Dynamics of feature categorization

Daniel Martí & John Rinzel

Abstract

In visual and auditory scenes, we are able to identify shared features among sensory objects and group them according to their similarity. This grouping is preattentive and fast, and is thought of as an elementary form of categorization by which objects sharing similar features are clustered in some abstract perceptual space. It is unclear what neuronal mechanisms underlie this fast categorization. Here we propose a neuro-mechanistic model of fast feature categorization based on the framework of continuous attractor networks. The mechanism for category formation does not rely on learning and is based on biologically plausible assumptions, like the existence of populations of neurons tuned to feature values, feature-specific interactions, and subthreshold evoked responses upon the presentation of single objects. When the network is presented with a sequence of stimuli characterized by some feature, the network sums the evoked responses and provides a running estimate of the distribution of features in the input stream. If the distribution of features is structured into different components or peaks (i.e., is multi-modal), recurrent excitation amplifies the response of activated neurons, and categories are singled out as emerging localized patterns of elevated neuronal activity (bumps), centered at the centroid of each cluster. The emergence of bump states through sequential, subthreshold activation, and the dependence on input statistics comprises a novel application of attractor networks. We show that the extraction and representation of multiple categories is facilitated by the rich attractor structure of the network, which can sustain multiple stable activity patterns for a robust range of connectivity parameters compatible with cortical physiology.

1 Introduction

A basic step for identifying structure in a sensory scene (visual, auditory, somatosensory) is to group related elements according to their similarity (Wertheimer, 1923; Koffka, 1999). Elements that are similar are grouped together, and groups that are dissimilar are segregated from one another. The process of grouping by similarity is easy, fast, and does not require attention or training when the stimuli are sufficiently simple. It has been consistently observed across visual (Beck, 1966, 1982; Olson and Attneave, 1970) and auditory (Heise and Miller, 1951; Bregman and Campbell, 1971; van Noorden, 1977) modalities and, at a cognitive level, is thought to play a pivotal role in the formation of concepts and abstract categories (Goldstone, 1994; Estes, 1994). Grouping as a basic form of categorization provides a simplified description of the scene, reducing redundancy through the use of a few statistical properties like mean values, ranges, or variances (Barlow, 1989). In psychophysical studies, subjects report mean values much more accurately than they can report the values of individual tokens (Ariely, 2001; Chong and Treisman, 2003), suggesting that at a perceptual level the scene is represented by its overall statistical properties (Hochstein and Ahissar, 2002).

Several conceptual models appeal to the existence of a *perceptual space* of features, in which individual objects are identified with points and where similarity relations correspond to distances (Shepard, 1987; see also Nosofsky, 1986; Kruschke, 1992). Categories are described as clusters in perceptual space, and clusters are regarded as samples from a probability density function over the space of perceived features (Fried and Holyoak, 1984; Ashby and Alfonso-Reese, 1995). In this view, the first stage of grouping by similarity is to solve the problem of estimating the distribution of feature values and identifying its modes—i.e., its peaks, centered at the category "prototypes". The neural mechanisms that underlie the formation of these categorical representations remain, however, unknown.

Learning constitutes a plausible mechanism with which to form categories that reflect the statistical structure of stimuli (Brunel et al, 1998; Rosenthal et al, 2001). In Hebbian-based models of categorization, the 'learned' synaptic modifications shape the internal representations of the categories, given by stable patterns of neuronal activity—or attractors (Amit, 1989). Learning is however a relatively slow process. While synaptic efficacies can be modified within a minute (Bi and Poo, 1998; O'Connor et al, 2005; Froemke et al, 2006), such timescales are much longer than those involved in preattentive phenomena. Slow synaptic modifications are therefore not likely involved in the categorization process of grouping, which is fast and contingent on the statistics of stimuli processed within a short temporal window.

Here we propose a simple network model of fast perceptual category formation that does not depend on learning but, rather, on the interplay between the ongoing activity of the network and the statistics of stimulation during a sequence of brief presentations. We apply the well-known and established framework of continuous attractor systems (Amari, 1977; Er-

mentrout, 1998). The network has neurons tuned to feature values and coupled to one another through feature-specific connections: neurons tuned to similar values strongly excite each other, while neurons tuned to disparate values either inhibit each other or do not interact at all, depending on the disparity. For a wide range of parameters, these network models have a stable uniform rest state and can also generate and sustain localized activity patterns in response to adequate transient inputs. In these localized patterns, or *bumps*, only a subgroup of neurons with similar selectivity properties are active. Because there is a continuous set of such patterns, the network activity can encode the value of a continuous variable. This property makes attractor networks an appealing model to describe a wide variety of phenomena involving the representation of analog quantities in the brain, like, e.g., the orientation selectivity of cells in primary visual cortex (Somers et al 1995; Ben-Yishai et al 1995; See Discussion for further examples and references).

Here we put this framework to novel use. We present the network with sequences of random, brief, weak stimuli and associate the emergence of a bump with the dynamic formation of a category. If enough samples, adequately similar, are presented within the network's integration time window a category forms due to the regenerative properties of recurrent excitation, which is strong enough to sustain the category representation after the stimulus stream is terminated. Categories do not form if samples are too disparate or too infrequently presented. Moreover, our network is capable of generating multiple, separated bumps if the input stream contains several clusters of feature values (Figure 1). The network implements a running estimate of the stimulus distribution, with emerging bumps providing the neuronal representations of the categories present in the input. The regenerative emergence, a temporal event, could be used to activate a short-term synaptic plasticity mechanism to identify and retain the location of the bump, and thereby freeze it for later classification of inputs, but we do not address this aspect here. We focus on the dynamics of category formation, acknowledging that after termination of a stimulus stream the bumps may slowly drift to become equally spaced in feature space (Figure 1). While the existence of multistable steady states with evenly-spaced bumps has been treated mathematically by various authors (Laing et al, 2002; Laing and Troy, 2003; Coombes et al, 2003; Guo and Chow, 2005), our main focus here is on the transient phase of bump emergence and its relation to the input statistics.

The paper is organized as follows. In Section 2 we present and motivate the network model. In Section 3 we review and describe the conditions that allow for the coexistence of multiple bump states. In Section 4 we present our main results. We first describe the network's response properties for individual stimuli. Then, using a simplified, sequential stimulation protocol, we illustrate the extraction of category prototypes carried out by the network. We also describe using semi-analytical arguments how the statistics of inputs determine the emerging activity pattern, as well as the accuracy and times of prototype extraction. The last section is the Discussion.

2 Description of the network model

The network model is essentially an adapted version of the ring model of orientation tuning (Somers et al, 1995; Ben-Yishai et al, 1995; Hansel and Sompolinsky, 1998).

Architecture

The network consists of a set of neurons selective to a one-dimensional feature, which we assume for concreteness is an orientation angle θ defined in the range $[-\pi/2, \pi/2)$, although any other continuous feature variable would be equally valid. Unlike models of orientation selectivity in the primary visual cortex, where θ is the physical angle of the stimulus, here θ represents the angle *perceived* by the subject. The fact that θ is periodic simplifies the analysis of the model, but it is not essential. Each neuron $i = 1, \ldots, N$ has a localized tuning curve and is labeled by its preferred (perceived) orientation angle θ_i at which the neuron responds with highest activity. For mathematical convenience, we assume that the set of preferred orientations of all neurons, $\theta_1, \ldots, \theta_N$, is evenly spaced and that the number of neurons is infinitely large. This assumption allows us to replace the discrete index $i = 1, \ldots, N$ by the continuous index θ, so that sums over neurons transform into integrals over preferred orientations of cells.

The connectivity kernel

All the neurons in the network are coupled with one another. The efficacy J_{ij} of the synapse between a presynaptic neuron j and postsynaptic neuron i is a function that depends on the distance in orientation space between their preferred orientations, θ_i and θ_j. That is, $J_{ij} = J(|\theta_i - \theta_j|)$, where the function $J(\theta)$ is the connectivity footprint. Given this dependence of synaptic efficacies on distances in feature space, it is natural to think of neurons as arranged by their preferred orientations, which gives rise to a ring layout (Fig. 2a).

We take the connectivity kernel to be a difference of an excitatory and an inhibitory footprint, $J_E(\theta)$ and $J_I(\theta)$, each of which has a Gaussian-like profile satisfying periodic boundary conditions. Specifically, the connectivity kernel is a difference of circular normal (or von Mises) functions

$$\begin{aligned} J(\theta) &= J_E(\theta) - J_I(\theta) \\ &= j_E \frac{\exp\big(m_E \cos(2\theta)\big)}{I_0(m_E)} - j_I \frac{\exp\big(m_I \cos(2\theta)\big)}{I_0(m_I)}, \end{aligned} \tag{1}$$

where $m_E > 0$ and $m_I > 0$ are the so-called concentration parameters of the circular normal distributions. These parameters are analogous to the inverse variance of a normal distribution. The other two free parameters are the coefficients j_E and j_I, which are positive and set the overall weight of excitation and inhibition, respectively. The normalization factor $I_0(m)$ is the zeroth-order modified Bessel function of the first kind,

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos\theta)\, \mathrm{d}\theta,$$

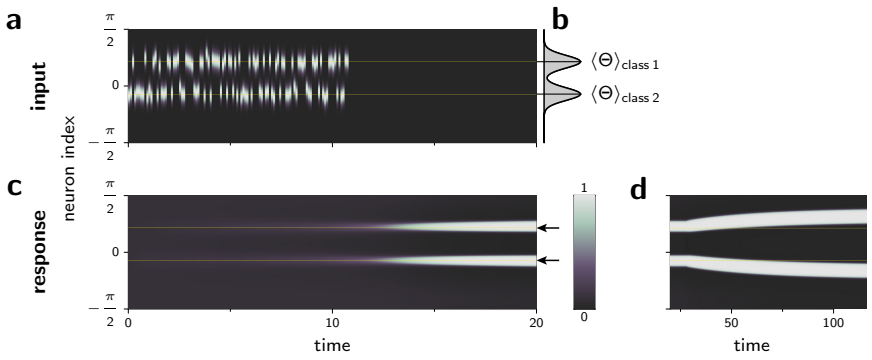

Figure 1. Grouping by similarity in sequential protocols. A sequence of stimuli is presented to the network. The presentation of a single item is modeled as a brief external input with a certain spatial profile, centered at the presented feature value (**a**). The features presented in the sequence are random and drawn according to some predefined feature distribution, which in this example has two discernible modes (**b**). Such a protocol induces the formation of localized patterns of activity centered at the category centers (**c**). The locations of the category centers are indicated by thin yellow lines (a and c) and black arrows (c). The localized patterns that emerge are not necessarily stable and may drift apart over long timescales (**d**).

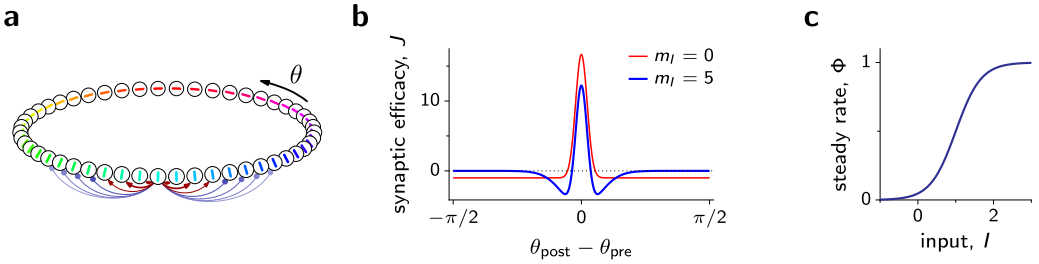

Figure 2. Continuous attractor model. **(a)** The network is composed of one population of neurons selective to orientation and coupled to one another through excitatory-inhibitory couplings. Each circle represents an iso-orientation column with preferred orientation angle indicated by the inner bar. For clarity, the diagram shows only the couplings of one particular column; red arrows represent excitatory connections, and disc-ended blue arrows represent inhibitory connections. **(b)** Synaptic efficacies depend on the difference of preferred orientation angles of pre- and postsynaptic cells, and are modeled as a weighted difference of two circular normals, see Eq. (1). The figure shows the connectivity profile for two different values of the concentration of the inhibitory footprint, m_I. The other connectivity parameters are $j_E = j_I = 1$ and $m_E = 50$. **(c)** Current-to-rate transfer function $\Phi(I)$, given by the sigmoid in Eq. (5) with parameters $\beta = 3$ and $x_0 = 1$.

and ensures that the von Mises functions multiplying the coefficients j_E and j_I in Eq. (1) are normalized under the measure $\pi^{-1}\int_{-\pi}^{\pi}\mathrm{d}\theta$, so that

$$\frac{1}{\pi}\int_{-\pi/2}^{\pi/2} J(\theta)\,\mathrm{d}\theta = j_E - j_I. \tag{2}$$

Circular normal functions become narrower the larger the concentration parameter m is. For values of m larger than $m \approx 4$, the profile of the circular normal is narrow relative to the width of orientation space, and the function can be approximated as a Gaussian function with standard deviation $\sigma = 1/\sqrt{2m}$ radians (see Appendix).

Dynamics

In the limit of an infinitely large network and a dense coverage of all preferred orientations, the firing activity of all the neurons in the network can be described by a continuous function $r(\theta, t)$. The temporal evolution of the activity profile $r(\theta, t)$ is governed by the integro-differential equation

$$\tau\frac{\partial}{\partial t}s(\theta,t) = -s(\theta,t) + r(\theta,t), \tag{3}$$

$$r(\theta,t) = \Phi\left[\frac{1}{\pi}\int_{-\pi/2}^{\pi/2} J\big(\theta-\theta'\big)s\big(\theta',t\big)\,\mathrm{d}\theta' + I(\theta,t)\right], \tag{4}$$

where $s(\theta, t)$ represents synaptic activation. Thus, firing rates respond instantaneously to changes in synaptic activation (Eq. (4)), and drive in turn synaptic activity through a temporal low-pass filter with time constant τ (Eq. (3)). The function $\Phi(\cdot)$ is the static current-to-rate transfer function (or *f-I* curve), which specifies how steady synaptic currents are transformed into firing rates. We take this function to be a sigmoid (Fig. 2c)

$$\Phi(x) = \frac{1}{1+\exp\big(-\beta[x-x_0]\big)}, \tag{5}$$

where β and x_0 are gain and location parameters, respectively. This particular choice for Φ is motivated by mathematical convenience and because it captures the most salient features of experimental *f-I* curves—namely, the vanishing at strong negative currents, the existence of a convex region at moderately low currents (subthreshold regime), and the presence of a concave region for higher currents (suprathreshold regime). We set $\beta = 3$ and $x_0 = 1$ throughout.

The synaptic currents in Eq. (4) consist of a recurrent contribution, given by the convolution of the rates with the connectivity profile $J(\theta)$, and an external input $I(\theta, t)$ that may depend on time and on the preferred orientation of each neuron. Note that firing rates are confined to the interval between 0 and 1 because of our choice for Φ. In the following we set $\tau = 1$, so that time is expressed in units of the synaptic time constant.

External inputs

The total external input $I(\theta, t)$ is

$$I(\theta,t) = I_{\mathrm{stim}}(\theta,t) + I,$$

where $I_{\mathrm{stim}}(\theta, t)$ are the feed-forward signals attributable to the stimuli, and where I is a constant input that accounts for background activity in areas outside the network.

When a single stimulus is presented, each neuron receives input in a magnitude that depends on the distance between the stimulus value and the neuron's preferred value. Specifically, the input to the network when stimulus Θ is presented is a circular Gaussian profile centered at Θ,

$$E(\theta-\Theta) = I_s \exp\left\{m_s\big[\cos(2\theta-2\Theta)-1\big]\right\}, \tag{6}$$

where I_s is the intensity and $m_s > 0$ determines how localized the input profile is in orientation space.

The steady homogeneous solution

In the absence of stimulation, the dynamical equation (3) may have one or several steady homogeneous solutions—i.e., solutions that do not depend on time or θ. These solutions are of the form $s(\theta, t) = R$, where R is a constant determined self-consistently. Imposing such a solution on Eqs. (3)–(4) leads to

$$R = \Phi(J_0 R + I), \tag{7}$$

where J_0 is 0-th order Fourier coefficient of the connectivity, whose expression is given by Eq. (2). Depending on the shape of Φ and the value of the parameters I and J_0, there may exist more than one R satisfying the fixed-point equation (7). We choose our parameters so that (i) there is only one such solution and (ii) the steady rate R is low. This way we can associate the steady homogeneous solution with the spontaneous activity of a cortical circuit. The value of R can always be chosen to be low by adjusting the baseline external input I. We adjust I so that $R = 0.1$.

