional rotations from multidimensional time series. Given data of the form $(\mathbf{y}(t), \dot{\mathbf{y}}(t)))$, where $\dot{\mathbf{y}}(t)$ denotes the temporal derivative of $\mathbf{y}(t)$, jPCA fits (through standard least-squares regression) a linear oscillatory model of the form

$$\dot{\mathbf{y}}(t) = \mathbf{M}_{\mathsf{skew}}\mathbf{y}(t) \tag{13}$$

where $\mathbf{M}_{\mathsf{skew}}$ is a skew-symmetric matrix, therefore one with purely imaginary eigenvalues. The two leading eigenvectors of the best-fitting $\mathbf{M}_{\mathsf{skew}}$ (associated with the largest conjugate pair of imaginary eigenvalues) define the plane in which the trajectory rotates most strongly. Here we computed the jPC projection exactly as prescribed in Churchland et al. (2012). Our model data consisted of the population responses $\Delta\mathbf{r}(t)$ during the first 200 ms following the go cue for each of our 27 initial conditions, sampled in 1 ms time steps. Note that the temporal derivatives are directly given by Equation 1, except in the spiking network (see below) where we estimated those derivatives using a finite-difference approximation. To make sure that the jPC projection captures enough of the data variance, that is, that the observed rotational dynamics (if any) are significant, the data was first projected down to its top 6 standard principal components (as in Churchland et al. (2012)).

Muscle activation through linear readouts

In Figure 5, a single pair of muscle readouts was learned from 200 training trials (100 trials for each of the "snake" and "butterfly" movements). We assumed the following linear model:

$$\mathbf{z}_t = (\mathbf{m}_1; \mathbf{m}_2)^T \, \Delta\mathbf{r}_t + \mathbf{b} + \boldsymbol{\varepsilon}_t \tag{14}$$

where $\mathbf{z}_t$ (size 2) denotes the vector of target muscle activations at discrete time t, $\Delta\mathbf{r}_t$ is the $N \times 1$ vector of momentary deviation from baseline firing rate in the network, and $\boldsymbol{\varepsilon}_t$ is the vector of residual errors (size 2). The readout weights (column vectors $\mathbf{m}_i$, $i = 1, 2$) are parameters which we optimized through simple least-squares regression, together with a pair of biases ($\mathbf{b}$). The snake (resp. butterfly) target trajectory was made of 58 points (resp. 26 points), equally spaced in time over 500 ms following the go cue. Those points defined the discrete time variable t in Equation 14. The activity vector $\Delta\mathbf{r}_t$ was sampled accordingly for each movement.

Spiking network simulations

We simulated a network of 15′000 neurons, composed of 12′000 excitatory and 3′000 inhibitory neurons, with parameters listed in Table 1. This network was divided into 200 subgroups of excitatory neurons and 200 subgroups of inhibitory neurons, which can be interpreted as the "rate units" we have focused on until here.

Single-neuron model Single cells were modelled as leaky integrate-and-fire (LIF) neurons (e.g. Gerstner and Kistler (2002), chap. 4) according to

$$\tau_m \frac{\mathrm{d}{V_\mathrm{m}}^{(i)}}{\mathrm{d}t} = -V_\mathrm{m}^{(i)} + V_\mathrm{rest} + h_\mathrm{exc.}^{(i)} + h_\mathrm{inh.}^{(i)} + h_\mathrm{ext.} \tag{15}$$

Neuron i emitted a spike whenever $V_\mathrm{m}^{(i)}$ crossed a threshold V_thresh from below. Following a spike, the voltage was reset to V_reset and held constant for an absolute refractory period of τ_r. The excitatory and inhititory synaptic inputs, $h_\mathrm{exc.}^{(i)}$ and $h_\mathrm{inh.}^{(i)}$ were sums of alpha-shaped post-synaptic currents (PSCs) of the form $c\left[\exp\left(-t/\tau^\mathrm{decay}\right) - \exp\left(-t/\tau^\mathrm{rise}\right)\right]$ where c is a synapse-type-specific scaling factor that regulates peak excitatory and inhibitory postsynaptic potential (PSP) amplitudes after further membrane integration through Equation 15 (Figure 7b).

