Bump attractor dynamics underlying stimulus integration in perceptual estimation tasks

Jose M. Esnaola-Acebes[1,2], Alex Roxin[1,2], Klaus Wimmer[1,2]

1. Centre de Recerca Matemàtica (CRM), Campus de Bellaterra, Edifici C, 08193 Bellaterra (Barcelona), Spain

2. Barcelona Graduate School of Mathematics, Barcelona, Spain

Abstract

Perceptual decision and continuous stimulus estimation tasks involve making judgments based on accumulated sensory evidence. Network models of evidence integration usually rely on competition between neural populations each encoding a discrete categorical choice. By design, these models do not maintain information of the integrated stimulus (e.g. the average stimulus direction in degrees) that is necessary for a continuous perceptual judgement. Here, we show that the continuous ring attractor network can integrate a stimulus feature such as orientation and track the stimulus average in the phase of its activity bump. We reduced the network dynamics of the ring model to a two-dimensional equation for the amplitude and the phase of the bump. Interestingly, these reduced equations are nearly identical to an optimal integration process for computing the running average of the stimulus orientation. They differ only in the intrinsic dynamics of the amplitude, which affects the temporal weighting of the sensory evidence. Whether the network shows early (primacy), uniform or late (recency) weighting depends on the relative strength of sensory stimuli compared to the amplitude of the bump and on the initial state of the network. The specific relation between the internal network dynamics and the sensory inputs can be modulated by changing a single parameter of the model, the global excitatory drive. We show that this can account for the heterogeneity of temporal weighting profiles observed in humans integrating a stream of oriented stimulus frames. Our findings point to continuous attractor dynamics as a plausible mechanism underlying stimulus integration in perceptual estimation tasks.

Introduction

Perceptual decisions and continuous stimulus estimation involve making perceptual judgements based on accumulated sensory evidence. While a categorical perceptual decision usually requires discriminating stimuli in order to select between several options (e.g. left vs. right motion of an ambiguous random dot stimulus (Britten et al., 1992), or clockwise vs. counter-clockwise motion compared to a reference (Jazayeri and Movshon, 2007)), in stimulus estimation tasks participants report a continuous stimulus feature, for example the stimulus orientation in degrees. Typical estimation tasks involve accumulating evidence about the average motion direction of a random dot pattern in degrees (Nichols and Newsome, 2002; Jazayeri and Movshon, 2007; Talluri et al., 2018; Aitken et al., 2020) or the average of of a sequence of oriented gratings (Wyart et al., 2012; Cheadle et al., 2014). In particular, we are interested here in experimental tasks that require estimating, as an analog quantity, the average orientation or direction of a potentially time-varying stimulus.

Neural network models of evidence integration usually rely on slow recurrent dynamics and competition between neural populations, each encoding one of the categorical choice options (Wang, 2002; Roxin and Ledberg, 2008; Wimmer et al., 2015; Wong and Wang, 2006). The architecture of these *discrete* attractor models thus directly reflects the categorical nature of the decision task. Similarly, phenomenological drift-diffusion models linearly integrate stimulus evidence for or against different choice options until one of the bounds is reached (Ratcliff and McKoon, 2008; Bogacz et al., 2006; Pardo-Vazquez et al., 2019; Kiani et al., 2008; Mazurek et al., 2003; Zhang et al., 2009). While the amount of accumulated evidence (i.e., the state of the decision variable) at the time of the decision is predictive of choice certainty or confidence (Kiani et al., 2014), by design these models do not maintain information about continuous features of the integrated stimulus (e.g. the average orientation; but see (Ratcliff, 2018)). They are thus not well suited to study stimulus estimation tasks that require continuous perceptual judgements or tasks that require both the report of a continuous estimate and a categorical choice (Jazayeri and Movshon, 2007; Luu and Stocker, 2018; Talluri et al., 2020).

The representation of continuous stimuli in neural population codes has been studied extensively using normative models (Luu and Stocker, 2018) or neural coding models (Jazayeri and Movshon, 2006, 2007). These models can explain how to optimally represent a stimulus in a neural population (e.g. an orientated stimulus in area V1), how to compute an estimate from a noisy sensory inputs (Latham et al., 2003), how to combine different cues in a single representation (Deneve et al., 2001), and how to read-out the population code in order to optimally discriminate stimuli (Pouget et al., 2003). The main focus of these theoretical studies has been on the representation of stimulus features in sensory cortices and not on perceptual decision making or stimulus estimation, operations that, while relying on sensory representations, involve additional computations. There is a lack of biophysical neural network models that can be used to study the underlying circuit mechanisms of these computations.

A natural candidate for such a neural network model is the well-known *continuous* attractor model (Amari, 1977; Ermentrout, 1998). Continuous bump attractor networks have a topological ring architecture and due to strong recurrent coupling they can show localized activity in a subset of neurons with a bell-shaped pattern (bump) that persists even in the absence of tuned external input. Because all states along the ring are energetically equal, the bump can be located at any location along the network. The bump location thus provides a substrate for encoding a continuous variable such as an orientation or a spatial location. Moreover, while intrinsically the bump position is stable, external inputs can drive the bump to move. Due to these features, bump attractor networks have been proposed as the neural mechanism underlying a variety of sensory and cognitive computations: the emergence of contrast-invariant orientation tuning in visual cortex (Ben-Yishai et al., 1995), the identification of a weak orientated stimulus (Liu and Wang, 2008), the maintenance of spatial working memory in prefrontal cortex (Compte, 2000; Wimmer et al., 2014), and the representation of spatial orientation and angular path integration in the mammalian head direction system (Skaggs et al., 1995; Zhang, 1996; Boucheny et al., 2005; Chaudhuri et al., 2019), the fly's heading direction system (Kim et al., 2017) and in grid cells (Burak and Fiete, 2009). None of these previous studies has investigated whether a bump attractor network can accumulate sensory information to estimate the time-average of a continuous stimulus as required in perceptual estimation tasks.

Here, we show that the ring attractor model can accurately compute the average of a time-varying orientated stimulus. Close to the dynamical regime where the bump of activity emerges the network approximates a *perfect vector integrator* such that the phase of the bump of activity at each time point corresponds to the running circular average of the stimulus. The temporal weighting of stimulus information is uniform in this case, as is characteristic of perfect integration. Moreover, the model can also exhibit primacy and recency regimes, i.e., it can integrate the stimulus giving higher weight to early or late parts of the stimulus in a fixed duration experiment. These integration regimes depend on the global excitatory drive to the network which is a single parameter controlling the amplitude and the dynamics of the bump. On average, the model estimates the mean stimulus orientation faithfully albeit with a higher trial-to-trial variability when the temporal weighting is not uniform. To illustrate the relevance of our theoretical results, we analyze data from psychophysical experiments in which human observers integrated a stream of eight oriented stimulus frames. We find that the temporal weighting of evidence across subjects is highly heterogeneous and can be quantitatively fit by our continuous attractor model. Our work suggests continuous attractor dynamics as a potential mechanism underlying stimulus integration in perceptual estimation tasks.

Results

Stimulus integration with bump attractor dynamics

We start by simulating the integration of stimuli in a continuous ring attractor model, similar to experimental tasks that require estimating, as an analog quantity, the average direction of motion of a noisy random dot stimulus (Nichols and Newsome, 2002; Jazayeri and Movshon, 2007) or of a sequence of oriented Gabor patches (Wyart et al., 2012; Cheadle et al., 2014). The network model has a ring architecture and Mexican-hat connectivity (Fig. 1A) and its firing rate dynamics are described by Eq. 4 (Methods).

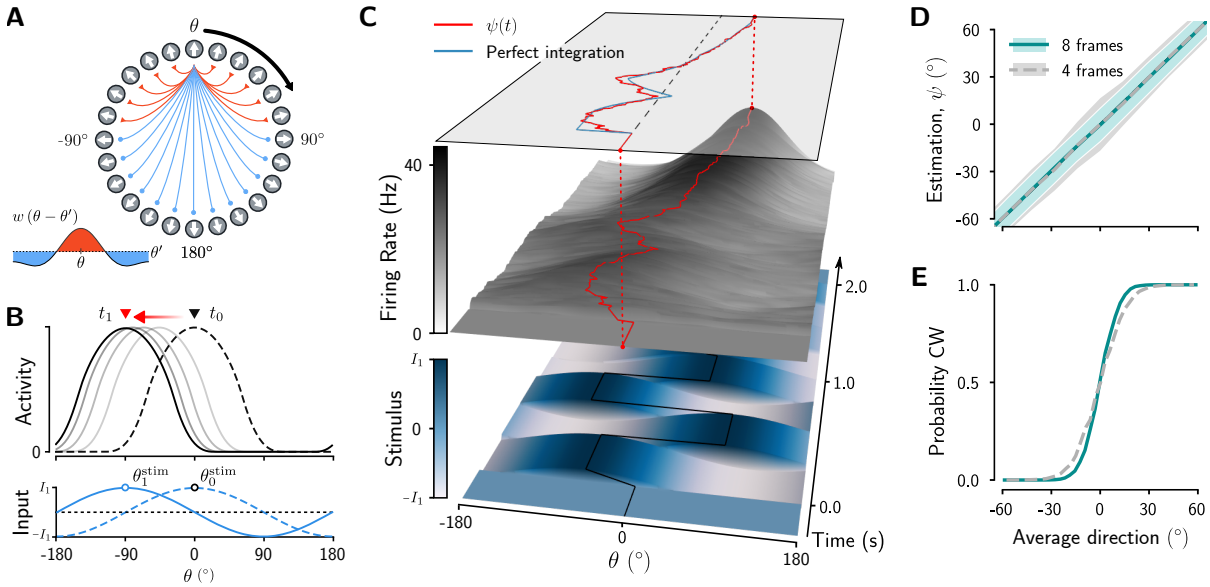

Figure 1: Stimulus integration in the bump attractor network.
(A) Ring network with Mexican-hat connectivity, that is, strong local excitation (red connections) and broader inhibitory coupling (blue). **(B)** Network activity for neurons arranged according to their position in the ring (A). Due to strong recurrent connectivity, a bump of activity emerges in this network at a position determined by the external input and persists when the input is removed. For time-varying inputs, the activity bump (dashed line) moves towards a new position (solid black line). **(C)** Network activity in a single trial. The initial activity of the network (middle panel) is spatially homogeneous and evolves into a bump while integrating a time-varying stimulus composed of eight oriented stimulus frames (bottom panel). Top: The phase of the bump (red) as a function of time closely tracks the running average of the orientations of the stimulus frames (blue). **(D, E)** Continuous stimulus estimate (D) and probability of clockwise choices (E) as a function of the average stimulus direction for stimulus durations of 1 s (4 stimulus frames of 250 ms) and 2 s (8 stimulus frames).

We studied stimulus integration in the regime in which, due to strong recurrent connectivity, a bump of activity emerges in the network at a position determined by the spatially modulated stimulus input, and this bump persists when the stimulus is removed (Fig. 1B). In this regime, a fast change of the stimulus to a new direction does in general not lead to an instantaneous extinction and re-appearance of the bump at the new direction, but rather initiates a continuous translation of the activity along the ring network, directed towards the new stimulus direction and with only small changes in the amplitude of the bump (Fig. 1B). This has been termed "virtual rotation" (Ben-Yishai et al., 1995). Thus, the bump position at a given time depends on the current input to the network and on the previous bump position. We wondered whether through this mechanism the position (phase, ψ) of the activity bump could track an estimate of the integrated stimulus direction. We simulated a task which required estimating the average direction of eight successive oriented stimulus frames with constant amplitude (e.g. stimulus contrast), and directions distributed between $-90°$ and $+90°$ (Fig. 1C). We found that the transient population response to changing stimulus input effectively computes this estimation (Fig. 1C). In the example trial, the evolution of the bump phase closely approximated the cumulative running average of the stimulus, that is, the time-average of the stimulus up to a given time point (Fig. 1C, top).

Indeed, estimation curves, obtained by simulating many trials, show that, on average, the bump phase closely tracks a continuous estimate of the averaged sensory input (Fig. 1D). The accuracy of this estimate improves with the stimulus duration, as expected from an integration process (errorbars in Fig. 1D). This improvement of accuracy can also be seen in psychometric curves obtained by converting the analog estimate to a categorical choice (with positive angles corresponding to clockwise reports and negative angles corresponding to counter-clockwise reports; Fig. 1E). In sum, we found that the phase of the activity bump can track an estimate of the integrated stimulus over times much larger than the intrinsic time constant of the network ($\tau = 20$ ms, stimulus duration T_{stimulus} = 2 s).

Dynamics of the bump

Motivated by the observation that the phase of the bump can track the average of the stimulus with striking precision (Fig. 1), we next sought to identify and characterize the neural network mechanisms underlying stimulus integration in the bump attractor network. In order to study the dynamics of the movement of the bump, we reduced the network model (Eq. 4) to a 2-dimensional differential equation for the amplitude, $R(t)$, and the phase of the bump, $\psi(t)$ (see for example Hansel and Sompolinsky, 1998; Ermentrout, 1998):

$$\tau \frac{\partial R}{\partial t} = \tilde{I}_1 \cos\left(\psi - \theta^{\text{stim}}\right) + \tilde{I}_0 R - cR^3 + \xi_1(t) \tag{1a}$$

$$\tau \frac{\partial \psi}{\partial t} = -\frac{\tilde{I}_1}{R} \sin\left(\psi - \theta^{\text{stim}}\right) + \frac{\xi_2(t)}{R}, \tag{1b}$$

where τ is the neural time constant and $\tilde{I}_0$ is proportional to the global excitatory drive to each neuron, relative to its critical value at the onset of the bump ($\tilde{I}_0 \propto I_0 = I_{\text{exc}} - I_{\text{crit}}$). The constant c depends on the neural transfer function $\Phi(\cdot)$ and the synaptic connectivity profile given by the connection weights w_k of Eq. 4. The time-varying stimulus input is characterized by its strength $\tilde{I}_1 \propto I_1$, and its orientation θ^{stim}. The noise terms $\xi_1(t)$ and $\xi_2(t)$ are related to the internal stochasticity and to fluctuations in the stimulus realized as independent noise inputs to each neuron in the full model (Eq. 4). A complete derivation of these equations is contained in Methods. We found that this reduced model accurately describes the bump dynamics close to the bifurcation point where the bump emerges, and that it is qualitatively valid for a wide parameter regime, allowing us to analytically study the dynamics of the bump phase. From Eq. 1b it directly follows that the rate of change of the phase (the angular speed) depends on the ratio of the stimulus strength I_1 and the bump amplitude R. Thus, for a given stimulus amplitude I_1, the higher the amplitude of the bump the smaller will be the angular speed $d\psi/dt$. In other words, larger bumps are more sluggish. The steady-state bump amplitude is determined by the global excitatory drive I_0, ($R_\infty \propto \sqrt{I_0}$; Methods), indicating the key role of this parameter for the integration dynamics.