3 Stable activity patterns

The network can sustain an unstructured, *homogeneous* state in which all neurons fire steadily at the same low firing rate. Such a state can be regarded as the spontaneous activity of a cortical circuit, and represents the absence of encoded categories. For a wide range of connectivity parameters the network can sustain additional stable states, in which the firing activity of each neuron is modulated by its preferred orientation. These so-called modulated activity patterns consist of one or more subgroups of active neurons, and can be evoked from the homogeneous state with finite-amplitude perturbations.

We have analyzed the conditions that the connectivity parameters should meet so that stable modulated patterns coexist with the stable homogeneous state (see Appendix). The analysis demonstrates the existence of a subcritical bifurcation and shows that activity patterns with *several* activity peaks arise naturally when the connectivity profile combines short-range excitation with mid-range (not global) inhibition—i.e., when the connectivity has a narrow Mexican-hat profile. Intuitively, this configuration allows for modulated states with several activity peaks because the spread of interaction between neurons is narrow with respect to the whole neuronal

space, so that the presence of one bump of activity in one particular region does not prevent other bumps from forming elsewhere.

Another condition necessary for the appearance of modulated patterns is that the overall amount of excitation, parametrized in our model by j_E, is broadly balanced with the overall amount of inhibition, parametrized by j_I. This balance prevents the appearance of homogeneous states with saturated activity, and ensures that the only homogeneous stable state is that associated with spontaneous activity. Importantly, the analysis shows that modulated patterns can coexist with the homogeneous state if, in addition to the conditions above, the firing rate of neurons in the homogeneous state is sufficiently low. These conditions guarantee that the homogeneous state destabilizes through a subcritical bifurcation, and thus provide an a priori set of necessary conditions for multistability. Numerical simulations confirm that modulated and homogeneous state indeed coexist when these conditions are met, as we show below.

We have picked a particular configuration of the connectivity profile that fulfills all these requirements, with synaptic strengths $j_E = j_I = 3.5$ and concentration parameters $m_E = 100$ and $m_I = 1$. Figure 3a shows the corresponding profile. If the network is initially in the homogeneous state, injecting a sufficiently strong transient input with spatial modulation induces a transition to a modulated activity pattern (Figs. 3b,c,d, thick black). This pattern persists after the offset of stimulation, and has the same number of peaks as the input profile, two in this case. Weak stimulation, on the other hand, fails to induce a transition between network states, but evokes instead a response that decays to the homogeneous state (Fig. 3b,c,d, thin blue).

A key feature of the network configuration is that it allows for the coexistence of patterns with different number of peaks. Figures 3c,d show the profiles of, respectively, the transient input and the ensuing steady activity pattern measured long after the offset of stimulation. Note that the emerging activity profile has the same angular phase as the input, i.e., activity bumps are aligned with the peaks of the stimulation profile. There is a continuum of such solutions; if the network is stimulated with the same input profiles, but shifted by an arbitrary phase, the emerging activity profiles are also shifted and keep the alignment. Thus for each pattern with a particular number of bumps, there exists a continuous set of equivalent stable states that are related to one another by an arbitrary shift of their spatial (angular) phase. Or, put differently, the set of modulated steady states with a given number of bumps defines a one-dimensional manifold parametrized by the orientation angle θ. The invariance under arbitrary shifts in the neuron index is the hallmark of continuous attractor networks, and is a direct consequence of the homogeneity of neurons and synapses (see Discussion).

Limit in the number of bumps accommodated by the network

The finite spread of excitation and inhibition imposes a limit in the number of bumps that the network can accommodate. As the number of bumps in the modulated pattern increases, the distance between any two neighboring bumps gets shorter and more strongly will the bumps inhibit one another due to surround inhibition. When the distance between two neighboring bumps is short enough, the competitive interaction between them is too strong to allow for the simultaneous coexistence of the neighboring bumps. This effect can be seen in Fig. 3e, which depicts the recurrent inputs received by every neuron when the network activity is in a particular stable state. Inhibition currents are strongest at the center of a bump, and die down at the flanks. As more bumps are added, the inhibition felt by neurons at bump flanks becomes stronger and eventually comparable to the inhibition at the centers of bumps. Such strong inhibition prevents the formation of new localized patterns. Incidentally, it is also responsible for the thinning of the bumps that one observes as more bumps are fit in the pattern (Fig. 3d).

Bumps drift slowly to a symmetric configuration

The patterns illustrated in Fig. 3d are special because they were induced by a very particular set of input profiles, consisting of peaks evenly spaced. We made this choice to maximize the chances of selecting an activity profile with a well defined wavenumber (i.e., number of peaks; see Appendix). In general, however, input profiles can have an arbitrary shape and may induce more complicated patterns that may not necessarily be stable states of the network (Coombes et al, 2003). For example, if the network is stimulated with a transient input profile with two peaks separated 40 degrees, the pattern that emerges consists of two bumps initially separated by 40 degrees apart. This activity pattern is not stable: the two bumps slowly drift away from each other until they are exactly 90 degrees apart (Fig. 4a). The same occurs for other distances between peaks (Fig. 4b).

More generally, once a given number of bumps is formed, the resulting activity pattern will slowly evolve to a configuration with equidistant bumps. This drift arises because the inputs feeding the flanks of each bump are not exactly balanced when bumps are not evenly spaced (Figs. 4c, and compare with Figs. 3e,f). The asymmetry in the inputs cause bumps to move in the direction where the asymmetry decreases (Ermentrout, 1998). Because the connectivity footprints are narrow and have exponentially suppressed tails, input asymmetries are, when they exist, small. As a result, drifts are slow, with a timescale orders of magnitude longer than the timescale of neuronal dynamics and bump formation. We can thus consider the emerging bumps, whether or not they are evenly spaced, as effectively stable states of the network over the timescales we are considering, and neglect the effect of slow drift (see Discussion).

4 Network response to stimulation

We now study how the network responds to stimulation, once the system is initialized in the homogeneous state and operates in the multistable regime described in the previous section.

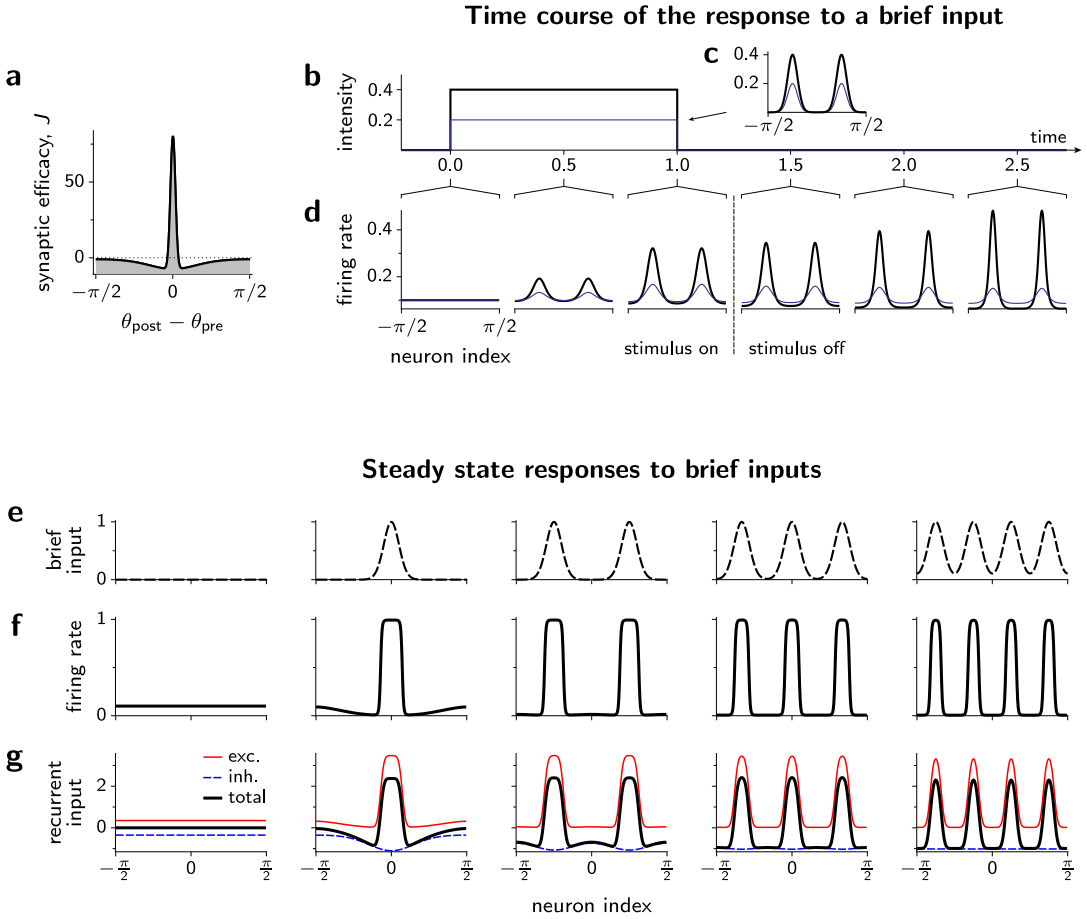

Figure 3. The network is multistable and can accommodate several bumps (**a**) Connectivity profile with parameters $j_E = j_I = 3.5$, $m_E = 100$, and $m_I = 1$. This parameter configuration allows for the coexistence of activity patterns with up to four bumps. (**b**)–(**d**) Sub and super-threshold inputs. The network is fed with external inputs during the time interval $[0, 1]$ (b) with an input profile consisting of a sum of two peaks centered at $\theta = -\pi/4$ and $\theta = \pi/4$ (c), each of which has concentration parameter $m_s = 10$ and intensity $I_s = 0.4$ (thick black), and $I_s = 0.2$ (thin blue). (**d**) Snapshots of the activity profile at different stages of a stimulation protocol. The strongest stimulation elicits the formation of stable modulated patterns. In contrast, the weakest stimulus evokes a response that slowly decays to the homogeneous state when input is switched off. For strong enough transient inputs, the shape of the input profile (**e**) determines the final steady activity pattern (**f**). The input profiles shown in (e) are injected to the network for 1 time unit. As a result, the network evolves to a stable modulated pattern of activity, shown in the corresponding plot in panel (f). Each plot in (f) is the activity pattern measured 50 time units after stimulus offset, when the network activity has practically stabilized. (**g**) Magnitude of the excitatory (red), inhibitory (blue) and total (black) recurrent input received by each neuron, for the stable pattern shown in (f). The current-to-rate transfer function is the sigmoid shown in Fig. 2, with same parameters x_0 and β.

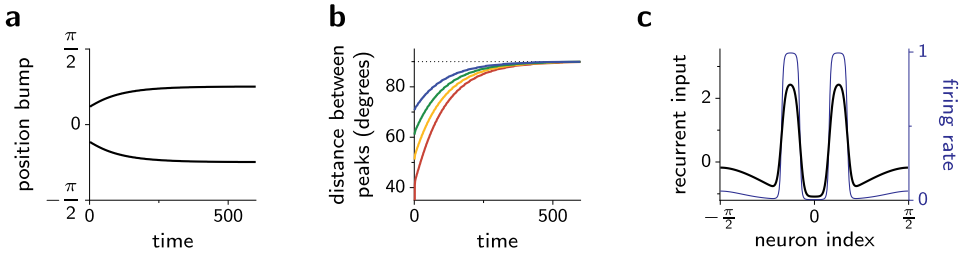

Figure 4. Slow repulsion between coexisting bumps. **(a)** Evolution of the location of two bumps initially separated 40 degrees apart. Notice the long timescale. **(b)** Time evolution of the distance between centers of two bumps initially separated 40, 50, 60, and 70 degrees. (**c**) Snapshot of the recurrent input (thick black) and activity profile (thin blue) at time $t = 10$ of the trial shown in panel (a). Notice the slight difference of recurrent inputs at both sides of each bump. Parameters as in Fig. 3.

Stimulation protocol

Stimulation is modeled as a sequential sampling of features, drawn from some particular distribution characterizing the statistical structure of the sensory scene. Such a random sampling can be thought of as the feed-forward signals sent from sensory areas during a random exploration of the scene, under the assumption that only one item can be processed at a time. This assumption is not fundamental for the results, but it simplifies the formulation of the problem.

According to this protocol model, the input stream feeding the network is a sequence of orientation angles $\Theta_1, \Theta_2, \ldots$, randomly drawn from some predefined distribution $P(\Theta)$ on the interval $[-\pi/2, \pi/2)$ (see below). The sequence is modeled as a superposition of single short-duration pulses of the form

$$I_{\text{stim}}(\theta, t) = \sum_{i=1}^{\infty} \alpha(t - t_i) E(\theta - \Theta_i), \tag{8}$$

where t_i is the time of occurrence of the ith stimulus of the sequence, $\alpha(t)$ is the temporal course of the pulse stimulation, and $E(\theta - \Theta_i)$ is the spatial profile given by Eq. (6). We choose the function $\alpha(t - t_i)$ as a switch that is turned on at t_i for a period d,

$$\alpha(t - t_i) = \begin{cases} 1 & \text{if } t_i \le t \le t_i + d, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The sequence of times t_i, $i = 1, 2, \ldots$, is an arithmetic progression given by $t_i = t_0 + i(d + d_0)$, where d_0 is a short inter-pulse time. Note that pulses do not overlap in time.

Implicit in our choice for $\alpha(t)$ is the assumption that inputs provide a moment-by-moment representation of the feature values being processed, an assumption that is consistent with the observation that lower areas respond fast to the presentation of stimuli (see, e.g., Ringach et al, 1997). The time course of the input described by Eq. (8) is in particular loosely inspired by the fast responses of V1 neurons observed during the presentation of a sequence of gratings with random orientations (Benucci et al, 2009). Of course, in real visual exploration tasks the inputs received by high level areas are definitely going to be far more complex than the simple sequential protocol proposed here, possibly involving the simultaneous representation of features present in broad regions in the visual field. The existence of such a parallel neural representation is indirectly supported by the presence of large receptive fields of high visual areas, as well as the observation that grouping phenomena are very sensitive to the spatial arrangement of stimuli (Wannig et al, 2011). This suggests that high visual areas may be receiving input signals that reflect some type of spatial integration among nearby objects. As a first approximation, these effects could be incorporated into the model by considering a more complex stimulation protocol in which input profiles convey an ongoing, spatially filtered version of the foveated image during exploration. While spatial effects are likely to play a key role in basic visual categorization, we have decided to ignore them here and restrict our investigation to how similarity is represented in high level areas, irrespective of how inputs are transformed in earlier stages.

Category structure of inputs

The feature distribution $P(\Theta)$ captures the full statistical structure of the sensory scene, and is given as follows. We assume that every object in the scene is the member of some particular category C_i, and that each category has probability of occurrence $P(C_i)$. Categories are specified by the conditional feature distribution $P(\Theta|C_i)$, which characterizes how the instances of category C_i are scattered around its prototype. Following our intuition that categories consist of dense regions of "typical instances surrounded by sparse regions of atypical ones" (Fried and Holyoak, 1984), we assume that the conditional distributions $P(\Theta|C_i)$ are unimodal and symmetric. The total distribution of features $P(\Theta)$ is then

$$P(\Theta) = \sum_{i=1}^{M} P(\Theta|C_i) P(C_i) \tag{10}$$

where M is the number of categories. Equation (10) can be regarded as a linear combination of Gaussian-like functions, or mixture of Gaussians, where each component $P(\Theta|C_i)$ is weighted by a coefficient given by the category prior $P(C_i)$.

To generate a random sequence of stimuli distributed according to Eq. (10), we select at each presentation a category according to the probabilities $P(C_1), \ldots, P(C_M)$, and then we draw a value from the probability density conditioned to the selected category $P(\Theta|C_i)$. For concreteness we take $P(\Theta|C_i)$ as a circular Gaussian distribution with mean μ_i and concentration parameter m_c, the latter assumed equal for all categories. We also restrict ourselves to easy categorization tasks and assume that there is no significant overlap between categories. We thus take the different conditional distributions $P(\Theta|C_i)$ as narrow relative to the whole range of feature values, and with means sufficiently separated from one another. In other words, the discriminabilities, defined as $d' = |\mu_i - \mu_j|/\sigma_c \simeq |\mu_i - \mu_j|\sqrt{2m_c}$, are all large for any $i \neq j$. The distribution $P(\Theta)$ is then multimodal and sharply peaked.

