Cognitive Neurodynamics
https://doi.org/10.1007/s11571-023-09979-3

RESEARCH ARTICLE

Robust working memory in a two-dimensional continuous attractor network

Weronika Wojtak[1,2] · Stephen Coombes[3] · Daniele Avitabile[4,5] · Estela Bicho[2] · Wolfram Erlhagen[1]

Received: 25 January 2023 / Revised: 6 April 2023 / Accepted: 1 May 2023

Abstract

Continuous bump attractor networks (CANs) have been widely used in the past to explain the phenomenology of working memory (WM) tasks in which continuous-valued information has to be maintained to guide future behavior. Standard CAN models suffer from two major limitations: the stereotyped shape of the bump attractor does not reflect differences in the representational quality of WM items and the recurrent connections within the network require a biologically unrealistic level of fine tuning. We address both challenges in a two-dimensional (2D) network model formalized by two coupled neural field equations of Amari type. It combines the lateral-inhibition-type connectivity of classical CANs with a locally balanced excitatory and inhibitory feedback loop. We first use a radially symmetric connectivity to analyze the existence, stability and bifurcation structure of 2D bumps representing the conjunctive WM of two input dimensions. To address the quality of WM content, we show in model simulations that the bump amplitude reflects the temporal integration of bottom-up and top-down evidence for a specific combination of input features. This includes the network capacity to transform a stable subthreshold memory trace of a weak input into a high fidelity memory representation by an unspecific cue given retrospectively during WM maintenance. To address the fine-tuning problem, we test numerically different perturbations of the assumed radial symmetry of the connectivity function including random spatial fluctuations in the connection strength. Different to the behavior of standard CAN models, the bump does not drift in representational space but remains stationary at the input position.

Keywords Continuous bump attractor · Two-dimensional neural field · Working memory · Memory fidelity · Robust neural integrator

✉ Weronika Wojtak
w.wojtak@dei.uminho.pt

✉ Wolfram Erlhagen
wolfram.erlhagen@math.uminho.pt

Stephen Coombes
stephen.coombes@nottingham.ac.uk

Daniele Avitabile
d.avitabile@vu.nl

Estela Bicho
estela.bicho@dei.uminho.pt

[1] Research Centre of Mathematics, University of Minho, Guimarães, Portugal

[2] Research Centre Algoritmi, University of Minho, Guimarães, Portugal

[3] Centre for Mathematical Medicine and Biology, School of Mathematical Sciences, University of Nottingham, Nottingham, UK

[4] Department of Mathematics, Vrije Universiteit, Amsterdam, The Netherlands

[5] MathNeuro Team, Inria Sophia Antipolis Méditerranée Research Centre, Sophia Antipolis, France

Published online: 29 May 2023

Introduction

Working memory (WM) is defined as the capacity to maintain and manipulate over shorter time spans information that is no longer present to the senses. It is crucial for many higher cognitive functions such as planning, decision making or learning. Stimulus-tuned persistent neural population activity is widely believed to represent a neural correlate of WM. It has been observed in many cortical areas in tasks in which a transient stimulus has to be memorized in order to perform a delayed behavioral response (Zylberberg and Strowbridge 2017). A class of neural circuits called continuous attractor networks (CAN, (Amari 1977; Wu et al. 2008; Brody et al. 2003)) has been widely used in the past to explain neural and behavioral findings in WM tasks ((Johnson et al. 2009; Wimmer et al. 2014), for overviews and discussion see (Schöner and Spencer 2016; Khona and Fiete 2021)), and to implement a working memory capacity in artificial agents (e.g., robots, (Erlhagen and Bicho 2006)). These circuits are characterized by recurrent excitatory and inhibitory connections between neurons tuned to continuous input features such as for instance the direction of heading during navigation or the position of an object in space. The recurrent network dynamics may settle in response to a transient input into a self-sustained activity pattern localized in feature space, also known as a "bump attractor". Moreover, since the interaction strength between neurons depends on their distance only, the network structure is translation invariant. As a consequence, the network can hold a continuous family of bumps, each representing the memory of a specific input value. Particularly compelling neurophysiological evidence in line with predictions of CAN models has been found in prefrontal cortex (PFC) of monkey in a WM task with movement direction as continuous dimension (Wimmer et al. 2014). Extrinsic noise in the direction of the attractor manifold causes a drift of the bump away from the initial state, with the characteristics of a diffusion process (Kilpatrick and Ermentrout 2013; Camperi and Wang 1998). Such a diffusing bump representation has been indeed observed in PFC during the delay period of the spatial WM task, and the read-out of the encoded movement direction predicted subsequent behavioral errors.

While successful in describing the phenomenology of many WM tasks, classical CAN models suffer from essential limitations. Firstly, the attractor state has a stereotyped shape exclusively determined by the recurrent interactions within the network. It is thus not possible to model the quality of WM representations which may depend on stimulus features (e.g., strength or duration) or may reflect an up-date to changing tasks demands (Wildegger et al. 2016). Secondly, the biologically unrealistic assumption of a perfect translational symmetry of the synaptic weights renders the network mechanism of memory formation structurally unstable. Any heterogeneity in the synaptic weight distribution destroys the continuity of the attractor, allowing only a few possible stationary activity profiles to which stimulus-induced neural population activity drifts (Zhang 1996; Itskov et al. 2011). Several additional mechanisms have been proposed over the last couple of years to gain the functionality of a neural integrator and/or to approximately restore the continuous attractor in the face of synaptic heterogeneity (e.g., bistable neurons (Koulakov et al. 2002), short-term synaptic facilitation (Itskov et al. 2011), homeostatic synaptic scaling (Renart et al. 2003), negative derivative feedback (Lim and Goldman 2013), for a review see (Barak and Tsodyks 2014)).

In this paper, we address both challenges in a model with a recurrent network architecture which combines a lateral-inhibition-type connectivity of classical CANs with locally balanced excitatory and inhibitory feedback loops. In previous work, we have shown that this architecture in a network representing a single dimension may stabilize a continuum of bump amplitudes (Wojtak et al. 2021a). Here we extend our investigation of a neural integrator capacity to two spatial dimensions (2D), representing neural populations tuned to two stimulus axes (e.g., spatial position and orientation, (Drucker et al. 2009)). We show how the integration of bottom-up and top-down information shapes WM representations. The model is formalized in the continuum limit of dynamic neural field (DNF) equations (Amari 1977). This allows us to analyze the existence, stability and bifurcation structure of radially symmetric 2D bumps. We then investigate numerically different perturbations of the radially symmetric connectivity function including systematic directional biases and noise-induced heterogeneities. The results show that the creation and maintenance of input-induced bumps is robust against these perturbations.

The paper is organized as follows. First, in section "The two-dimensional CAN model", we introduce the two-field model with two spatial dimensions, discuss the network architecture and explain some assumption we make for the modeling work. In section "Existence and stability of radially symmetric bump solutions", we present the main mathematical results about the existence, linear stability and bifurcation structure of radially symmetric bump solutions. In section "Input-induced bumps", we show model simulations with transient external inputs. We discuss the impact of the results on modeling WM and highlight the differences to standard CAN models. The numerical investigation of the network with different perturbations of the radially symmetric weight function is presented in section "Non-radially symmetric connectivity

functions". We finish with a critical discussion of our results in relation to other modeling approaches to WM and highlight some topics for future work. Details of the mathematical analysis are presented in three appendixes.

The two-dimensional CAN model

We study the extension of the CAN model presented in (Wojtak et al. 2021a) to two spatial dimensions. It is formalized by two coupled field equations of Amari type (Amari 1977) describing the activity (e.g., membrane potential) of two neural populations, u and v, at time t at a position $\mathbf{r} \subseteq \Omega \subset \mathbb{R}^2$:

$$\tau_u \frac{\partial u(\mathbf{r},t)}{\partial t} = -u(\mathbf{r},t) + v(\mathbf{r},t) + \int_\Omega w(|\mathbf{r}-\mathbf{r}'|) f(u(\mathbf{r}',t)-\theta)\mathrm{d}\mathbf{r}' + I(\mathbf{r},t), \tag{1a}$$

$$\tau_v \frac{\partial v(\mathbf{r},t)}{\partial t} = -v(\mathbf{r},t) + u(\mathbf{r},t) - \int_\Omega w(|\mathbf{r}-\mathbf{r}'|) f(u(\mathbf{r}',t)-\theta)\mathrm{d}\mathbf{r}'. \tag{1b}$$

The nonlinearity f denotes the firing rate function which is assumed to be bounded and positive monotonic. A typical choice is a sigmoid with threshold θ and gain parameter η

$$f(u) = \frac{1}{1+\mathrm{e}^{-\eta(u-\theta)}}.$$

For $\eta \to \infty$, f approximates the Heaviside function $H_\theta(u) = 1$ for $u > \theta$ and $H_\theta(u) = 0$ otherwise. Following Amari's original analysis, we use H_θ to show the existence and linear stability of 2D bumps, and a sigmoid with high gain for the numerical bifurcation analysis which requires the nonlinearity to be differentiable. The qualitative model behavior is robust to changes in the neural gain. The radially symmetric Mexican hat coupling function $w(\mathbf{r})$ is given by the difference of two Gaussians

$$w_{mex}(x,y) = A_{ex}\mathrm{e}^{\left(-\left(\frac{x^2}{2\sigma_{ex_x}^2}+\frac{y^2}{2\sigma_{ex_y}^2}\right)\right)} - A_{in}\mathrm{e}^{\left(-\left(\frac{x^2}{2\sigma_{in_x}^2}+\frac{y^2}{2\sigma_{in_y}^2}\right)\right)} - w_{inh} \tag{2}$$

where $A_{ex} > A_{in} > 0$ and $\sigma_{in_x}^2 = \sigma_{in_y}^2 > \sigma_{ex_x}^2 = \sigma_{ex_y}^2$ and $w_{inh} > 0$. As shown by Amari (1977), it describes the effective interactions of two separate excitatory and inhibitory populations when inhibition is assumed to act instantaneously. More recent experimental studies investigating the neural circuits supporting spatial WM have described the existence of diverse types of inhibitory interneurons (Constantinidis and Wang 2004). They receive input from nearby excitatory neurons and also show spatially tuned persistent activity. Synaptic interactions between a specific class of interneurons are thought to implement a disinhibition mechanism which effectively unmasks more excitatory input to excitatory neurons. Interestingly, it has been shown that a dynamic field model featuring a connectivity function with local recurrent inhibition and surround excitation can support sustained, spatially patterned solutions (Rubin and Troy 2004). In our mechanistic two-field model, a neuron of the v-population integrates the activity from the u-population with the inverted Mexican hat profile and projects its activity back locally. The linear feedback loop between the two populations guarantees a tight balance of local excitation and inhibition. It supports the capacity of the network to stabilize input-induced bumps with a continuum of amplitudes (Wojtak et al. 2021a). Note that introducing nonlinearities in the feedback loop (e.g., using a piecewise linear transfer function) would introduce a saturation limit for a range of possible bump amplitudes.