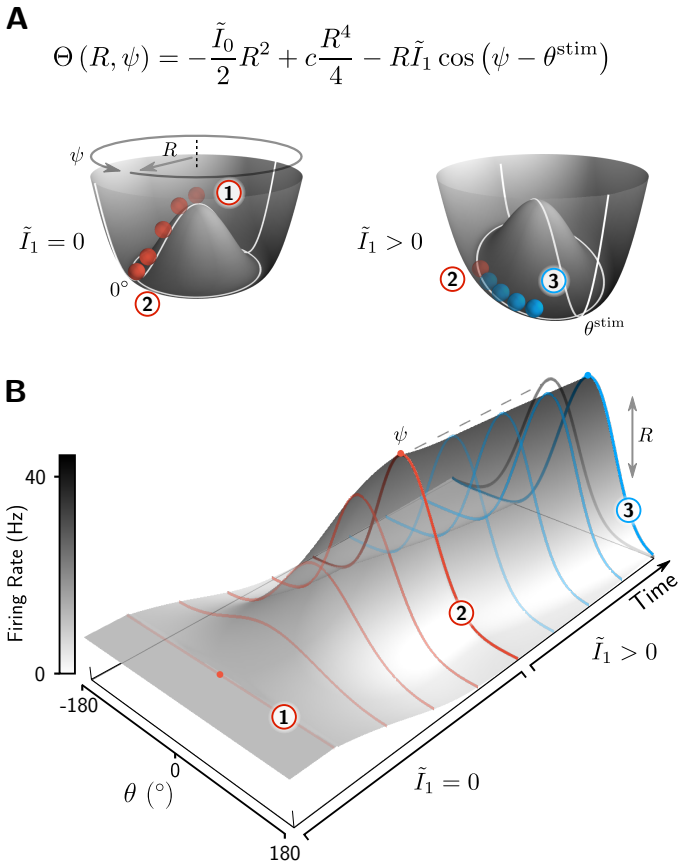

Figure 2: Dynamics of the bump attractor network in the potential landscape.
(A) The equation of the potential corresponding to the dynamics of the bump attractor (Eq. 1) and its geometrical representation. The movement of the red ball in each potential corresponds to the dynamics of bump formation and translation shown in (B).
(B) Evolution of the network activity from the unstable homogeneous state ① to the stable bump state ② in the absence of external stimuli ($\tilde{I}_1 = 0$), followed by a translation of the bump (transition from ② to ③) towards the location of the stimulus θ^{stim} when the external stimulus is present ($\tilde{I}_1 > 0$).

An intuition about how a time-varying stimulus impacts the phase of the bump can be gained if we visualize the dynamics of the system, Eq. 1, by means of the physical idea of the potential energy. We can interpret the dynamics of the bump as a heavily damped particle sliding down the walls of a 3-dimensional potential well, where a potential $\Theta\left(R,\psi\right)$ of the system described in Eq. 1 is given by

$$\Theta\left(R,\psi\right) = -R\tilde{I}_1 \cos\left(\psi - \theta^{\mathrm{stim}}\right) - \frac{\tilde{I}_0}{2}R^2 + \frac{c}{4}R^4. \tag{2}$$

The surface of this potential represents the dynamical manifold of Eq. 1, such that a point on this surface is a possible state of the activity bump, characterized by its amplitude R and phase ψ (Fig. 2A). By definition of the potential, we have that $d\Theta/dt \leq 0$ (Methods), which means that $\Theta(t)$ decreases over time, so that the particle always moves towards the minima of the potential with only the fluctuation terms $\xi_1(t)$ and $\xi_2(t)$ causing possible motion upwards and along the neutrally stable manifold. This interpretation allows us to describe the mechanism underlying the evolution of the bump graphically (Fig. 2).

In the absence of stimulus input ($I_1 = 0$), the potential $\Theta\left(R,\psi\right)$ resembles a juice-squeezer-shaped surface with an unstable fixed point at the center, corresponding to a bump with zero amplitude ($R = 0$), i.e. the unstable spatially homogeneous network state (Fig. 2A,B; ①). The circular well of the manifold represents bump attractor states ($R > 0$) that are neutrally stable along the angular dimension ψ. The formation of the bump is described by the movement of the particle that, starting at the center of the juice squeezer, rolls downwards towards the stable manifold while increasing its distance to the center, R (Fig. 2A, transition from ① to ②). The corresponding evolution of the network activity is shown in Fig. 2B, where the initially homogeneous network activity evolves into a bump state, reaching its steady-state amplitude at point ②. Note that because here we consider the bump formation in the absence of stimulus input, the bump forms at an arbitrary angle ψ, determined by noise fluctuations while in the presence of a stimulus the bump will emerge close to the initial stimulus orientation θ^{stim} (Fig. 1C).

When presenting a stimulus to the network ($I_1 > 0$), the radial symmetry of the potential is broken and a deeper region arises at the location of the stimulus θ^{stim} (Fig. 2A, right). If the particle had already reached the stable circular manifold at the bottom of the potential, the stimulus will force the particle to move towards this deeper state along the manifold (Fig. 2A, transition from ② to ③). In terms of the network activity, this corresponds to a movement of the bump towards θ^{stim}, accompanied by only slight changes in bump amplitude (Fig. 2B). Solving Eq. 1 for small stimulus amplitudes, we found that the evolution of the phase of the bump depends approximately linearly on the difference of the bump phase and the stimulus direction (Fig. S2C and Methods). In contrast, if the particle has not yet reached the stable manifold, i.e. when a stimulus is applied during the initial bump formation, it will cause both a change in bump phase and, together with the internal dynamics, a change in bump amplitude.

In general, successive stimulus frames of a time-varying stimulus will drag the bump towards their respective orientations in a continuous fashion as shown in the example trial of Fig. 1C. Importantly, for a given stimulus strength I_1 the angular displacement produced by the stimulus depends on the radius R of the potential, that is, $Rd\psi \propto I_1$ (the length of the arc is proportional to the stimulus strength). Thus, the larger the bump, the smaller the resulting change in the bump phase ψ (Fig. S1). In sum, we can distinguish three distinct stages of the evolution of the bump during which it becomes increasingly difficult for the stimulus to impact the phase of the bump. Initially, starting from an homogeneous network activity, the stimulus orientation will strongly impact the phase of the emerging bump. Second, the bump will grow in amplitude, driven by the internal dynamics and the stimulus, and thus successive stimulus frames become less and less effective in moving the bump. Third, when the bump amplitude reaches its steady state, it cannot grow further and the impact of subsequent stimuli becomes constant over time. Equipped with these theoretical results, we can now study different stimulus integration regimes in the network.

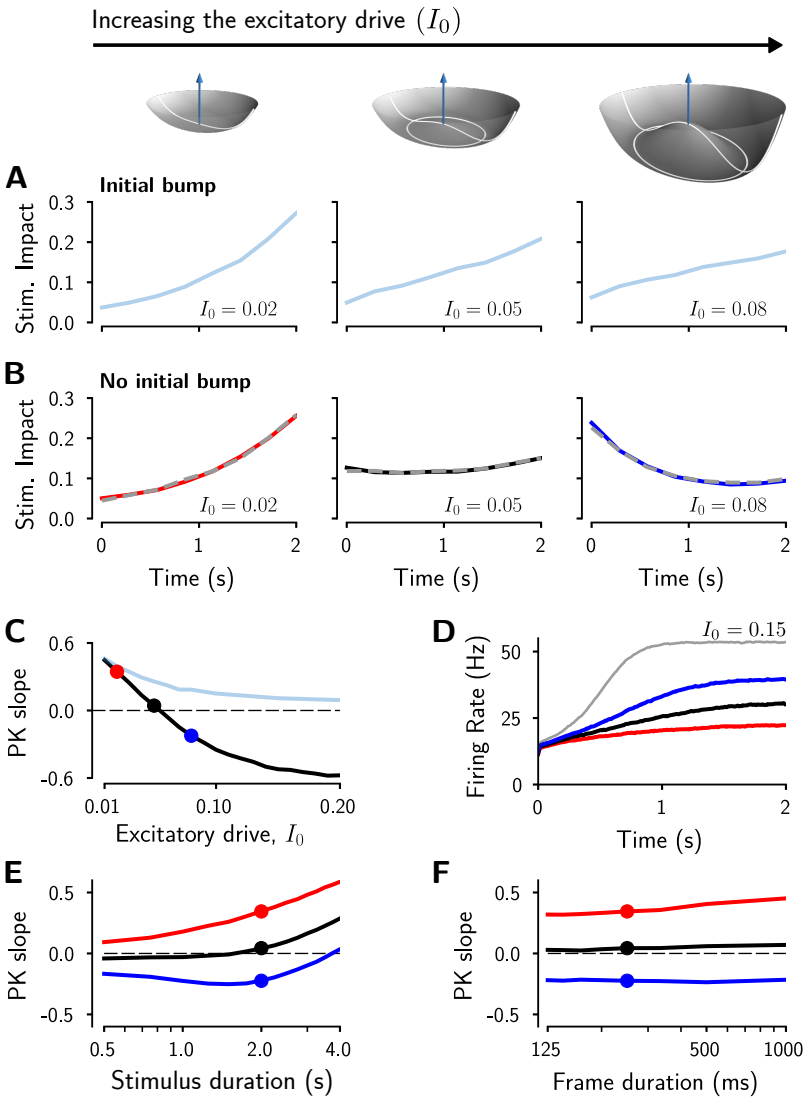

Figure 3: The global excitatory drive I_0 determines the integration regime of the network. **(A)** Psychophysical kernels (PKs) for increasing values of I_0 (bottom, from left to right, $I_0 = 0.02, 0.05$ and 0.08) with a fully formed bump as the initial condition. The PKs describe the impact of each stimulus frame on the estimated average orientation. Top: For small values of I_0 the potential takes the form of a steep paraboloid where the bump has little space to grow (top left). Intermediate values of I_0 transform the potential into a wide surface similar to a dish with a relative flat surface at its central area (top middle). Increasing I_0 widens the potential and gives rise to an increasingly deeper circular manifold at the bottom of the juice squeezer (top right). **(B)** PKs as in panel (A) but with spatially homogeneous initial conditions (flat network activity). PKs were obtained from numerical simulations of the ring attractor network (colored lines) and from simulating the amplitude equation (dashed gray lines). **(C)** PK slope as a function of the excitatory drive I_0. Solid black and faint blue curves correspond to the spatially homogeneous (B) and inhomogeneous (A) initial conditions, respectively. **(D)** Evolution of the bump amplitude for different values of I_0. Colored lines correspond to values of I_0 used in panel (B). **(E)** PK slope as a function of the stimulus duration, with stimuli composed of frames with a fixed duration of 250 ms. **(F)** PK slope for a fixed stimulus duration of T = 2 s as a function of the duration of a stimulus frame. The number of frames ranges from 2 (frame duration 1000 ms) to 16 (frame duration 125 ms).

The ring model shows a variety of temporal evidence integration regimes

Our theoretical results showed that the changes in bump phase caused by a time-varying stimulus critically depend on the bump amplitude, suggesting its key role in the integration of sensory stimuli. To investigate this, we ran simulations in which networks with different steady-state bump amplitudes had to estimate the average direction of a sequence of oriented frames, distributed between $-90°$ and $+90°$ as in Fig. 1C. To quantify the dynamics of evidence integration, we used the psychophysical kernel (PK) which measures the influence of sensory evidence during the course of the stimulus (see Methods). We first considered the case where a trial starts with an initial bump (i.e., the particle is at a position on the circular attractor well of the potential; point ② in Fig. 2), whose amplitude we set by varying the global excitatory drive to the network I_0 and keeping all other parameters fixed. The PKs we obtained for different I_0 are qualitatively similar (Fig. 3A): they all rise, indicating that the stimulus frames later in the trial have more impact on the final estimate (i.e., the bump phase ψ at the end of the trial) than early stimulus frames. This commonly called "recency" effect becomes weaker as I_0 increases, yielding a decrease of the PK slope (Fig. 3C, blue line). The general over-weighting of late stimulus information occurs because each successive stimulus frame leads to an angular displacement of the bump towards the current stimulus orientation θ^{stim} with updates proportional to the difference $\psi - \theta^{\mathrm{stim}}$ (Eq. 14) which leads to a partly discounting of the previous estimate of the stimulus average (leaky integration).