Recurrent synapses Each neuron received input from 1′500 excitatory and 1′500 inhibitory network neurons. For 50% of those recurrent connections (750 exc. and 750 inh. synapses), the presynaptic partner was drawn randomly and uniformly from the corresponding population (exc. or inh.), providing a sea of unspecific, random synapses that was instrumental in maintaining the network in a regime of asynchronous and irregular firing. These connections were thought to target proximal dendritic zones and therefore to evoke fast PSCs (parameter $\tau_\mathrm{fast}^\mathrm{decay}$ in Table 1). The other half of the network synapses were used to mirror the structure of the network of rate units described throughout the paper, and were therefore drawn according to probabilities jointly determined by i) the subgroups that the pre- and post-synaptic neurons belonged to, ii) an optimized SOC matrix $\mathbf{W}$ of size 400×400 that described the connectivity between subgroups.

We first normalized the excitatory and inhibitory parts for each row of $\mathbf{W}$, obtaining a matrix $\widehat{\mathbf{W}}$ of connection *probabilities*. Then, for any cell i in group m ($1 \leq m \leq 400$, exc. or inh.), each of 750 exc. partners were chosen in two steps: first, a particular group n was picked with probability $\widehat{W}_{mn}$, and second, a presynaptic neuron was picked at random from this group n. We applied the same procedure to generate the second half of the inhibitory synapses (750 per neuron). These structured SOC connections were given a slower PSC decay time constant $\tau^{\text{decay}}_{\text{slow}}$ (c.f. Table 1), and can be interpreted as targeting more distal dendritic parts.

Sample PSPs are shown in Figure 7b for all four types of synapses. The ratio between exc. and inh. synaptic efficacies was set to achieve a stable background firing state of 5 Hz. Note that because of the amplifying behavior of SOCs and the superlinear nature of the input-output function of LIF neurons, the network ended up with a mean of ~ 6 Hz instead (Figure 7c).

Each neuron also received a constant positive external input current $h^{\text{ext.}}$, which was set to the mean current a cell would receive from $5'000$ independent Poisson sources at 5 Hz with fast synapses. We boiled this input down to its mean to motivate that the slow, seemingly stochastic rate fluctuations we observed in the spiking SOC (Figure 8a) did not require any external source of noise.

Generation of W SOC matrices for spiking networks were generated in a similar manner as described above for rate-based networks, except for a few simple variations to account for the effective gains of the excitatory and inhibitory synaptic pathways between subgroups. To calculate these gains, we isolated two subgroups of neurons, n (sender) and m (receiver), and assumed that the rest of the network fired with Poisson statistics at 5 Hz. We then numerically estimated the effect of small firing rate deviations in group n, Δ_n, on the firing rate Δ_m of the receiver (thus linearizing the joint dynamics of subgroups around the background state). Due do the E/I balance, $\frac{\mathrm{d}\Delta_m}{\mathrm{d}\Delta_n}$ is a roughly linear function of the connection probability $\widehat{W}_{mn}$ from group n to group m. Using a simple least-squares fit, we extracted the coefficients of proportionality $\beta_{\text{exc.}} \simeq 5.2$ and $\beta_{\text{inh.}} \simeq -21.9$ for the two types of sender groups (exc. or inh.). These constants were then used to rescale the excitatory and inhibitory parts of the normalized matrix of connection probabilities, $\widehat{\mathbf{W}}$, yielding an effective connectivity matrix $\mathbf{W}$ such that linear dynamics of the form

$$\tau \frac{\mathrm{d}\Delta\mathbf{r}}{\mathrm{d}t} = -\Delta\mathbf{r} + \mathbf{W}\Delta\mathbf{r} + \text{noise} \tag{16}$$

approximated the joint dynamics of our spiking subgroups. We used the smoothed spectral abscissa method described above to modify the connections probabilities $\widehat{W}_{mn}$ ($200 < n \leq 400$) that involved inhibitory presynaptic subgroups to achieve stability in an initially unstable, random sparse matrix. In each gradient step, the inh. part of each row of $\widehat{\mathbf{W}}$ was re-normalized to sum to 1, to preserve the probabilistic interpretation of $\widehat{\mathbf{W}}$. Second, we used a straightforward application of the chain rule to propagate the gradient of the smoothed spectral abscissa of $\mathbf{W}$ to $\widehat{\mathbf{W}}$ given the values of $\beta_{\text{exc.}}$ and $\beta_{\text{inh.}}$.