The time-dependent external input $I(\mathbf{r},t)$ to the u-population is modeled as one or more Gaussians centered at positions $\mathbf{r}_{c_j}$:

$$I(\mathbf{r},t) = (H_{t_0}(t) - H_{t_e}(t)) \sum_{j=1}^{n} A_{I_j}\mathrm{e}^{\left(-(\mathbf{r}-\mathbf{r}_{c_j})^2/2\sigma_{I_j}^2\right)}, \tag{3}$$

where $A_{I_j} > 0$ controls the input strength and $H_{t_0}(t)$ represents the Heaviside step function with threshold $t_0 \geq 0$ controlling the start and the end of the input at times t_0 and t_e, respectively. In the following we use for the input duration the notation $d_I = t_e - t_0$.

Existence and stability of radially symmetric bump solutions

We first derive the necessary conditions for the existence of radially symmetric 2D bumps of the two field model (1) and analyze their stability using Fourier methods and properties of Bessel functions. Following the approach presented in (Bressloff 2012; Bressloff and Coombes 2013), we consider a wizard hat weight distribution given by a combination of modified Bessel functions of the second kind

$$w(r) = \frac{2}{3\pi}(K_0(r) - K_0(2r) - A(K_0(r/\sigma) - K_0(2r/\sigma))). \tag{4}$$

An example of function (4) is depicted in Fig. 1.

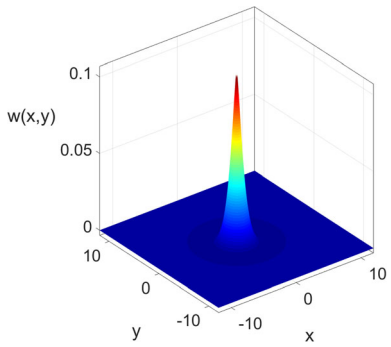

Fig. 1 Weight distribution given by a combination of modified Bessel functions of the second kind as defined in (4). Weight parameters are $A = 1/4$ and $\sigma = 2$

Existence of bumps

We study the model (1) with initial condition $u(\mathbf{r},0) + v(\mathbf{r},0) = K$, where $K > 0$ is a constant. We consider a circularly symmetric bump of radius R such that $u(\mathbf{r},t) = U(r)$ with $U(R) = \theta$, $U(r) > \theta$ for $r < R$, $U(r) < \theta$ for $r > R$ and $U(r) \to 0$ as $r \to \infty$. A stationary solution of system (1) then gives

$$U(r) = V(r) + \int_0^{2\pi}\int_0^R w(\sqrt{r^2 + r'^2 - 2rr'\cos\phi})r'\mathrm{d}r'\mathrm{d}\phi, \tag{5a}$$

$$V(r) = U(r) - \int_0^{2\pi}\int_0^R w(\sqrt{r^2 + r'^2 - 2rr'\cos\phi})r'\mathrm{d}r'\mathrm{d}\phi. \tag{5b}$$

In Appendix A, we show that the double integral in (5) can be calculated using the Fourier transforms and Bessel function identities to obtain

$$U(r) = V(r) + 2\pi R\int_0^\infty \widehat{w}(\rho)J_0(\rho r)J_1(\rho R)\mathrm{d}\rho, \tag{6a}$$

$$V(r) = U(r) - 2\pi R\int_0^\infty \widehat{w}(\rho)J_0(\rho r)J_1(\rho R)\mathrm{d}\rho, \tag{6b}$$

where $\widehat{w}(\rho)$ is the Fourier transform of w.

Knowing that $U(R) = \theta$ and $U(r) + V(r) = K$, we obtain the following necessary condition for the existence of a bump with radius R

$$\theta = \frac{K}{2} + \pi R\int_0^\infty \widehat{w}(\rho)J_0(\rho r)J_1(\rho R)\mathrm{d}\rho. \tag{7}$$

Like for the 2D Amari model, with the wizard hat coupling function there exist a maximum of two bump solutions for a given value of threshold θ, as shown in Fig. 2.

Stability of bumps

In the following we determine the linear stability of radially symmetric solutions of (1) with respect to different possible perturbations of the circular boundary exhibiting D_n symmetry.

In order to determine the linear stability of a stationary bump $U(r)$, we substitute $u(\mathbf{r},t) = U(r) + \psi(\mathbf{r})\mathrm{e}^{\lambda t}$ and $v(\mathbf{r},t) = V(r) + \zeta(\mathbf{r})\mathrm{e}^{\lambda t}$ into (1) and expand to first order in ψ and ζ using (5). This leads to the system of eigenvalue equations

$$\lambda\psi(\mathbf{r}) = -\psi(\mathbf{r}) + \zeta(\mathbf{r}) + \int_\Omega w(|\mathbf{r} - \mathbf{r}'|)\delta(U(r') - \theta)\psi(\mathbf{r}')\mathrm{d}\mathbf{r}', \tag{8a}$$

$$\lambda\zeta(\mathbf{r}) = -\zeta(\mathbf{r}) + \psi(\mathbf{r}) - \int_\Omega w(|\mathbf{r} - \mathbf{r}'|)\delta(U(r') - \theta)\psi(\mathbf{r}')\mathrm{d}\mathbf{r}'. \tag{8b}$$

In Appendix B, we show that solving (8) gives the following eigenvalues

$$\lambda_{-1} = 0, \tag{9}$$

$$\lambda_n = -2 + 2\frac{\int_0^\infty \widehat{w}(\rho)J_n(\rho R)J_n(\rho R)\rho\mathrm{d}\rho}{\int_0^\infty \widehat{w}(\rho)J_1(\rho R)J_1(\rho R)\rho\mathrm{d}\rho}, \tag{10}$$

where index n corresponds to the number of modes of the boundary perturbation exhibiting D_n symmetry. Similar to the 2D Amari model, we have $\lambda_1 = 0$, the bump of radius R is thus linearly stable if $\lambda_n < 0$ for all $n \neq 1$. Figure 2a depicts the branches of stable (solid line) and unstable (dashed line) bump solutions with radius R as a function of threshold θ. For $n = 2,\ldots,7$, we also plot the points of azimuthal instability to planar perturbations with D_n symmetry determined by the condition $\lambda_n = 0$. Examples of a stable (P_1) and an unstable (P_2) bump are shown in Fig. 2b. Note that in the real biological system, an excitation pattern with the smaller radius will not persist. Any perturbation of the circular boundary due to noise in the network will destroy it.

Bumps with initial condition $u(\mathbf{r},0) + v(\mathbf{r},0) = K(\mathbf{r})$

Figure 3 shows examples of bump solutions for the initial conditions

$$u(\mathbf{r},0) = K(\mathbf{r}), \quad v(\mathbf{r},0) = 0, \quad K(\mathbf{r}) = A_K\mathrm{e}^{\left(-\mathbf{r}^2/2\sigma_K^2\right)}, \tag{11}$$

which represent a homogeneous initial state for the v-population and a spatially structured state for the u-population. As can be clearly seen when comparing the

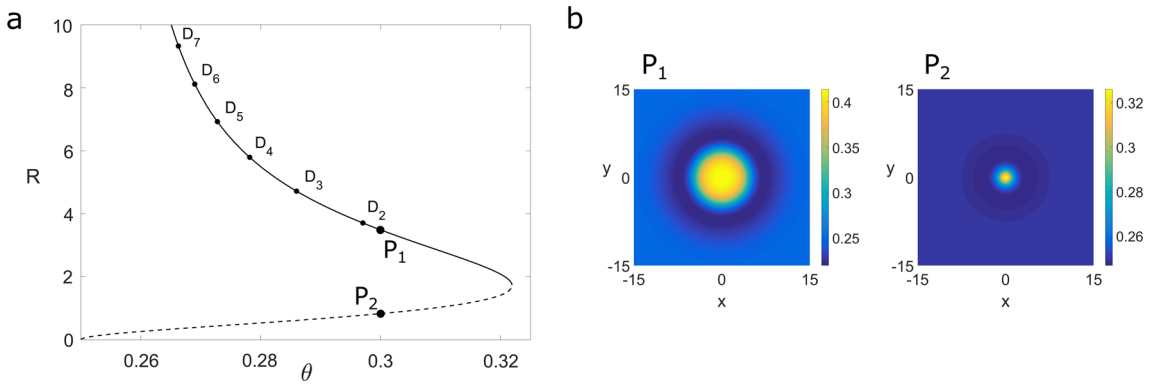

Fig. 2 **a** Bump radius R of stable (solid line) and unstable (dashed line) solutions as a function of θ for the two-field model (1) with $K = 0.5$ and a wizard hat weight distribution (4) with $A = 1/4$ and $\sigma = 2$. Dots show points along the stable branch where bumps become unstable to planar perturbations with $D_2, \ldots, D_7$ symmetry. **b** Top view of a stable bump with radius $R = 3.49$ (P_1) and an unstable bump with radius $R = 0.83$ (P_2) for $\theta = 0.3$

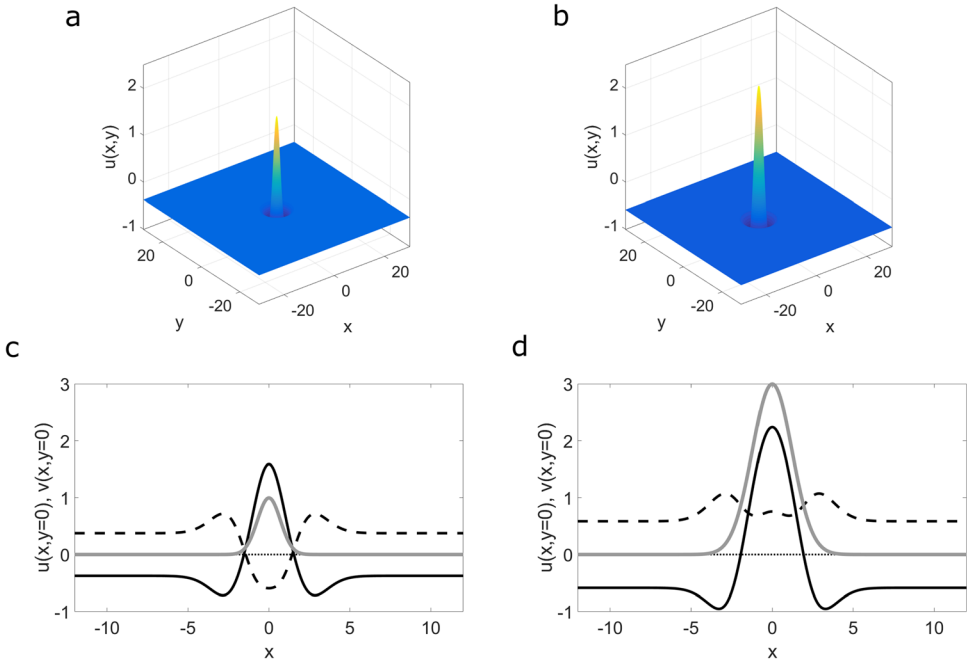

Fig. 3 One-bump solutions at time $t = 50$ of the two-field model (1) for two different initial profiles (11) with $A_K = 1$, $\sigma_K = 1$ (**a** and **c**) and $A_K = 3$, $\sigma_K = 3$ (**b** and **d**). Solid grey lines represent initial profiles of the u-population. The kernel w is given by (2) with $A_{ex} = 2$, $A_{in} = 1$, $\sigma_{ex} = 1$, $\sigma_{in} = 1.5$ and $w_{inh} = 0.1$. Threshold $\theta = 0$

activation patterns in the two panels, the bump shape correlates perfectly with the shape of the initial activation profile of the u-population.