We stress that in our model there is no stochasticity associated with the neuronal responses. The presentation of a particular stimulus Θ elicits a reliable network response that is completely determined by the input profile $E(\theta - \Theta)$, Eq. (6). The only source of stochasticity is the random sampling from the feature distribution $P(\Theta)$.

Subthreshold stimulation

A key assumption of the model is that inputs are subthreshold. The pulse associated with the presentation of a single object is not strong enough or peaked enough to induce the transition to a stable modulated pattern. More precisely, applying a single pulse stimulus perturbs the system, but it does not drive the system out of the basin of attraction of the homogeneous state, defined as the set of all initial activity profiles that converge to the homogeneous state after stimulation is switched off.

We illustrate the subthreshold nature of single pulse stimulation in Fig. 5a, where we show the temporal course of both the stimulation and the elicited response. Because the

pulse intensity I_s is chosen small, the network responds to the stimulation with a weak, slowly decaying trace in the network activity (see the response panel of Fig. 5a). The relatively long timescale of the decay results from the reverberant activity brought about by recurrent excitation, which reinjects the activity into the network and compensates for the dissipation caused by neuronal leakage, thereby slowing the decay. This effect will be more apparent the stronger the perturbation, as more excitation is recruited from collateral neurons (Figs. 5b,c). The slowing is more prominent at intensity values that are close to trigger a transition to a modulated pattern, where the positive feedback from recurrent excitation is comparable to the negative feedback from leakage (see curve $I_s = 0.20$ in Fig. 5b). For intensities exceeding a critical value, the induced positive feedback dominates over negative feedback, and the network evolves to a modulated activity profile (see curve $I_s = 0.25$ in Fig. 5b).

To push the system out of the basin of attraction of the homogeneous state, perturbations have to be not only sufficiently strong, but also sufficiently modulated. A stimulus with a broad profile activates a large number of neurons that inhibit one another due to surround inhibition. These mutual inhibitory interactions quickly wash out the activation induced by the stimulus, speeding up the decay to the uniform state. In contrast, inputs with a narrow profile activate groups of neurons that fall within the range of local excitation, inducing cooperative interactions that are reflected in higher reverberation and longer decay times. These points are illustrated in Fig. 5e. Note that the decay becomes slower as m_s increases, i.e., as the spatial profile of the input gets narrower.

In summary, transient activity profiles persist for longer the more modulated (i.e., the more narrow and intense) they are. This is because modulated patterns are configurations that fall close to the boundary of the basin of attraction associated with the homogeneous state. Intuitively, the system slows down when it is pushed near the boundary between basins of attraction because that's where the attracting drive to the initial attractor is weaker. As we show below, this property plays a key role in the emergence of category representations when the inputs feeding the network are distributed non-uniformly.

Category formation conditioned on the statistics of the stimuli

The network singles out the non-uniformities in the distribution of inputs by slowing the decay of modulated patterns and, eventually, forcing the transition to a stable activity pattern that will reflect the structure of stimulation. When the time between successive stimulus presentations is short enough, the evoked activity traces sum up and provide a spatial trace of the presented stimuli. An input stream containing sequences of similar features will produce successive hits in one or more localized regions of feature space, causing the activity profile to become modulated enough to trigger one or more bumps of activity. This will happen if the temporal window of integration set by the activity decay is long enough to sample enough objects, so that the statistical structure of the inputs can be revealed.

The mechanism of category formation is illustrated in Fig. 6, which summarizes the response of the network to different input streams characterized by a particular feature distribution $P(\Theta)$. When the input stream consists of a long sequence of pulses of similar orientations, the traces evoked by a single stimulation accumulate over time and eventually force the system to select a particular modulated state—in this case, a state with one bump (Fig. 6a). The emerging bump is centered around the mean orientation of the pulses in the sequence, and it hence encodes implicitly the first-order statistics of the exemplars belonging to the category. The curve next to the input panel in Fig. 6a shows a schematic of the average input profile that one would have if there were infinitely many pulses (see next section for details). Note the alignment of the bump with the maximum of the input distribution, indicated by the arrow on the right of the response panel.

For the creation of bump states it is necessary that the overall distribution of feature values of the input stream be sufficiently peaked, showing one or more modes, or "clusters" of similar items. Consequently, feeding the network with sequences of uniformly distributed orientations fails to induce a bump (Fig. 6b). The network responds in this case with a sequence of weak, scattered traces that do not accumulate because the surround inhibition evoked by subsequent presentations tends to efface the traces of prior dissimilar stimuli (see white inset in the response panel of Fig. 6b). In psychophysical parlance, sequences of statistically unstructured stimuli induce *masking* between successive presentations.

Crucially, the network is not limited to the encoding of one category. When the feature values are drawn from a mixture of two or more narrow distributions, bumps emerge around the mean value of each orientation cluster (Figs. 6c-g). The fact that a variable number of bumps can be sustained at arbitrary locations in feature space results from the rich attractor structure of the network, which combines multistability between multipeaked activity patterns with invariance under phase shifts. This attractor configuration enables the network to represent in an elementary way a wide range of mixture distributions. Distributions are encoded in this representation as $r(\theta) = \sum_{i=1}^{M} U^{(M)}(\theta - \mu_i)$, where μ_i is the prototypical value of each category, and where the function $U^{(M)}(\theta)$ is the profile of an individual bump when the whole network profile contains M bumps. We use the superscript (M) to emphasize that the shape of the individual profile varies according to the number of bumps contained in the activity profile (see Fig. 3f).

The average input profile determines the emerging activity pattern

The emergence of bumps at peaks of the distribution of features can be understood in terms of the average input profile received by the network. When the duration of individual pulses is short compared to the typical integration time of the network, the barrage of pulses is well approximated by

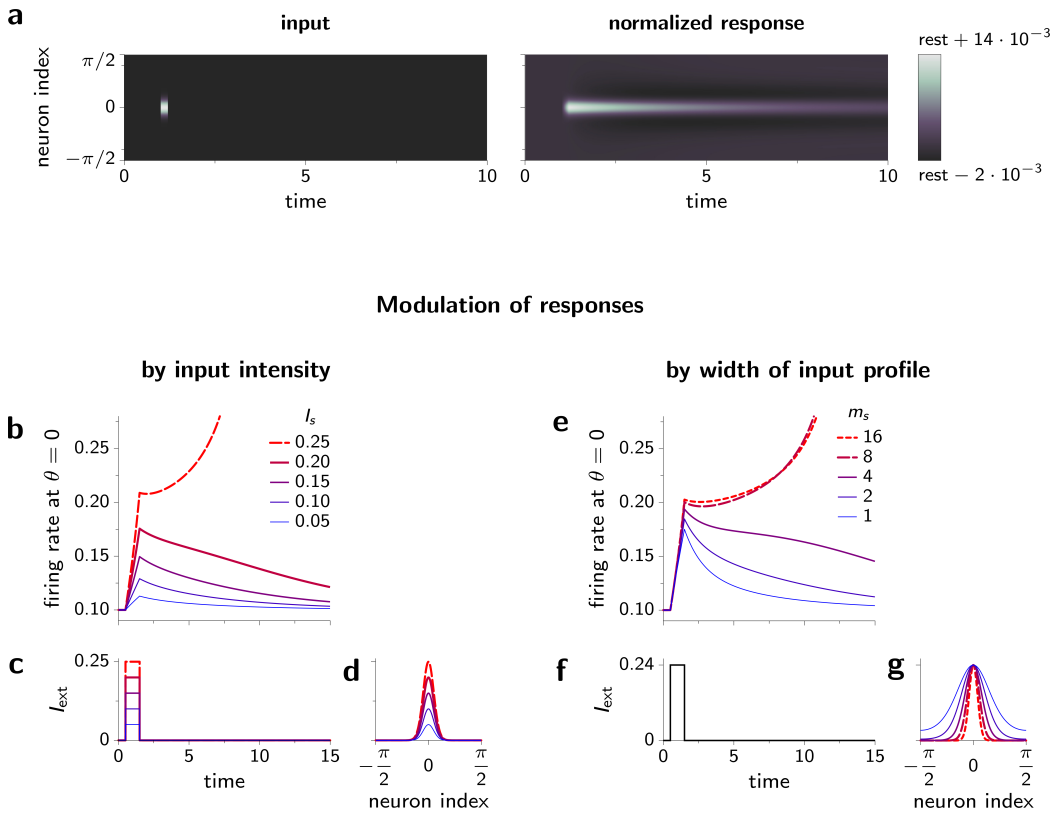

Figure 5. Network response to a brief input (**a**) Time course of the input (left) and the corresponding network response (right). Notice that the colormap spans a narrow range around the rest value $R = 0.1$. The pulse has intensity $I_s = 0.2$ and duration $d = 0.1$ (**b**) Temporal evolution of the peak of the activity profile for different values of I_s, indicated by the key. (**c**) temporal course and (**d**) spatial profile of the input pulse. In all plots solid and dashed curves represent respectively sub- and supra-threshold responses. Input parameters: $d = 1$, $m_s = 10$. Panels (**e**)–(**g**) are analogous to panels (b)–(d), except that the intensity fixed at $I_s = 0.24$ and curves correspond to different spreads of the input profile, parameterized by m_s. Network parameters as in Fig. 3.

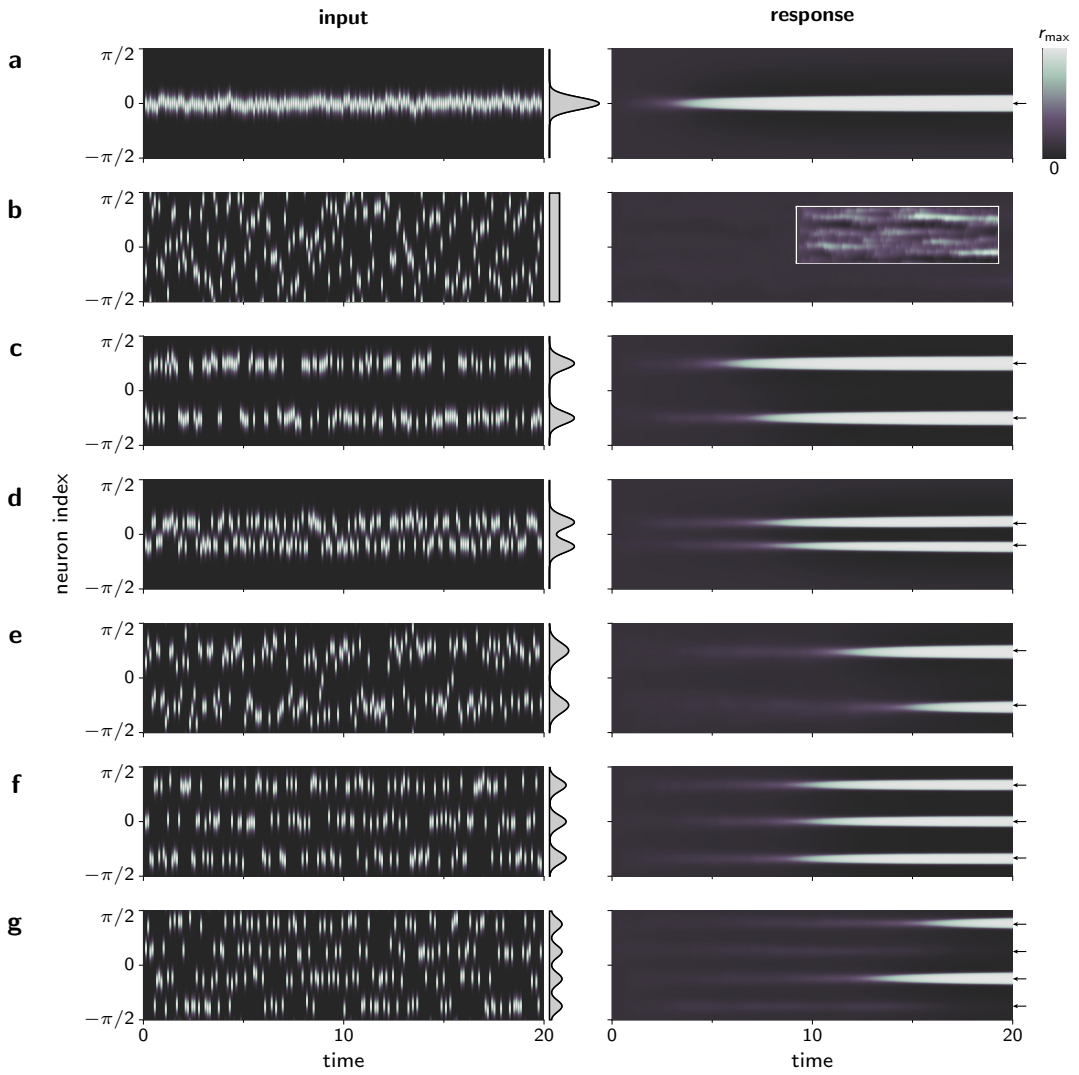

Figure 6. Activity of the multistable network in response to input patterns with different feature distributions. **(a)** Input protocol (left) and network response (right) shown in a color-coded map spanning the neurons' dynamic range. Inputs consist of a sequence of pulses whose location is a von Mises variate with parameters $\mu = 0$ and $m_c = 50$. The schematic next to the input panel shows the average input profile. The arrow at the right of the response panel indicates the value of μ. (**b**–**e**) Same as (a), for different distributions $P(\Theta)$. **(b)** Pulses distributed uniformly over all orientations. No response is observed using the original colormap. The inset in the response panel shows the evoked response in a colormap spanning the narrower range $[0.1, 0.125]r_{\max}$. **(c)** Pulses distributed bimodally with components that are narrow and far apart ($m_c = 50$ for both peaks, $\mu = \pm\pi/4$). **(d)** As in (c), but with peaks lying closer to each other ($\mu = \pm\pi/10$). **(e)** As in (c), but with a broader distribution per component ($m_c = 5$). **(f)** As in (c), but with three components centered at $\mu = \pm\pi/3, 0$. **(g)** As in (c), with four components centered at $\mu = \pm 3/8, \pm\pi/8$. In all panels single pulses have a duration of $d = 0.2$ time units, and are separated by a inter-pulse gap $d_0 = 0.01$ time units. Each pulse has a circular Gaussian profile with $m_s = 10$ and $I_s = 0.2$ (see Eq. (6)). Network parameters as in Fig. 3.

the time-independent input profile

$$\langle I_{\text{stim}}(\theta)\rangle_\Theta = \nu\langle E(\theta-\Theta)\rangle_\Theta \int_{-\infty}^{\infty} \alpha(t)\,\mathrm{d}t\,. \tag{11}$$

Here $\langle\cdot\rangle_\Theta$ and denotes the average over the distribution of features $P(\Theta)$, and ν is the frequency of pulse presentation, given in our particular stimulation protocol by $\nu = 1/(d+d_0)$. The factor $\langle E(\theta-\Theta)\rangle_\Theta$ is the average size of the input kicks that neuron θ receives every time an item is presented. This kick size is a random variable because it depends on the location Θ of the pulse, which is itself a random variable. Finally, the integral factor in Eq. (11) quantifies the overall contribution of a single pulse, which for our choice of $\alpha(t)$ (Eq. (9)) evaluates to d.

The shape of the average input profile is determined by $\langle E(\theta-\Theta)\rangle_\Theta$, defined by

$$\langle E(\theta-\Theta)\rangle_\Theta \equiv \int_{-\pi/2}^{\pi/2} E(\theta-\Theta)\,P(\Theta)\,\mathrm{d}\Theta. \tag{12}$$

This is the convolution of the feature distribution with the input profile. It follows from the properties of the convolution that the peaks of the input profile $\langle E(\theta-\Theta)\rangle_\Theta$ are located at the modes of the distribution $P(\Theta)$. In other words, the average input is highest at the category centers. This results in an inhomogeneous feed-forward drive that is amplified by the recurrent network and that, if the modulation is pronounced enough, triggers the formation of bumps at the category prototypes.

Another consequence of Eq. (12) is that the peaks seen in the average input profile are narrower the less variability there is within a category (see the average input profiles shown between the input and response panels in Figs. 6, specially those in panels c and e). We discuss below the implications of this sharpening of the input peaks.

At early stimulation stages the network is essentially feed-forward-driven.