To fully understand the recency effect, it is necessary to consider how the running average of the stimulus directions can be tracked optimally. Mathematically, we want to keep track of the circular mean of the time-varying stimulus direction $\theta^{\mathrm{stim}}(t)$, an operation involving the computation of the vector sum (Methods). We derived a dynamical system that optimally tracks the cumulative circular average, the *perfect vector integrator (PVI)*. The corresponding equations (Eqs. 21a and 21b in Methods) turn out to be identical to the amplitude equation but without the intrinsic amplitude dynamics (the two terms depending on R in Eq. 1a). To gain an intuition on how the PVI tracks the mean orientation, we can again consider the movement of a particle in a potential which for the PVI takes the form of an infinite plane going through the origin and inclined towards θ^{stim}

$$\Theta^{\mathrm{PVI}}(R, \psi) = -RI \cos\left(\psi - \theta^{\mathrm{stim}}\right). \tag{3}$$

In response to each stimulus the particle will move in both the radial R direction and the angular ψ direction, thus integrating stimulus vectors defined by their amplitude I and their direction θ^{stim}. This is in contrast to the bump attractor network with a formed bump in which the bump amplitude is fixed. How does R evolve over time in the PVI? This depends on the particular combination of the directions of stimulus frames in each trial, but, on average, R will grow over time because of a net movement in the direction of the mean of the stimulus distribution $\bar{\theta}^{\mathrm{stim}}$ (Fig. S3A, B shows how the growth rate depends on the width of the stimulus distribution). Because R is growing, the angular updates become smaller with each stimulus frame (in other words, as the number of samples grows, the running direction average increasingly relies on the already accumulated information). The bump attractor network, in contrast, gives equal weight to the angular updates over time, thus partly erasing the accumulated stimulus information while over-weighting late stimulus information. The strength of the recency effect depends on the bump amplitude. As described in the previous section, the smaller is I_0 and thus the bump amplitude, the larger is the angular displacement for a given stimulus strength I_1 and the faster the bump can track the orientation of incoming stimuli, resulting in a PK with increasing recency effect (Fig. 3A,C; Fig. S1).

Strikingly, when we ran simulations in which the network had to estimate the average stimulus direction starting the trials in the homogeneous network state, with no bump formed (i.e., the particle is at the unstable center position on the potential; point ① in Fig. 2), we observed three qualitatively distinct integration regimes depending on the global excitatory drive to the network I_0 (Fig. 3B).

First, small I_0 yields again a recency PK (Fig. 3B, left). In this regime, the steady-state bump amplitude is relatively small and the bump can only grow very slightly over time (Fig. 3D, red line). Thus, stimulus integration is limited by the small bump amplitude and the network behaves similarly to the previously considered case with the bump formed from the beginning of the trial and with constant amplitude throughout the trial (Fig. 3A, left).

Second, for large I_0, the networks showed a "primacy" effect, that is, early stimulus information now had a higher impact on the final estimation than later stimulus information (Fig. 3B, right). This can be understood by taking into account that in this regime the bump amplitude is small at the beginning of the trial but then grows to a large amplitude (Fig. 3D, blue line). Thus, the bump can quickly follow the stimulus direction early in the trial during the formation of the bump, but while growing it becomes increasingly sluggish and stimulus information becomes comparably much less effective in displacing the bump. Compared to the PVI, the bump amplitude grows too quickly because it is driven by the internal dynamics in addition to the stimulus (i.e., the potential is not flat as for the PVI but has a peak in the center; Fig. 3A, right inset). The resulting down-weighting of later stimulus frames only occurs while the bump is growing (transition from ① to ② in Fig. 2). Once the steady-state bump amplitude has been reached (corresponding to the attractor well of the potential), the model will start to overweight incoming evidence and, as a consequence, for very long stimulus durations the model will eventually show a recency PK (Fig. 3D, blue line).

Third, for intermediate values of I_0, primacy and recency effects balance out and we obtained an almost perfectly uniform PK (Fig. 3B, middle panel). Thus, in this regime the directions of all stimulus frames had the same impact on the final estimate over the whole trial (PK slope $= 0$; Fig. 3C). In this case, the potential has only a relatively small peak at the center, approximating the flat potential of the PVI (Fig. 3A, middle inset). The growth rate of the bump amplitude is mainly driven by the stimulus as in the PVI (Fig. S3B). The value of I_0 in which the model best approximates the PVI depends on the stimulus amplitude and the distribution of the stimulus directions (Fig. S3C). Approximately perfect integration is only possible as long as the bump amplitude can increase (before the well of the potential is reached), such that for long stimulus durations integration becomes again leaky and we obtain again a recency PK (Fig. 3E, black line). Thus, both primacy and uniform PKs are robustly observed for small and intermediate stimulus durations (Fig. 3E; $T \leq 2$ s) while for long durations the recency regime is recovered. The results in Fig. 3 were obtained with time-varying stimuli that change their orientation every 250 ms (frame duration). The PKs remain qualitatively the same when keeping the stimulus duration constant and presenting a variable number of stimulus frames of different duration (Fig. 3F).

Accuracy in stimulus estimation and categorization

Since the continuous attractor model showed different evidence integration regimes when starting the integration without an initial bump of activity, we next quantified the impact of the integration regimes on estimation and choice accuracy. We computed stimulus estimation curves, by reading out the bump phase from the network activity at the end of the trial, averaging the estimates across trials and plotting them as a function of the actual mean orientation $\bar{\theta}^{\mathrm{stim}}$ (Fig. 4A). We noticed that independent of the bump amplitude, that was controlled by varying the excitatory drive I_0, the average of the estimates was always close to the true mean direction. We quantified this by computing the estimation bias for $\bar{\theta} = 45°$ and obtained $b_{\mathrm{est}}(45°) < 1°$ for all values of I_0. Consistent with this, the slopes of the estimation curve obtained from regression analysis (Methods) were $k_1 \approx 1$ in all cases. The estimates computed by the ring attractor network were thus approximately unbiased, independent of the temporal weighting of stimulus evidence. The unbiased estimates are a direct consequence of starting the integration process in the neutral center of the potential (Fig. 2, point ①) and the symmetry of the potential (Fig. 2). The direction of the first stimulus frame determines the initial phase of the forming bump in an unbiased manner. Subsequently, the bump will move, but since the directions of the stimulus frames are drawn i.i.d. from some underlying distribution, clockwise and counter-clockwise bump displacements are equally likely and the estimate always remains unbiased independent of the integration regime (Fig. 5D, red line).

However, the trial-to-trial variability of the estimates (Fig. 4A, shaded areas and Fig. 4C, black line) depended on the integration regime. Several factors could give rise to estimation errors in the attractor model and may thus contribute to this dependence: (1) internal noise and stimulus fluctuations of magnitude σ (Methods), (2) sub-optimal evidence integration due to non-uniform evidence integration (i.e., non-uniform PKs), and (3) additional impact of the non-linear internal dynamics on evidence integration not captured by the PK. We first considered the impact of the noise. We ran simulations of

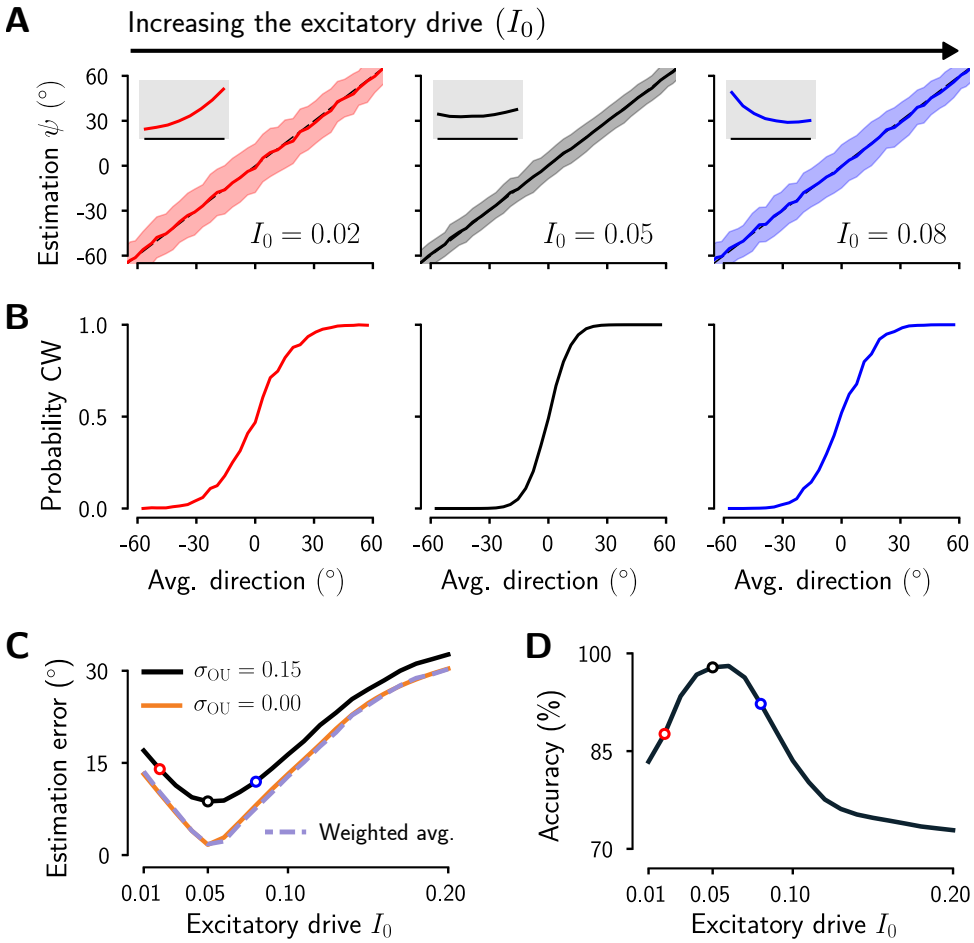

Figure 4: Dependence of estimation and categorization accuracy on the global excitatory drive $\mathbf{I_0}$.

(A) Estimated average direction as a function of the true average of a stimulus composed of 8 oriented frames, for increasing values of I_0 (from left to right, $I_0 = 0.02, 0.05$ and 0.08). Solid lines indicate the average across trials and the shaded regions correspond to the standard deviation. Insets: corresponding PKs from Fig. 3B. **(B)** Probability of a clockwise choice versus average stimulus orientation, for increasing values of I_0. A clockwise/counterclockwise choice corresponds to a positive/negative phase of the bump at the end of the trial. **(C)** Root-mean-square error (Methods) of the ring attractor network with noise ($\sigma_{\mathrm{OU}} = 0.15$) and without noise ($\sigma_{\mathrm{OU}} = 0$) as a function of the global excitatory drive I_0. The estimation error for $\sigma_{\mathrm{OU}} = 0.15$ (black line) corresponds to the estimations shown in (A). The estimation error for a perfect weighted average of the stimulus directions (with the same temporal weighting as the network model but otherwise optimal; Methods) is shown for comparison. Note that the curve corresponding to $\sigma_{\mathrm{OU}} = 0$ has been shifted such that for both conditions the critical value of I_{exc} at the bifurcation point is aligned (see Methods). **(D)** Categorization accuracy as a function of the excitatory drive I_0 (for a fixed average stimulus orientation $\bar{\theta}^{\mathrm{stim}} = 23°$).

the network without noise ($\sigma_{\mathrm{OU}} = 0$) and obtained a general decrease of the estimation errors (Fig. 4C, orange line). We next tested to what degree the remaining estimation errors can be explained by the impact of non-uniform evidence integration in the primacy and recency regimes, that is, the expected decrease in estimation accuracy when not all stimulus frames are taken into account with an equal weight. To show this we compared the dependency of the estimation error on I_0 of the attractor model with the estimation error that was obtained when computing the true average of the stimulus directions with the PVI but weighted with the PKs from the bump attractor model (Fig. 4C, purple dashed line). We found that the obtained estimation errors are almost indistinguishable from the errors of the noise-less bump attractor network (Fig. 4C). Thus, the increase of estimation errors for non-optimal I_0 (corresponding to values of I_0 different from 0.05 in Fig. 4C) can be attributed to the suboptimal, non-uniform, temporal weighting captured by the PK. Because the non-linear internal dynamics of the ring model can only yield an approximately uniform PK, the estimation errors are non-zero even for the best I_0 for a given stimulus distribution, in particular when the stimulus distribution is wide (Fig. S3C). Together, these analyses show that the estimation accuracy in the bump attractor network is limited by noise and by the non-linear internal dynamics causing non-uniform evidence integration, whereas the further contributions of non-linearities (that are not captured by the PK) are negligible.

A direct consequence of the trial-to-trial variability of the estimates is that they limit the accuracy of a categorical decision that is based on the read-out of the bump phase. This can be seen when computing psychometric curves by interpreting a positive bump phase at the end of the trial as a clockwise choice and a negative bump phase as a counter-clockwise choice, respectively (Fig. 4B). As expected from the stimulus estimation results, the psychometric curve for the uniform integration regime had the highest slope (Fig. 4B, middle panel). Indeed, decision accuracy is maximal for an I_0 that gives rise to a uniform PK (Fig. 4D).

Impact of the initial condition on stimulus estimation.

We next investigated in more detail the accuracy of stimulus estimation in the network with an initially formed bump (Fig. 3A), in particular the impact of the initial bump phase on the estimation process. In psychophysical experiments this initial bump position could be determined by a reference line shown before the onset of the stimulus (Jazayeri and Movshon, 2007; Talluri et al., 2018) or it could be related to the subject's prior expectation before the start of the stimulus (Aitken et al., 2020). Single trials are illustrative of the effect of the different initial conditions (Fig. 5A,B). When starting a trial in the homogeneous network state, the first stimulus frame determines the phase where the bump emerges (Fig. 5A) and, as the trial progresses, the network integrates the stimulus and the bump phase moves successively closer to the true average direction. Across trials, the network's estimates were unbiased (Fig. 5C,D; red line), as already described above. In contrast, when starting with a fully formed bump as initial condition, the bump phase was biased towards the initial phase of the bump (Fig. 5B,C). As the network integrates the stimulus, the bump phase moves successively closer to the true average orientation, leading thus to a decrease of the systematic deviations from the true average with stimulus duration (Fig. 5C, blue lines). The magnitude of the estimation bias for a fixed stimulus duration of 2 s increased with the global excitatory drive I_0 to the network (Fig. 5D). The reason is that a higher I_0 yields larger bumps that are more sluggish and thus need more time to overcome the bias caused by the initial bump position. While the variability of the estimates in the biased case is generally lower than in the unbiased case, the overall estimation error (bias + variance) is higher (Fig. 5E). To confirm that the bias is actually caused by the pre-determined phase of the activity bump at the beginning of the trial and not a consequence of starting the trials with a fully formed bump, we ran simulations with a fully formed initial bump with a phase determined by the first stimulus frame and indeed observed no bias (Fig. S4). In sum, starting with an initial bump led to recency PKs (Fig. 5F) and, when setting the initial phase to some pre-determined value it yielded estimates that were biased towards the initial bump phase.