Numerical continuation is a powerful tool that has been widely used to track specific solutions of neural field equations as model parameters vary (Rankin et al. 2014). We extend here the numerical scheme developed to track bump solutions of the scalar two-field model (Wojtak et al. 2021a) to the 2D case. Figure 4a and c compares bifurcation curves for the bumps created with the narrower and the wider initial condition of Fig. 3 with threshold θ as continuation parameter. Independent of the initial condition, the bifurcation curves maintain the same qualitative behavior. Stable and unstable bumps coexist in a certain range of threshold values. Importantly, and different to the Amari model, there exists also a branch of stable subthreshold solutions defined by the balanced local interactions between the two populations. Figure 4b and d depicts for both initial conditions top views of stable, localized activity patterns above and below threshold. The pairs (P_1, P_2) respectively (P_3, P_4) are obtained with the same threshold value.

Input-induced bumps

In classical CAN models, memory encoding and maintenance is modeled as an "all-or-none" phenomenon regardless of the input strength above threshold. The stereotyped bump shape is exclusively determined by the

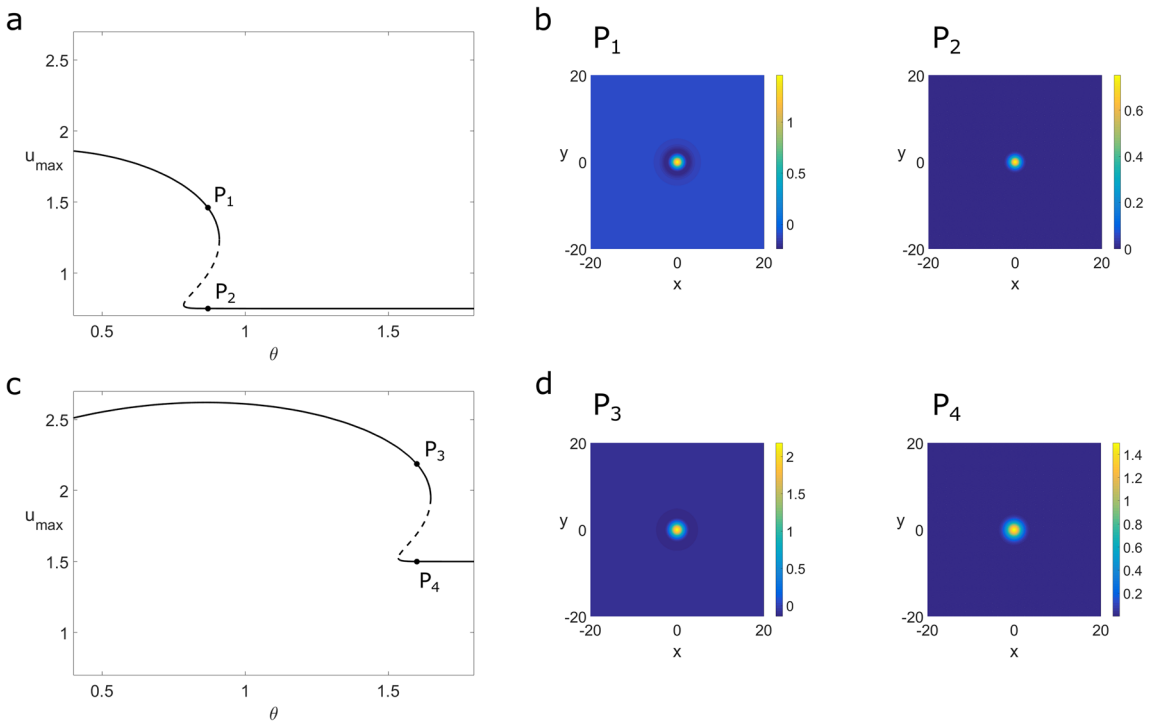

Fig. 4 **a** and **c** Bifurcation curves showing single bump solutions of (1) with the initial condition (11) with $A_K = 1.5$, $\sigma_K = 1.5$ (**a**) and $A_K = 3$, $\sigma_K = 3$ (**c**) as the parameter θ is varied. Examples of solutions at the points P_1 - P_4 for a narrower and a wider profile of $K(\mathbf{r})$ are shown in panels (**b**) and (**d**), respectively. Parameters of the kernel: $A_{ex} = 2$, $A_{in} = 1$, $\sigma_{ex} = 1$, $\sigma_{in} = 1.5$ and $w_{inh} = 0.1$

recurrent interactions within the network. Converging lines of evidence from behavioral and neurophysiological experiments suggests that not only the content of stimuli but also their quality or fidelity is represented in WM (Ma et al. 2014). Input characteristics such as stimulus contrast/strength and stimulus duration are known to affect WM performance. A possible neural substrate of stimulus saliency is the level of neural activity in a population representing a memorized item (Brody et al. 2003; Constantinidis et al. 2001). Figure 5 illustrates the capacity of the two-field model to continuously integrate inputs over time. The bumps represent the population response to a briefly presented stimulus (a, b) for which the strength (c, d) or the duration (e, f) has been changed. Since in this example the manipulation of strength and duration results in the same total input applied, the bumps in (c) and (e) have equal amplitude. In recent works, we have used this integrator property to model WM for serial order of sensory events represented by a multi-bump pattern with an activation gradient (Wojtak et al. 2021b) and to elucidate the neural underpinning of interval timing (Wojtak et al. 2019). The latter study models findings of an experiment in which monkeys were trained to measure different sample intervals and immediately afterwards reproduce it by a proactive saccade to a predefined target. In line with the model predictions, longer sample intervals resulted in higher firing rates at the end of the measuring period. The observed monotonic increase of population activity to a fixed threshold associated with saccade onset during the production phase can be explained by the continuous integration of an input with a strength inversely proportional to the bump amplitude at the end of the measuring period.

Although there is broad agreement that persistent activity is central to maintenance in WM, there is a debate regarding how many items can be simultaneously represented in an active state. Recent neuroimaging data has been interpreted as evidence that only a single item in the "focus of attention" is held in a prioritized active state at a time (Lewis-Peacock et al. 2012; Rose et al. 2016). The concept of "activity-silent" WM (Stokes 2015) assumes that additional items are stored by a stimulus-specific pattern of synaptic facilitation in the recurrent connections. Other experimental studies, however, report evidence for a concurrent storage of multiple active neural representations in WM (Sutterer et al. 2019; Scotti et al. 2021).

It is well known that classical CANs with lateral inhibition kernel have difficulties generating and maintaining input-induced multi-bump patterns. The mutually inhibitory interactions between neural representations of individual items lead to memory drift or complete memory loss (Amari 1977; Mégardon et al. 2015). Figure 6a–d shows an example of the Amari model with three identical inputs presented sequentially at three different positions. Only the

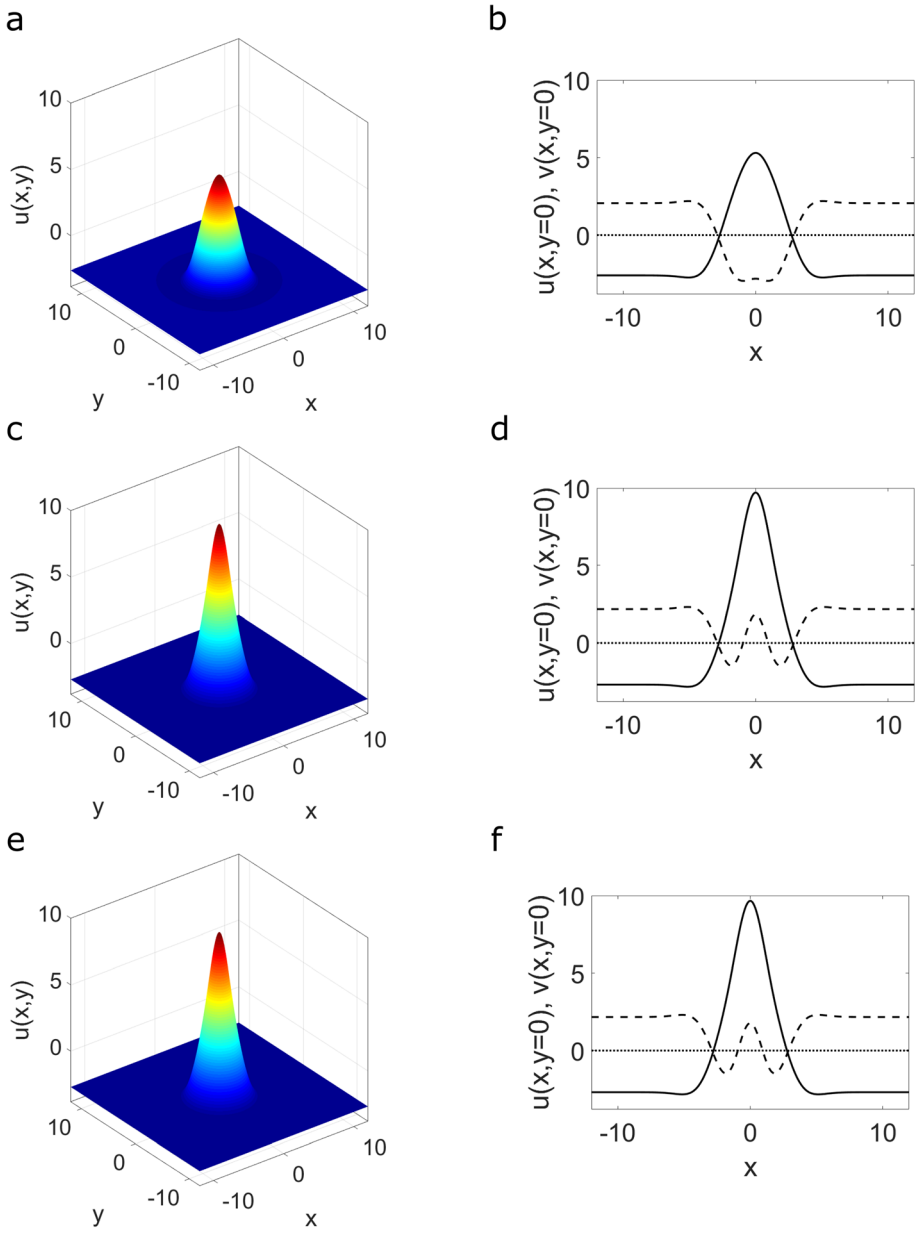

Fig. 5 Left column: Bump solutions at time $t = 50$ of the model (1) created with transient inputs $I(\mathbf{r}, t)$ given by (3) with variation of (**c**) input strength A_I, and (**e**) input duration d_I. Right column: Cross sections of the bump solutions in x-direction at $y = 0$. Solid and dashed black lines represent $u(x)$ and $v(x)$, respectively. Parameters of the inputs: (**a** and **b**) $A_{I_1} = 3, \sigma_{I_1} = 1, d_I = 1$ (**c** and **d**) $A_{I_1} = 12, \sigma_{I_1} = 1, d_I = 1$ (**e** and **f**) $A_{I_1} = 3, \sigma_{I_1} = 1,$ $d_I = 4$. The kernel w is given by (2) with $A_{ex} = 3$, $A_{in} = 1$, $\sigma_{ex} = 1.2$, $\sigma_{in} = 1.6$ and $w_{inh} = 0.2$. Threshold $\theta = 0$, $K = -0.5$