We used the effective weight matrix $\mathbf{W}$ to compute the preferred initial states of the spiking SOC, akin to the rate-based case with $\Delta r_i \propto x_i$ (i.e. assuming linear dynamics among subgroups given by

description	name	value	unit
membrane time constant	τ_m	20	ms
refractory period	τ_r	2	ms
axonal transmission delay	–	0.5	ms
resting potential	V_{rest}	-70	mV
spiking threshold	V_{rest}	-55	mV
voltage reset	V_{reset}	-60	mV
PSC rise time	τ^{rise}	1	ms
fast PSC decay time	$\tau^{\text{decay}}_{\text{fast}}$	10	ms
slow PSC decay time	$\tau^{\text{decay}}_{\text{slow}}$	100	ms

Table 1: Parameters used in the spiking model.

Equation 16). During the preparatory period, each LIF neuron received a ramping input current that depended only on the subgroup it belonged to.

Control random network The random network used for comparison in Figure 8 was identical in every respect to the SOC, except that presynaptic partners for slow synapses were drawn completely randomly (there was no notion of neuronal subgroups).

Simulations were custom-written in OCaml and parallelized onto 8 compute cores following the strategy developed in Morrison et al. (2005), taking advantage of a finite axonal propagation delay which we set to 0.5 ms. We used simple Euler integration of Equation 15 with a time step of 0.1 ms.

Acknowledgments

Research was supported by the Swiss National Science Foundation grant no. 200020 13287 (Coding Characteristics), ERC grant no. 268689 (MultiRules), and European FP7 grants no. 215910 (BIOTACT) and 269921 (BrainScaleS). Additionally, G.H. was supported by a Swiss National Science Foundation Postdoctoral fellowship no. PBELP3.146554, and T.P.V. was supported by the European Community's Seventh Framework Marie Curie International Reintegration grant no. 268436 and by a Sir Henry Dale Wellcome Trust and Royal Society research fellowship. The authors wish to thank Danilo Rezende, Alex Seeholzer and Friedemann Zenke for useful discussions, and Máté Lengyel for pointing us to Churchland et al. (2012).

References

Afshar, A., Santhanam, G., Yu, B., Ryu, S., Sahani, M., and Shenoy, K. (2011). Single-trial neural correlates of arm movement preparation. *Neuron*, 71(3):555–564.

Ahmadian, Y., Rubin, D. B., and Miller, K. D. (2013). Analysis of the stabilized supralinear network. *Neural Comput.*, 25:1994—-2037.

Amit, D. J. (1992). *Modeling brain function: The world of attractor neural networks.* Cambridge Univ. Pr.

Amit, D. J. and Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex*, 7:237–252.

Bartels, R. H. and Stewart, G. W. (1972). Solution of the matrix equation AX+XB=C. *Communications of the ACM*, 15:820–826.

Ben-Yishai, R., Bar-Or, R. L., and Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proc. Natl. Acad. Sci. USA*, 92:3844–3848.

Branco, T. and Häusser, M. (2011). Synaptic integration gradients in single cortical pyramidal cell dendrites. *Neuron*, 69:885–892.

Brunel, N. (2000). Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *J. Comput. Neurosci.*, 8:183—-208.

Cafaro, J. and Rieke, F. (2010). Noise correlations improve response fidelity and stimulus encoding. *Nature*, 468:964–967.

Churchland, M. M. and Abbott, L. F. (2012). Two layers of neural variability. *Nat. Neurosci.*, 15:1472—-1474.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., and Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, 487:51–56.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Ryu, S. I., and Shenoy, K. V. (2010). Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron*, 68:387–400.

Churchland, M. M. and Shenoy, K. V. (2007). Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *J. Neurophysiol.*, 97:4235–4257.

Dayan, P. and Abbott, L. F. (2001). *Theoretical neuroscience.* Cambridge, MA: MIT Press.

DePasquale, B., Churchland, M. M., and Abbott, L. F. (2013). Low-rank connectivity induces firing rate fluctuations in a chaotic spiking model. In *Cosyne abstracts 2013*, Salt Lake City, USA.