How does the network respond to the constant profile specified by Eq. (11)? At this point it is important to realize that at the first stages of stimulation the network responds as if there were no lateral connections between neurons and if currents were transformed linearly into rates (i.e., as if $J(\theta) = 0$ and $\Phi(x) \sim x$ in the dynamical Eq. (3)). This is because at early stimulation stages the network is still close to the baseline state $r(\theta,t) \simeq R$ and has a nearly flat profile. The recurrent input has a result a nearly flat profile, given by $J * s(\theta) \simeq J_0 R$, where the asterisk denotes angular convolution. When the overall amount of excitation is balanced with that of inhibition, $J_0 = 0$, and recurrent inputs vanish (see also Fig. 3g, leftmost panel). The network is at that point purely driven by feed-forward inputs. On the other hand, weak stimulation together with the absence of recurrent inputs cause neuronal activities to confine to a low and narrow dynamical range. This allows us to approximate $\Phi(x)$ by the same linear function for all neuron indices. In this approximation, the dynamical Eq. (3) becomes a linear differential equation

$$\begin{aligned}\frac{\partial}{\partial t}s(\theta,t) &\simeq -s(\theta,t) + \Phi\big(0 + I + \langle I_{\text{stim}}(\theta)\rangle_\Theta\big),\\ &\simeq -s(\theta,t) + \Phi(I) + \Phi'(I)\langle I_{\text{stim}}(\theta)\rangle_\Theta,\end{aligned}$$

where Φ' denotes the derivative of Φ. The activity profile $s(\theta,t)$ converges to

$$\bar{r}(\theta) = R + \Phi'(I)\nu d\langle E(\theta-\Theta)\rangle_\Theta, \tag{13}$$

where we have used $\Phi(I) = R$ (Eq. 7) and Eq. (11). The factor $\nu d = d/(d_0+d)$ represents the fraction of time over which the network is being activated by the inputs, and the derivative $\Phi'(I)$ is a gain factor. Equation (13) states that, in the regime where the linear approximation is valid, the activity profile mimics the profile of the input except for the offset R and the multiplicative factor $\Phi'(I)\nu d$.

If the external inputs drive the system to a configuration $\bar{r}(\theta)$ that is sufficiently modulated, recurrent inputs will start dominating and localized positive feedback will arise, eventually rendering the above approximation invalid and causing the network to relax to a modulated state.

The time of category formation reflects the statistical structure of inputs

The extraction of prototypes takes longer the more categories are included in the input stream. This can be seen comparing Figs. 6a,c,f,g, which show the responses to four input streams that differ only in the number M of categories they contain. The increase of formation time with M is a consequence of the constraint that only one pulse is presented at a time. Since the overall presentation rate $\nu = 1/(d+d_0)$ is held fixed, the presentation rate for items of a particular category necessarily scales down by a factor $1/M$. This results in a decreased average intensity $\langle E(\theta-\Theta)\rangle_\Theta$ at the category centers, and therefore in an increase in the formation time of bumps.

Another factor that delays, or even prevents, the formation of bumps is the proximity between categories in feature space (Fig. 6d). Categories that are close to one another give rise to competitive interactions by virtue of surround inhibition, which counterbalances the drive from stimulation. This suppressive effect is specially pronounced when the distance between category centers is around the spatial scale of surround inhibition. In this case the total average input at the category centers is weaker than it would be if centers were far apart, becoming less effective at triggering bumps. Competitive interactions also impair the extraction of prototypes when categories are close (Fig. 6d, right; note the slight mismatch between arrows and bump locations). In this case, the location of the peaks in the net input profile do not coincide with the location of the peaks in the external drive, but they are slightly shifted away from each other due to surround inhibition. As a result, bumps emerge slightly away from the real category centers. When the centers are even closer, the network behaves as a winner-take-all system if the two competing categories are equally likely (not shown). The bump will form at either of the two category centers with

equal probability. The final outcome will depend on which category happens to be activated by the particular realization of the stimulation protocol.

Input statistics dictate the distribution of pulse intensities received by each neuron

The coarse representation of categories provided by the network does not allow for an explicit encoding of variances. Although the location of a bump clearly reflects the mean of a given category, the spread of instances around a category center cannot be captured by the bump profile, which is completely determined by the network's recurrent dynamics. Variances are however implicitly encoded in the time it takes to elicit a bump: categories displaying larger variability among its members take longer to form (see for instance Fig. 6e, and compare it to Fig. 6c). This increase in formation time is a consequence of the blunting of the average input profile that occurs when items are widely spread around prototypes.

To gain qualitative insight into how input statistics govern the input profile, we have studied the dependence between the variability within categories, or category spread, and the distribution of input intensities felt by each neuron.

When the network is stimulated with a sequential protocol, every neuron receives a kick every time a stimulus is presented. The size of this kick is a random variable because it depends on the exact value of the stimulus, which is itself a random variable. The distribution of kick sizes is typically non-Gaussian and depends on the preferred value of the neuron as well as on the stimulus statistics (Figs. 7a,b; see also the Appendix for its derivation). When the preferred feature value of a neuron lies close to a category center, the distribution of kick sizes received by the neuron will be skewed toward high values (Fig. 7a, right). The opposite happens for neurons with preferred values falling far from category centers (Fig. 7a, left). Neurons with preferred values lying neither too far nor too close to the category centers receive intensities that are distributed over the whole range of intensities. The shape of this distribution is strongly modulated by category spread (Fig. 7c).

Importantly, for neurons located near a category center, increasing the category spread decreases both the median and the mean of the intensity distribution (Fig. 7c). This is because as the spread within the category increases, so does the probability of receiving small kicks caused by instances located away from the prototype. Larger dispersion around prototypes also causes the mean input intensity (though not so much the median) to increase for neurons at the flanks of the category.

The dispersion of pulse intensities around their typical values also increases with category spread. This is illustrated in Fig. 7c, which shows that the interquartile ranges of the pulse distribution widen as the category spread increases. The effect is more pronounced at the category centers.

Input statistics determine the speed of category formation and accuracy of prototype extraction

The link between category spread and dispersion of pulse intensities suggests that categories with higher variability are expected not only to take longer to form on average, but also to display more variability in their formation times. To gain a quantitative feel of how the statistics of formation times vary with the variability within a category, we have simulated a set of sample trials using the input protocol shown in Fig. 6a, for different levels of category spread and different noise realizations. Figure 8a summarizes the estimates of the mean and standard deviation of the simulated formation times. Note that, as suggested above, the average time of bump formation gets longer and becomes more variable as the category spread σ_c increases.

Category spread also affects the accuracy with which prototypes are extracted (Fig. 8c). While the sample mean of the extracted prototypes coincides with the mean of the original distribution of features ($\mu = 0$), the variability of the extracted prototypes (as measured by the sample standard deviation), grows with the category spread (Fig. 8b).

Another factor that modulates the variability of both formation times and prototype extraction is the duration of single pulses. By shortening the duration d of each pulse, more items can be sampled per unit time, leading to more accurate estimates of the feature distribution. Better estimates of $P(\Theta)$ are reflected in turn in lesser variability in the input profiles and, therefore, in less variable formation times. This effect is shown in the two panels of Fig. 8(a),(b), which illustrate how the standard deviation of both formation times and final bump locations are reduced when the duration d is shortened from $d = 0.10$ to $d = 0.05$. Note that shortening d reduces the variability of the input profile, but it does not modify its average. This explains why mean formation times remain unaffected as d is varied (Fig. 8d).

To elucidate how pulse durations determine the variability in formation times and accuracies, we have run another set of trials using a fixed category spread and varying the pulse duration d. The dependence of formation time on d, shown in Fig. 8c, confirms the points highlighted above. First, mean formation times are independent of the pulse duration except for very short pulses, for which formation times are slightly longer. This mild dependence arises from the nonlinearity of the network. Second, formation times become more variable the longer the pulse duration. The increase of variability with pulse duration is a consequence of the effective reduction of sample size that brought about by lengthening d. The same trend is seen in the dependence on d of the extracted prototype value, given by the bump position (Fig. 8d).

Relation with kernel density estimation

The results presented above suggest that the network combines two basic operations. At first stages of the stimulation protocol, when the activity profile differs little from the homogeneous state and the effect of nonlinearities are negligible, the network operates as a running estimator of the feature distribution. More precisely, the presentation of an item

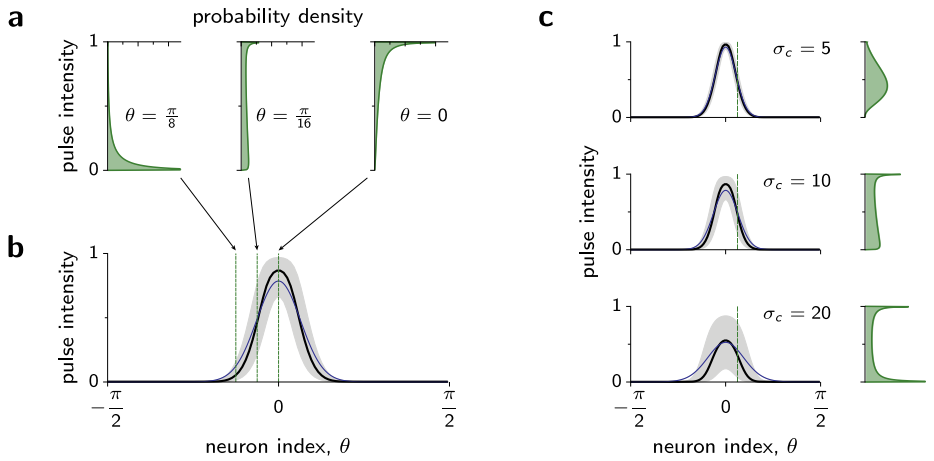

Figure 7. The distribution of pulse intensities varies across neurons (**a**) Distribution of pulse intensities for three representative neurons, when the input stream contains one single component with parameters $\mu = 0$ and $\sigma_c = 10$ ($m_c = 16.4$). (**b**) Mean (thick black curve), median (thin blue curve) and the inter-quartile range (shaded region) of the distribution of pulse intensities for each neuron. The input stream is the same as in (a). Here we show both the median and the mean as location parameters because the distribution of pulse intensities is highly non-Gaussian and does not have an unambiguous metric for typical values. (**c**) Statistical properties of the input profiles for three different degrees of intra-class variability, measured by σ_c. Conventions as in (b). As an illustration, we show next to each panel the distribution of pulse intensities at $\theta = \pi/16$ (indicated by a dashed green line in left panels). Input parameters as in Fig. 6.

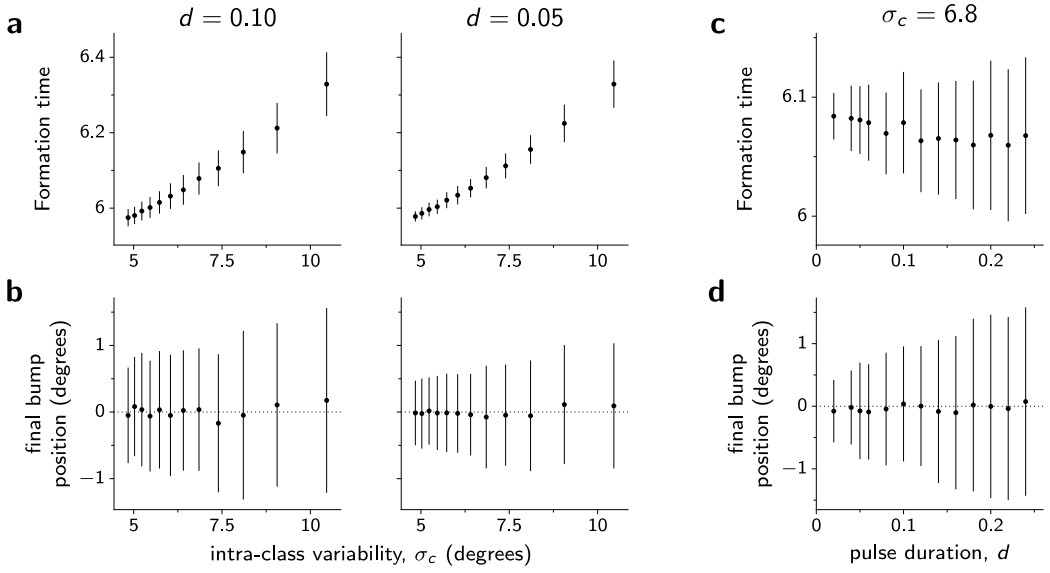

Figure 8. Speed and accuracy of category formation depends on the variability within a class and on pulse duration. (**a**) Time required to trigger a bump and (**b**) final peak location of the emerging bump as a function of σ_c. Left and right panels differ in the duration d of the pulses in the input stream. Dots indicate the sample mean of each data set, formed by $n = 200$ samples. Error bars represent the sample standard deviation. When d is varied, the inter-pulse gap d_0 is also varied to keep constant the activation rate, given by $\nu\, d = d/(d + d_0)$ (see Eq. 13). (**c**)–(**d**) Same as (a)–(b), but keeping variability within categories constant and using the duration of single pulses as independent variable. Network and input parameters as in Fig. 6.

with attribute Θ contributes to the initial activity profile with a term proportional to the pulse profile $E(\theta - \Theta)$. This contribution is approximately additive and wanes in the absence of subsequent stimulation. This can be seen from the approximate dynamical equation in this regime,

$$\frac{\partial}{\partial t} r(\theta, t) \simeq -r(\theta, t) + R + \Phi'(I) \left\{ \sum_{i=1}^{\infty} E(\theta - \Theta_i) \alpha(t - t_i) \right\},$$

where we use the explicit form of $I_{\text{stim}}(\theta, t)$ given in Eq. (8). This equation is essentially a low-pass filter transformation of the ongoing stream of inputs. At this stage the network performs a running kernel density estimation, with the input profiles of individual pulses, $E(\theta - \Theta_i)$, playing the role of Parzen windows, or kernel functions (Izenman, 1991; Duda et al, 2000). The low-pass filtering of inputs carried out by the network, on the other hand, keeps the stimulation values as activity traces and provides an effective time window over which inputs are to be integrated. Although the operation is not a proper probability density estimation because it does not include any normalization step, it does well in capturing the overall statistical structure of the inputs, like estimating the presence and the location of modes.

In a second stage of the stimulation process, when the network activity profile becomes sufficiently modulated, nonlinearities come into play and recurrent inputs dominate over external inputs. The modulations in the activity profile triggered by stimulation are amplified by the recurrent circuit, and localized patterns of activity emerge, thus signaling the presence of a particular cluster of features—a category.

5 Discussion

In this work we have proposed a simplified network model of fast category formation inspired by the framework of continuous attractor networks. The model can identify the categories contained in a stream of inputs without modifying the synaptic strengths between neurons. Category prototypes are encoded as self-sustaining, localized activity patterns of feature-selective neurons, with highest-peaks of activation indicating the location of the prototypical value of each category. These category representations arise from the summation of the activity traces elicited by the presentation of single items, each of which elicits a weak, transient response proportional to a smooth tuning curve centered at the stimulus' feature value. By summing the responses to a sufficient number of sample stimuli, the network's activity profile provides a running estimate of the feature distribution. Input streams that are structured in clusters of features give rise to multimodal activity profiles, whose peaks sustain themselves through reverbatory activity. These self-sustaining activity patterns persist after the stimulus train is terminated, and may drift slowly toward a periodic pattern (the truly stable activity patterns of the system).

Relation to previous work

Models based on continuous attractor networks have been invoked to account of the representation of analog variables in the brain. Examples include the tuning of neurons in the primary visual cortex (Somers et al, 1995; Ben-Yishai et al, 1995), the persistent activity in prefrontal and parietal cortices during spatial working memory tasks (Camperi and Wang, 1998; Compte et al, 2000; Laing and Chow, 2001; Gutkin et al, 2001; Fall et al, 2005), or the responses of head-direction cells (Skaggs et al, 1995; Zhang, 1996; Redish et al, 1996; Boucheny et al, 2005; Song and Wang, 2005), eye-position cells (Seung, 1996; Seung et al, 2000), and hippocampal place cells (Tsodyks and Sejnowski, 1995; Samsonovich and McNaughton, 1997; Battaglia and Treves, 1998) and grid cells (Burak and Fiete, 2009). In these works, stable neural representations of the relevant analog quantity are provided by persistent, localized patterns of activity that are mediated by effective local excitation and long-range inhibition. Such a connectivity pattern is known to lead to spatially periodic (Ermentrout and Cowan, 1980; Tass, 1995; Bressloff et al, 2001) and localized (Kishimoto and Amari, 1979; Laing et al, 2002) patterns of activity—in particular, patterns with multiple bumps.