first transient input is able to trigger the evolution of a bump since it causes a significant suppression of neural activity at the positions of the two successive stimuli. As a consequence, their neural representation remains subthreshold and quickly falls back to resting state. The loss of input information is irreversible since it cannot be recovered with any type of additional processing. The suppressive effect of lateral inhibition appears to be mitigated to some extent in the two-field model due to the spread of excitation mediated by the inverted Mexican hat connectivity of the v-population. Fig. 6e–h shows that in response to the identical series of transient inputs, the local feedback mechanisms are able to stabilize a three-bump solution.

It is important to stress however that bump competition is still visible in the reduced bump amplitudes compared to a single item memory. This can be directly seen when applying a series of inputs with reduced strength (Fig. 7a – c). Now, only the neural representations of the first and the second stimulus reach a suprathreshold level, the representations of the third remains subthreshold (a). However, different to the Amari model, the input information is not completely lost since the activation patterns are self-sustained. The integrator capacity of the network can be tested by applying a second subthreshold input ("post-cue") at the pre-activated site long after the first is gone (b). The peak position of the evolving bump again faithfully represents a memory of the two input dimensions (c). This

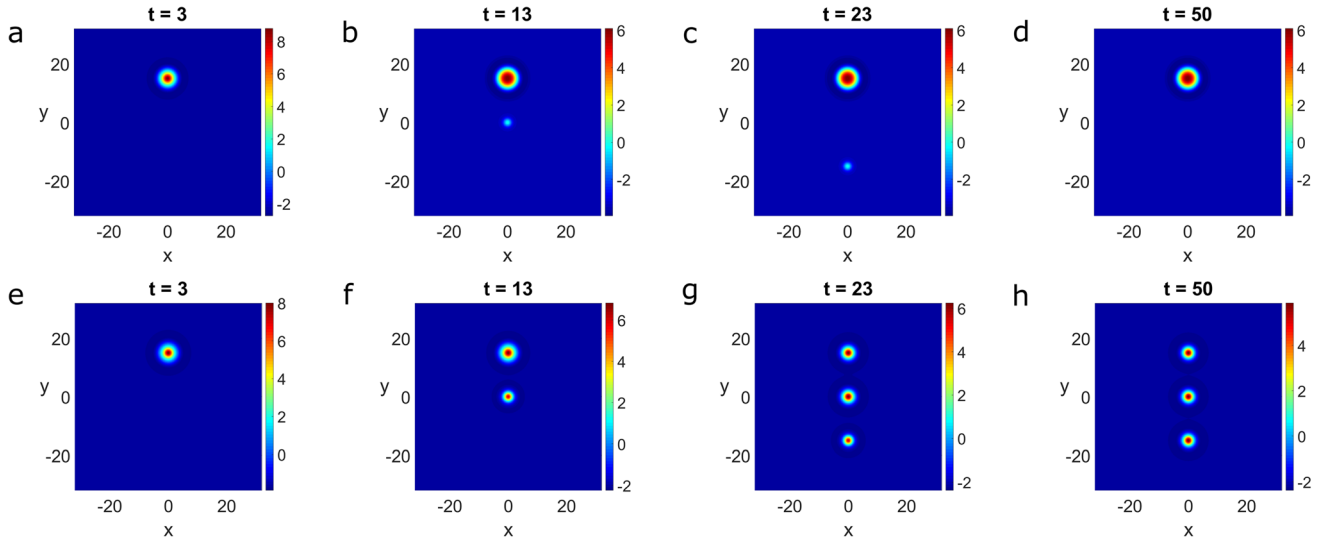

Fig. 6 Solutions of the Amari model (**a**–**d**) and the two-field model (**e**–**h**) created with sequential inputs. Inputs $I(\mathbf{r}, t)$ are applied at times $t_1 = 1$, $t_2 = 11$ and $t_3 = 21$. Snapshots taken at times: $t = 3$ (**a** and **e**), $t = 13$ (**b** and **f**), $t = 23$ (**c** and **g**), $t = 50$ (**d** and **h**). Parameters of the inputs are $A_I = 4$, $\sigma_I = 1$, $d_I = 2$. The kernel w is given by (2) with $A_{ex} = 2$, $A_{in} = 1$, $\sigma_{ex} = 1.6$, $\sigma_{in} = 2$ and $w_{inh} = 0.1$. Threshold $\theta = 0$, $K = 0$

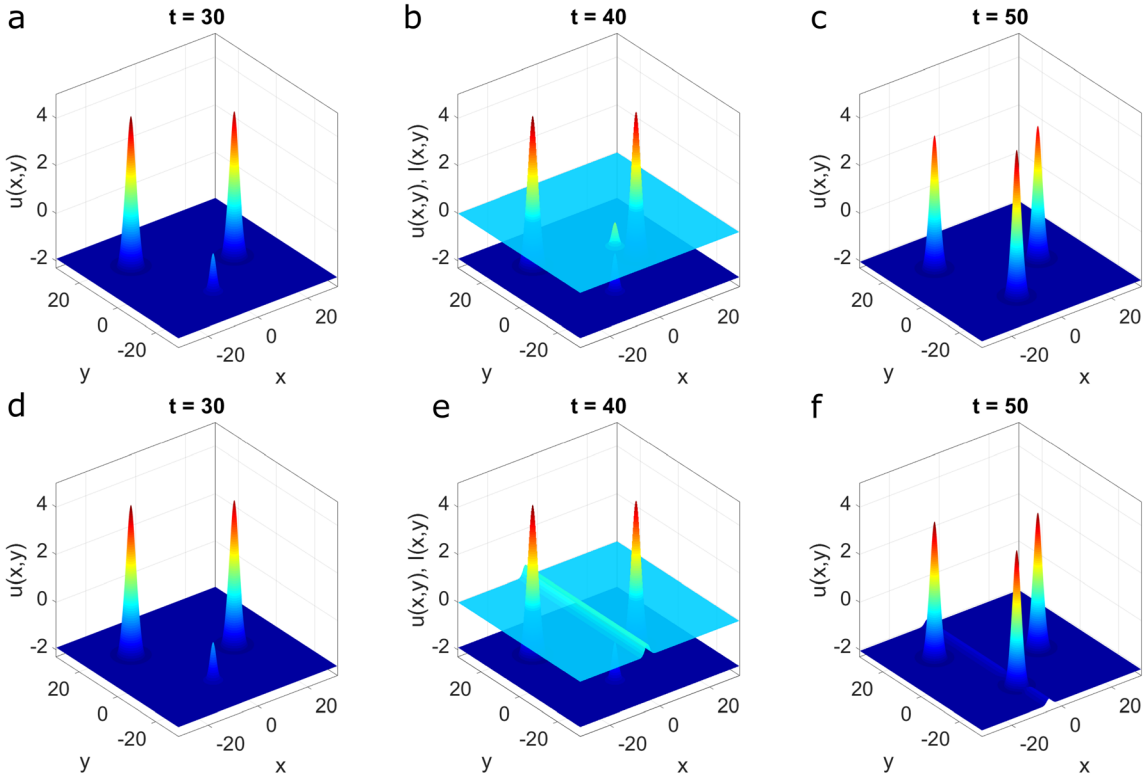

Fig. 7 Recovery of a subthreshold memories with specific (**a**–**c**) and unspecific input (**d**–**f**). Inputs $I(\mathbf{r}, t)$ with $A_I = 1.6$ are applied at times $t_1 = 1$, $t_2 = 11$ and $t_3 = 21$. (**b**) At time $t_4 = 40$, an additional input with $A_I = 1$ and $d_I = 1$ is applied at one of the positions. (**e**) At time $t_4 = 40$, a ridge input along the y-dimension with $A_I = 0.5$ and $d_I = 1$ is applied. Snapshots taken at times: $t = 30$ (**a** and **d**), $t = 40$ (**b** and **e**) and $t = 50$ (**c** and **f**). The remaining parameters of the inputs are $\sigma_I = 1$ and $d_I = 2$. The kernel w is given by (2) with $A_{ex} = 2$, $A_{in} = 1$, $\sigma_{ex} = 1.6$, $\sigma_{in} = 2$ and $w_{inh} = 0.1$. Threshold $\theta = 0$, $K = 0$

model behavior is conceptually in line with experimental findings showing that a long-lasting visual memory trace ($\geq 16s$) of a briefly presented subthreshold signal facilitates the detection of a target signal (Tanaka and Sagi 1998).

Current research not only investigates the impact of stimulus characteristics on the quality of WM representations but also addresses the active role for a top-down modulation during WM maintenance. Findings in so-called "retro-cuing" paradigms show that top down signals may prioritize items being held in WM even after encoding to cope with changing task demands (for review see (Gazzaley and Nobre 2012)). Retro-cues differ from post-cues in that they do not prompt the retrieval of a specific item but instead direct attention to a given object category in visual WM. A particularly interesting result comes from a study investigating retro-cues directing attention to the position of a near-threshold stimulus which otherwise would pass unseen. Spatial attention greatly improved the viewers' capacity to discriminate stimulus orientation, along with a drastic increase in subjective visibility (Sergent et al. 2013). Figure 7d–f shows model simulation with the two stimulus dimensions representing horizontal position (x-axis) and orientation (y-axis) of a stimulus array. Like in the simulation shown in the upper row, a sequence of three inputs is first presented (d). The retro-cue directing attention to the position of the third input is modeled as a transient input with the shape of a ridge along the orientation dimension, that is, all conjunctive neurons encoding the same location but different orientations receive the identical input (e). Despite the lack of specification in the input pattern, the transition from a subthreshold to an active state represented by the evolving bump allows a down-stream network to read-out the encoded orientation information (f).