Dorrn, A. L., Yuan, K., Barker, A. J., Schreiner, C. E., and Froemke, R. C. (2010). Developmental sensory experience balances cortical excitation and inhibition. *Nature*, 465:932–936.

Ecker, A. S., Berens, P., Keliris, G. A., Bethge, M., Logothetis, N. K., and Tolias, A. S. (2010). Decorrelated neuronal firing in cortical microcircuits. *Science*, 327(5965):584–587.

Froemke, R. C., Merzenich, M. M., and Schreiner, C. E. (2007). A synaptic memory trace for cortical receptive field plasticity. *Nature*, 450:425—-429.

Fuster, J. M. and Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, 173:652–654.

Ganguli, S., Huh, D., and Sompolinsky, H. (2008). Memory traces in dynamical systems. *Proc. Natl. Acad. Sci. USA*, 105:18970–18975.

Gerstner, W. and Kistler, W. M. (2002). *Spiking neuron models: Single neurons, populations, plasticity.* Cambridge Univ. Pr.

Goldberg, J. A., Rokni, U., and Sompolinsky, H. (2004). Patterns of ongoing activity and the functional architecture of the primary visual cortex. *Neuron*, 42:489—500.

Goldman, M. S. (2009). Memory without feedback in a neural network. *Neuron*, 61:621–634.

Haider, B., Häusser, M., and Carandini, M. (2013). Inhibition dominates sensory responses in the awake cortex. *Nature*, 493:97–100.

Hennequin, G., Vogels, T. P., and Gerstner, W. (2012). Non-normal amplification in random balanced neuronal networks. *Phys. Rev. E*, 86:011909.

Hoerzer, G. M., Legenstein, R., and Maass, W. (2012). Emergence of complex computational structures from chaotic neural networks through reward-modulated hebbian learning. *Cerebral Cortex*, pages 1–14.

Kriener, B., Tetzlaff, T., Aertsen, A., Diesmann, M., and Rotter, S. (2008). Correlations and population dynamics in cortical networks. *Neural Comput.*, 20:2185–2226.

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

Laje, R. and Buonomano, D. V. (2012). Complexity without chaos: Plasticity within random recurrent networks generates robust timing and motor control. *arXiv:1210.2104*.

Laje, R. and Buonomano, D. V. (2013). Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nat. Neurosci.*, Advance Online Publication.

Latham, P. E. and Nirenberg, S. (2004). Computing and stability in cortical networks. *Neural Comput.*, 16:1385–1412.

Litwin-Kumar, A. and Doiron, B. (2012). Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nat. Neurosci.*, 15(11):1498–1505.

Luz, Y. and Shamir, M. (2012). Balancing feed-forward excitation and inhibition via hebbian inhibitory synaptic plasticity. *PLoS Comput. Biol.*, 8:e1002334.

Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Comput.*, 14:2531–2560.

Mariño, J., Schummers, J., Lyon, D. C., Schwabe, L., Beck, O., Wiesing, P., Obermayer, K., and Sur, M. (2005). Invariant computations in local cortical networks with balanced excitation and inhibition. *Nat. Neurosci.*, 8(2):194–201.

Miller, K. D. and Fumarola, F. (2011). Mathematical equivalence of two common forms of firing rate models of neural networks. *Neural Comput.*, 24:25–31.

Morrison, A., Mehring, C., Geisel, T., Aertsen, T. G. A., and Diesmann, M. A. (2005). Advancing the boundaries of high-connectivity network simulation with distributed computing. 17:1776–1801.

Murphy, B. K. and Miller, K. D. (2009). Balanced amplification: A new mechanism of selective amplification of neural activity patterns. *Neuron*, 61:635–648.

Okun, M. and Lampl, I. (2008). Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nat. Neurosci.*, 11:535–537.

Ozeki, H., Finn, I. M., Schaffer, E. S., Miller, K. D., and Ferster, D. (2009). Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62:578–592.

Rajan, K. and Abbott, L. F. (2006). Eigenvalue spectra of random matrices for neural networks. *Phys. Rev. Lett.*, 97:188104.

Rajan, K., Abbott, L. F., and Sompolinsky, H. (2010). Stimulus-dependent suppression of chaos in recurrent neural networks. *Physical Review E*, 011903:1–5.

