CrossMark

Synaptic efficacy shapes resource limitations in working memory

Nikhil Krishnan[1] · Daniel B. Poll[2] · Zachary P. Kilpatrick[1] 🔴

Abstract

Working memory (WM) is limited in its temporal length and capacity. Classic conceptions of WM capacity assume the system possesses a finite number of slots, but recent evidence suggests WM may be a continuous resource. Resource models typically assume there is no hard upper bound on the number of items that can be stored, but WM fidelity decreases with the number of items. We analyze a neural field model of multi-item WM that associates each item with the location of a bump in a finite spatial domain, considering items that span a one-dimensional continuous feature space. Our analysis relates the neural architecture of the network to accumulated errors and capacity limitations arising during the delay period of a multi-item WM task. Networks with stronger synapses support wider bumps that interact more, whereas networks with weaker synapses support narrower bumps that are more susceptible to noise perturbations. There is an optimal synaptic strength that both limits bump interaction events and the effects of noise perturbations. This optimum shifts to weaker synapses as the number of items stored in the network is increased. Our model not only provides a circuit-based explanation for WM capacity, but also speaks to how capacity relates to the arrangement of stored items in a feature space.

Keywords Bump attractor · Working memory · Capacity · Limited resource · Interface methods

1 Introduction

Working memory (WM) is defined by both its short timescale and its capacity limitations (Ma et al. 2014). Detailed behavioral and electrophysiological recordings have demonstrated that WM is associated with persistent neural activity in a number of cortical regions (Constantinidis and Klingberg 2016). Neural and synaptic activity fluctuations account for commonly observed errors accumulated during the delay-period of typical WM tasks (Compte et al. 2000; Wimmer et al. 2014). However, there is

controversy surrounding the origin of errors arising from limitations of WM capacity. Classic models contend that item-limits are best defined by a 'slot model,' placing a hard upper bound on the number of items that can be stored (Luck and Vogel 1997; Cowan 2010). On the other hand, recent evidence suggests a 'resource model,' with no hard item-number limit, in which a fixed continuous resource is spread across an arbitrary number of items to be remembered (Wilken and Ma 2004; Bays and Husain 2008; van den Berg et al. 2012; Keshvari et al. 2013).

Both the slots model and the resource model reproduce some gross statistics from WM tasks equally well. For example, the recall variability tends to increase in both models for a WM task with a high number of items (Luck and Vogel 2013; Bays et al. 2009). However, for lower item counts, response variability is flat in a slots model whereas it increases in a resource model. In addition, more task-relevant cues can be stored with higher precision in a continuous resource model, but not in a slots model. Recent experiments have demonstrated that WM precision varies for low item counts (Bays et al. 2009) and for cued versus uncued items (Gorgoraptis et al. 2011). These results have been recapitulated in several human and non-human primate studies, suggesting flexibility in the allocation of WM

Action Editor: A. Compte

✉ Zachary P. Kilpatrick
zpkilpat@colorado.edu

Nikhil Krishnan
Nikhil.Krishnan@colorado.edu

Daniel B. Poll
daniel.poll@northwestern.edu

[1] Department of Applied Mathematics, University of Colorado Boulder, Boulder, CO 80309, USA

[2] Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208, USA

(Buschman et al. 2011; Lara and Wallis 2012). Resource models of WM allow for such flexibility, suggesting many possibilities for how storage precision varies across task parameters (van den Berg et al. 2012; Fougnie et al. 2012).