Non-radially symmetric connectivity functions

In our analysis of rotationally invariant bump solutions of the planar two-field model, we closely followed previous work on stationary, localized activation patterns in field models of Amari type. A crucial assumption in all these studies is that the coupling function depends on the Euclidian distance, $|\mathbf{r} - \mathbf{r}'|$, between interacting neurons at positions $\mathbf{r}$ and $\mathbf{r}'$. Any deviation from this perfect circular symmetry can substantially perturb the continuous attractor, preventing the network from stabilizing stationary bumps at any location of the plane. How the brain deals with the problem of fine-tuning and maintenance of a continuous symmetry across neurons remains unknown and is a matter of considerable debate (Zylberberg and Strowbridge 2017). Here we test in numerical model simulations to which extent the local feedback mechanisms introduced in the two-field model allow us to relax the biologically implausible symmetry assumption. We consider three distinct perturbation of the radially symmetric weight distribution: 1) a coupling function with elliptic shape, 2) a systematic directional bias, and 3) a noise-induced heterogeneity.

Elliptic connectivity function

Neurons with receptive fields conjointly tuned to two stimulus features have been described (Drucker et al. 2009). This does not necessarily mean however that the tuning width is the same, regardless of which dimension is considered. Since narrow (broad) tuning curves imply that a small (large) fraction of neurons is active for a given stimulus parameter, the spatial spread of the neural population activity in the 2D feature space will differ along the two stimulus dimensions. Modeling such a distribution in WM applications as a self-sustained activity pattern requires a connectivity function with non-radially symmetric shape. Figure 8 depicts an example of a Mexican hat connectivity (2) for which the spatial ranges of excitation and inhibiton are larger for one dimension compared to the other, that is, $\sigma^2_{ex_x} > \sigma^2_{ex_y}$ and $\sigma^2_{in_x} > \sigma^2_{in_y}$. Figure 9 illustrates that the two-field model is able to stabilize a localized activity pattern with an elliptic contour which was initially triggered by a transient Gaussian input. In the Amari case, this activation pattern appears to be destabilized with a speed depending on the degree of the symmetry-breaking (not shown). Note that the bump shape not only reflects the asymmetry in the coupling function but may be also affected by input characteristics as shown in Fig. 5.

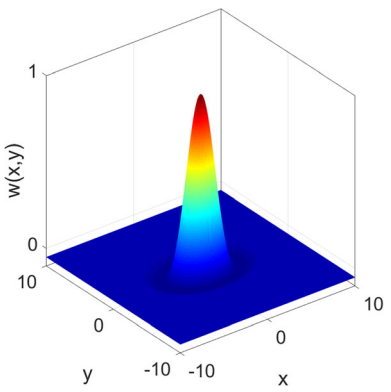

Fig. 8 Mexican hat connectivity with elliptic shape. The parameters are $A_{ex} = 2$, $A_{in} = 1.5$, $\sigma_{ex_x} = 1.6$, $\sigma_{ex_y} = 1.07$, $\sigma_{in_x} = 2$, $\sigma_{in_y} = 1.34$ and $w_{inh} = 0.05$

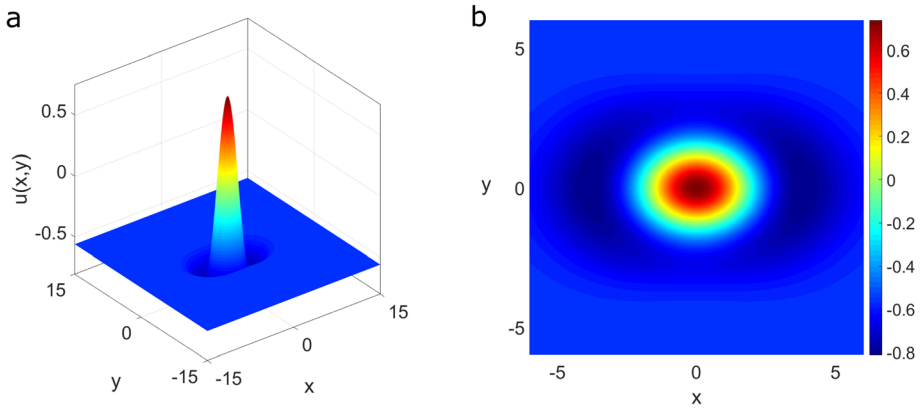

Fig. 9 Side view (**a**) and top view (**b**) of the persistent activity pattern created with the connectivity function shown in Fig. 8. Parameters of the input are $A_I = 1, \sigma_I = 2, d_I = 1$

Directional bias in the connectivity function

A second type of perturbation is a coupling with a systematic directional bias. In 1D CAN models, it has been shown that such a recurrent network architecture supports bumps freely traveling in the biased direction. This mechanism has been applied to model the direction selectivity of cortical neurons (Xie and Giese 2002; Zhang 1996; Bressloff and Wilkerson 2012) and to track moving objects through occlusion (Erlhagen and Bicho 2006). Here we generalize the mechanism to the 2D case to compare it with the behavior of the two-field model. The asymmetry can be implemented by adding to the standard Mexican hat kernel its directional derivative as an antisymmetric component (Zhang 1996)

$$w_{asym}(\mathbf{r}) = w(\mathbf{r}) + \eta D_v w(\mathbf{r}), \tag{12}$$

where w is the weight function given by (2), $D_v w$ is the derivative of w in the direction $v = (x, y)$ with a scaling factor $\eta > 0$ (compare Fig. 10). Alternatively, the Mexican hat kernel can be simply shifted by an offset r_0 (Xie and Giese 2002), that is, $w(\mathbf{r} - \mathbf{r_0}) = w(-(\mathbf{r} - \mathbf{r_0}))$. Figures 11 and 12 compare the pattern formation process of the Amari model (top row) and the two-field model (bottom row) at different times after the presentation of a transient input at time $t = 0$ at position $\mathbf{r} = (0, 0)$. A bias in y-direction using the derivative mechanism causes the Amari bump to travel in this direction whereas the bump in the two-field model remains stationary at the initial position (Fig. 11). With the offset mechanism ($\mathbf{r_0} = (0.5, 0.5)$), the two-field bump again remains stationary whereas the Amari bump moves in the biased direction (Fig. 12). The balance of excitation and inhibition mediated by the additional local feedback mechanism thus leads to a qualitatively different model behavior.

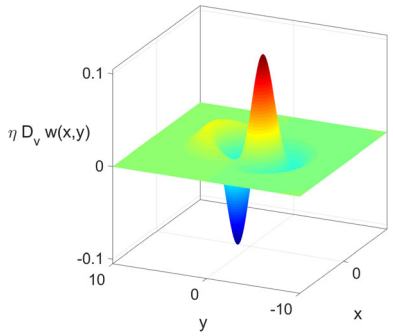

Fig. 10 Plot of $\eta D_v w(\mathbf{r})$ from (12) with $\eta = 5$

Heterogeneous connectivity function

For a theory of working memory based on persistent activity, weak random spatial fluctuations in the connection strength are perhaps the most worrisome perturbation of the assumed continuous symmetry. These fluctuations are to be expected when learning the coupling function with Hebbian plasticity (Zhang 1996; Zou et al. 2017). In the 1D case, it is well known that the presence of such synaptic heterogeneity causes a drift of an input-induced activity pattern to one of a finite number of attractor positions which are randomly spread over representational space. Additional processing mechanisms have been proposed to slow this memory drift and reduce the quantization error (Itskov et al. 2011; Renart et al. 2003). Following (Itskov et al. 2011), we add small noise to the weight profile which breaks the radial symmetry

$$w_{pert}(\mathbf{r}) = w(\mathbf{r}) + \epsilon^{1/2} \mathrm{d}\mathcal{Y}(\mathbf{r}, t), \tag{13}$$

where $\mathrm{d}\mathcal{Y}(\mathbf{r}, t)$ is a spatially white noise process and $\epsilon \ll 1$ is the noise amplitude. Figure 13 shows cross sections in x-direction (a) and y-direction (b) of the Mexican hat connectivity with fluctuations in the connection strengths. The simulation of the 2D Amari model essentially replicates the findings for the 1D case. The activity pattern drifts from the input location, $\mathbf{r} = (0, 0)$, towards a position close to the field boundary (c). The line indicates the center-of-mass

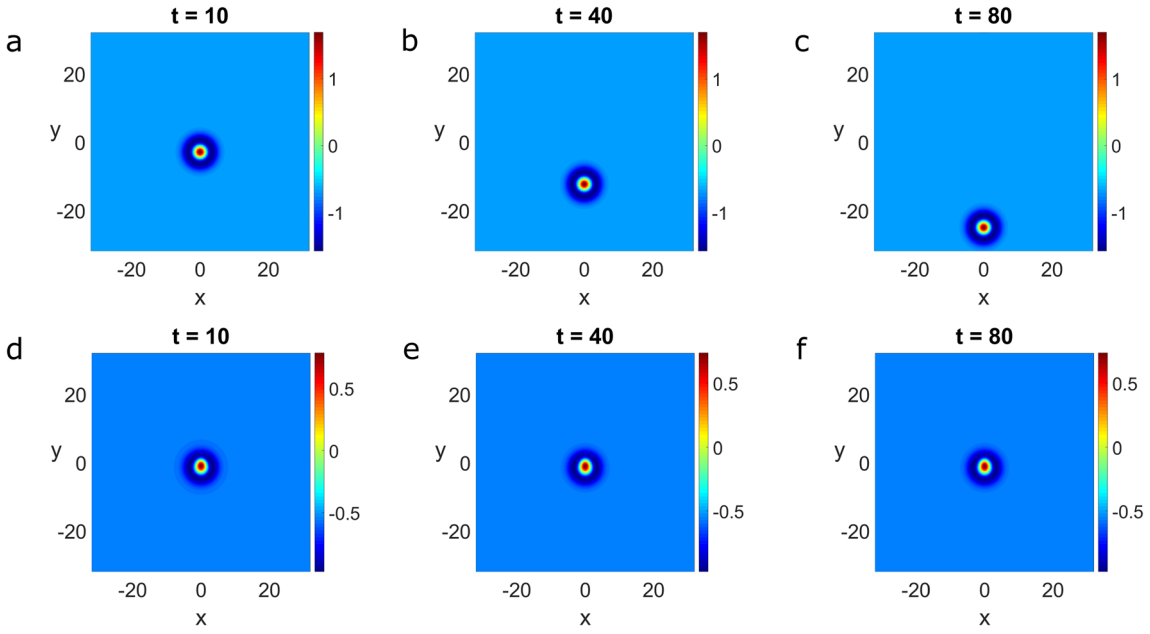

Fig. 11 Solutions of the Amari model (**a**–**c**) and the two-field model (**d**–**f**) created with the connectivity function shown in Fig. 10. Input $I(\mathbf{r}, t)$ with $A_I = 1$, $\sigma_I = 1$ and $d_I = 1$ is applied at time $t = 1$ at position $\mathbf{r}_{c_1} = (0, 0)$