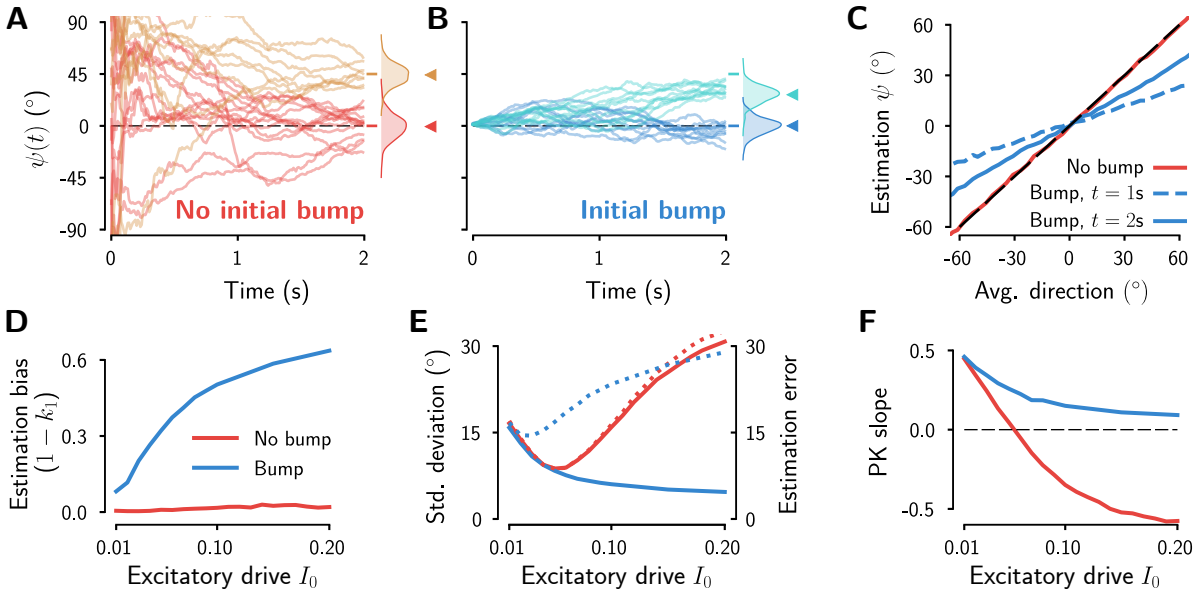

Figure 5: Impact of the initial condition on the bias and variance of the stimulus estimates. **(A)** Bump phase in example trials with 0° and 45° average stimulus directions ($N = 8$ stimulus frames, stimulus duration $T = 2$ s) when starting the trial with homogeneous network activity (i.e., no bump formed). Inset: distribution of the phase of the bump at the end of the trial. **(B)** Bump phase for the same network and the same stimuli as in (A), but starting the trial with a fully formed bump centered at 0°. **(C)** Continuous estimations as a function of the mean stimulus direction $\bar{\theta}^{\text{stim}}$ for two different stimulus durations (no bump condition in red, bump condition in blue). For the no-bump condition, the average estimation curve does not depend on stimulus duration (red line) and it overlaps with the diagonal (dashed black line). **(D)** Systematic bias of the estimations as a function of the global excitatory drive I_0 to the network (stimulus duration $T = 2$ s). The estimation bias is defined as one minus the slope of the estimation curve (k_1 in Eq. 28, Methods). **(E)** Standard deviation (solid lines, left axis) and root-mean-square estimation error (dotted lines, right axis) of the final bump position as a function of the global excitatory drive I_0 (same trials as in D). **(F)** Slope of the psychophysical kernel versus the global excitatory drive I_0.

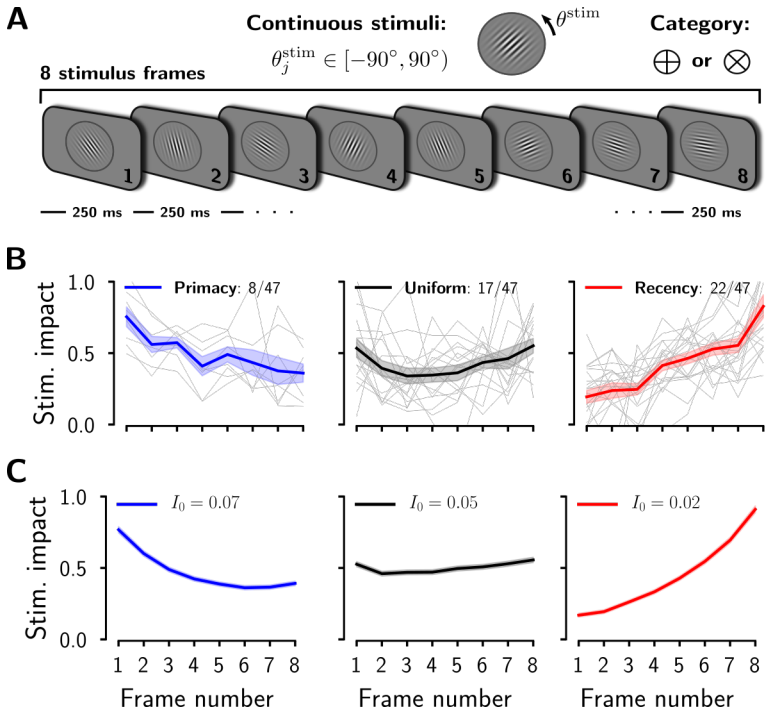

Figure 6: The ring attractor model accounts for experimentally observed psychophysical kernels. **(A)** Category-level averaging task (Wyart et al., 2012; Cheadle et al., 2014; Wyart et al., 2015). Participants reported whether, on average, the tilt of eight oriented Gabor patterns fell closer to the cardinal or diagonal axis. **(B)** PKs obtained from human subjects were heterogeneous. Thin lines represent the PKs of individual subjects and colored lines the group averages. **(C)** PKs obtained by fitting the ring model to the experimental data by varying the excitatory drive I_0.

The ring model explains heterogeneity in integration dynamics in humans

Finally, we tested whether the ring model could account for different integration dynamics in human subjects performing a categorization task (Fig. 6). In this category-level averaging task, Fig. 6A (Wyart et al., 2012; Cheadle et al., 2014; Wyart et al., 2015), a visual stream of eight oriented Gabor patters were presented to participants. Each frame had an orientation between $-90°$ and $90°$ and a fixed duration of 250 ms, conforming a stimulus with total duration of $T = 2$ s. At the end of the stream participants reported whether, on average, the tilt of the eight frames fell closer to the cardinal $\{0, 90\}$ or diagonal $\{-45, 45\}$ axis. On average, participants were able to combine the evidence favouring either option improving their accuracy in trials with higher evidence (not shown).

We merged the data from three different experiments and classified subjects according to their temporal weighting behavior (primacy, uniform, recency) by computing their individual PK (Methods). As previously reported (Cheadle et al., 2014), across participants there was a tendency for late temporal weighting (recency). However, subject-by-subject analysis revealed a broad range of integration dynamics (Fig. 6B). The majority of participants weighted more the sensory evidence during the late periods of the stimulus (22/47 participants), yet a substantial proportion of the participants weighted the evidence approximately uniformly (17/47 participants). Finally, a minority of participants (8/47 participants) weighted early frames of the stimulus more that late frames, showing a primacy PK. Qualitatively, the ring model could capture the different temporal integration dynamics observed in the psychophysical experiments (Fig. 3). To asses whether the model could explain the data in a quantitative manner, we performed simulations with the same stimulus statistics as in the behavioral experiment. We systematically adjusted both the global excitatory drive to the network, I_0, and the noise level, σ_{OU}, such that the ring network reproduced, on average, the experimental results (Fig. 6B,C), characterized by the slope of the across-subjects PKs (Fig. 6B) and the average performance (not shown) for each integration regime (primacy, uniform and recency). The inter-subject heterogeneity in temporal evidence weighting could thus parsimoniously be explained by varying a single parameter in the ring model, the overall excitatory drive that determines the amplitude of the bump.

Discussion

We investigated how the classical continuous ring attractor model (Ben-Yishai et al., 1995; Zhang, 1996; Compte, 2000) integrates a time-varying stimulus in the amplitude and the phase of the network's activity bump. We found that the model can approximately compute the time integral of stimulus vectors defined by the amplitude and the direction of the stimulus. This operation can be understood in terms of a population vector (Georgopoulos et al., 1986). But while a population vector is typically used to read out direction information from a neural population (by weighting the activity of individual neurons according to their preferred direction), the population vector encoded in the amplitude and the phase of the activity bump results from a temporal integration process and is continuously updated in response to incoming stimulus evidence.

We studied the neural network mechanism underlying this non-trivial integration process in detail and identified how the model integrates a sequence of stimulus frames of varying direction and constant amplitude, e.g. a sequence of visual gratings with different orientation but the same contrast. Through simulations we confirmed that the phase of the activity bump tracks the running circular average of the stimulus orientations (for variable stimulus amplitudes we would get a weighted circular average). Concurrently, the amplitude of the bump evolves depending on the dispersion of the stimulus directions (Fig. S3B). The dynamical regime in which the model closely approximates perfect vector integration depends on the stimulus amplitude and the stimulus statistics (Fig. S3C). It is characterized by a relatively wide and shallow potential so that the dynamics of the bump is dominated by the stimulus and not by the non-linear internal dynamics (Fig. 3). The limiting factor imposed by the internal dynamics of the network is that the bump amplitude cannot grow beyond a maximum value that depends on the model parameters and is proportional to the overall excitatory drive to the network. When the maximum bump

amplitude is reached, which will eventually happen for long stimulus durations, the model has reached the neutrally stable ring attractor and can still track the average stimulus phase. However, the model is then over-weighting later stimulus frames, similar to a leaky integrator. In addition to this recency effect, the model can also show primacy temporal weighting when the internal network dynamics contributes to a rapid initial growth of the bump and initial stimulus frames have a relatively higher impact on the evolution of the bump phase. Independent of the temporal weighting regime, the orientation estimates of the model were always unbiased; the variance of the estimates is however minimal when the model acts as perfect vector integrator and increases for non-uniform temporal weighting (Fig. 4).

In sum, we have uncovered the fundamental mechanism underlying stimulus integration in the phase and the amplitude of the bump of the ring attractor model and the corresponding computation of the running circular mean. Our results suggest ring attractor dynamics as a potential neural mechanism for the estimation of average circular stimulus features (such as the average stimulus direction) and for perceptual discrimination based on the same features.

The role of the global excitatory drive

A key parameter that controls the integration dynamics in the bump attractor model is the global excitatory drive to all neurons in the network, I_0 (Fig. 3). The reason is that, keeping all other parameters fixed, I_0 determines the depth and the width of the potential (Fig. S1) and thus the maximal bump amplitude and the intrinsic network dynamics. Just by varying I_0, the model could account for the heterogeneity in how human observers weighted sensory evidence across a stream of oriented stimulus frames (Fig. 6). While different subjects may indeed differ in their level of global excitatory drive to the circuit involved in evidence integration, their neural circuits are likely to show other differences as well, for example a different E/I ratio or different neuronal gain. Effectively, in the model these inter-subject differences can be captured by different values of I_0. In fact, a modulation of other parameters can have qualitatively similar effects in the model, in particular a modulation of the neural gain changing the neuronal transfer function $\Phi(\cdot)$. Beyond potential subject-to-subject differences, the global cortical gain can also be modulated across time in individual subjects, for example controlled by neuromodulators such as noradrenalin originating from the locus coeruleus (Aston-Jones and Cohen, 2005; Eckhoff et al., 2009; Niyogi and Wong-Lin, 2013). A time-varying neural gain or excitatory drive in the model can switch the network from an integration regime into a working memory regime. During stimulus estimation, the excitatory drive should be moderate so that the network would have a relatively shallow potential and would act as a perfect vector integrator. Subsequently, after stimulus presentation, an increase of I_0 would make the potential wider and deeper and thus increase the stability of the bump against noise and against further incoming stimuli as is needed for distractor-resistant working memory (Compte, 2000; Murray et al., 2017). This transition could also be gradual, for example caused by a ramping I_0 (Finkelstein et al., 2019). Moreover, it has been shown in combined discrimination and estimation tasks that stimulus estimation is influenced by a categorical decision causing post-decision biases (Jazayeri and Movshon, 2007; Luu and Stocker, 2018; Talluri et al., 2018, 2020). One hypothesis is that these bias effects are mediated through attention signals and through changes in global gain (Talluri et al., 2018, 2020). Our network model provides a comprehensive computational framework for investigating the neural mechanisms underlying stimulus estimation and perceptual categorization and their interaction in future studies.

Comparison with other models

The investigation of neural networks that can integrate a continuous stimulus feature has a long history (Goldman et al., 2009). These neural integrators are usually categorized as being either rate based (Cannon et al., 1983; Usher and McClelland, 2001; Machens et al., 2005), accumulating the signal in the graded firing rate of the neurons, or location based, in which the integrated stimulus is encoded in the localized activity of the neurons (Zhang, 1996; Boucheny et al., 2005; Song and Wang, 2005). The models in the latter category studied the encoding of the head direction in the phase of an activity bump. While this representation is the same as we have studied here in the context of estimating the average

direction of a sensory stimulus, its updating mechanism followed a different integration process. Namely, integrating self-motion information in the form of angular head velocity signals, and using visual cues only to calibrate the head direction system, by rotating the entire bump attractor manifold towards the direction of the visual cue. In contrast, here we focused our attention on the integration of a time-varying stimulus as the basic operation to estimate its average direction. Crucially, we have shown that this computation is mediated by the transient interplay between the location of the bump and also its firing rate amplitude, a feature that, to the best of our knowledge, has remained unexplored until now.