Computational models that capture behavioral patterns in multi-item WM are an active area of research (Barak and Tsodyks 2014). It remains an open question what neural mechanisms underlie these trends in response variability. Recent studies have extended the framework of continuous attractor networks, successful in capturing error accumulation in single-item WM tasks (Wimmer et al. 2014), to account for errors observed in multi-item WM (Edin et al. 2009; Wei et al. 2012; Almeida et al. 2015). These models are well-suited to store memoranda drawn from a continuous space, such as locations and colors (See Fig. 1). Recurrent networks comprised of a locally excitatory population coupled to a broadly tuned inhibitory population produce "bumps" of persistent neural activity (Amari 1977; Compte et al. 2000). Bumps encode the remembered location of a presented angle during the WM delay period, and fluctuations arising from stochastic spiking or synaptic transmission degrade memory of the initial position (Compte et al. 2000; Kilpatrick et al. 2013). Multi-item WM errors arise in these models via the interactions of multiple bumps, each bump encoding a distinct angle (Edin et al. 2009; Wei et al. 2012; Almeida et al. 2015). Bumps can repel, merge, or annihilate one another via nonlocal synaptic interactions of the network. For randomly chosen angles, the relative precision of recall decreases with set size according to a power law (Wei et al. 2012), as in Bays and Husain (2008). Thus, a multiple bumps model of WM appears to reconcile observed behavioral trends with known neural circuit mechanisms for storing WM using persistent activity.

These previous studies were performed using large-scale spiking simulations, however, and could not draw clear connections between parameters and the model's WM performance. An advantage of using neural field equations

to describe large-scale network interactions is that they are analytically tractable, and their dynamics can often be approximated by low-dimensional systems that solely describe variables of interest (Bressloff 2012). For instance, previous neural field studies of bump attractor models of single-item WM have developed explicit expressions for the relationship between network connectivity and the response variability (Kilpatrick 2013; Carroll et al. 2014; Kilpatrick 2017). Other work has explored the interaction of multiple bumps in neural field equations, but in special cases which are not relevant to the problem of storing an arbitrary set of memoranda (Laing and Troy 2003a; Bressloff 2005; Laing et al. 2002; Lu et al. 2011). A robust model for storing multiple items would allow for multiple bumps to be stored at arbitrary locations around a network. Our study explores tradeoffs in the strength of neural architecture as it impacts bumps' response to fluctuations, as well as interactions between neighboring bumps.

We utilize interface methods, originally applied to single bump solutions (Amari 1977; Coombes et al. 2012), to project the dynamics of multiple bumps in a neural field to a low-dimensional system of differential equations for the edges of the bumps. This approximate system can be analyzed in order to uncover the relationship between the architecture of the network and the robustness of multi-item WM. In particular, we examine how bumps interact with one another, and how they respond to external fluctuations that model the known stochastic evolution of persistent activity during the WM delay-period (Wimmer et al. 2014). Interestingly, increasing the strength of synaptic coupling makes networks with bumps that are robust to noise, but at the cost of producing stronger interactions between bumps. As a result, networks with the lowest response variability have an intermediate value of synaptic strength, which trades off the robustness of wide bumps to noise with the increased precision of networks containing narrower bumps.

2 Neural field model of visuospatial working memory

Most bump attractor models of working memory (WM) focus on tasks where a subject must remember a single orientation each trial (Kilpatrick and Ermentrout 2013; Wimmer et al. 2014). However, WM capacity can be probed by testing subjects' ability to recall multiple items (Bays and Husain 2008; Zhang and Luck 2008). We analyze a recurrent, spatially-organized network, which can represent multiple orientations during the delay period of a visuospatial WM task (Almeida et al. 2015). The model is similar to the case of single-item WM, but the network architecture plays an important role in shaping memory capacity (Bays 2015).
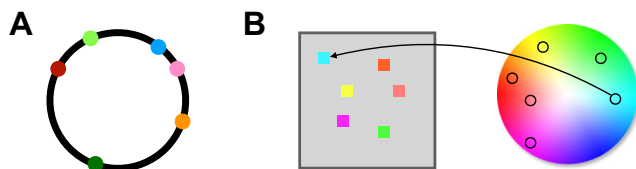

Fig. 1 Examples of multi-item visual stimuli used in working memory (WM) tasks (Zhang and Luck 2008; Bays et al. 2009; Ma et al. 2014). **a** Memoranda here are angles on a circle corresponding to the dot locations, identified by their color. Subject will be required to memorize all objects and then just recall one item; e.g., the location of the blue dot. **b** Alternatively, subjects may have to memorize the color of each item. For example, a subject may be asked what the color of the top left square was