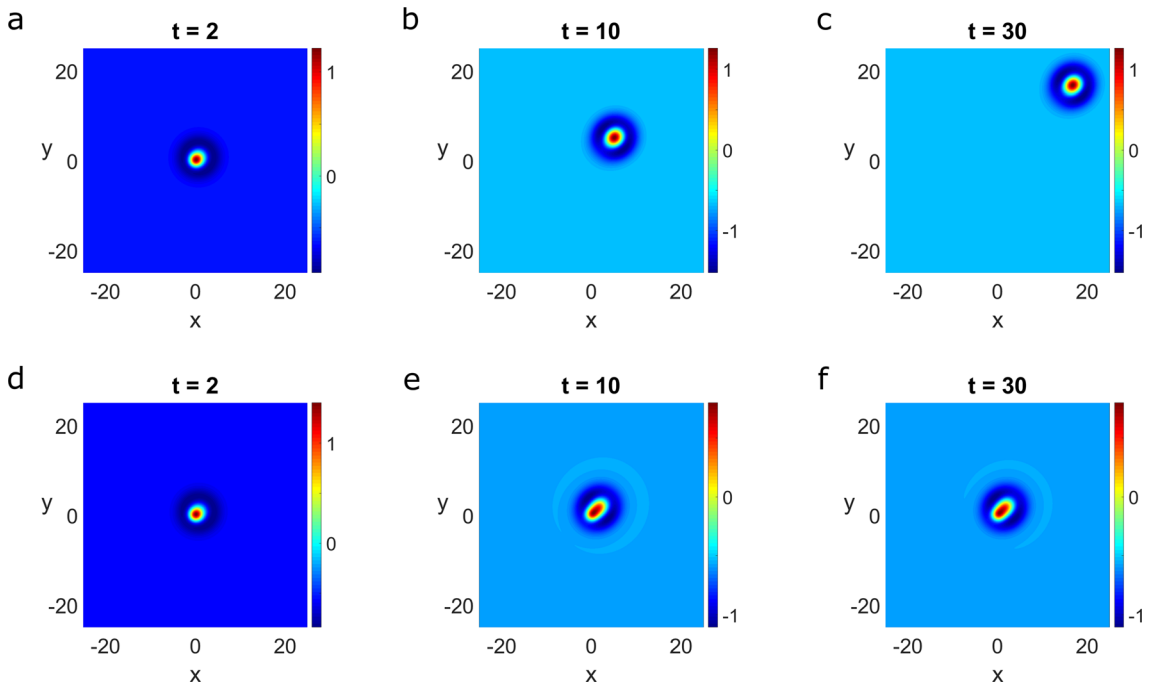

Fig. 12 Solutions of the Amari model (**a**–**c**) and the two-field model (**d**–**f**) created with the connectivity function centered at (0.5, 0.5). Input $I(\mathbf{r}, t)$ with $A_I = 2$, $\sigma_I = 1$ and $d_I = 1$ is applied at time $t = 1$ at position $\mathbf{r}_{c_1} = (0, 0)$

trajectory of the activity profile. Panel (d) depicts these center-out trajectories for 100 model simulations, each with the same initialization but with different random fluctuations of the Mexican hat coupling. The results can be directly compared with the simulations of the two-field model in panels (e) and (f). Here, the bump remains centered at the initial position. As can be seen in the top view of the heat map, the spatial fluctuations in the connection strengths cause a slight perturbation of the bump shape. This explains why the center-of-mass trajectories show

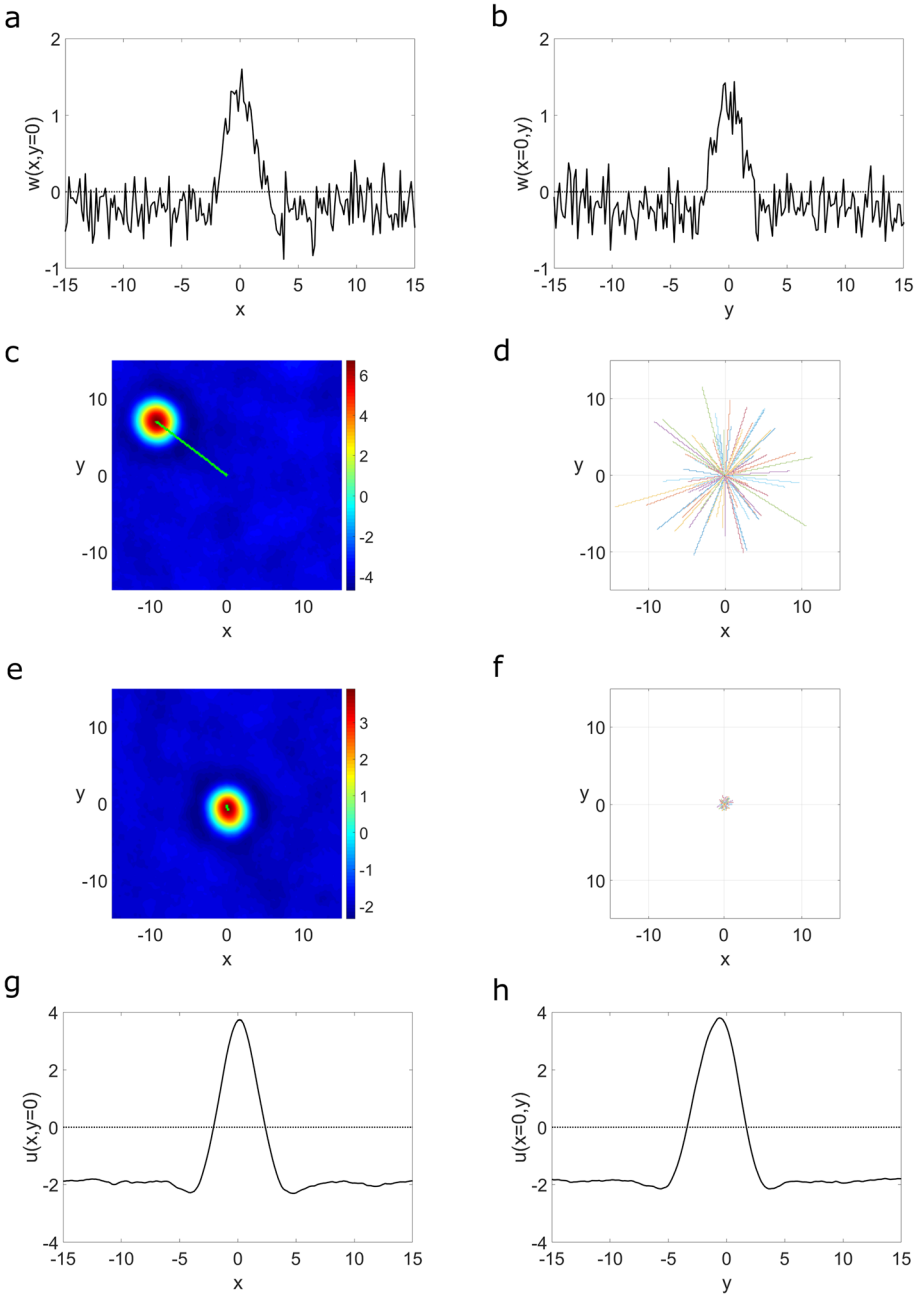

Fig. 13 **a** and **b**: Cross sections of the connectivity function given by (13) with $\epsilon = 0.05$. **c**–**f**: Bump drift in the Amari model (**c** and **d**) and in the two-field model (**e** and **f**). Input $I(\mathbf{r}, t)$ with $A_I = 1.5$, $\sigma_I = 1$ and $d_I = 1$ was applied at time $t = 1$ at position $\mathbf{r}_{c_1} = (0, 0)$. Panels (**c**) and (**e**) show single realizations at time $t = 200$, with the green line tracking the centroid of the bump. Panels (**d**) and (**f**) show the trajectories of the bump centroid over 100 trials. **g** and **h**: Cross sections of the solution of the two-field model from panel (**e**)

small fluctuations around the input location. As an example, panels (g) and (h) depict cross sections of a bump. Its peak position appears to be slightly shifted in the y-direction.

Discussion

In this work, we have analyzed and tested a bump attractor network which is able to robustly maintain in self-sustained activity not only the content but also the quality of continuous-valued information. The bump activity is stabilized equally well at any location of the two-dimensional feature space and the bump amplitude reflects a perfect temporal integration of inputs. This WM functionality is based on pure network mechanisms which in contrast to standard CAN models do not require a biologically unrealistic degree of fine-tuning of recurrent connections and nonlinearities. In previous modeling work, one approach to address this fine tuning problem and to increase the robustness of WM representations to extrinsic noise has been to introduce discrete "wells" in the continuous attractor manifold. This can be done for instance by using bistable neurons with different thresholds for a neural

integrator network (Koulakov et al. 2002) or a periodic heterogeneous coupling function for a neural field model of spatial WM (Kilpatrick and Ermentrout 2013). In both cases, a discretisation error is introduced since the neural integrator loses its sensitivity to weaker inputs and bumps are pinned to a finite number of discrete positions in representational space. The model simulations show that despite different types of substantial perturbations of the assumed coupling symmetry, a bump representing the accumulated evidence remains stationary at the stimulated site. Extrinsic noise is a second source affecting memory precision in CAN models. It randomly displaces encoded memories along the continuum of states. We have shown in our previous work on the one-dimensional two-field model that it also shows this diffusive drift pattern (Wojtak et al. 2021a). However, the variance of the displacement decreases drastically with increasing bump amplitude, suggesting that the fidelity of WM representations correlates with their precision (Klyszejko et al. 2014). We plan in future work to further explore this model prediction in 2D feature spaces.