One previous work has studied the identification of a noisy oriented stimulus in a bump attractor network and how the stimulus representation is affected by perturbations mimicking a cortical microstimulation experiment (Liu and Wang, 2008). While the network architecture is similar, their network operates in a different regime in which two competing localized bumps can emerge, due to the visual stimulus and the microstimulation, giving rise to winner-take-all or averaging depending on the orientation difference of the two bumps. This is similar to the effect of distractors in working memory models (Compte, 2000) and different from the computation of the average orientation of a time-varying stimulus in a single bump that we studied here.

The bump attractor model is a generalization of the double-well model of perceptual decision making

Our continuous bump attractor model can be viewed as a generalization of the discrete attractor model of perceptual decision making (Wang, 2002; Wong and Wang, 2006; Roxin and Ledberg, 2008; Prat-Ortega et al., 2021). This can be seen by considering input stimuli $I_{\mathrm{stim}}(t)$ of an arbitrary fixed orientation θ^{stim} and a time-varying amplitude $I_1(t)$. If we disregard the noise term, Eq. 1b reduces to $\frac{\partial \psi}{\partial t} = 0$, that is, a stationary bump will form at $\psi = \theta^{\mathrm{stim}}$. The evolution of the bump is then fully described by changes of its amplitude (Eq. 1a). By assuming that the stimulus amplitude $I_1(t)$ can take positive and negative values, this equation can yield either a positive R, corresponding to a bump at θ^{stim}, or a negative R, corresponding to a bump at $\theta^{\mathrm{stim}} + 180°$. Formally, the bump amplitude obeys then exactly the same dynamics as the decision variable in the winner-take-all double well model (Roxin and Ledberg, 2008; Prat-Ortega et al., 2021). Intuitively, the one-dimensional double well potential can directly be obtained from the two-dimensional juice squeezer potential for a fixed phase ψ (Fig. 2A, white lines).

However, in general evidence integration in the two models is different. Here, we found that the bump attractor model can compute the average stimulus direction in the phase of the bump with a variety of temporal weighting profiles, including primacy, uniform and recency regimes (Fig. 3B). In the double-well model we have recently described similar temporal weighting profiles for a categorical decision (Prat-Ortega et al., 2021) that rely, however, on a different mechanism than in the bump attractor model. The two attractor states of the double well model impose a categorization dynamics that leads to a primacy PK because once the system settles into one of the attractors it remains there until the end of the trial. More uniform and recency PKs are obtained when fluctuations in the stimulus together with the internal noise are strong enough to overcome the categorization dynamics and cause transitions between the attractor states (Prat-Ortega et al., 2021). In contrast, in the bump attractor model a continuous integration process – without abrupt transitions – occurs in the neutrally stable ring manifold and the stimulus always always impacts the bump phase, yielding recency without the need for strong fluctuations (Fig. 3A).

Low-dimensional ring attractor dynamics in neuronal circuits

Is bump attractor dynamics realized in a cortical circuit a plausible neural mechanism for stimulus estimation? A bump attractor realized in a physical ring layout has recently been found in the fly's heading direction system (Seelig and Jayaraman, 2015; Green et al., 2017; Kim et al., 2017). Taking advantage of simultaneous recordings of multiple neurons it has been confirmed that the high-dimensional neural population activity in the mammalian head direction system evolves along a one-dimensional ring attractor manifold (Chaudhuri et al., 2019) consistent with the dynamics of ring models (Skaggs et al.,

1995; Zhang, 1996; Boucheny et al., 2005). We hypothesize that a similar topological ring manifold may be embedded in the neural population activity of cortical areas that are involved in stimulus estimation. In order to test experimentally whether bump attractor dynamics underlies orientation averaging, it would be interesting to record neural activity from parietal and frontal areas of non-human primates that perform an estimation task and apply manifold identification approach as in Chaudhuri et al. (2019). Such data would also allow for testing whether trial-to-trial variability of neural firing rates and the estimates reported by the subject are related as predicted by the bump attractor model, as in prefrontal cortical neurons during spatial working memory (Wimmer et al., 2014). In fact, our work suggests that the same neural circuits that are involved in working memory may also be able to carry out stimulus averaging in perceptual estimation tasks. Bump attractor dynamics may thus be a unifying neural mechanism underlying both working memory and evidence integration over prolonged timescales.

Methods

Ring model

The dynamics of the bump attractor model are described in terms of the effective firing rate, $r(\theta, t)$, of a neural population arranged in a ring, $\theta \in [-\pi, \pi)$ (Fig. 1A), obeying the integro-differential equation (Ben-Yishai et al., 1995; Hansel and Sompolinsky, 1998; Ermentrout, 1998)

$$\tau \frac{\partial r}{\partial t} = -r + \Phi\left(\frac{\tau}{2\pi}\int_{-\pi}^{\pi} w\left(\theta - \theta'\right) r\left(\theta', t\right) d\theta' + I_{\text{exc}} + I_{\text{stim}}\left(\theta, t\right) + \xi\left(\theta, t\right)\right), \tag{4}$$

where τ is the neural time constant and $\Phi(\cdot)$ is the current-to-rate transfer function. The synaptic input to a neuron with preferred orientation θ consists of a recurrent current due to the presynaptic activity at a location θ' with a weight $w(\theta - \theta')$, and external currents $I_{\text{exc}} + I_{\text{stim}}(\theta, t)$ plus fluctuations $\xi(\theta, t)$. The connectivity profile $w(\theta)$, is written in terms of its Fourier coefficients, w_k $(k = 0, 1, 2, \dots)$ and represents the effective excitatory/inhibitory coupling, and can therefore include both positive and negative interactions. The inset in Fig. 1A shows an example of Mexican-hat type connectivity with strong recurrent excitation and broad inhibition. External inputs are divided into a global net excitatory drive I_{exc} that modulates the excitability of the network and a time-varying input I_{stim} that represents the sensory stimulus. In general, the sensory stimulus can be written in terms of its Fourier coefficients I_k, $k = 1, 2, \dots$ with directions $\theta_k(t)$. Throughout this work we will focus on an input of the form $I_{\text{stim}}(\theta, t) = I_1 \cos\left(\theta - \theta^{\text{stim}}(t)\right)$ to model a stimulus with constant amplitude I_1 and a time-varying orientation $\theta^{\text{stim}}(t)$ but our derivations are equally valid for time-varying stimulus amplitudes I_1. A detailed description of the stimuli used in the simulations is given below (Model simulations). Finally, the fluctuation term $\xi(\theta, t)$ reflects the joint effect of the internal stochasticity of the network and temporal variations in the stimulus. For simplicity, we model these fluctuations as independent Ornstein-Uhlenbeck processes for each neuron, with amplitude $\sigma_{\text{OU}} = 0.15$ and time constant $\tau_{\text{OU}} = 1$ ms. Tab. 1 summarizes the values of the model parameters used in this work.

The effective network, Eq. 4, assumes that excitatory and inhibitory neurons follow the same dynamics, and thus can be grouped together, greatly simplifying the mathematical analysis. Nevertheless, a mathematical analysis of the more general model with separated pools of excitatory and inhibitory neurons is possible and is expected to give the same results (Tass, 1995; Bressloff, 2012; Esnaola-Acebes et al., 2017; Bressloff, 2019).

The bump attractor

In the following we describe the conditions for which the ring network shows a stable bump of activity. We consider the quadratic/square root transfer function (Brunel, 2003; Boucheny et al., 2005)

$$\Phi(x) = \frac{1}{\tau} \cdot \begin{cases} 0 & , \quad x < 0 \\ x^2 & , \quad 0 \le x \le 1 \\ 2\sqrt{x - 3/4}, & \quad x > 1 \end{cases} \tag{5}$$

where x is the total synaptic input arriving to the neurons. Other continuous non-linear transfer functions will lead to the same qualitative behavior provided they are a monotonically increasing function of the input.

In the absence of sensory stimuli, $I_{\text{stim}} = 0$, the ring network can show different spatial activity profiles depending on the form of the connectivity profile and the global excitatory drive I_{exc}. For $I_{\text{exc}} \leq 0$ the network has a unique equilibrium solution with zero activity, $r(\theta) = 0$ Hz. Increasing the excitatory drive, $I_{\text{exc}} > 0$, will increase the activity of this equilibrium solution uniformly along the ring $r(\theta) = r_\infty = \Phi(\tau w_0 r_\infty + I_{\text{exc}})/\tau > 0$. By studying the linear stability of this solution we can find a critical value of the global excitatory drive, $I_{\text{exc}} = I_{\text{crit}}$, for which the spatially homogeneous state destabilizes into a spatially inhomogeneous state via a pattern forming instability known as Turing bifurcation. This new stable state will have a single bump of activity as shown in Fig. 1B for a connectivity profile that satisfies $w_1 > w_k$ for all $k \neq 1$. For our choice of the transfer function (Eq. 5) and the connectivity profile (see Tab. 1) this critical value is given by

$$w_1^{\text{crit}} = \frac{1}{\tau \Phi'_\infty} \Rightarrow I_{\text{crit}} = \frac{1}{4w_0}\left[1 - \left(\frac{w_0 - w_1}{w_1}\right)^2\right] = 1, \tag{6}$$

where $\Phi'_\infty = \Phi'(\tau r_\infty w_0 + I_{\text{crit}})$ is the slope of the transfer function evaluated at the bifurcation point.

Table 1: List of the parameters of the ring network and the stimulus.

Ring network (Eq. 4)	*Symbol*	Value
Time constant of the network	τ	20 ms
Fourier modes of the connectivity	w_k, $k = 0, 1, 2$	$-2, 1, 0.5$
Global net excitatory drive	I_{exc}	$\in [1.01, 1.2]$
Noise term	$\xi(\theta, t)$	Eq. 4
Noise amplitude	σ_{OU}	0.15
Noise time constant	τ_{OU}	1 ms
Transfer function	$\Phi(\cdot)$	Eq. 5
Critical excitatory drive	I_{crit}	1
Stationary FR at $I_{\text{exc}} = I_{\text{crit}}$	r_∞	12.5 Hz
Stimulus (Eqs. 23 and 24)	*Symbol*	Value
Stimulus amplitude	I_1	$5 \cdot 10^{-3}$
Duration of the stimulus	T	2 s
Number of stimulus frames	N	8

Amplitude equation

The bump solution described in the previous section is analytically intractable. However, we can study its dynamics near the onset of the bump, i.e. in the vicinity of the bifurcation point $I_{\text{exc}} = I_{\text{crit}}$, by performing a multiple-scale analysis based on a perturbation method (Kuramoto, 1984). We chose I_{exc} as our bifurcation parameter and define the perturbation

$$I_{\text{exc}}(\theta) = I_{\text{crit}} + \varepsilon^2 I_0 + \varepsilon^3 2 I_1 \cos\left(\theta - \theta^{\text{stim}}\right), \tag{7}$$

and the corresponding expansion of the firing rate

$$r = r_\infty + \sum_{n=1}^{\infty} \varepsilon^n r_n. \tag{8}$$

The multiple scale method assumes that the dynamics of the network can be separated into two independent time scales, such that we can define the slow time $T = \varepsilon^2 t$ in which the dynamics of the bump evolve. Plugging all these definitions into the equation of the ring network, Eq. 4, we obtain the equation that describes the evolution of the amplitude of the bump, $A(t)$, close to the bifurcation point:

$$\tau \frac{\partial A}{\partial T} = \tilde{I}_1 e^{-i\theta^{\text{stim}}} + \tilde{I}_0 A - c\,|A|^2 A + \zeta(t). \tag{9}$$

where $\tilde{I}_1 = \Phi'_\infty I_1$ is the scaled stimulus amplitude, $\tilde{I}_0$ corresponds to the scaled distance to the bifurcation point, which is related to the global excitatory drive:

$$\tilde{I}_0 = \frac{\Phi''_\infty}{\Phi'_\infty\left(1 - \tau w_0 \Phi'_\infty\right)} \left(I_0 - 2\sigma^2\right), \tag{10}$$

and the cubic factor c is

$$c = -\left[\frac{\left(\Phi''_\infty\right)^2}{\left(\Phi'_\infty\right)^3}\left(\frac{\tau w_0}{1 - \tau w_0 \Phi'_\infty} + \frac{\tau w_2}{2\left(1 - \tau w_2 \Phi'_\infty\right)}\right) + \frac{\tau^2 \Phi'''_\infty}{2\left(\Phi'_\infty\right)^3}\right]. \tag{11}$$

The last term in eq. 9 corresponds to the first Fourier component of the noise process $\xi(\theta, t)$ in Eq. 4, that can be written as

$$\zeta(t) = \frac{\Phi'_\infty}{\sqrt{2n}}\left(\xi_a(t) + i\xi_b(t)\right) = \frac{\Phi'_\infty}{\sqrt{2n}}\sqrt{\xi_a^2 + \xi_b^2}\; e^{-i\cdot \text{atan2}(\xi_b, \xi_a)}, \tag{12}$$

where, n is the discrete number of Fourier modes of the noise process $\xi(\theta, t)$, and $\xi_a(t)$ and $\xi_b(t)$ are independent Ornstein-Uhlenbech processes with the same amplitude, σ_{OU}, and time constant, τ_{OU} as $\xi(\theta, t)$. Additionally, the non-linear filtering of the noise process, $\xi(\theta, t)$, through the transfer function, introduces spatial correlations that effectively shift the bifurcation point to $I_{\text{crit},\sigma} = I_{\text{crit}} + 2\sigma_{\text{OU}}^2$, which explains the correction term in Eq. 10 (Hutt et al., 2007, 2008).