Laing and Troy (2003) thoroughly studied the conditions that the connectivity has to satisfy to sustain two bumps, using a Heaviside transfer function, a difference of exponentials as connectivity kernel (or Wizard-had function), and a voltage-based description of neuronal activity. These results were confirmed and extended by Coombes et al (2003), who used a more realistic sigmoid transfer function and incorporated the effects of axonal signal propagation. We have provided an alternative analysis for a similar, simpler system, using instead a difference of circular Gaussians (von Mises functions) as connectivity kernel, and a ring network topology rather than the real line. Our analysis is consistent with previous findings, and reveals the regions in parameter space where patterns with multiple bumps appear and coexist with one another, a regime in which our network operates.

The relation between continuous attractor models and clustering has been recently investigated by Jin et al (2011). They showed that the time evolution of the Amari model defines a hierarchical clustering sequence, where different clusters of activity are progressively merged and diffused until the system settles into a final steady state. The model, together with algorithm defining the clustering sequence, is shown to highlight clusters of arbitrary shape, and to do so in a robust way. Despite the overall similarities with our work, and despite the shared underlying idea that similarity relations can be mapped to cooperative or competitive interactions, the work by Jin et al uses static, two-dimensional data, and a simpler network model based on a Heaviside transfer function, and places the emphasis on the numerical analysis of complex data structures rather than on the biological mechanisms of fast categorization.

Multistability among modulated patterns

The coexistence of selective activity patterns has been proposed as a neural correlate of multi-item memory, both in the context of discrete items (Amit et al, 2003; Amit and Mongillo, 2003) as for items characterized by a continuous

parameter (Edin et al, 2009). In this work we have hypothesized that such a coexistence may also underlie a coarse, transient encoding of probability distributions involved in grouping by similarity, in situations where items are defined in a continuous space. Specifically, for some parameter configurations the network can sustain modulated states with one or more peaks in feature space. The truly stable states of the network are patterns with evenly spaced bumps, but the network can also accommodate slow transients consisting of patterns with several bumps at arbitrary locations. These transient patterns slowly arrange into a pattern with evenly spaced peaks, with a timescale that is vastly longer than the timescale of neurons.

Stable states can be divided into different families according to the number of peaks in the activity profile, with each family being formed by a continuum set of modulated patterns related to one another by arbitrary phase shifts. All these properties, together with the existence of slow transients, confer the network with a flexible set of basis activity patterns with which to represent a wide range of distributions. We reiterate that our interest here is less in the long-term, steady properties of bumps but rather in their properties at emergence—the pseudo-steady states, or slow transients. The conditions for existence of true steady states provide us with the parameter values that not only maximize the number of categories, but also allow for the coexistence between different number of categories.

The mathematical analysis carried out in this work reveals the conditions necessary for the coexistence of (evenly-spaced) multimodal and homogeneous stable states. Multistability can be robustly achieved by imposing a few physiologically plausible properties on the connectivity profile and the firing activities: (i) rough balance between the overall amount of excitation and inhibition; (ii) Mexican-hat connectivity profile with a narrow spread of excitation and a wider (though not global) spread of inhibition; and (iii) low firing rate in the spontaneous state—more precisely, the net input received by each neuron in the spontaneous state should fall in the subthreshold regime of its input-output curve.

These conditions enjoy a variable degree of experimental support. The scenario where excitation is balanced with inhibition has long been favored by theoretical studies (Gerstein and Mandelbrot, 1964; Amit and Brunel, 1997; Shadlen and Newsome, 1998; van Vreeswijk and Sompolinsky, 1998; Renart et al, 2010) and has been recently supported by *in vivo* experiments (Shu et al, 2003; Wehr and Zador, 2003; Haider et al, 2006; Higley and Contreras, 2006; Okun and Lampl, 2008). Likewise, the claim that neurons operate at a subthreshold regime during spontaneous activity is supported by the observation of low and highly irregular neuronal discharge in the absence of stimulation (Shadlen and Newsome, 1998; Softky and Koch, 1993), which suggests that the firing of neurons is essentially driven by noise. As it turns out, noise-driven activity arises naturally in scenarios where excitation and inhibition are balanced at the neuronal level (Gerstein and Mandelbrot, 1964; Amit and Brunel, 1997; Shadlen and Newsome, 1998)

The plausibility of Mexican-hat connectivity profiles in high-level cortices is rather speculative and arguable. While high-level cortices have been known to be implicated in the categorical and invariant representations of sensory objects (at least in vision studies; see, e.g., Maunsell and Newsome, 1987; Logothetis and Sheinberg, 1996; Freedman et al, 2001), neurons in these areas are often tuned for multiple objects and features, and show complex tuning properties (Desimone et al, 1984; Kreiman et al, 2006; Rust and DiCarlo, 2010). It is therefore difficult to assess the selectivity properties of neurons in these areas, let alone determine how synaptic interactions are shaped by neuronal tuning. Given this lack of experimental constraints, we have speculated that the interaction between any two neurons in high-level areas is structured in a similar way as in the primary visual cortex, where synaptic strengths depend on distances of preferred orientation values (Gilbert and Wiesel, 1989; Malach et al, 1993). We remark that such a synaptic structure arises dynamically from an unstructured network if synapses change according to Hebbian learning prescription and if neurons have overlapping tuning curves. In that case, neurons responding to similar stimuli will end up connected with potentiated synapses, while neurons with non-overlapping tuning curves (i.e., responding to dissimilar stimuli) will see the connections between them depressed.

Subthreshold stimulation

We have assumed that inputs are subthreshold. In this way, the activity traces induced by single-object stimulation sum up and provide an estimate of the distribution of features in the input stream. Subthreshold stimulation is indirectly supported by the finding that higher level areas accumulate information over time and, therefore, are weakly driven by the instantaneous features of particular stimuli (Hasson et al, 2008). Even if the need for sufficiently weak inputs may seem an ad-hoc requirement, one could think of plausible modulatory signals controlling the magnitude of inputs depending on ongoing demands. These signals would control the contribution of single tokens to the network activity, reducing it when an accurate estimate of the input distribution is needed, at the cost of longer processing time, or, conversely, increasing it when a coarse and fast estimate is preferable. This modulation would therefore control the trade-off between speed and accuracy, a well-known constraint that arises in tasks requiring some sort of decision making (see, e.g., Gold and Shadlen, 2007).

Another way to imagine an input being subthreshold is that the perception of an individual token, say for feature extraction, is challenged. If the token stream were embedded in a noisy background then each presentation would have some perceptual uncertainty associated with it, due to the activation of sensory neurons responding to noise rather to the signal. The net effect of adding noise is to broaden the neuronal representation of the stimulus. In our model, a broad sensory representation corresponds to a blunt input profile, which cannot elicit a bump by itself and is therefore subthreshold by definition. In general, inputs can become subthreshold with the addition of a sufficient amount of noise.

The requirement that inputs be subthreshold has an analogous counterpart in plasticity-based network models of categorization based on realistic synapses (Brunel et al, 1998; Rosenthal et al, 2001). In these models, the statistical structure of stimuli can be captured by the synaptic matrix when learning is sufficiently slow, in the sense that the presentation of a single stimulus induces, on average, small synaptic changes. Small synaptic modifications ensure that the synaptic matrix changes appreciably only after a reasonable number of stimuli have been sampled, forcing it to reflect the average statistical properties of the stimulus stream, and making it less sensitive on the particular order of presentation.

Limitations and further directions

We have assumed that the network is homogeneous in its neuronal and synaptic parameters. Making these parameters heterogeneous disrupts the stability of the continuous set of modulated states. Bumps could no longer persist at arbitrary locations in the space of preferred features, but they would slowly drift toward particular hot spots where the local excitability is higher than average (Zhang, 1996; Koulakov et al, 2002; Renart et al, 2003). This would reduce the number of stable states from a continuum to a small number of discrete attractors, thereby compromising the network's ability to indefinitely maintain the value of a continuous variable. Category formation is, however, more robust to the impact of heterogeneities because the timescales involved in the formation of bumps are considerably shorter than the timescales of bump drift. Such a disparity in timescales renders the formation of perceptual categories insensitive to the presence of inhomogeneities in the network.

On the other hand, let's consider the longer term effects of heterogeneity from a positive viewpoint, say as a stabilizing mechanism for 'remembering' a category, for 'pinning' its location. This can be achieved with a slow, activity-dependent positive feedback mechanism like, e.g., short-term synaptic facilitation (Tsodyks et al, 1998; Mongillo et al, 2008), whereby synaptic interactions are temporarilly facilitated between cells discharging at high rates. Such a mechanism increases the local excitability of neurons at the bump location, thus preventing the bump from drifting and, therefore, effectively pinning it at the original location (Itskov et al, 2011). This could in principle hold or pin a bump in place even after the stimulus sequence is terminated or if the frequency of tokens for the corresponding category were greatly reduced. Even if the bump itself disappears (say, by a clearance or resetting mechanism), there remains a memory trace in the form of facilitated synapses (Mongillo et al, 2008) that prime new stimuli according to previously formed categories. These notions could be likewise implemented if we wished to develop a secondary phase to categorization, say, 'classification'—the identification of subsequent tokens as members of an existing and pinned category. An alternative scenario for the short-term memory of categories would be NMDA-mediated recurrent excitation. The slow timescale of decay for NMDA synaptic excitation, together with the voltage-gating of NMDA receptors, would greatly slow the tendency for bumps to drift (Compte et al, 2000).

While the regenerative aspect of bump emergence is important for the selectivity of category formation, and possible pinning, the persistence of activity patterns is by itself not fundamental for the mechanism of category formation discussed here. The mechanism that singles out a category is the emergence of slowly decaying, structured pattern of activity. Whether the emerging pattern is persistent or not is inessential for the signaling of categorical information, which should reflect the ongoing statistical structure of the input stream rather than maintaining it. In this view, the formation of a modulated activity pattern would signal to higher brain areas the presence of structured ensembles of low-level signals.

Testable predictions

The model makes some predictions that can be tested in behavioral experiments. One could devise a task in which a subject is presented with a rapid sequence of simple stimuli characterized by some continuous feature. For example, the subject might view a sequence of low-contrast dots appearing at random angles and constant excentricity, while fixating at a central spot. Dots would appear one at a time, at angular locations drawn from some distribution that changes in each trial to vary, also randomly, the number and the location of its modes. The subject would then be asked, after the sequence presentation, to determine the number of classes (clusters of features) and to report the corresponding prototypical locations. The sequence would have a finite total duration, but it would be long enough to include a sufficient number of samples, so that it is a priori possible to get a reasonable estimation of the feature distribution. The background could be noisy to guarantee that the presentation of each token is not sufficient to trigger a response—i.e., to ensure subthreshold inputs.

According to the model, the subject will on average need longer presentation times the more categories are present, given a fixed accuracy threshold. Another related prediction is that, also for a fixed accuracy level, the subject will need longer presentation times as the intra-class variability increases (Fig. 8). In both cases the model accounts for these dependences in terms of a decrease in the modulation of the average input profiles, which leads to a slower recruitment of selective activity and hence a longer time for category extraction (Fig. 7). Moreover, one expects a degradation of the accuracy of prototype extraction as the intra-class variability increases, because categories that are more spread give rise to less reliable distribution estimates.

The similarity between classes is also expected to affect accuracies and processing times. Categories that are sufficiently close in feature space will take longer to extract than categories that are far apart. One also expects that for two classes lying sufficiently close, the estimates of the prototypes will be biased away from the real prototypes (repulsion effect). Both predictions stem from the effect of surround inhibition exerted by neighboring bumps. Finally, the model predicts that only one category representation arises when classes are too close; in that case, the subject is expected to report the

presence of one single category with prototypical value lying between the true prototypical values (any of the two original prototypes, or the average of both).

Analogous predictions have been formulated by Wang and collaborators in the context of multi-choice decision making in a random-dot discrimination task (Furman and Wang, 2008; Liu and Wang, 2008). In their model the synaptic footprint is wide enough to guarantee the stability of only unimodal activity patterns. Thus, while their stimulation protocol consists of a steady, multipeaked, and suprathreshold input profile, the network activity decays always to a pattern with one single bump. This is in contrast to our model, where the connectivity footprint is sufficiently narrow to allow for the coexistence of several steady, or quasi-steady, bumps. The fact that architecture is essentially the same in both cases suggests this type of networks may constitute a computational building block used in different areas for different purposes. Sensory areas may discriminate signals with such an architecture, using strong inputs and wide synaptic footprints. The same architecture, but with narrow footprints and subthreshold stimulation, can be used in higher areas to encode statistical information.

Appendix

Analysis of stability

The stability analysis of an extended system like ours proceeds as follows. We first assume that in the absence of stimulation there is a stationary homogeneous solution. This is a reasonable assumption to make when the connectivity footprint is weak and not heavily modulated. We then add a small and smooth perturbation to the homogeneous solution and study its temporal evolution using linear stability analysis, as it would be normally done in a finite-dimensional system (Hirsch and Smale, 1974; Strogatz, 1994). Since our dynamical variables are not finite-dimensional, but defined on the infinite-dimensional space of periodic functions, perturbations are functions as well. As such, they can be decomposed in a sum of spatial Fourier modes, and the stability of each mode can be analyzed separately. By studying which mode gets destabilized first when we vary some connectivity parameter, we can determine the pattern of activity that will emerge and therefore the activity profiles that we expect to find in one particular region of parameter space.

Linear stability We study the temporal dependence of a small perturbation around the homogeneous state, $r(\theta,t) = R + \rho(\theta,t)$, with $|\rho(\theta,t)| \ll R$. The perturbation is a real function, and is periodic under the transformation $\theta \to \theta + \pi$. It can thus be expanded in a Fourier series of the form

$$\rho(\theta,t) = \rho_0 + \left\{ \sum_{k=1}^{\infty} \rho_k(t) \mathrm{e}^{\mathrm{i}2k\theta} \right\} + \text{c.c.}, \qquad (14)$$

where $\rho_k(t)$ the k-th Fourier coefficient of the perturbation. The symbol c.c. means that the Fourier series also includes the complex conjugate of all the terms inside the curly brackets.

The time evolution of the perturbation is obtained by plugging the perturbed equation $R + \rho(\theta,t)$ into the dynamical equation (3), Taylor-expanding $\Phi(\cdot)$ around the stationary solution, and using the Fourier expansion in Eq. (14). We obtain an infinite set of decoupled differential equations, each of which describes the time evolution of one particular mode. To linear order, these equations read

$$\frac{\mathrm{d}\rho_k}{\mathrm{d}t} = -\rho_k + \Phi' \, J_k \rho_k, \qquad (15)$$

where Φ' is the derivative of Φ, evaluated at the fixed point given by the solution of Eq. (7). The factor J_k is the k-th Fourier coefficient of the connectivity profile,

$$J_k = \frac{1}{\pi} \int_{-\pi/2}^{+\pi/2} J(\theta) \mathrm{e}^{-\mathrm{i}2k\theta} \, \mathrm{d}\theta = \frac{1}{\pi} \int_{-\pi/2}^{+\pi/2} J(\theta) \cos(2k\theta) \, \mathrm{d}\theta,$$

where in the second equality we have used the fact that $J(\theta)$ is a real and even function. Note that the coefficients J_k are real and that $J_k = J_{-k}$.

The differential equation (15) is linear in ρ_k and has therefore a solution proportional to $\mathrm{e}^{\lambda_k t}$ with

$$\lambda_k = -1 + \Phi' J_k, \quad k = 1, 2, \ldots \qquad (16)$$

This is the dispersion relation, which relates the rate of growth of the perturbation with its wavenumber k. For the homogeneous solution to be stable, all the eigenvalues in Eq. (16) should be negative, and thus the Fourier coefficients should obey $J_k < 1/\Phi'$ for all k. The uniform state destabilizes when at least one of the eigenvalues crosses the 0 value. If there is only one such eigenvalue, the pattern of activity that will emerge will show a modulation in θ dictated by the critical wavenumber k^c satisfying $\lambda_{k^c} = 0$.

Critical curves The region of stability of the homogeneous state is defined by the set of all connectivity parameters for which $\lambda_k < 0$, for all wavenumbers k. We can delimit this region by calculating the critical curve of each mode, defined as the parameter values for which $\lambda_k = 0$, i.e., as those satisfying the criticality condition $J_k = 1/\Phi'$. In the following we study in detail the critical condition for our particular model.

First, note that the derivative Φ' is a constant that does not depend of the connectivity parameters. This is because the external input I is adjusted so that the homogeneous rate takes a fixed value independently of the connectivity parameters. When Φ is a sigmoid, $\Phi'(x) = \beta\Phi(x)(1-\Phi(x))$, which takes the value $\beta R(1-R)$ at the fixed point. The critical curve for the k-th mode is then given by $J_k = 1/(\beta R(1-R))$.