The mathematical analysis of radially symmetric bump solutions revealed that the two-field model shows qualitatively the same behavior as the 2D Amari model. Stable and unstable bumps occur in pairs and non-radially symmetric perturbations of the stationary bump may lead to azimuthal instabilities. A major difference is the existence of persistent subthreshold activity patterns that only depend on a balanced local interaction between the two populations. As shown in the example of Fig. 7, the 2D case offers new perspective for modeling WM storage since it allows us to address binding information (e.g., which stimulus has been in which location, or in which sequential position of a series of stimuli) represented by the persistent activity of conjunctive neurons (Schneegans and Bays 2017; Wojtak et al. 2021b). DNF models of multi-item memory assume that feature binding is completely linked to space (Johnson et al. 2008). Different non-spatial features characterizing a single object (e.g., orientation, color, size) are represented by stable activation patterns of distinct populations of conjunctive feature-space neurons bound to location. A separate working memory population stores the locations of all items in the scene. During recall, the location information is used to couple the various features defining the specific object.

For the subthreshold memory trace of the third stimulus in Fig. 7e and f, a transient ridge input which specifies the location but not the orientation recovers a high fidelity memory of the combined location and orientation information. Without the content-specific pre-activation, the same input would lead to a homogeneous increase of activity. Any working-memory model thus needs to accommodate a representational state where information can be maintained without ever being inside the focus of attention (Bergström and Eriksson 2018).

Our purely activation based account differs from "activity silent" approaches postulating that WM storage may be instead accomplished by weight-based changes in synaptic connectivity even in the absence of sustained activity (Stokes 2015). A popular example is synaptic facilitation which is thought to temporarily amplify connections between neurons that are activated by a stimulus (Mongillo et al. 2008). The decaying synaptic memory trace can be reignited into activity by an unspecific input. Recent experimental evidence suggests that both mechanism are not mutually exclusive but may play different roles in specific WM tasks (Barbosa et al. 2020). Concerning the example in Fig. 7e and f, it remains an open question whether a stimulus that fails to cross the threshold for active reverberation (and subjective visibility) would still induce enough activity in the network to trigger a long lasting synaptic memory trace (e.g., > 16 s (Tanaka and Sagi 1998), see also the discussion in (Bergström and Eriksson 2018)).

The robustness of the two-field model to global changes in the neural gain and perturbations of the coupling symmetry opens new perspectives for learning the weights using Hebbian plasticity mechanisms (Zou et al. 2017; Zhang 1996). Even when stimuli are drawn from a continuous family, irregular training will introduce heterogeneities and/or directional biases into the synaptic connections. We plan to address the challenge of learning WM representations in attractor networks in future work. In this context it is important to notice that the performance of the two-field model is sensitive to changes in the local feedback loop. Introducing nonlinearities such as saturation by using for instance a piecewise linear transfer function will limit the continuous integrator capacity to a maximum bump amplitude. Any change that severely perturbs the local balance of excitation and inhibition will disrupt WM performance.

A WM model based on persistent population activity must necessarily include a forgetting mechanism. Since the two-field model works as a perfect neural integrator, an existing bump cannot be destabilized by simply applying a strong inhibitory input. To address this forgetting problem, we have proposed a plausible gating mechanism for the local feedback loop controlling the continuous input integration (Wojtak et al. 2021a). Functionally, the gate can be considered to be like a threshold that the input to a specific location must exceed in order to start the accumulation of bottom-up and top-down evidence for the specific feature value(s). The application of a sufficiently strong inhibition will drive any persistent activity in the u-field below this threshold. The decoupled dynamics of the two populations is then governed by the Amari field equation with a

stable homogeneous resting state. In recent works, we have used the capacity of attractor networks to stabilize multiple bumps with a continuum of amplitudes to address the problem of memorizing the temporal order and the relative timing of sequential events (Ferreira et al. 2020; Wojtak et al. 2021b). The information is stored in a gradient of activation strength such that the neural representation of each item is stronger than its successor. The event memories are autonomously recalled in the correct order and at the expected time using a competitive dynamics of a decision field which receives the activation gradient as subthreshold input.

Persistent elevated firing rates of populations of neurons may have also an oscillatory character. Indeed, the observation of stimulus-selective oscillatory dynamics in WM tasks has led to the hypothesis that neural oscillations in different frequency bands play an important role in the maintenance of information (Roux and Uhlhaas 2014). Computational studies investigating this hypothesis typically use neural mass models with no inherent spatial structure (Ursino et al. 2023; Pina et al. 2018). Ensemble activity of cortical microcircuits comprising distinct cell types shows stable oscillations. Network architectures of such microcircuits are then used to address salient attributes of working memory such as maintaining multiple items and their serial order or to establish feature binding through synchronous relationships. While sharing many of the basic research questions, the spatially structured, distance-dependent recurrent interactions implemented in DNF models provide explanatory power for many experimentally observed metric effects in WM and other cognitive tasks (e.g., memory precision, feature misbinding errors, for an overview of experimental and modeling studies see (Schöner and Spencer 2016)). It will be interesting to explore in future work the potentially complementary roles that persistent population activity with stationary or oscillatory character might have. This includes WM tasks for which the metric dimension is not as clear or for which a spatial-binding model is not sufficient since the stimuli are presented sequentially at a single location.

Appendix A

The double integral in (5) can be calculated using the Fourier transforms and Bessel function identities (Bressloff 2012). We start with expressing $w(r)$ as a 2D Fourier transform using polar coordinates

$$\begin{aligned} w(r) &= \frac{1}{2\pi}\int_{\mathbb{R}^2} \mathrm{e}^{i\mathbf{r}\cdot\mathbf{k}}\widehat{w}(\mathbf{k})\mathrm{d}\mathbf{k} \\ &= \frac{1}{2\pi}\int_0^\infty \left(\int_0^{2\pi} \mathrm{e}^{ir\rho\cos\phi}\widehat{w}(\rho)\mathrm{d}\phi\right)\rho\mathrm{d}\rho, \end{aligned} \tag{14}$$

where $\widehat{w}$ denotes the Fourier transform of w and $\mathbf{k} = (\rho, \phi)$. Using the integral representation

$$\frac{1}{2\pi}\int_0^{2\pi} \mathrm{e}^{ir\rho\cos\phi}\mathrm{d}\phi = J_0(\rho r), \tag{15}$$

where J_0 is the Bessel function of the first kind, we express w in terms of its Hankel transform of order zero

$$w(r) = \int_0^\infty \widehat{w}(\rho)J_0(\rho r)\rho\mathrm{d}\rho, \tag{16}$$

which, when substituted into (5), gives

$$\begin{aligned} U(r) = V(r) &+ \int_0^{2\pi}\int_0^R \left(\int_0^\infty \hat{w}(\rho)J_0(\rho|\mathbf{r}-\mathbf{r}'|)\rho\mathrm{d}\rho\right) r'\mathrm{d}r'\mathrm{d}\zeta', \end{aligned} \tag{17a}$$

$$\begin{aligned} V(r) = U(r) &- \int_0^{2\pi}\int_0^R \left(\int_0^\infty \hat{w}(\rho)J_0(\rho|\mathbf{r}-\mathbf{r}'|)\rho\mathrm{d}\rho\right) r'\mathrm{d}r'\mathrm{d}\zeta'. \end{aligned} \tag{17b}$$

We reverse the order of integration and use the addition theorem

$$\begin{aligned} &J_0(\rho\sqrt{r^2+r'^2-2rr'\cos\zeta'}) \\ &\quad = \sum_{m=0}^\infty \epsilon_m J_m(\rho r)J_m(\rho r')\cos m\zeta', \end{aligned} \tag{18}$$

where $\epsilon_0 = 1$ and $\epsilon_n = 2$ for $n \geq 1$. Then using the identity $J_1(\rho R)R = \rho\int_0^R J_0(\rho r')r'\mathrm{d}r'$, we obtain (6). Note that the Fourier transform of (4) is easily calculated using the result that the Fourier transform of $K_0\left(\frac{r}{\sigma}\right) = \frac{2\pi}{r^2+\sigma^2}$.

Appendix B

Using polar coordinates we can rewrite system (8) as

$$\begin{aligned} \lambda\psi(r,\phi) &= -\psi(r,\phi) + \zeta(r,\phi) \\ &+ \int_0^{2\pi}\mathrm{d}\phi'\int_0^\infty r'\mathrm{d}r' \\ &w(\sqrt{r^2+r'^2-2rr'\cos\phi})\delta(U(r')-\theta)\psi(r',\phi-\phi'), \end{aligned} \tag{19a}$$

$$\lambda\zeta(r,\phi) = -\zeta(r,\phi) + \psi(r,\phi) - \int_0^{2\pi} \mathrm{d}\phi' \int_0^{\infty} r'\mathrm{d}r' w(\sqrt{r^2 + r'^2 - 2rr'\cos\phi})\delta(U(r') - \theta)\psi(r', \phi - \phi'). \tag{19b}$$

We look for solutions of the form

$$(\psi(r,\phi), \zeta(r,\phi)) = \mathrm{e}^{in\phi}(\psi(r), \zeta(r)), \tag{20}$$

where n is the number of modes of the boundary perturbation. System (19) then takes the form

$$\lambda\psi(r)\mathrm{e}^{in\phi} = -\psi(r)\mathrm{e}^{in\phi} + \zeta(r)\mathrm{e}^{in\phi} + \int_0^{2\pi} \mathrm{d}\phi' \int_0^{\infty} r'\mathrm{d}r' w(\sqrt{r^2 + r'^2 - 2rr'\cos(\phi - \phi')})\delta(U(r') - \theta)\psi(r')\mathrm{e}^{in(\phi-\phi')}, \tag{21a}$$

$$\lambda\zeta(r)\mathrm{e}^{in\phi} = -\zeta(r)\mathrm{e}^{in\phi} + \psi(r)\mathrm{e}^{in\phi} - \int_0^{2\pi} \mathrm{d}\phi' \int_0^{\infty} r'\mathrm{d}r' w(\sqrt{r^2 + r'^2 - 2rr'\cos(\phi - \phi')})\delta(U(r') - \theta)\psi(r')\mathrm{e}^{in(\phi-\phi')}. \tag{21b}$$

We set $r = R$ and after dividing both sides by $\mathrm{e}^{in\phi}$ we get

$$\lambda\psi(R) = -\psi(R) + \zeta(R) + \int_0^{2\pi} \mathrm{d}\phi R w(R\sqrt{2 - 2\cos\phi})\frac{\psi(R)\mathrm{e}^{-in\phi}}{|U'(R)|}, \tag{22a}$$

$$\lambda\zeta(R) = -\zeta(R) + \psi(R) - \int_0^{2\pi} \mathrm{d}\phi R w(R\sqrt{2 - 2\cos\phi})\frac{\psi(R)\mathrm{e}^{-in\phi}}{|U'(R)|}. \tag{22b}$$

The system (22) can be written as

$$A\begin{bmatrix}\psi(R)\\ \zeta(R)\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix},$$

where the matrix A is given by

$$A = \begin{bmatrix}\lambda + 1 - S_n & -1\\ -1 + S_n & \lambda + 1\end{bmatrix},$$

with

$$S_n = \frac{R}{|U'(R)|}\int_0^{2\pi} w(R\sqrt{2 - 2\cos\phi})\mathrm{e}^{-in\phi}\mathrm{d}\phi. \tag{23}$$

Then, we find that

$$(\lambda + 1 + S_n)(\lambda + 1) - (-1 + S_n)(-1 + S_n) = 0. \tag{24}$$

Hence the eigenvalues of A are

$$\lambda_{-1} = 0, \tag{25}$$

$$\lambda_n = -2 + S_n. \tag{26}$$

Note that λ_n is real, since after setting $\sqrt{2 - 2\cos\phi} = 2\sin\left(\frac{\phi}{2}\right)$ and rescaling ϕ we have

$$\mathrm{Im}\{\lambda_n\} = -\frac{2R}{|U'(R)|}\int_0^{\pi} w(2R\sin(\phi))\sin(2n\phi)\mathrm{d}\phi = 0, \tag{27}$$

i.e., the integrand is odd-symmetric about $\frac{\pi}{2}$. Hence,

$$\lambda_n = \mathrm{Re}\{\lambda_n\} = -2 + \frac{R}{|U'(R)|}\int_0^{2\pi} w(2R\sin(\phi/2))\cos(n\phi)\mathrm{d}\phi, \tag{28}$$

with the integrand even-symmetric about $\frac{\pi}{2}$.