Finally, by considering the polar form of $A(t) = Re^{-i\psi}$, Eq. 9 results in the system described by Eqs. 1a and 1b with the noise terms ξ_1 and ξ_2 being

$$\xi_1 = \frac{\Phi'_\infty}{\sqrt{2n}}\sqrt{\xi_a^2 + \xi_b^2}\cos\left(\psi - \text{atan2}\left(\xi_b, \xi_a\right)\right),\; \xi_2 = \frac{\Phi'_\infty}{\sqrt{2n}}\sqrt{\xi_a^2 + \xi_b^2}\sin\left(\psi - \text{atan2}\left(\xi_b, \xi_a\right)\right). \tag{13}$$

The values of the parameters used for the simulation of the amplitude equation are listed in Tab. 2.

The amplitude equation (Eq. 9) derived here corresponds to the normal form of a Turing bifurcation where the first spatial mode of the spatially homogeneous state becomes unstable. Similar derivations can be found in the literature (see for example, Tass, 1995; Hansel and Sompolinsky, 1998; Ermentrout, 1998), which also study the general case of separated excitatory-inhibitory populations.

Dynamics of the phase for a stationary bump amplitude

An analytical expression for the approximate evolution of the phase, $\psi(t)$, can be obtain by assuming an almost static bump amplitude ($\dot{R} \sim 0$), and then solving the ordinary differential equation Eq. 1b,

$$\psi(t) = \theta^{\text{stim}} - 2\arctan\left[\tan\left(\frac{\theta^{\text{stim}} - \psi_0}{2}\right) e^{-\frac{\tilde{I}_1}{\tau R_\infty} t}\right], \quad \psi(0) = \psi_0, \quad \psi_\infty = \theta^{\text{stim}}, \tag{14}$$

where ψ_0 is the initial phase of the bump, and the stationary amplitude, R_∞, is obtained by solving the third order polynomial that results from equating Eq. 1a to zero. The solution to this polynomial equation

Table 2: List of the parameters of the amplitude equation, and parameters of the numerical integration, used to simulate both Eqs. 1 and 4.

Amplitude equation (Eq. 1)	*Symbol*	Value
Additional excitatory drive, $I_{\text{exc}} - I_{\text{crit}}$	I_0	$\in [0.01, 0.2]$
Slope of $\Phi(\cdot)$ at $I_{\text{exc}} = I_{\text{crit}}$	Φ'_∞	50
2nd derivative of $\Phi(\cdot)$ at $I_{\text{exc}} = I_{\text{crit}}$	Φ''_∞	100
3rd derivative of $\Phi(\cdot)$ at $I_{\text{exc}} = I_{\text{crit}}$	Φ'''_∞	0
1st order prefactor	$\tilde{I}_0$	$\in [-7/300, 31/300]$
Cubic prefactor	c	$1/3750$
Scaled stimulus amplitude	$\tilde{I}_1$	0.125
Noise terms	$\xi_i(t)$, i=1,2	Eq. 13
Numerical integration	*Symbol*	Value
Integration time step	Δt	0.2 ms
Spatial discretization	n	200
Number of trials	K	10^5

depends on the value of $\tilde{I}_1$, which if set to 0, gives $R_\infty = \sqrt{\frac{\tilde{I}_0}{c}}$. Fig. S2A shows the evolution of the phase of the bump for an initial condition in which the bump is completely formed and centered at 0° for a stimulus $\theta^{\text{stim}} = 90°$. The approximation (shaded lines), Eq. 14, is plotted along with the corresponding simulation of the amplitude equation showing good agreement. As expected, the approximation works better for larger values of $\tilde{I}_0$, that make the amplitude of the bump less sensitive to the stimulus.

A first order approximation of Eq. 14, as a function of θ^{stim} (Fig. S2C), is given by

$$\Delta\psi\left(\theta^{\text{stim}}\right) \approx \left(1 - e^{-\frac{\tilde{I}_1}{\tau R_\infty} t}\right) \Delta\theta^{\text{stim}}. \tag{15}$$

The error incurred by this approximation is small (approx. 10% for $\Delta\theta^{\text{stim}} = 90°$, t = 250 ms and $\tilde{I}_0 = 0.02$), confirming that stimulus integration in the bump attractor model is relatively linear across a wide range of $\Delta\theta^{\text{stim}}$.

Potential of the amplitude equation

Here we show that the system $\dot{\mathbf{x}} = \mathbf{f}(R, \psi)$ describing the evolution of the amplitude equation (Eq. 1), is indeed a gradient system, $\dot{\mathbf{x}} = -\nabla\Theta(R, \psi)$, with a compatible potential function $\Theta(R, \psi)$ given by Eq. 2:

$$\begin{aligned}
\nabla\Theta &= \left(\frac{\partial}{\partial R}, \frac{1}{R}\frac{\partial}{\partial \psi}\right)\Theta(R, \psi), \\
&= -\left(\tilde{I}_1 \cos\left(\psi - \theta^{\text{stim}}(t)\right) + \tilde{I}_0 R - cR^3,\ -\tilde{I}_1 \sin\left(\psi - \theta^{\text{stim}}(t)\right)\right) \\
&= -\left(\frac{dR}{dt}, R\frac{d\psi}{dt}\right) = -\mathbf{f}(R, \psi)
\end{aligned} \tag{16}$$

where we have used the polar form of the gradient operator and the differential of $\mathbf{x} = (R, \psi)$ is $(dR, Rd\psi)$. Similarly, we can compute its time-evolution

$$\begin{aligned}\frac{d\Theta}{dt} &= \frac{\partial\Theta}{\partial R}\frac{dR}{dt} + \frac{\partial\Theta}{\partial\psi}\frac{d\psi}{dt} \\ &= -\left(\frac{\partial\Theta}{\partial R}\right)^2 - \left(\frac{1}{R}\frac{\partial\Theta}{\partial\psi}\right)^2 \\ &= -\left(\nabla\Theta\right)^2 \leq 0, \end{aligned} \tag{17}$$

that shows that for all initial conditions $\mathbf{x}_0 = (R_0, \psi_0)$, the potential will always decrease towards a stationary stable state (R_∞, ψ_∞).

Dynamical system for the cumulative circular average (perfect vector integrator)

In order to understand the integration process carried out by the ring attractor model, we derived a dynamical system that perfectly integrates a circular variable to compute the circular average. In general, the optimal strategy to keep track of the average, $\bar{z}(t)$, of a time-varying stimulus, $z(t)$, is to compute the cumulative running average. For discrete-time stimuli the cumulative running average can be computed by the following iterative equation:

$$\bar{z}_t = \bar{z}_{t-1} + \frac{1}{t}\left(z_t - \bar{z}_{t-1}\right),\ t = 1, 2, \ldots \tag{18}$$

where t represents discrete time points. If we assume that the updates $\bar{z}_t - \bar{z}_{t-1}$ are small enough, we can write the above equation as a continuous process in time:

$$\frac{d\bar{z}}{dt} = \frac{1}{t}\left(z(t) - \bar{z}\right),\ t \geq 1; \quad \bar{z}(1) = z(1). \tag{19}$$

In order to avoid dealing with a non-autonomous system we can rewrite this equation by absorbing the time dependence into a new variable $\alpha(t) = t\bar{z}(t)$ which leads to the integral form of the perfect integrator:

$$\frac{d\alpha}{dt} = z(t). \tag{20}$$

According to this equation, $\alpha(t)$ is just the sum of all $z(t)$ (integral over time), and the average is obtained by dividing the summation by t:

$$\bar{z}(t) = \frac{1}{t}\int_{t_0}^{t} z(t')dt'.$$

However, even if the result of the above integral is equivalent to that obtained by means of the iterative method, Eq. 19, the computational approach is fundamentally different. Here we would be required to memorize the entire time-varying stimulus up to the instant t, and divide its summation by the duration of the stimulus. In contrast, using the iterative method, Eq. 19, only requires keeping track of the average quantity at the last time point and update that value by weighting successive evidence in a time-dependent manner.

Eq. 20 holds for any time-varying process $z(t)$, in particular, here we are interested in $z(t) \in \mathbb{C}$. In this case we can interpret $z(t)$ as stimulus vectors defined by the stimulus amplitude and the stimulus direction and $\bar{z}(t)$ as the vector average up to time t. The resulting complex quantities can be written in polar form as $\alpha(t) = R(t)e^{i\psi(t)}$ and $z(t) = I(t)e^{i\theta(t)}$ which gives the two equations

$$\frac{dR}{dt} = I(t)\cos\left(\psi - \theta(t)\right), \tag{21a}$$

$$\frac{d\psi}{dt} = -\frac{I(t)}{R}\sin\left(\psi - \theta(t)\right). \tag{21b}$$

We refer to these equations as the *perfect vector integrator (PVI)*. As described previously in this work, our main goal is to compute the circular average of the stimulus orientations, i.e. the argument of $\bar{z}(t)$. From the definition of $\alpha(t) = t\bar{z}(t)$, we can already see that $\arg(\bar{z}) = \arg(\alpha) = \psi$. Therefore Eq. 21b exactly describes the evolution of the circular average of the orientations of the stimulus frames, i.e. the PVI computes the cumulative circular average. Moreover, note that Eq. 21b will compute the circular cumulative running average for any given normalization function $f(t)$ such that $\alpha(t) = f(t)\bar{z}(t)$.

The variable R, needed in the computation of the circular average, has several interpretations. First, the average magnitude of $\bar{z}(t)$, can be computed as $|\bar{z}(t)| = \frac{1}{t}R(t)$. In other words, $R(t)$ is the magnitude of the integrated vectors (i.e., the length of the resulting vector). Second, from this geometric interpretation it is easy to see that $R(t)$ measures the dispersion of the directions θ, with $R(t)$ growing linearly with t if $\theta(t)$ is a constant. If $\theta(t)$ is sampled from some underlying distribution, the growth will be slower and depends on the width of the distribution (Fig. S3A,B). The circular variance of the samples can be computed from the ratio of $R(t)$ and the maximum R obtained for $\theta(t) = \text{const.}$ as $\text{var} = 1 - \frac{R(t)}{\int I(t)dt}$. Third, if we assume that $\theta(t)$ is sampled from a von Mises distribution with mean μ and concentration κ, the running circular mean $\psi(t)$ will also be distributed according to a von Mises distribution, with concentration proportional to $R(t)\kappa$. In a Bayesian framework, $\psi(t)$ and $R(t)$ thus track the mean and the concentration of the posterior distribution (Kutschireiter et al., 2021).

Potential landscape for the perfect vector integrator

We follow the same approach used in the main text to graphically picture the dynamics of the perfect vector integrator (Eqs. 21a and 21b) by defining its potential landscape,

$$\Theta^{\text{PVI}}(R, \psi) = -RI\cos\left(\psi - \theta^{\text{stim}}\right). \tag{22}$$

Proceeding in the same way as with the amplitude equation (Eq. 16) it is easy to see that the gradient of this potential gives the system Eq. 21.

Model simulations

Stimuli

The stimulus in each trial consists of a sequence of N i.i.d. oriented frames $\theta_j^{\text{stim}}, j = 1, \ldots, N$ that are sampled from some probability distribution. We define the stimulus orientation as a piecewise constant function of time

$$\theta^{\text{stim}}(t) = \sum_{j=1}^{N} \theta_j^{\text{stim}} H\left(t - \frac{j-1}{N}T\right) H\left(\frac{j}{N}T - t\right), \tag{23}$$

where $H(t)$ is the Heaviside step function and T is the stimulus duration. The input current to each neuron on the ring was modeled as

$$I_{\text{stim}}(t) = I_1(t)\cos\left(\theta - \theta^{\text{stim}}(t)\right), \tag{24}$$

where I_1 is the stimulus amplitude and θ is the position of the neuron on the ring (Eq. 4). In most of the simulations we run $K = 10^5$ trials each consisting of $N = 8$ stimulus frames with a constant amplitude I_1 and stimulus orientations sampled from a uniform distribution in the range $\theta_j^{\text{stim}} \in [-90°, 90°)$, but randomly selecting combinations $\left\{\theta_1^{\text{stim}}, \ldots, \theta_N^{\text{stim}}\right\}$ such that their circular mean, $\bar{\theta}^{\text{stim}}$, is uniformly distributed in the range $\bar{\theta}^{\text{stim}} \in [-80°, 80°]$.

Initial conditions

In order to simulate the **No initial bump** condition (Figs. 1C-E, 2B, 3B-F, 4 and 5A) the activity of the network at the beginning of each trial is set at a spatially homogeneous state ($I_{\text{exc}} = I_{\text{crit}}$) with an average

activity around $r_\infty = 12.5$ Hz. The bump state is formed by increasing the global excitatory drive to the network, I_{exc}, such that the bump starts to emerge at the beginning of the stimulus presentation.

On the other hand, for the **Initial bump** condition (Figs. 3A,C, 5B and S4B) the activity of the network at the beginning of each trial is set at a stationary bump. We compute this state taking into account both the amplitude of the stimulus, I_1, and the effect of the noise (see the derivation of the Amplitude equation), by waiting long enough such that the system reaches the steady state. For the *biased* condition (Figs. 3A,C, and 5B), the bump is centered at $0°$, while for the *unbiased* condition (Fig. S4B), the initial phase of the bump is shifted to match the orientation of the first stimulus frame. This shift is taken to be equal to the position of the bump in the *No initial bump* condition at the end of the first stimulus frame, which still captures the impact of fluctuations at the beginning of the trial.

Numerical integration

We simulated Eqs. 4 and 9 following the Euler-Maruyama scheme with a time step $\Delta t = 0.5$ ms, and discretizing the angular space θ into $n = 200$ evenly distributed locations, $\theta_k = k \cdot 360°/n - 180°$. All simulations and analysis were done in Python 3.7.3, using standard libraries including among others: `NumPy`, `ScyPy` and `Pandas`.

Statistical measures and fits

Estimation bias and estimation error

To characterize the accuracy with which an estimate $\bar{\theta}_{\text{est}}$ describes the true stimulus average $\bar{\theta}$, we define the following quantities.