Renart, A., de la Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A., and Harris, K. (2010). The asynchronous state in cortical circuits. *Science*, 327:587.

Shenoy, K. V., Kaufman, M. T., Sahani, M., and Churchland, M. M. (2011). A dynamical systems view of motor preparation: Implications for neural prosthetic system design. *Progr. Brain Res.*, 192:33.

Sompolinsky, H., Crisanti, A., and Sommers, H. J. (1988). Chaos in random neural networks. *Phys. Rev. Lett.*, 61:259–262.

Stern, M., Sompolinsky, H., and Abbott, L. F. (2013). Dynamics of random clustered networks. In *Cosyne abstracts 2013*, Salt Lake City, USA.

Sussillo, D. and Abbott, L. (2009). Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63(4):544–557.

Terauchi, A. and Umemori, H. (2012). Specific sets of intrinsic and extrinsic factors drive excitatory and inhibitory circuit formation. *The Neuroscientist*, 18(3):271–86.

Tsodyks, M. V., Skaggs, W. E., Sejnowski, T. J., and McNaughton, B. L. (1997). Paradoxical effects of external modulation of inhibitory interneurons. *J. Neurosci.*, 17:4382–4388.

van Vreeswijk, C. and Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274:1724.

Vanbiervliet, J., Vandereycken, B., Michiels, W., Vandewalle, S., and Diehl, M. (2009). The smoothed spectral abscissa for robust stability optimization. *SIAM Journal on Optimization*, 20:156–171.

Vogels, T. P. and Abbott, L. F. (2005). Signal propagation and logic gating in networks of integrate-and-fire neurons. *J. Neurosci.*, 25(46):10786–10795.

Vogels, T. P. and Abbott, L. F. (2009). Gating multiple signals through detailed balance of excitation and inhibition in spiking networks. *Nat. Neurosci.*, 12:483–491.

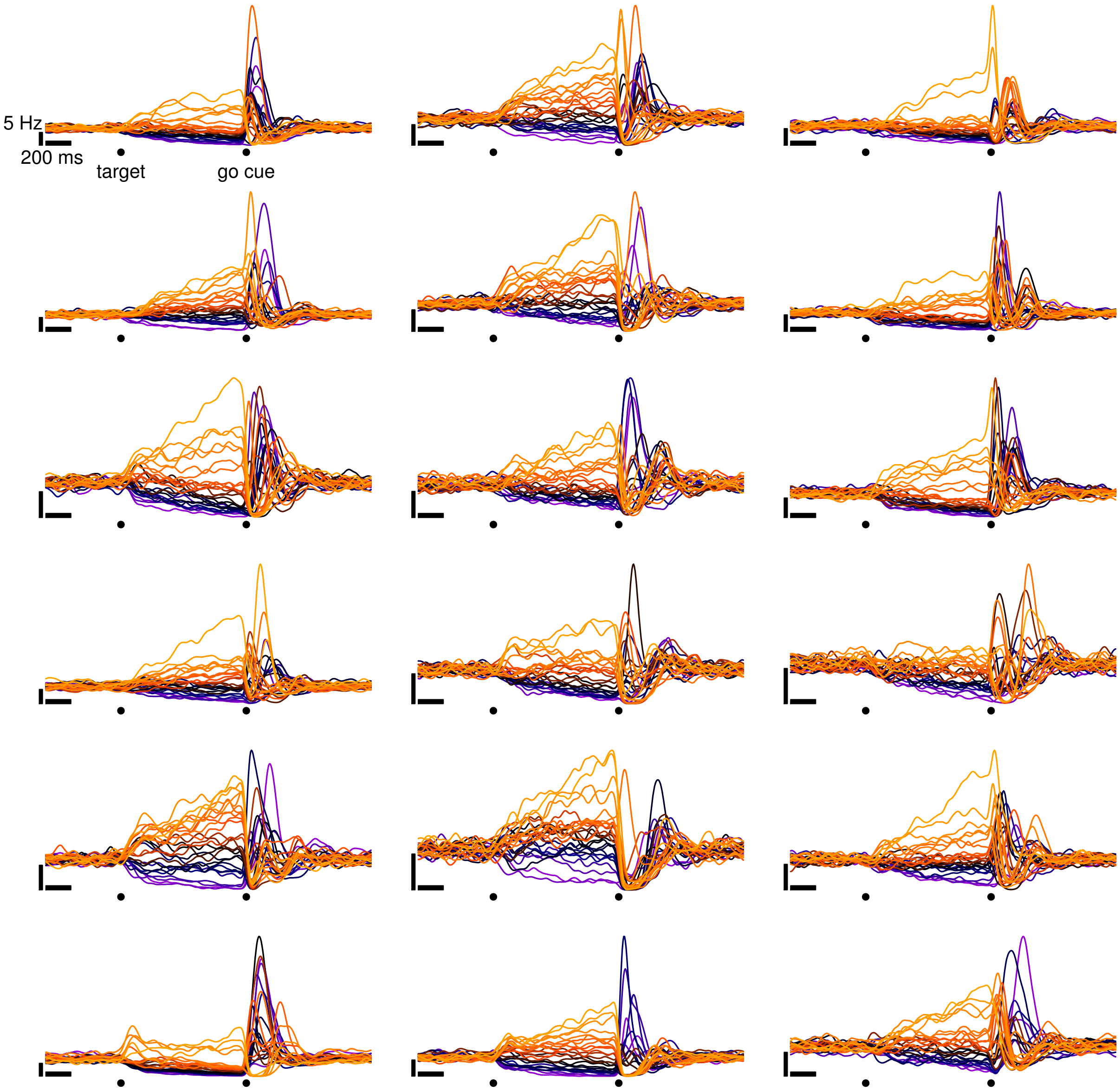
Vogels, T. P., Rajan, K., and Abbott, L. F. (2005). Neural network dynamics. *Annu. Rev. Neurosci.*, 28:357–376.