We then evaluate the integral in (28) using Bessel functions

$$\begin{aligned}\int_0^{2\pi} w(2R\sin(\phi'/2))\cos(n\phi')\mathrm{d}\phi' &= \int_0^{2\pi}\left(\int_0^{\infty}\hat{w}(\rho)J_0(\rho(2R\sin(\phi'/2)))\rho\mathrm{d}\rho\right)\cos\phi'\mathrm{d}\phi'\\ &= 2\pi\int_0^{\infty}\hat{w}(\rho)J_n(\rho R)J_n(\rho R)\rho\mathrm{d}\rho.\end{aligned} \tag{29}$$

We differentiate (6a) with respect to r, and, knowing that $U(r) + V(r) = K$ we have

$$U'(R) = -\pi R\int_0^{\infty}\widehat{w}(\rho)J_1(\rho R)J_1(\rho R)\rho\mathrm{d}\rho. \tag{30}$$

We can now write the eigenvalues of A as (9) and (10).

Appendix C

Numerical simulations of the model were done in MATLAB using a forward Euler method with uniform spatial mesh with dx = 0.05 and time step dt = 0.01. To compute the two-dimensional spatial convolution of w and f we employ a two-dimensional fast Fourier transform (2D FFT), using MATLAB's in-built functions `fft2` and `ifft2` to perform the Fourier transform and the inverse Fourier transform, respectively. Periodic boundary conditions are used. By choosing a sufficiently large domain size, we make sure that the localized patterns evolve sufficiently far from the boundaries.

For performing numerical continuation, we use the method described in (Rankin et al. 2014) and adapt MATLAB code available in (Avitabile 2016). The main advantage of this method is that it can be applied directly to the full integral model. This is possible due to the usage of Newton-GMRES solvers combined with a fast Fourier transform (FFT) employed for computing the convolution term (Rankin et al. 2014).

Acknowledgments The work received financial support from FCT through the PhD fellowship PD/BD/128183/2016, the project "Neurofield" (PTDC/MAT-APL/31393/2017), the Project I-CATER: Intelligent robotic Coworker Assistant for industrial Tasks with an Ergonomics Rationale (Ref[a] PTDC/EEI-ROB/3488/20211), R&D Units Project Scope: UIDB/00319/2020" - ALGORITMI Research Centre and the Research Centre CMAT within the project UID/MAT/00013/2020.

Data Availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

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

Avitabile D (2016) Numerical computation of coherent structures in spatially-extended systems. Second International Conference on Mathematical Neuroscience, Antibes Juan-les-Pins, 2016

Barak O, Tsodyks M (2014) Working models of working memory. Curr Opin Neurobiol 25:20–24

Barbosa J, Stein H, Martinez RL et al. (2020) Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. Nat Neurosci 23(8):1016–1024

Bergström F, Eriksson J (2018) Neural evidence for non-conscious working memory. Cereb Cortex 28(9):3217–3228

Bressloff PC (2012) Spatiotemporal dynamics of continuum neural fields. J Phys A Math Theor 45(3):033001

Bressloff PC, Coombes S (2013) Neural bubble dynamics revisited. Cognit Comput 5(3):281–294

Bressloff PC, Wilkerson J (2012) Traveling pulses in a stochastic neural field model of direction selectivity. Front Comput Neurosci 6:90

Brody CD, Romo R, Kepecs A (2003) Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. Curr Opin Neurobiol 13(2):204–211

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

Constantinidis C, Wang XJ (2004) A neural circuit basis for spatial working memory. Neuroscientist 10(6):553–565

Constantinidis C, Franowicz MN, Goldman-Rakic PS (2001) The sensory nature of mnemonic representation in the primate prefrontal cortex. Nat Neurosci 4(3):311–316

Drucker DM, Kerr WT, Aguirre GK (2009) Distinguishing conjoint and independent neural tuning for stimulus features with FMRI adaptation. J Neurophysiol 101(6):3310–3324

Erlhagen W, Bicho E (2006) The dynamic neural field approach to cognitive robotics. J Neural Eng 3(3):R36

Ferreira F, Wojtak W, Sousa E et al. (2020) Rapid learning of complex sequences with time constraints: a dynamic neural field model. IEEE Trans Cogn Develop Syst 13(4):853–864

Gazzaley A, Nobre AC (2012) Top-down modulation: bridging selective attention and working memory. Trends Cogn Sci 16(2):129–135

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

Johnson JS, Spencer JP, Schöner G (2008) Moving to higher ground: the dynamic field theory and the dynamics of visual cognition. New Ideas Psychol 26(2):227–251

Johnson JS, Spencer JP, Luck SJ et al. (2009) A dynamic neural field model of visual working memory and change detection. Psychol Sci 20(5):568–577

Khona M, Fiete IR (2021) Attractor and integrator networks in the brain. arXiv preprint arXiv:2112.03978

Kilpatrick ZP, Ermentrout B (2013) Wandering bumps in stochastic neural fields. SIAM J Appl Dyn Syst 12(1):61–94

Klyszejko Z, Rahmati M, Curtis CE (2014) Attentional priority determines working memory precision. Vision Res 105:70–76

Koulakov AA, Raghavachari S, Kepecs A et al. (2002) Model for a robust neural integrator. Nat Neurosci 5(8):775–782

Lewis-Peacock JA, Drysdale AT, Oberauer K et al. (2012) Neural evidence for a distinction between short-term memory and the focus of attention. J Cogn Neurosci 24(1):61–79

Lim S, Goldman MS (2013) Balanced cortical microcircuitry for maintaining information in working memory. Nat Neurosci 16(9):1306–1314

Ma WJ, Husain M, Bays PM (2014) Changing concepts of working memory. Nat Neurosci 17(3):347–356

Mégardon G, Tandonnet C, Sumner P et al. (2015) Limitations of short range Mexican hat connection for driving target selection in a 2d neural field: activity suppression and deviation from input stimuli. Front Comput Neurosci 9:128

Mongillo G, Barak O, Tsodyks M (2008) Synaptic theory of working memory. Science 319(5869):1543–1546

Pina JE, Bodner M, Ermentrout B (2018) Oscillations in working memory and neural binding: a mechanism for multiple memories and their interactions. PLoS Comput Biol 14(11):e1006517

Rankin J, Avitabile D, Baladron J et al. (2014) Continuation of localized coherent structures in nonlocal neural field equations. SIAM J Sci Comput 36(1):B70–B93

Renart A, Song P, Wang XJ (2003) Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. Neuron 38(3):473–485

Rose NS, LaRocque JJ, Riggall AC et al. (2016) Reactivation of latent working memories with transcranial magnetic stimulation. Science 354(6316):1136–1139

Roux F, Uhlhaas PJ (2014) Working memory and neural oscillations: alpha-gamma versus theta-gamma codes for distinct WM information? Trends Cogn Sci 18(1):16–25

Rubin JE, Troy WC (2004) Sustained spatial patterns of activity in neuronal populations without recurrent excitation. SIAM J Appl Math 64(5):1609–1635

Schneegans S, Bays PM (2017) Restoration of fMRI decodability does not imply latent working memory states. J Cogn Neurosci 29(12):1977–1994

Schöner G, Spencer JP (2016) Dynamic thinking: a primer on dynamic field theory. Oxford University Press

Scotti PS, Hong Y, Leber AB et al. (2021) Visual working memory items drift apart due to active, not passive, maintenance. J Exp Psychol Gen 150(12):2506

Sergent C, Wyart V, Babo-Rebelo M et al. (2013) Cueing attention after the stimulus is gone can retrospectively trigger conscious perception. Curr Biol 23(2):150–155

Stokes MG (2015) Activity-silent working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn Sci 19(7):394–405

Sutterer DW, Foster JJ, Adam KC et al. (2019) Item-specific delay activity demonstrates concurrent storage of multiple active neural representations in working memory. PLoS Biol 17(4):e3000239

Tanaka Y, Sagi D (1998) A perceptual memory for low-contrast visual signals. Proc Natl Acad Sci 95(21):12729–12733

Ursino M, Cesaretti N, Pirazzini G (2023) A model of working memory for encoding multiple items and ordered sequences exploiting the theta-gamma code. Cogn Neurodyn 17:489–521

Wildegger T, Humphreys G, Nobre AC (2016) Retrospective attention interacts with stimulus strength to shape working memory performance. PloS One 11(10):e0164174

Wimmer K, Nykamp DQ, Constantinidis C et al. (2014) Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat Neurosci 17(3):431–439

Wojtak W, Ferreira F, Bicho E, et al. (2019) Neural field model for measuring and reproducing time intervals. In: International conference on artificial neural networks, Springer, pp 327–338

Wojtak W, Coombes S, Avitabile D et al. (2021) A dynamic neural field model of continuous input integration. Biol Cybern 115(5):451–471

Wojtak W, Ferreira F, Vicente P et al. (2021) A neural integrator model for planning and value-based decision making of a robotics assistant. Neural Comput Appl 33(8):3737–3756

Wu S, Hamaguchi K, Si Amari (2008) Dynamics and computation of continuous attractors. Neural Comput 20(4):994–1025

Xie X, Giese MA (2002) Nonlinear dynamics of direction-selective recurrent neural media. Phys Rev E 65(5):051904

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

Zou X, Ji Z, Liu X, et al. (2017) Learning a continuous attractor neural network from real images. In: International conference on neural information processing, Springer, pp 622–631

Zylberberg J, Strowbridge BW (2017) Mechanisms of persistent activity in cortical circuits: possible neural substrates for working memory. Annu Rev Neurosci 40:603