Second, for our particular choice for the connectivity kernel, given by Eq. (1), the Fourier coefficients read

$$J_k = j_E \frac{I_k(m_E)}{I_0(m_E)} - j_I \frac{I_k(m_I)}{I_0(m_I)}, \qquad (17)$$

with $I_k(m)$ being the k-th order modified Bessel function of the first kind

$$I_k(m) = \frac{1}{\pi} \int_0^{\pi} \exp(m \cos\theta) \cos(k\theta) \, \mathrm{d}\theta.$$

Notice that in Eq. (17), the functions $I_k(m)$ enter as quotients of the form $I_k(m)/I_0(m)$. These quotients take values

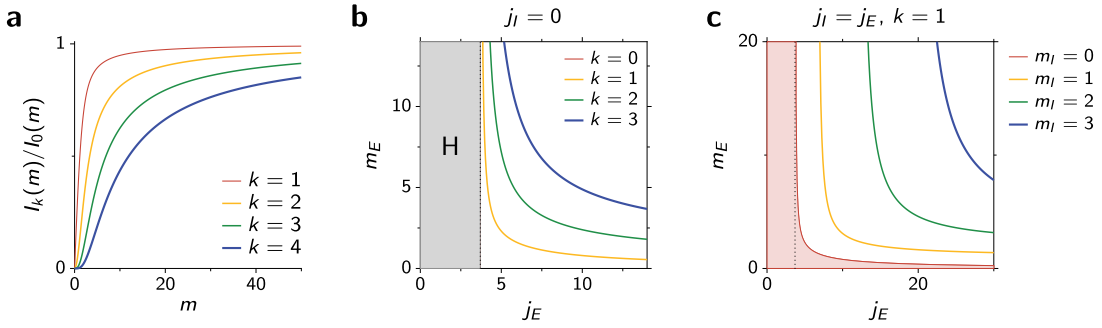

Figure 9. Stability of the homogeneous state. (**a**) Dependence of the quotient $I_k(m)/I_0(m)$ on m, for different values of k. **(b)** Critical curves in the plane (j_E, m_E), in the absence of inhibition ($j_I = 0$) and for four different modes, indicated in the key. The gray area is the region where the homogeneous state (H) is stable. The dotted vertical line is $j_E = 1/\Phi'$. **(c)** Same as (b), but here excitation is balanced with inhibition, and the different critical curves correspond to the wavenumber $k = 1$ for different levels of inhibitory modulation. The shaded area in red is the region where $\lambda_1 < 0$ when the inhibition is global ($m_I = 0$). For clarity the regions $\lambda_1 < 0$ for the remaining three values of m_I are not shown.

between 0 and 1 due to the properties of modified Bessel functions of the first kind, which satisfy $I_0(m) > I_1(m) > I_2(m) > \cdots$ for any m (Fig. 9a). The Fourier coefficients J_k are therefore of the order of the excitatory and inhibitory weights, j_E and j_I. Also relevant is the property that $I_k(m)/I_l(m)$ tends to 1 as m increases for any given k and l. Note also that $I_k(0) = 0$ for any $k > 0$, while $I_0(0) = 1$.

Inhibition prevents all neurons from saturating

We start considering the case where there is no inhibition present in the system, by setting $j_I = 0$. The values of the parameters j_E and m_E for which the homogeneous state is stable are represented as a shaded area in Fig. 9b. Unsurprisingly, the homogeneous state is stable as long as the strength of excitation j_E is sufficiently low. Recurrent excitation is in that case insufficient to disrupt the stability of homogeneous state, regardless of how narrow the excitatory footprint is. The diagram also shows that without inhibition the uniform state will always destabilize through the 0-th mode—there is no way to leave the stability region (shown in gray in Fig. 9b) other than by crossing the critical curve associated with $k = 0$. This implies that the activity pattern that arises when the homogeneous state destabilizes is flat, and consists of all neuron firing at saturated rates.

Saturation can be prevented with inhibition. In the model inhibition can be added parametrically by increasing the value of the overall inhibitory weight j_I. We see from Eq. (17) that increasing j_I reduces the value of all the J_k and, therefore, decreases the associated eigenvalues λ_k (Eq. (16)). This is clearly true for the 0-th mode, whose coefficient $J_0 = j_E - j_I$ does not depend on the parameters m_E and m_I. If we impose in particular a balance condition between excitation and inhibition by setting $j_E = j_I$, then $J_0 = 0$, which implies $\lambda_0 = -1$. That is, the zero mode is always stable. In such case the homogeneous state will always destabilize through some mode $k > 0$, and the emerging activity profile will no longer be flat but modulated in θ. This is the reason why we set $j_E = j_I$ in the article.

Narrowing the spread of inhibition stabilizes the homogeneous state

Having fixed the relation between excitatory and inhibitory couplings, the relevant parameters that determine which patterns will emerge are the widths of the excitatory and inhibitory footprints, m_E and m_I. When excitation and inhibition are balanced, narrowing down the range of inhibition helps stabilize the homogeneous state. This is a byproduct of the balance constraint: as m_I approaches the value of m_E the excitatory and inhibitory profiles become more similar and eventually cancel each other, leading to a vanishing connectivity profile and therefore to the suppression of recurrent feedback.

The effect of narrowing the inhibition range can be seen in Fig. 9c, which shows the critical curves of the first mode in the plane (j_E, m_E) for different values of m_I. The critical curves move further away from the origin as m_I gets larger, thereby broadening the region of stability of the homogeneous state. Although not shown in the figure, the same trend occurs for modes higher than $k = 1$. Figures 9b,c also illustrate the fact that the homogeneous state can always be destabilized by increasing synaptic strengths sufficiently enough.

Selection of patterns by the spreads of excitation and inhibition

The spread of inhibition determines the pattern that appears when the uniform state destabilizes. When inhibition is global ($m_I = 0$) and sufficiently strong, the first Fourier mode is always the first to become unstable. This is because, when $m_I = 0$, $J_1 > J_2 > \cdots$, as one can see from Eq. (17) and from Fig. 9a. On the other hand, J_0 can be always set to an arbitrary small or negative number by increasing the inhibition weight j_I, so that $J_0 < J_1$. These two properties explain why only single bump states appear in the original models of orientation tuning and working memory, where inhibition is strong and global (Ermentrout, 1998).

When inhibition is not global but modulated, patterns with more than one bump can arise. If we vary smoothly the connectivity parameters until the homogeneous state desta-

bilizes, the number of bumps of the emerging pattern will be given by the mode that destabilizes first. This is illustrated in Fig. 10. The region of stability of the homogeneous state is shown in gray in Figs. 10a,b, which are like Fig. 9b,c, except that the plane is now spanned by m_E and m_I, and that synaptic strengths $j_E = j_I$ are fixed. Two aspects are noteworthy. First, the region of stability of the homogeneous state shrinks as the synaptic strength increases. Second, the stability region is bounded by different critical curves, meaning that the destabilization can occur through modes with wavenumber $k > 1$ and can therefore give rise to multipeaked modulated patterns.

We have confirmed that this is indeed the case by simulating the network activity for several connectivity configurations that fall slightly off the stability region of the homogeneous state. We have picked two such configurations, one destabilizing the first mode only (diamond in Fig. 10a) and another destabilizing the fourth mode only (square, in the same figure). The corresponding connectivity profiles and the eigenvalues of each mode are shown in Fig. 10c and d, respectively, while the resulting simulated activity is summarized in Fig. 10e. As expected, the wavenumber of the unstable mode determines the number of peaks in the emerging pattern.

Effect of nonlinearities and multistability

The analysis of the previous section shows that a modulated pattern of activity arises when one crosses a critical curve in parameter space. More generally, it demonstrates the ability of the network to sustain modulated activity patterns, which provide the basis for the neural representation of categories. Yet the examples of pattern formation that we have considered so far, although illustrative for the analysis of stability, are unrealistic because parameters were finely tuned to lie near a bifurcation point. This special choice explains why modulations were weak and had well-defined wavenumbers (see Fig. 10e). In general the emerging patterns will become more complex as one goes further away from the instability curves, with nonlinear effects taking over and with other modes possibly getting destabilized. Although a full analytical treatment of the system becomes prohibitive in this nonlinear regime, we can use approximations to gain some insight into the behavior of the network in more general configurations.

In our categorization model, it is particularly important to determine the conditions for the coexistence of homogeneous state with the set of modulated stable patterns. We have used weakly nonlinear analysis for this purpose. Weakly nonlinear analysis provides a low-dimensional description of the slow dynamics that arise near a bifurcation point, by studying the temporal evolution of perturbations beyond linear order (Bender and Orszag, 1999). In this particular case, the method can reveal whether the bifurcation responsible for the emergence of a modulated pattern is sub- or supercritical, and, therefore, whether or not the uniform state coexists with another stable state—presumably, a state with a modulated profile (Ermentrout and Cowan, 1980; Roxin and Montbrió, 2011; Bressloff and Kilpatrick, 2008). While the analysis is local and thus strictly valid only at bifurcation points, it provides us with a more refined description of how the system behaves when the homogeneous state loses its stability. The full weakly nonlinear analysis can be obtained from the authors upon request.

In short, the analysis reveals that the amplitude a of any critical mode, say $k = 1$, will evolve in time according to the nonlinear differential equation:

$$\frac{\mathrm{d}a}{\mathrm{d}T} = a(\alpha + \gamma a^2), \tag{18}$$

where T is a slow time variable proportional to $\epsilon^2 t$, with $\epsilon \ll 1$. The coefficients of the linear and cubic terms in Eq. (18) are

$$\alpha = \Phi'(J_1 - J_1^c), \tag{19}$$

$$\gamma = \frac{1}{2\Phi'^3}\left(\Phi''' + (\Phi'')^2\left[\frac{J_2}{1-\Phi' J_2} + \frac{2J_0}{1-\Phi' J_0}\right]\right), \tag{20}$$

where Φ', Φ'', and Φ''' denote the first, second, and third derivatives of Φ evaluated at the homogeneous solution. The coefficient α measures the distance to the bifurcation point; for values of J_1 below the critical value J_1^c, $\alpha < 0$, and then $a = 0$ is a stable fixed point. In that case the amplitude of the perturbations of the homogeneous state decay with time and, therefore, the homogeneous state is stable. This is just the result of the linear stability analysis. The novel insight given by weakly nonlinear analysis is that there is another fixed point at $a = \sqrt{-\alpha/\gamma}$ whenever $\alpha/\gamma < 0$. This fixed point, when it exists, is unstable when $\gamma > 0$ while it is stable when $\gamma < 0$ (Fig. 11a). The coefficient γ determines therefore whether the bifurcation is sub- or supercritical in the vicinity of the bifurcation.

We can now determine the sign of γ for the connectivity given in Eq. (1). First, note that the Fourier coefficients are all positive when $j_E = j_I$ and $m_E > m_I$, i.e., when excitation and inhibition are balanced and when the range of inhibition is broader than that of excitation (Mexican-hat type connectivity). Also, since we are studying the system in the vicinity of the critical curve and we assume that all modes but the first are stable, the denominators $1-\Phi' J_k$, with $k = 0, 2$ of Eq. (20) are both positive, and therefore the terms within the square brackets of Eq. (20) are both positive. This implies that, given that Φ' is positive everywhere, γ will always be positive unless Φ''' is negative and has an absolute value sufficiently large (Ermentrout, 1998). For our particular choice of transfer function given by Eq. (5), the third derivative can be expressed as $\Phi'''(R) = \beta^3(1-\Phi(R))\Phi(R)\big(6\Phi(R)^2 - 6\Phi(R) + 1\big)$, which takes positive values for low values of R and, in particular, for $R = 0.1$. In general the third derivative of a plausible transfer function is positive—or negative with small absolute value—at the lower range of the subthreshold regime. This is precisely the regime where the inputs induced by spontaneous activity fall.

Taken together, we have that $\gamma > 0$ and hence the bifurcation is subcritical when (i) there is an approximate balance between excitation and inhibition, (ii) the connectivity has Mexican-hat profile, and (iii) the firing rate of the homogeneous stable state is low. This is a sufficient, not necessary,

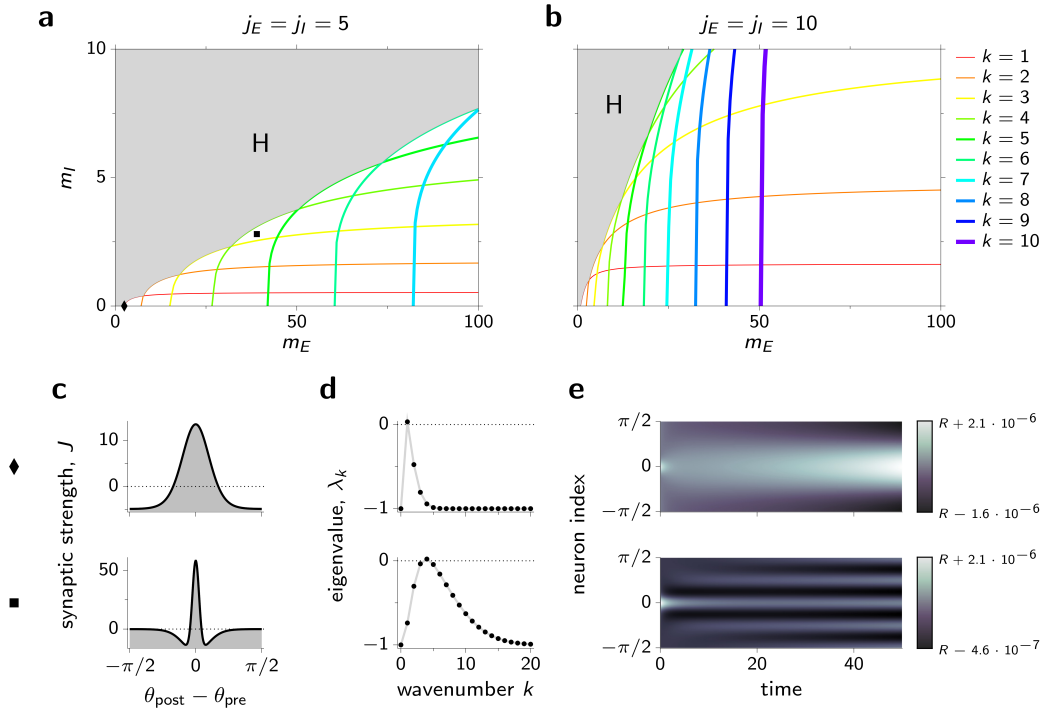

Figure 10. Narrowing the inhibitory footprint allows for patterns with more than one peak (a) Critical curves in the plane (m_E, m_I) for the first ten wavenumbers for $j_E = j_I = 5$. The gray shaded area is the stability region of the homogeneous state (H). The diamond and square indicate the parameter values used in panels c, d, and e. **(b)** Same as (a), for $j_E = j_I = 10$. **(c)** Connectivity profile, **(d)** eigenvalues of the first 20 modes and **(e)** temporal course of the activity profile, for the connectivity configurations indicated by the diamond (top row) and square (bottom row) in panel (a). In the simulations shown in **(e)** the network is perturbed at $t = 0.01$ with a weak pulse of duration $d = 0.01$, intensity $I_s = 10^{-7}$, and concentration parameter $m_s = 10$. Notice the slim range spanned by the color map.

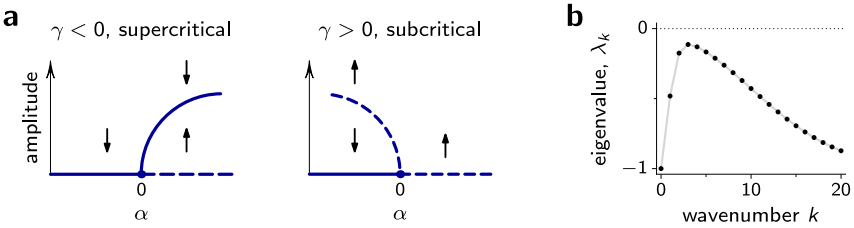

Figure 11. Connectivity determines how the homogeneous state destabilizes. (**a**) bifurcation diagrams of Eq. (18), for positive and negative values of γ. Solid curve: stable branch; dashed curve: unstable branch. The sign of the parameter α, Eq. (19), determines whether or not the homogeneous state is stable, whereas the sign of γ, Eq. (20), dictates whether the bifurcation is sub- or supercritical. If subcritical, the homogeneous state can coexist with another state. (**b**) Eigenvalues of the Fourier modes for the connectivity profile used in the main text (Fig. 3a), with parameters $j_E = j_I = 3.5$, $m_E = 100$, $m_I = 1$.

set of conditions, which is motivated by the known physiology of the cortex.