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.

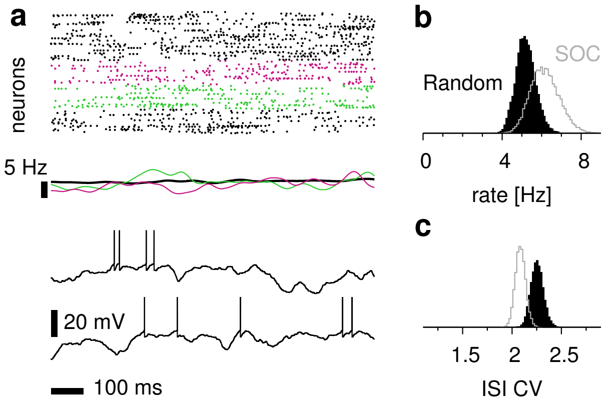
Wang, X. (1999). Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *J. Neurosci.*, 19:9587–9603.

Wang, X. J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36:955–968.

Wehr, M. and Zador, A. (2003). Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature*, 426:442–446.



Supplementary Figure 1: Diversity of single-cell transient responses in the spiking SOC. Shown here are trial-averaged responses (1'000 trials) of the first cell for each of the first 18 subgroups (one panel per subgroup). The preparatory period drives the network in one of 27 random combinations (color coded) of its top 2 preferred initial conditions. The color code is meaningless and not consistent across panels, but only intended to ease visualization. Rate traces were smoothed using a Gaussian kernel of width 20 ms.



Supplementary Figure 2: Control random network. **(a)** Top: raster plot of spiking activity in the control network. Middle: average momentary network activity (black), and activities of the two colored subgroups. Note that the concept of subgroups is absent from the connectivity; subgroups are completely arbitrary here, and only meant to provide a comparison with Figure 8. Bottom: voltage traces for two randomly chosen cells. **(b)** Distribution of average firing rates in the random network (black) and in the SOC (gray). Averate rates were computed from 1'000 second-long simulation. **(c)** Same as in (b), for the CVs of the ISI distributions.

Supplementary Movie 1 The eigenvalues of the original random balanced network are shown as fixed gray dots in the complex plane. Roughly 50% of them lie to the right of the critical line defined by $\mathsf{Re}(\lambda) = 1$ (purple line), thus corresponding to unstable eigenmodes. The moving dots show the evolution of the eigenvalues as the inhibitory connectivity is being refined to produce a SOC. Each frame corresponds to one step of the gradient descent (cf. Experimental Procedures). Purple dots indicate unstable eigenmodes.