Bias of the estimator. Usually the estimation bias is define as $b_{\text{est}}\left(\bar{\theta}\right) = \langle\bar{\theta}_{\text{est}}\rangle - \bar{\theta}$, i.e., the difference of the trial-average of the estimates $\bar{\theta}_{\text{est}}$ and the true value $\bar{\theta}$. For our simulations (Fig. 5C), the bias defined in this way would be zero for $\bar{\theta} = 0°$ and increase approximately linearly with $\bar{\theta}$. Thus, we preferred to define the estimation bias as one minus the slope of the estimation curve obtained from a regression fit (k_1 in Eq. 28). An unbiased estimator would have a slope of 1 and thus zero bias ($1 - k_1 = 0$), while slopes smaller than 1 correspond to increasingly large values of the estimation bias.

Variance of the estimator. The variance of the estimator is defined as $\sigma^2_{\text{est}}\left(\bar{\theta}\right) = \langle\left(\bar{\theta}_{\text{est}} - \langle\bar{\theta}_{\text{est}}\rangle\right)^2\rangle$, i.e. the variation of the estimates about their mean value. In Fig. 5E, we plot the square root of this quantity, the standard deviation of the estimates.

Estimation error. To assess the total deviations of estimated average orientations $\bar{\theta}_{\text{est}}$ from the true stimulus average $\bar{\theta}$, due to estimation bias and variance, we compute the root-mean-square error

$$\text{RMSE}\left(\bar{\theta}\right) = \sqrt{\langle\left(\bar{\theta}_{\text{est}} - \bar{\theta}\right)^2\rangle}, \tag{25}$$

i.e., the square root of the trial-average squared estimation error. For Figs. 4C and 5E, we computed the mean of the RMSE across all average orientations $\bar{\theta}$.

Weighted circular average. To check whether other factors, besides the intrinsic nonlinearities of the model, are contributing to the estimation error, we compute the weighted circular average by simulating the PVI (Eq. 21), with the same exact stimulus $I^{\text{stim}}(t)$ used to simulate the ring network (Eqs. 23 and 24) but with a time-varying amplitude obeying

$$I_1\left(t\right) = I_1 \sum_{j=1}^{N} \beta_j H\left(t - \frac{j-1}{N}T\right) H\left(\frac{j}{N}T - t\right), \tag{26}$$

where β_j correspond to the weights of the PK obtained for the ring network, and $H(t)$ and T are the Heaviside step function and the stimulus duration, respectively. This measure corresponds to the purple line in Fig. 4C.

Regression models for circular data

We use regression models to assess the impact of each stimulus frame on the final estimate of the stimulus average provided by the bump attractor model. Standard linear regression is not well suited for this purpose because we are estimating the average of a circular quantity. We thus derived a regression model for circular data. We define the weighted circular mean as the argument of a weighted sum of complex numbers:

$$\bar{\alpha} = \phi\left(\boldsymbol{\theta}; \boldsymbol{\beta}\right) = \arg \sum_{j=1}^{N} \beta_j \exp\left(i \cdot \theta_j\right), \tag{27}$$

where N in the number of stimulus frames, θ_j the orientation of the jth frame and β_j the corresponding weight. In order to account for systematic deviations of estimated average orientations, we include a bias term k_0 and a scale factor k_1 (determining the slope of the estimation curve) and model the average stimulus estimation in trial i as:

$$\mu_i = k_0 + k_1 \phi\left(\boldsymbol{\theta_i}; \boldsymbol{\beta}\right). \tag{28}$$

Assuming that the data follows a von Mises distribution with concentration parameter κ, we can derive the log-likelihood of observing the data $\boldsymbol{\psi}$ (i.e., the reported average angles) given the weights $\boldsymbol{\beta}$ as

$$LL\left(\boldsymbol{\beta}, k_0, k_1, \kappa; \boldsymbol{\psi}\right) = \sum_{i=1}^{K} \kappa \cos\left(\psi_i - \mu_i\right) - K \log\left(2\pi I_0\left(\kappa\right)\right), \tag{29}$$

where K is the number of trials and $I_0(\kappa)$ is the modified Bessel function of order 0. To uniquely define the weights $\boldsymbol{\beta}$ we include the constraint that the sum of $\sum \beta_j = 1$. The parameters $(\boldsymbol{\beta}, k_0, k_1, \kappa)$ that maximize the log-likelihood were estimated using the `minimize` function from `scipy.optimize` module in Python 3.7.3.

Psychophysical kernel

For the model simulations, we obtain the PK as the stimulus weights of a circular regression model. To quantify the shape of a PK, we define the PK slope (Prat-Ortega et al., 2021). It is the slope of a linear regression of the PK, normalized between -1 (decaying PK, primacy) and $+1$ (increasing PK, recency). Formally, we first fit the PK with a linear function of time,

$$\mathrm{PK}(t) = \beta_0 + k\beta_1 t, \tag{30}$$

where β_1 is the PK slope and $k = \frac{1}{2\mathrm{Var}(t)}$ is a factor that normalizes the PK slope to the interval $(-1, 1)$.

Psychophysical data and data analysis

Psychophysical experiment and datasets

We used published data from three studies carrying out a psychophysical experiment in which subjects had to categorize the average direction of a stream of Gabor patterns Fig. 6A, (Wyart et al., 2012; Cheadle et al., 2014; Wyart et al., 2015). All three experiments shared a similar task consisting on a stream of eight oriented high-contrast (50%) Gabor patches with orientations $\theta_j^{\mathrm{stim}},\ j = 1, 2, \ldots, 8$ on the range between $-90°$ and $90°$. The duration of each sample was of 250 ms in the first and second experiments (Wyart et al., 2012; Cheadle et al., 2014) and of 333 ms on the third experiment (Wyart et al., 2015), and their orientation, θ_j^{stim}, was distributed uniformly across trials.

In each trial participants reported whether, on average, the orientation of the eight samples fell closer to the cardinal or diagonal axes. The category level of each trial was computed by means of a folded decision-mapping rule consisting on a w-shaped profile that mapped the orientation of each sample to a decision variable ranging from -1 (diagonal) to 1 (cardinal), and then averaged across samples. Participants responded by pressing two keys of a standard keyboard with their left or right index finger.

Auditory feedback was given at the end of each trial depending on the agreement between the response and the sign of the average decision value.

In total, we analyzed data from 47 participants (15 from Wyart et al. (2012), 15 from Cheadle et al. (2014) and 17 from Wyart et al. (2015)). The average number of trials performed by each participant in these three experiments were 565, 440 and 185, respectively.

Psychophysical kernels for behavioral data

Psychophysical kernels of each participant are obtained as the weights of logistic regression (Fig. 6B). In order to characterize the shape of individual PKs, we use logistic regression with the constraint that the PK is either uniform (constant) or a linear function of time. We compare the model fits using AIC and classify each PK as uniform (best model has a constant PK), primacy (linear model with negative slope) or recency (linear model with positive slope).

Code availability

The code to generate the figures of the paper will be uploaded in `github`.

References

Aitken, F., Turner, G., and Kok, P. (2020). Prior expectations of motion direction modulate early sensory processing. *J Neurosci*, 40(33):6389–6397.

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

Aston-Jones, G. and Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu Rev Neurosci*, 28:403–450.

Ben-Yishai, R., Lev Bar-Or, R., and Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 92(9):3844–3848.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev*, 113(4):700–765.

Boucheny, C., Brunel, N., and 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.

Bressloff, P. C. (2012). Spatiotemporal dynamics of continuum neural fields. *Journal of Physics A: Mathematical and Theoretical*, 45(3):033001.

Bressloff, P. C. (2019). Stochastic neural field model of stimulus-dependent variability in cortical neurons. *PLoS Computational Biology*, 15(3).

Britten, K. H., Shadlen, M. N., Newsome, W. T., and Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J Neurosci*, 12(12):4745–4765.

Brunel, N. (2003). Dynamics and Plasticity of Stimulus-selective Persistent Activity in Cortical Network Models. *Cerebral Cortex*, 13(11):1151–1161.

Burak, Y. and Fiete, I. (2009). Accurate path integration in continuous attractor network models of grid cells. *PLoS Comput Biol*, 5(2).

Cannon, S. C., Robinson, D. A., and Shamma, S. (1983). A proposed neural network for the integrator of the oculomotor system. *Biological Cybernetics*, 49(2):127–136.

Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A., and Fiete, I. (2019). The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nature Neuroscience*, 22(9):1512–1520.

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., de Gardelle, V., Herce Castañón, S., and Summerfield, C. (2014). Adaptive Gain Control during Human Perceptual Choice. *Neuron*, 81(6):1429–1441.

Compte, A. (2000). Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. *Cerebral Cortex*, 10(9):910–923.

Deneve, S., Latham, P. E., and Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nat Neurosci*, 4(8):826–831.

Eckhoff, P., Wong-Lin, K. F., and Holmes, P. (2009). Optimality and robustness of a biophysical decision-making model under norepinephrine modulation. *J Neurosci*, 29(13):4301–4311.

Ermentrout, B. (1998). Neural networks as spatio-temporal pattern-forming systems. *Reports on Progress in Physics*, 61(4):353–430.

Esnaola-Acebes, J. M., Roxin, A., Avitabile, D., and Montbrió, E. (2017). Synchrony-induced modes of oscillation of a neural field model. *Physical Review E*, 96(5).

Finkelstein, A., Fontolan, L., Economo, M. N., Li, N., Romani, S., and Svoboda, K. (2019). Attractor dynamics gate cortical information flow during decision-making. *BioRxiv*.

Georgopoulos, A. P., Schwartz, A. B., and Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419.

Goldman, M. S., Compte, A., and Wang, X. J. (2009). Neural Integrator Models. In *Encyclopedia of Neuroscience*, pages 165–178. Elsevier Ltd.

Green, J., Adachi, A., Shah, K. K., Hirokawa, J. D., Magani, P. S., and Maimon, G. (2017). A neural circuit architecture for angular integration in drosophila. *Nature*, 546(7656):101–106.

Hansel, D. and Sompolinsky, H. (1998). Modeling Feature Selectivity in Local Cortical Circuits. *Methods in Neuronal Modeling: From Ions to Networks*, pages 499–567.

Hutt, A., Longtin, A., and Schimansky-Geier, L. (2007). Additive Global Noise Delays Turing Bifurcations. *Physical Review Letters*, 98(23):230601.

Hutt, A., Longtin, A., and Schimansky-Geier, L. (2008). Additive noise-induced turing transitions in spatial systems with application to neural fields and the swift-hohenberg equation. *Physica D: Nonlinear Phenomena*, 237(6):755–773.

Jazayeri, M. and Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience*, 9(5):690–696.

Jazayeri, M. and Movshon, J. A. (2007). A new perceptual illusion reveals mechanisms of sensory decoding. *Nature*, 446(7138):912–915.

Kiani, R., Corthell, L., and Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, 84(6):1329–1342.

Kiani, R., Hanks, T. D., and Shadlen, M. N. (2008). Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *J Neurosci*, 28(12):3017–3029.

Kim, S. S., Rouault, H., Druckmann, S., and Jayaraman, V. (2017). Ring attractor dynamics in the drosophila central brain. *Science*, 356(6340):849–853.

Kuramoto, Y. (1984). *Chemical Oscillations, Waves, and Turbulence*, volume 19 of *Springer Series in Synergetics*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kutschireiter, A., Rast, L., and Drugowitsch, J. (2021). Angular path integration by projection filtering with increment observations. *arXiv*, 2102.09650.

Latham, P. E., Deneve, S., and Pouget, A. (2003). Optimal computation with attractor networks. *J Physiol Paris*, 97(4-6):683–694.

Liu, F. and Wang, X.-J. (2008). A Common Cortical Circuit Mechanism for Perceptual Categorical Discrimination and Veridical Judgment. *PLoS Computational Biology*, 4(12):e1000253.

Luu, L. and Stocker, A. A. (2018). Post-decision biases reveal a self-consistency principle in perceptual inference. *eLife*, 7.

Machens, C. K., Romo, R., and Brody, C. D. (2005). Flexible control of mutual inhibition: a neural model of two-interval discrimination. *Science*, 307(5712):1121–1124.

Mazurek, M. E., Roitman, J. D., Ditterich, J., and Shadlen, M. N. (2003). A role for neural integrators in perceptual decision making. *Cereb Cortex*, 13(11):1257–1269.

Murray, J. D., Jaramillo, J., and Wang, X.-J. (2017). Working memory and decision-making in a frontoparietal circuit model. *J Neurosci*, 37(50):12167–12186.

Nichols, M. J. and Newsome, W. T. (2002). Middle temporal visual area microstimulation influences veridical judgments of motion direction. *Journal of Neuroscience*, 22(21):9530–9540.

Niyogi, R. K. and Wong-Lin, K. (2013). Dynamic excitatory and inhibitory gain modulation can produce flexible, robust and optimal decision-making. *PLoS Comput Biol*, 9(6):e1003099.

Pardo-Vazquez, J. L., Castiñeiras-de Saa, J. R., Valente, M., Damião, I., Costa, T., Vicente, M. I., Mendonça, A. G., Mainen, Z. F., and Renart, A. (2019). The mechanistic foundation of Weber's law. *Nature Neuroscience*, 22(9):1493–1502.

Pouget, A., Dayan, P., and Zemel, R. S. (2003). Inference and computation with population codes. *Annu Rev Neurosci*, 26:381–410.

Prat-Ortega, G., Wimmer, K., Roxin, A., and de la Rocha, J. (2021). Flexible categorization in perceptual decision making. *Nat Commun*, 12(1):1283.

Ratcliff, R. (2018). Decision making on spatially continuous scales. *Psychol Rev*, 125(6):888–935.

Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput*, 20(4):873–922.

Roxin, A. and Ledberg, A. (2008). Neurobiological Models of Two-Choice Decision Making Can Be Reduced to a One-Dimensional Nonlinear Diffusion Equation. *PLoS Computational Biology*, 4(3):e1000046.