Mexican-hat connectivities with narrow spreads allow for the coexistence of sets of modulated states

We can now justify our choice of connectivity parameters in the main text. The parameters were chosen so that the homogeneous state be stable, while satisfying the conditions for subcriticality (and, therefore, potential multstability) derived in the previous section. The parameters also allowed for the multistability among modulated states with different number of activity peaks. Such a coexistence can be understood in terms of the spectrum of eigenvalues, which depend on the Fourier transform of the connectivity kernel (Eq. (16)). In general the Fourier transform of a tall and narrow function is a short and broad function in the space of wavenumbers, and vice versa. The Fourier transform of a Mexican-hat-like function will therefore be the difference of a spread and a peaked functions —i.e., a volcano-like function with mild slopes (Fig. 11b can be seen as a volcano-like profile split in half). For a function with such a shape there exists a relatively wide range of wavenumbers for which the function takes values close to its maximum (Fig. 11b). These modes fulfill the subcriticality conditions and have the highest eigenvalues so that, loosely speaking, they can be selectively amplified equally likely. In this configuration, external inputs can destabilize the uniform state and select one particular modulated pattern. Which stable pattern is selected depends then on the mode that contributes the most to the applied input profile.

Equivalence of Gaussian and circular Gaussian profiles for large values of the concentration parameter

For the one-dimensional system, the connectivity kernel is given by a weighted difference of functions of the form

$$g(\theta) = \frac{\exp\big(m\cos(2\theta)\big)}{I_0(m)}. \tag{21}$$

When m is large, the numerator is exponentially suppressed for values of θ away from $\theta = 0$. The normalization by $I_0(m)$ keeps the value of $g(\theta)$ under control despite the presence of exponentially large factors in the numerator. It thus makes sense to neglect the contributions of $g(\theta)$ away from $\theta = 0$ and to keep the Taylor expansion $\cos(2\theta)$ up to second order only. On the other hand, the $I_0(m)$ in the denominator can be replaced by its large-m asymptotic expansion (Abramowitz and Stegun, 1965). In the limit of large m

$$g(\theta) = \frac{\exp\Big(m\big[1 - \frac{(2\theta)^2}{2!} + \cdots\big]\Big)}{\frac{\mathrm{e}^m}{\sqrt{2\pi m}}\Big(1 + \frac{1}{8m} + \cdots\Big)} \approx \sqrt{2\pi m}\exp(-2m\theta^2),$$

and hence $g(\theta)$ approximates to a Gaussian function with variance $\sigma^2 = 1/(2m)$, in units of radians^2 and normalized under the measure $\pi^{-1}\int_{-\infty}^{\infty}\mathrm{d}\theta$.

Distribution of pulse intensities

In this section we derive the probability density function of the input amplitudes, for each neuron θ and when the features are distributed according a mixture distribution given in Eq. (10). Recall that the location of a pulse is a random variable X with probability density function

$$\begin{aligned} p_X(x) &= \sum_{i=1}^{M} p(C_i)p_X(x|C_i) \\ &= \sum_{i=1}^{M} p(C_i)\frac{\exp(m\cos(2x - 2\mu_i))}{I_0(m)}, \end{aligned}$$

where we have assumed for simplicity that all the components C_i have the same variability, determined by the concentration parameter m.

Being the pulse location X a random variable, so must be the intensity of the input felt by each neuron every time a stimulus is presented. We denote such intensity by Y_θ, where the subindex makes explicit the dependence of Y on the particular neuron we consider. We want to know the probability density function of Y_θ, given that Y_θ depends on the pulse location X through

$$Y_\theta = f_\theta(X) = I_s\exp\big(m_s[\cos(2\theta - 2X) - 1]\big). \tag{22}$$

To lighten the notation we will drop in the following the subscript θ, keeping in mind that the random variable Y always refer to a particular neuron.

The probability density of Y is formally given by

$$p_Y(y) = \int \delta(f(x) - y)p_X(x)\,\mathrm{d}x = \sum_j \frac{p_X(f_j^{-1}(y))}{|f'(f_j^{-1}(y))|},$$

where the sum is over the preimages of y. For the function f in Eq. (22) there are two preimages

$$f^{-1}(y) = \theta \pm \frac{1}{2}\arccos\left(1 + \frac{1}{m_s}\log\frac{y}{I_s}\right), \equiv \theta \pm \frac{1}{2}\alpha(y),$$

which exist for any y in the interval $[I_s\mathrm{e}^{-2m_s}, I_s]$. The derivatives of f at the preimages read $f'(f^{-1}(y)) = \pm 2m_s y\sin\alpha(y)$. Putting everything together we have $p_Y(y) = \sum_{i=1}^{N} p(C_i)p_Y(y|C_i)$ where

$$p_Y(y|C_i) = \frac{1}{I_0(m)}\frac{A_i(y)B_i(y)}{m_s y\sin\alpha(y)},$$

and where we have defined the functions

$$\begin{aligned} A_i(y) &= \left[\mathrm{e}\left(\frac{y}{I_s}\right)^{1/m_s}\right]^{m\cos(2\theta - 2\mu_i)}, \\ B_i(y) &= \cosh\big(m\sin(2\theta - 2\mu_i)\sin\alpha(y)\big). \end{aligned}$$

Acknowledgments

We are grateful to Rita Almeida, Michael Graupner, Ernest Montbrió, Michiel Remme, Mattia Rigotti, and Alex Roxin

for very useful discussions. We are also indebted to Michael Graupner for his critical reading of the manuscript, and to two anonymous reviewers for their useful comments. DM gratefully acknowledges the financial support provided by the Leonard Bergstein Award from the Swartz Foundation.

References

Abramowitz M, Stegun IA (1965) Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables. Dover Publications

Amari S (1977) Dynamics of pattern formation in lateral-inhibition type neural fields. Biol. Cybern. 27(2):77–87, 10.1007/bf00337259

Amit DJ (1989) Modeling Brain Function: The World of Attractor Neural Networks. Cambridge University Press

Amit DJ, Brunel N (1997) Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cereb. Cortex 7(3):237–252, 10.1093/cercor/7.3.237

Amit DJ, Mongillo G (2003) Selective delay activity in the cortex: Phenomena and interpretation. Cereb. Cortex 13:1139–1150

Amit DJ, Bernacchia A, Yakovlev V (2003) Multiple-object working memory – A model for behavioral performance. Cereb. Cortex 13:435–443

Ariely D (2001) Seeing sets: Representation by statistical properties. Psychol. Sci. 12(2):157–162, 10.1111/1467-9280.00327

Ashby FG, Alfonso-Reese LA (1995) Categorization as probability density estimation. J. Math Psychol. 39(2):216–233, 10.1006/jmps.1995.1021

Barlow HB (1989) Unsupervised learning. Neural Comput. 1(3):295–311, 10.1162/neco.1989.1.3.295

Battaglia FP, Treves A (1998) Attractor neural networks storing multiple space representations: A model for hippocampal place fields. Phys. Rev. E 58(6):7738, 10.1103/physreve.58.7738

Beck J (1966) Effect of orientation and of shape similarity on perceptual grouping. Perception & Psychophysics 1(5):300–302, 10.3758/bf03207395

Beck J (1982) Textural segmentation. In: Beck J (ed) Organization and Representation in Perception, Lawrence Erlbaum Associates, pp 285–317

Ben-Yishai R, Bar-Or RL, Sompolinsky H (1995) Theory of orientation tuning in visual cortex. P. Natl. Acad. Sci. USA 92(9):3844–3848, 10.1073/pnas.92.9.3844

Bender CM, Orszag SA (1999) Advanced Mathematical Methods for Scientists and Engineers. Springer Verlag

Benucci A, Ringach DL, Carandini M (2009) Coding of stimulus sequences by population responses in visual cortex. Nat. Neurosci. 12(10):1317–1324, 10.1038/nn.2398

Bi GQ, Poo MM (1998) Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J. Neurosci. 18(24):10,464

Boucheny C, Brunel N, Arleo A (2005) A continuous attractor network model without recurrent excitation: Maintenance and integration in the head direction cell system. J. Comput. Neurosci. 18(2):205–227, 10.1007/s10827-005-6559-y

Bregman AS, Campbell J (1971) Primary auditory stream segregation and perception of order in rapid sequences of tones. J. Exp. Psychol. 89(2):244–249, 10.1037/h0031163

Bressloff PC, Kilpatrick ZP (2008) Nonlocal ginzburg-landau equation for cortical pattern formation. Phys. Rev. E 78:041,916, 10.1103/physreve.78.041916

Bressloff PC, Cowan JD, Golubitsky M, Thomas PJ, Wiener MC (2001) Geometric visual hallucinations, euclidean symmetry and the functional architecture of striate cortex. Philos. T. Roy. Soc. B 356(1407):299–330, 10.1098/rstb.2000.0769

Brunel N, Carusi F, Fusi S (1998) Slow stochastic hebbian learning of classes of stimuli in a recurrent neural network. Network-Comp. Neural 9(1):123–152, 10.1088/0954-898x/9/1/007

Burak Y, Fiete IR (2009) Accurate path integration in continuous attractor network models of grid cells. PLoS Comput. Biol. 5(2):e1000,291, 10.1371/journal.pcbi.1000291

Camperi M, Wang X-J (1998) A model of visuospatial working memory in prefrontal cortex: recurrent network and cellular bistability. J. Comput. Neurosci. 5(4):383–405, 10.1023/a:1008837311948

Chong SC, Treisman A (2003) Representation of statistical properties. Vision Res. 43(4):393–404, 10.1016/s0042-6989(02)00596-5

Compte A, Brunel N, Goldman-Rakic PS, Wang X-J (2000) Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cereb. Cortex 10(9):910–923, 10.1093/cercor/10.9.910

Coombes S, Lord GJ, Owen MR (2003) Waves and bumps in neuronal networks with axo-dendritic synaptic interactions. Physica D 178(3-4):219–241, 10.1016/s0167-2789(03)00002-2

Desimone R, Albright T, Gross C, Bruce C (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. J. Neurosci. 4(8):2051–2062, http://www.jneurosci.org/content/4/8/2051.abstract

Duda RO, Hart PE, Stork DG (2000) Pattern Classification, 2nd edn. Wiley-Interscience

Edin F, Klingberg T, Johansson P, McNab F, Tegnér J, Compte A (2009) Mechanism for top-down control of working memory capacity. P. Natl. Acad. Sci. USA 106(16):6802–6807, 10.1073/pnas.0901894106

Ermentrout GB (1998) Neural networks as spatio-temporal pattern-forming systems. Rep. Prog. Phys. 61(4):353–430, 10.1088/0034-4885/61/4/002

Ermentrout GB, Cowan JD (1980) Large scale spatially organized activity in neural nets. SIAM J. Appl. Math 38(1):1–21, 10.1137/0138001

Estes WK (1994) Classification and cognition. Oxford University Press US

Fall CP, Lewis TJ, Rinzel J (2005) Background-activity-dependent properties of a network model for working memory that incorporates cellular bistability. Biol. Cybern. 93(2):109–118, 10.1007/s00422-005-0543-5

Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. Science 291(5502):312–316, 10.1126/science.291.5502.312

Fried LS, Holyoak KJ (1984) Induction of category distributions: a framework for classification learning. J. Exp. Psychol. Learn. 10(2):234–257, 10.1037/0278-7393.10.2.234

Froemke RC, Tsay IA, Raad M, Long JD, Dan Y (2006) Contribution of individual spikes in Burst-Induced Long-Term synaptic modification. J. Neurophysiol. 95(3):1620–1629, 10.1152/jn.00910.2005

Furman M, Wang X-J (2008) Similarity effect and optimal control of multiple-choice decision making. Neuron 60(6):1153–1168, 10.1016/j.neuron.2008.12.003

Gerstein GL, Mandelbrot B (1964) Random walk models for the spike activity of a single neuron. Biophys. J. 4(1):41–68, 10.1016/s0006-3495(64)86768-0

Gilbert CD, Wiesel TN (1989) Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. J. Neurosci. 9(7):2432–2442, http://www.ncbi.nlm.nih.gov/pubmed/2746337

Gold JI, Shadlen MN (2007) The neural basis of decision making. Annu. Rev. Neurosci. 30:535–574, 10.1146/annurev.neuro.29.051605.113038

Goldstone RL (1994) The role of similarity in categorization: providing a groundwork. Cognition 52(2):125–157, 10.1016/0010-0277(94)90065-5

Guo Y, Chow CC (2005) Existence and stability of standing pulses in neural networks: I. existence. SIAM J. Dyn. Syst. 4(2):217–248, 10.1137/040609471

Gutkin BS, Laing CR, Colby CL, Chow CC, Ermentrout GB (2001) Turning on and off with excitation: The role of spike-timing asynchrony and synchrony in sustained neural activity. J. Comput. Neurosci. 11:121–134, 10.1023/a:1012837415096

Haider B, Álvaro Duque, Hasenstaub AR, McCormick DA (2006) Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. J. Neurosci. 26(17):4535–4545, 10.1523/jneurosci.5297-05.2006

Hansel D, Sompolinsky H (1998) Modeling Feature Selectivity in Local

Cortical Circuits, MIT press, chap 13, pp 1–25

Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N (2008) A hierarchy of temporal receptive windows in human cortex. J. Neurosci. 28(10):2539–2550, 10.1523/jneurosci.5487-07.2008

Heise GA, Miller GA (1951) An experimental study of auditory patterns. Am. J. Psychol. 64(1):68–77, 10.2307/1418596

Higley MJ, Contreras D (2006) Balanced excitation and inhibition determine spike timing during frequency adaptation. J. Neurosci. 26(2):448–457, 10.1523/jneurosci.3506-05.2006

Hirsch MW, Smale S (1974) Differential equations, dynamical systems, and linear algebra. Academic Press

Hochstein S, Ahissar M (2002) View from the Top: Hierarchies and reverse hierarchies in the visual system. Neuron 36(5):791–804, 10.1016/s0896-6273(02)01091-7

Hoyle RB (2006) Pattern formation: an introduction to methods. Cambridge University Press

Itskov V, Hansel D, Tsodyks M (2011) Short-term facilitation may stabilize parametric working memory trace. Front. Comput. Neurosci. 5(40), 10.3389/fncom.2011.00040

Izenman AJ (1991) Recent developments in nonparametric density estimation. J. Am. Stat. Assoc. 86(413):205–224, 10.2307/2289732

Jin D, Peng J, Li B (2011) A new clustering approach on the basis of dynamical neural field. Neural Comput. 23(8):2032–2057, 10.1162/neco_a_00153

Kishimoto K, Amari S (1979) Existence and stability of local excitations in homogeneous neural fields. J. Math. Biol. 7:303–318, 10.1007/bf00275151

Koffka K (1999) Principles of Gestalt Psychology. Routledge

Koulakov AA, Raghavachari S, Kepecs A, Lisman JE (2002) Model for a robust neural integrator. Nat. Neurosci. 5(8):775–782, 10.1038/nn893

Kreiman G, Hung CP, Kraskov A, Quiroga RQ, Poggio T, DiCarlo JJ (2006) Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. Neuron 49(3):433–445, 10.1016/j.neuron.2005.12.019

Kruschke JK (1992) ALCOVE: an exemplar-based connectionist model of category learning. Psychol. Rev. 99(1):22–44, 10.1037/0033-295x.99.1.22

Laing CR, Chow CC (2001) Stationary bumps in networks of spiking neurons. Neural Comput. 13(7):1473–1494, 10.1162/089976601750264974

Laing CR, Troy WC (2003) Two-bump solutions of amari-type models of neuronal pattern formation. Physica D 178(3-4):190–218, 10.1016/s0167-2789(03)00013-7

Laing CR, Troy WC, Gutkin B, Ermentrout GB (2002) Multiple bumps in a neuronal model of working memory. SIAM J. Appl. Math 63(1):62–97, 10.1137/s0036139901389495

Liu F, Wang X-J (2008) A common cortical circuit mechanism for perceptual categorical discrimination and veridical judgment. PLoS Comput. Biol. 4(12), 10.1371/journal.pcbi.1000253

Logothetis NK, Sheinberg DL (1996) Visual object recognition. Annu. Rev. Neurosci. 19:577–621, 10.1146/annurev.ne.19.030196.003045

Malach R, Amir Y, Harel M, Grinvald A (1993) Relationship between intrinsic connections and functional architecture revealed by optical imaging and in vivo targeted biocytin injections in primate striate cortex. P. Natl. Acad. Sci. USA 90(22):10,469–10,473, http://www.ncbi.nlm.nih.gov/pubmed/8248133

Maunsell JHR, Newsome WT (1987) Visual processing in monkey extrastriate cortex. Annu. Rev. Neurosci. 10(1):363–401, 10.1146/annurev.ne.10.030187.002051

Mongillo G, Barak O, Tsodyks M (2008) Synaptic theory of working memory. Science 319(5869):1543, 10.1126/science.1150769

van Noorden LPAS (1977) Minimum differences of level and frequency for perceptual fission of tone sequences ABAB. J. Acoust. Soc. Am. 61(4):1041–1045, 10.1121/1.381388

Nosofsky RM (1986) Attention, similarity, and the identification-categorization relationship. J. Exp. Psychol. Gen. 115(1):39–61, 10.1037/0096-3445.115.1.39

O'Connor DH, Wittenberg GM, Wang SS (2005) Graded bidirectional synaptic plasticity is composed of switch-like unitary events. P. Natl. Acad. Sci. USA 102(27):9679–9684, 10.1073/pnas.0502332102

Okun M, Lampl I (2008) Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. Nat. Neurosci. 11(5):535–537, 10.1038/nn.2105

Olson RK, Attneave F (1970) What variables produce similarity grouping? Am. J. Psychol. 83(1):1, 10.2307/1420852

Redish AD, Elga AN, Touretzky DS (1996) A coupled attractor model of the rodent head direction system. Network-Comp. Neural 7(4):671–685, 10.1088/0954-898x_7_4_004

Renart A, Song P, Wang X-J (2003) Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. Neuron 38:473–485

Renart A, de la Rocha J, Bartho P, Hollender L, Parga N, Reyes A, Harris KD (2010) The asynchronous state in cortical circuits. Science 327(5965):587–590, 10.1126/science.1179850

Ringach DL, Hawken MJ, Shapley R (1997) Dynamics of orientation tuning in macaque primary visual cortex. Nature 387(6630):281–284, 10.1038/387281a0

Rosenthal O, Fusi S, , Hochstein S (2001) Forming classes by stimulus frequency: Behavior and theory. P. Natl. Acad. Sci. USA 98(7):4265, 10.1073/pnas.071525998

Roxin A, Montbrió E (2011) How effective delays shape oscillatory dynamics in neuronal networks. Physica D 240(3):323–345, 10.1016/j.physd.2010.09.009

Rust NC, DiCarlo JJ (2010) Selectivity and tolerance ("Invariance") both increase as visual information propagates from cortical area V4 to IT. J. Neurosci. 30(39):12,978–12,995, 10.1523/jneurosci.0179-10.2010

Samsonovich A, McNaughton BL (1997) Path integration and cognitive mapping in a continuous attractor neural network model. J. Neurosci. 17(15):5900–5920

Seung H, Lee DD, Reis BY, Tank DW (2000) Stability of the memory of eye position in a recurrent network of conductance-based model neurons. Neuron 26(1):259–271, 10.1016/s0896-6273(00)81155-1

Seung HS (1996) How the brain keeps the eyes still. P. Natl. Acad. Sci. USA 93(23):13,339, doi:10.1073/pnas.93.23.13339

Shadlen MN, Newsome WT (1998) The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. J. Neurosci. 18(10):3870

Shepard RN (1987) Toward a universal law of generalization for psychological science. Science 237(4820):1317–1323, 10.1126/science.3629243

Shu Y, Hasenstaub A, McCormick DA (2003) Turning on and off recurrent balanced cortical activity. Nature 423(6937):288–293, 10.1038/nature01616

Skaggs WE, Knierim JJ, Kudrimoti HS, McNaughton BL (1995) A model of the neural basis of the rat's sense of direction. In: Tesauro DSTG, Leen TK (eds) Advances in Neural Information Processing Systems, MIT Press Cambridge, MA, USA, vol 7, pp 173–180

Softky WR, Koch C (1993) The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. J. Neurosci. 13:334–350, http://www.jneurosci.org/cgi/content/abstract/13/1/334

Somers DC, Nelson SB, Sur M (1995) An emergent model of orientation selectivity in cat visual cortical simple cells. J. Neurosci. 15(8):5448–5465

Song P, Wang X-J (2005) Angular path integration by moving "hill of activity": A spiking neuron model without recurrent excitation of the head-direction system. J. Neurosci. 25(4):1002–1014, 10.1523/jneurosci.4172-04.2005

Strogatz SH (1994) Nonlinear Dynamics and Chaos. Addison-Wesley Reading, MA

Tass P (1995) Cortical pattern formation during visual hallucinations. J. Biol. Phys. 21(3):177–210, 10.1007/bf00712345

Tsodyks M, Sejnowski TJ (1995) Associative memory and hippocampal place cells. Int. J. Neur. Syst. 6:81–86

Tsodyks M, Pawelzik K, Markram H (1998) Neural networks with dynamic synapses. Neural Comput. 10(4):821–835, 10.1162/089976698300017502

van Vreeswijk C, Sompolinsky H (1998) Chaotic balanced state in a model of cortical circuits. Neural Comput. 10(6):1321–1371, 10.1162/089976698300017214

Wannig A, Stanisor L, Roelfsema PR (2011) Automatic spread of attentional response modulation along gestalt criteria in primary visual cortex. Nat Neurosci 14(10):1243–1244, 10.1038/nn.2910

Wehr M, Zador AM (2003) Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. Nature 426(6965):442–446, 10.1038/nature02116

Wertheimer M (1923) Untersuchungen zur Lehre von der Gestalt. II. Psychologische Forschung 4(1):301–350, 10.1007/bf00410640, reprinted in part in W. D. Ellis (Ed.), *A Source Book of Gestalt Psychology* (pp. 71–88), The Gestalt Journal Press, 1997.

Zhang K (1996) Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. J. Neurosci. 16(6):2112–2126

A Supplementary Information: Weakly Nonlinear analysis

In this section we derive the amplitude equation (A.14), closely following the general derivation given in (Roxin and Montbrió, 2011). To keep the analysis simple, we assume that the connectivity is such that only one spatial mode k^c reaches the critical value $1/\Phi'$, and that all the other spatial modes are sufficiently below the critical value so that they do not contribute with additional instabilities. We also assume without loss of generality that the first mode to destabilize is the first Fourier mode, $k^c = 1$. We refer to the connectivity kernel that results from such a configuration as the *critical* connectivity kernel, denoted by $J^c(\theta)$.

There exist several, more or less equivalent methods to derive the amplitude equation. One of them is based on multi-scale analysis (see, e.g., Bender and Orszag (1999); Hoyle (2006)), a method that exploits the disparity between the timescales of stable and critical modes near a bifurcation, where dynamics are slow. The basic idea is to introduce a slow timescale $T = \epsilon^2 t$, with $\epsilon \ll 1$, and treat it as an independent variable (even though it is clearly not). We can then assume that the solution $r(t,\theta)$ has a perturbation expansion of the form

$$r(t,\theta) = R + \epsilon r_1(t,T,\theta) + \epsilon^2 r_2(t,T,\theta) + O(\epsilon^3), \tag{A.1}$$

so that

$$\frac{\partial}{\partial t} r(t,\theta) = \epsilon \frac{\partial r_1}{\partial t} + \epsilon^2 \frac{\partial r_2}{\partial t} + \epsilon^3 \left(\frac{\partial r_1}{\partial T} + \frac{\partial r_3}{\partial t} \right) + O(\epsilon^4). \tag{A.2}$$

A natural perturbative parameter in this context is some measure of the distance of the system to the bifurcation point, which in our case could be the difference of the first-order Fourier coefficient of the connectivity kernel and the critical value $1/\Phi'$. Because we want the effect of changing the parameter to be of the same order of the nonlinearities of the system, we scale the difference as

$$J_1 - J_1^c = \epsilon^2 \Delta J_1. \tag{A.3}$$

This is why we chose our slow timescale to scale as $\epsilon^2 t$. The connectivity kernel that results from using J_1 as first Fourier coefficient instead of J_1^c is the *perturbed* kernel, which differs from the critical kernel by a perturbation ΔJ (a function).

Nonlinearities enter through the current-to-rate transfer function Φ. To see the effect of these nonlinearities we first expand the argument of Φ using equations (A.1) and (A.3),

$$J * r + I = (J_0 R + I) + \epsilon \left(J^c * r_1 \right) + \epsilon^2 \left(J^c * r_2 \right) + \epsilon^3 \left(J^c * r_3 + \Delta J * r_1 \right) + O(\epsilon^4),$$

where the asterisk denotes convolution over the angular variable, and where we have grouped terms by powers in ϵ. In this expansion we have omitted the terms containing $\Delta J * R$, which vanish because the perturbation is orthogonal to the uniform (0-th) mode. The Taylor expansion of the transfer function around the homogeneous solution is then

$$\Phi(J * r + I) = \Phi(J_0 R + I) + \epsilon f_1(r_1) + \epsilon^2 f_2(r_1, r_2) + \epsilon^3 f_3(r_1, r_2, r_3) + O(\epsilon^4), \tag{A.4}$$

where

$$\begin{aligned} f_1(r_1) &= \Phi' J^c * r_1, \\ f_2(r_1, r_2) &= \Phi' J^c * r_2 + \frac{\Phi''}{2} (J^c * r_1)^2, \\ f_3(r_1, r_2, r_3) &= \Phi' J^c * r_3 + \frac{\Phi'''}{6} (J^c * r_1)^3 + \Phi''(J^c * r_1)(J^c * r_2). \end{aligned}$$

If we plug the expansions (A.1), (A.2), (A.4) into the dynamical equation (4) and collect powers in ϵ, we end up with a hierarchy of equations. At zeroth order we simply get the fixed point condition given by Eq. (7). The equations at higher orders in ϵ take the form

$$\text{order } \epsilon: \quad L_0 r_1 = 0, \tag{A.5}$$

$$\text{order } \epsilon^2: \quad L_0 r_2 = N_2(r_1), \tag{A.6}$$

$$\text{order } \epsilon^3: \quad L_0 r_3 = L_2 r_1 + N_3(r_1, r_2), \tag{A.7}$$

$$\vdots$$

where we have defined the linear operators

$$\begin{aligned} L_0 &= \frac{\partial}{\partial t} + 1 - \Phi' J^c * (\cdot), \\ L_2 &= \frac{\partial}{\partial T} - \Phi' \Delta J * (\cdot), \end{aligned}$$

and the nonlinear forcing terms

$$\begin{aligned} N_2 &= \frac{1}{2} \Phi''(J^c * r_1)^2, \\ N_3 &= \frac{1}{6} \Phi'''(J^c * r_1)^3 + \Phi''(J^c * r_1)(J^c * r_2). \end{aligned}$$

Note that forcing terms depend on solutions of lower order. This property allows us to solve the hierarchy of equations (A.5)–(A.7) iteratively.

First order The homogeneous equation (A.5) is equivalent to the dispersion relation in Eq. (16), with $\lambda_k = 0$, and it is therefore equivalent to imposing criticality. Because the linear growth rate of the critical mode is zero, the first order correction r_1 must be independent of the fast natural time scale t. This in turn implies that the remaining higher order terms will also be independent of t. Thus perturbations depend on time through the slow timescale T, i.e., $r_i = r_i(T)$.

From the previous linear analysis, we know that the first order correction r_1 must be of the form:

$$r_1(T,\theta) = A(T) \mathrm{e}^{2\mathrm{i}\theta} + \bar{A}(T) \mathrm{e}^{-2\mathrm{i}\theta}, \tag{A.8}$$

where we have used the fact that the critical wavenumbers are $k = 1, -1$, and where the overbar denotes complex conjugation. The temporal dependence of the perturbation enters via the coefficient $A(T)$, called the *aplitude*. The factor 2 in the phase comes from the periodic boundary conditions on the interval $[-\pi/2, \pi/2]$ of orientations. In the following we shall drop the explicit dependence of A on T to lighten the notation.

Second order Plugging the first-order correction (A.8), into the expression for $N_2(r_1)$ on the right hand side of Eq. (A.6) leads to

$$L_0 r_2(T,\theta) = \frac{\Phi''}{2} J_1^{c\,2} \left(A^2 \mathrm{e}^{4\mathrm{i}\theta} + \bar{A}^2 \mathrm{e}^{-4\mathrm{i}\theta} + 2|A|^2 \right). \tag{A.9}$$

where we have used $J_{-k} = J_k$. From this expression we can derive the second order correction to the solution, $r_2(T,\theta)$. The correction r_2 should be independent of the time scale t, and must be such that when the linear operator $1 - \Phi' J^c * (\cdot)$ operates on it, it gives rise to the r.h.s. of Eq. (A.9). The solution is therefore a combination of the modes $k = -2, 0, 2$ present in the r.h.s. of (A.9), with coefficients chosen to match both sides of the equation. If we replace J_1^c with its explicit value $1/\Phi'$, the second order correction reads

$$r_2(T,\theta) = \frac{\Phi''}{2\Phi'^2} \left(\frac{A^2 \mathrm{e}^{4\mathrm{i}\theta} + \bar{A}^2 \mathrm{e}^{-4\mathrm{i}\theta}}{1 - \Phi' J_2} + \frac{2|A|^2}{1 - \Phi' J_0} \right). \tag{A.10}$$

Third order Substituting equations (A.8) and (A.10) into Eq. (A.7) we obtain

$$\begin{aligned} L_0 r_3 = \Big\{ &- \left(\frac{\mathrm{d}A}{\mathrm{d}T} - \Phi' \Delta J_1 A \right) \mathrm{e}^{2\mathrm{i}\theta} \\ &+ \frac{\Phi'''}{6\Phi'^3} \left(A^3 \mathrm{e}^{6\mathrm{i}\theta} + 3A|A|^2 \mathrm{e}^{2\mathrm{i}\theta} \right) \\ &+ \frac{(\Phi'')^2}{2\Phi'^3} \left(J_2 \frac{A^3 \mathrm{e}^{6\mathrm{i}\theta} + A|A|^2 \mathrm{e}^{2\mathrm{i}\theta}}{1 - \Phi' J_2} \right. \\ &\left. + 2 J_0 \frac{A|A|^2 \mathrm{e}^{2\mathrm{i}\theta}}{1 - \Phi' J_0} \right) \Big\} + \text{c.c.} \end{aligned} \tag{A.11}$$

The general solution of Eq. (A.11) is the sum of the general solution of the homogeneous equation, $L_0 r_3 = 0$, and any particular solution of the nonhomogeneous equation. The homogeneous equation has the same form as Eq. (A.5). It has therefore the same solution, given by a linear combination of $\mathrm{e}^{2\mathrm{i}\theta}$ and $\mathrm{e}^{-2\mathrm{i}\theta}$. On the other hand, the forcing terms on the right-hand side of Eq. (A.7) include terms also proportional to $\mathrm{e}^{2\mathrm{i}\theta}$ and $\mathrm{e}^{-2\mathrm{i}\theta}$. These terms are resonant and generate secular terms that must vanish in order for the expansion to make sense (Bender

and Orszag, 1999). The differential equation for the amplitude $A(T)$ is derived from the requirement that secular terms vanish at current order. Gathering the secular terms containing $e^{2i\theta}$, we obtain the following amplitude equation

$$\frac{dA}{dT} = \alpha A + \gamma A|A|^2 + O(\epsilon^4), \tag{A.12}$$

where α and γ are given in equations (19)–(20). If we repeat the same step for the secular terms containing $e^{-2i\theta}$ we obtain the complex conjugate of Eq. (A.12), which does not convey any additional information. To gain more insight into Eq. (A.12), we can solve it for $A(T)$ using the polar form $A(T) = a(T)e^{i\varphi(T)}$, where both a and φ take real values and represent, respectively, the amplitude and the phase of the perturbation. Substituting the polar form in the amplitude equation and equating real and imaginary parts yields

$$\frac{d\varphi}{dT} = 0, \tag{A.13}$$

$$\frac{da}{dT} = a(\alpha + \gamma a^2). \tag{A.14}$$

Equation (A.13) reflects the invariance of our system under angular translations. It states that the phase of the perturbation does not change with time and that, therefore, the phase is determined entirely by the initial conditions. This is why the phase of a pattern coincides with the phase of the input profile used to trigger it. The implications of Eq. (A.14) are discussed in the Appendix.

Other critical modes In the derivation presented above we assumed that the critical mode was the first Fourier mode, $k^c = 1$. The outcome is essentially the same if one assumes an arbitrary critical mode k^c. It is not difficult to show that the resulting amplitude equation is (A.12) with parameters given by equations (20), but replacing J_1 by J_{k^c}, ΔJ_1 by ΔJ_{k^c}, and J_2 by J_{2k^c}.