Seelig, J. D. and Jayaraman, V. (2015). Neural dynamics for landmark orientation and angular path integration. *Nature*, 521(7551):186–191.

Skaggs, W. E., Knierim, J. J., Kudrimoti, H. S., and McNaughton, B. L. (1995). A model of the neural basis of the rat's sense of direction. *Adv Neural Inf Process Syst*, 7:173–180.

Song, P. and Wang, X. J. (2005). Angular path integration by moving "hill of activity": A spiking neuron model without recurrent excitation of the head-direction system. *Journal of Neuroscience*, 25(4):1002–1014.

Talluri, B. C., Urai, A., Bronfman, Z., Brezis, N., Tsetsos, K., Usher, M., and Donner, T. (2020). Choices Change the Temporal Weighting of Decision Evidence. *bioRxiv*, page 2020.03.06.979690.

Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., and Donner, T. H. (2018). Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence. *Current Biology*, 28(19):3128–3135.e8.

Tass, P. (1995). Cortical pattern formation during visual hallucinations. *Journal of Biological Physics*, 21(3):177–210.

Usher, M. and McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108(3):550–592.

Wang, X.-J. J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36(5):955–968.

Wimmer, K., Compte, A., Roxin, A., Peixoto, D., Renart, A., and De La Rocha, J. (2015). Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. *Nature Communications*, 6:1–13.

Wimmer, K., Nykamp, D. Q., Constantinidis, C., and Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat Neurosci*, 17(3):431–439.

Wong, K.-F. and Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(4):1314–28.

Wyart, V., de Gardelle, V., Scholl, J., and Summerfield, C. (2012). Rhythmic Fluctuations in Evidence Accumulation during Decision Making in the Human Brain. *Neuron*, 76(4):847–858.

Wyart, V., Myers, N. E., and Summerfield, C. (2015). Neural Mechanisms of Human Perceptual Choice Under Focused and Divided Attention. *Journal of Neuroscience*, 35(8):3485–3498.

Zhang, J., Bogacz, R., and Holmes, P. (2009). A comparison of bounded diffusion models for choice in time controlled tasks. *J Math Psychol*, 53(4):231–241.

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.

Acknowledgements

We thank Valentin Wyarth and Christopher Summerfield for sharing the experimental data and for fruitful discussions, and Tobias H. Donner, Bharath Chandra Talluri and Federico Devalle for excellent discussions. The research leading to these results has received funding from the Spanish Ministry of Science and Innovation together with the European Regional Development Fund (RYC-2015-17236 and BFU2017-86026-R to K.W, MTM2015-71509-C2-1-R and RTI2018-097570-B-I00 to A.R.) and from the Generalitat de Catalunya (grant AGAUR 2017 SGR 1565 to A.R. and K.W.).

Supplementary figures

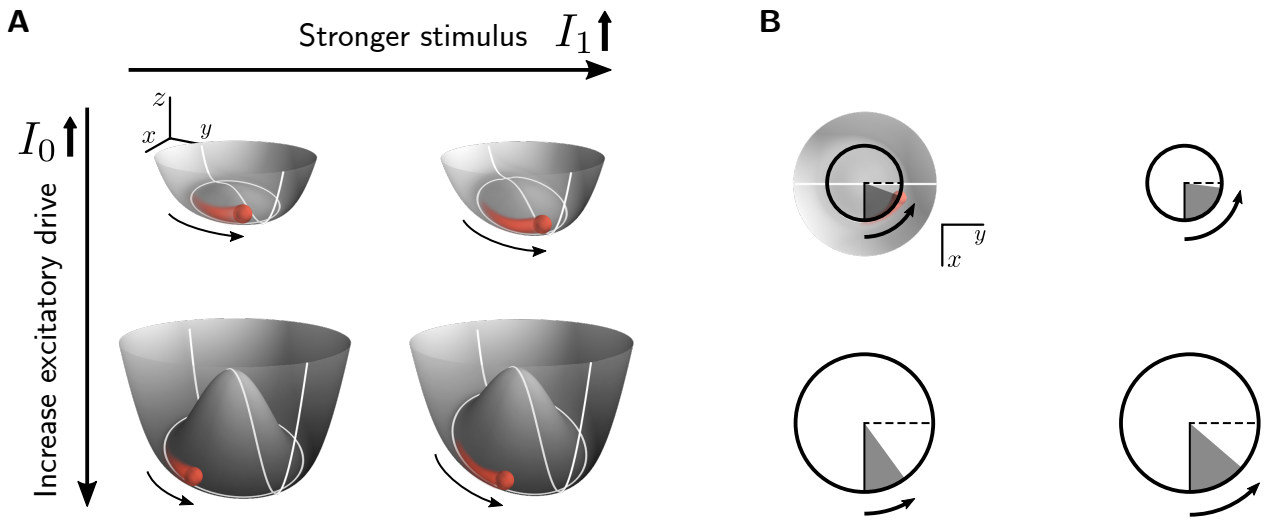

Figure S1: The global excitatory drive to the network determines the speed of movement of the activity bump.
(A) Displacement of the bump represented as the movement of a particle in the potential (Eq. 2) corresponding to the amplitude equation, Eq. 1, for different values of the excitatory drive and input strength. On one hand, increasing the excitatory drive to the network, I_0, deepens the potential and increases the radius of the stable manifold (white circle), which in turn elongates the path that the particle has to travel in order to achieve a given angular displacement. Therefore, for the same stimulus strength, increasing the excitatory drive will result in smaller displacements of the bump. On the other hand, increasing the stimulus input I_1 will achieve the opposite. **(B)** Top view of the potentials depicted in (A).

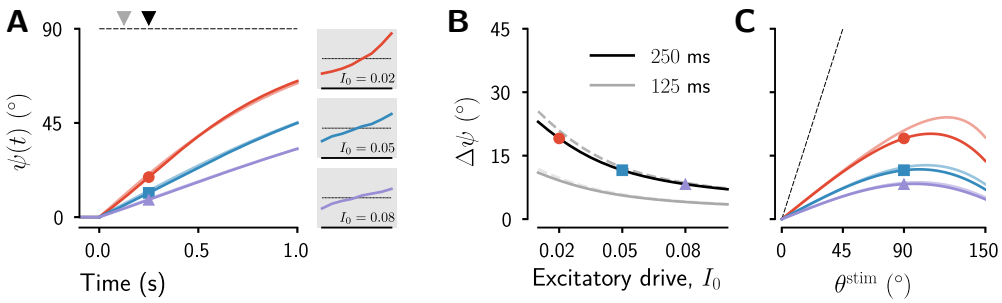

Figure S2: Impact of a single stimulus frame on the translation of the fully formed bump.
(A) Evolution of the phase of the bump for a single stimulus frame of strength $I_1 = 0.005$ and duration $T_{\text{stim}} = 1$ s centered at $\theta^{\text{stim}}(t) = 90°$, for three different values of the excitatory drive I_0. The solid lines correspond to simulations of the amplitude equation, Eq. 1, while the superimposed translucent lines have been analytically obtained using Eq. 14 which assumes a stationary bump amplitude. At the beginning of the trial the bump is fully formed and centered at 0°. Insets: PKs for each value of the excitatory drive (same as Fig. 3A). **(B)** Change of phase as a function of the global excitatory drive to the network for a stimulus as in (A) but with a duration of $T_{\text{stim}} = 250$ ms (solid black) and $T_{\text{stim}} = 125$ ms (solid gray). The dashed lines correspond to the analytical approximation Eq. 14. **(C)** Change of phase as a function of the orientation of the stimulus frame for a stimulus of duration $T = 250$ ms.

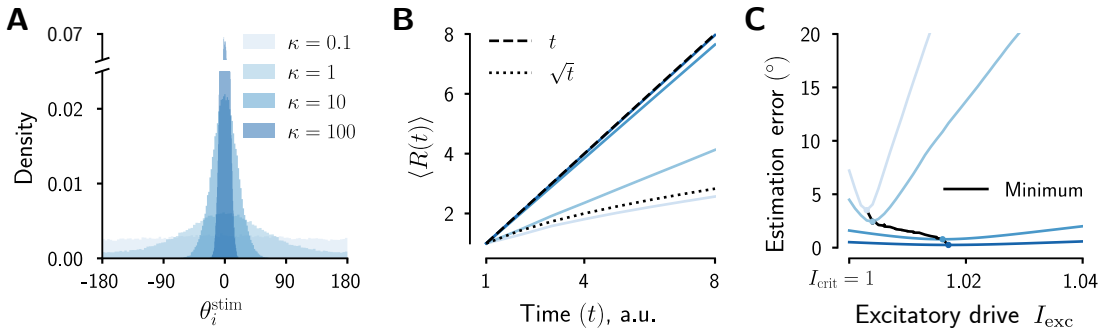

Figure S3: Effect of the stimulus distribution on the integration dynamics.
(A) The directions of the individual frames that compose the stimulus (Eq. 23) can be taken from a von Mises distribution with concentration parameter κ. **(B)** Evolution of the trial-averaged modulus $\langle R(t) \rangle$ of the PVI (Eq. 21b) for the stimulus described in Eqs. 23 and 24 with $N = 8$, $I_1 = 1$ and orientations θ_j^{stim} sampled from the distributions shown in (A). When the concentration parameter κ is large, i.e. the dispersion of the stimulus directions is small, $\langle R(t) \rangle$ grows approximately linearly in time. On the other hand, when the dispersion is large (e.g. $\kappa = 0.1$), $\langle R(t) \rangle$ grows much slower (for comparison we show a growth with $\sqrt{t}$). **(C)** The root-mean-square estimation error (Methods) of the amplitude equation as a function of the global excitatory drive I_{exc} for the stimulus described in the Methods (Eqs. 23 and 24) but with directions θ_j sampled from the distributions of panel (A). The model performs best (black line) for a value of the excitatory drive, I_{exc}, close to the bifurcation point I_{crit} for wide direction distributions. The optimal value of I_{exc} increases as the distribution of directions becomes narrower. Note that $I_0 = I_{\text{exc}} - I_{\text{crit}}$.

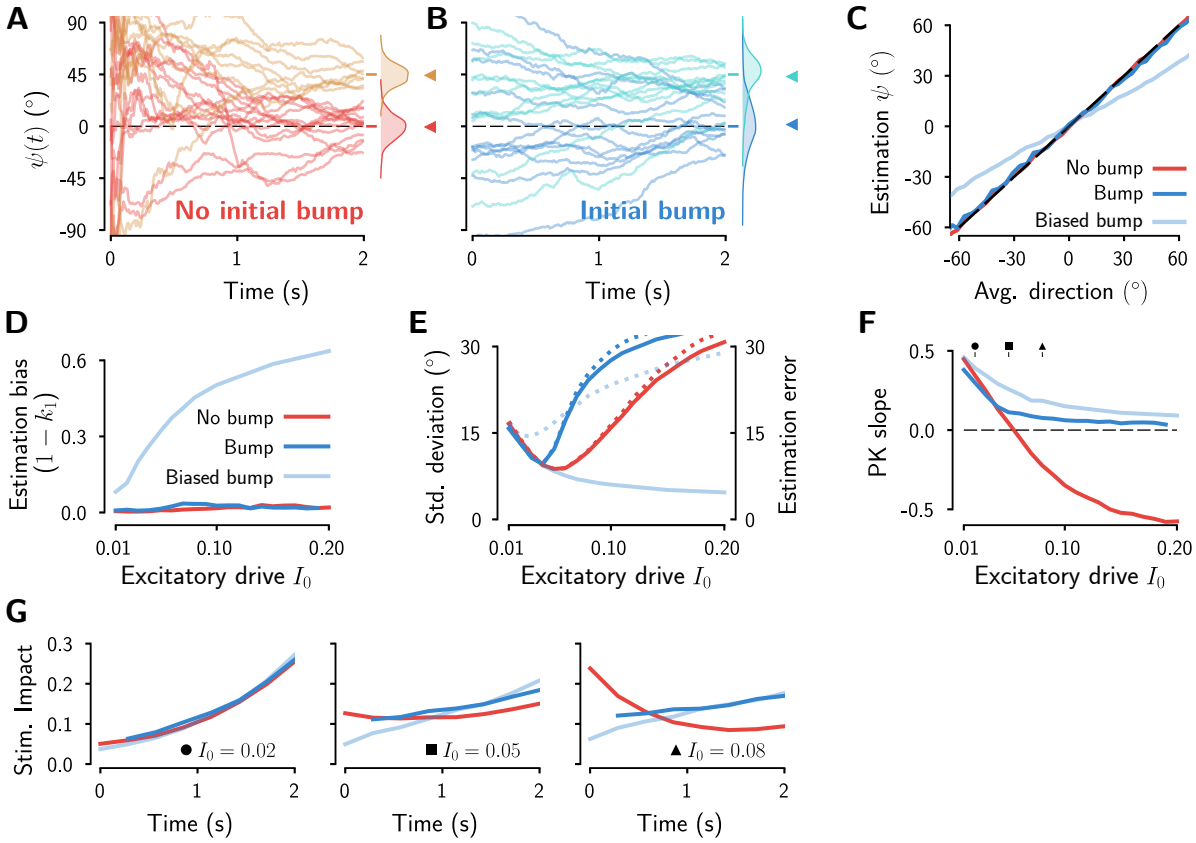

Figure S4: Bias and variance of the stimulus estimates with an initially formed bump and unbiased initial condition.

Panels **(A)** to **(F)** as in Fig. 5. The initial position of the bump in the fully formed condition (panel B and blue lines) is now aligned to the orientation of the first stimulus frame, taking the position of the bump at the end of the first stimulus frame in the *no initial bump* condition as a reference (Methods). The faint blue lines in panels (C) to (G) correspond to the biased condition described in Fig. 5. **(G)** PKs (Methods) for the three conditions: no initial bump condition (red), initially biased bump condition (faint blue), and initially unbiased bump condition (solid blue). To compute the PKs of the initially unbiased condition the contribution of the first stimulus frame has been discarded.