2.1 Model definition

We study a neural field model in which locations of neurons correspond to their preferred stimulus orientation, organized in a ring architecture with slow local excitation and broad inhibition (Ermentrout 1998):

$$
\begin{aligned}
\mathrm{d}u(x,t) = & [-u(x,t) + w(x) * H(u(x,t) - \theta)] \, \mathrm{d}t \\
& + \sqrt{\epsilon \cdot |u(x,t)|} \, \mathrm{d}Z(x,t).
\end{aligned} \tag{2.1}
$$

The variable $u(x,t)$ represents synaptic input to spatial location $x \in [-L, L]$ at time t, which is periodic so $u(L,t) = u(-L,t)$. The weight function $w(x - y)$ represents the synaptic connectivity from neurons at location y to location x via the convolution $w * H(u - \theta) = \int_{-L}^{L} w(x-y)H(u(y,t) - \theta)\mathrm{d}y$. Note, we assume the weight function is even $w(x) = w(-x)$ and satisfies periodicity $w(-L) = w(L)$.

We consider the weight function,

$$
w(x - y) = \mathcal{A}(1 - |x - y|_L)e^{-|x-y|_L}, \tag{2.2}
$$

with local excitation and broad inhibition (Coombes and Owen 2005), where $\mathcal{A}$ parameterizes the synaptic strength, and $|x|_L = \min(|x - y|, |2L - |x - y||)$ is the distance on the ring (Fig. 2). Note, neurons with similar orientation will tend to activate one another, while neurons with dissimilar orientation tend to inhibit one another. Integrating $\int_{-L}^{L} w(x - y)\mathrm{d}y = 2\mathcal{A}(1 - e^{-L}) - 2\mathcal{A}(1 - (1 + L)e^{-L}) =$

Fig. 2 Recurrent network model of multiple-item WM. **a** The neural field, Eq. (2.1) is comprised of local populations (green triangles) organized on a ring with distance-dependent connectivity. This single layer describes the activity $u(x,t)$ of a combined excitatory/inhibitory neural field, derived in the limit of fast inhibition (Amari 1977; Pinto and Ermentrout 2001; Carroll et al. 2014). Strong effective excitation (blue dots) is narrow whereas weaker effective inhibition (red squares) is wide. **b** Weight function, Eq. (2.2), scaled by different maximal synaptic strengths $\mathcal{A} = \max_{x \in [-L,L]} w(x)$

$2\mathcal{A}Le^{-L}$, we find the total excitation excitation and inhibition is approximately balanced ($\int_{-L}^{L} w(x - y)\mathrm{d}y \approx 0$) for $L \gg 1$. Since we are interested in focusing on how scaling $\mathcal{A}$ impacts WM capacity for angles, we fix $L := 180$ consistent with typical oculomotor delayed-response tasks (Funahashi et al. 1989; Goldman-Rakic 1995; Wimmer et al. 2014; Constantinidis and Klingberg 2016). The weight function Eq. (2.2) is one example in a class of synaptic kernels that arises in the limit of fast inhibition, such that separate excitatory and inhibitory populations can be combined into a single population Eq. (2.1) (Carroll et al. 2014). The nonlinearity in Eq. (2.1) is a Heaviside step function

$$
H(u - \theta) = \begin{cases} 1, & u > \theta, \\ 0, & u < \theta, \end{cases} \tag{2.3}
$$

representing the input-output relationship between synaptic activation and population firing rate at x. Smooth sigmoids are also often used, but the qualitative dynamics of Eq. (2.1) remain similar for steep sigmoids (Coombes and Schmidt 2010; Bressloff 2012). We exploit the fact that the output of $H(u - \theta)$ is binary ($\{0, 1\}$) to develop interface equations for the dynamics of bumps in Eq. (2.1), adapting methods originally developed by Amari (1977) and extended by Coombes et al. (2012).

Rather than modeling instantiation of bumps in Eq. (2.1) using a spatiotemporal input as in Compte et al. (2000), Kilpatrick et al. (2013), and Almeida et al. (2015), we consider bump initiation implemented with initial conditions. Initiating bumps with external input does not significantly alter our results. Stochasticity is modeled by weak and multiplicative noise $\sqrt{\epsilon \cdot |u(x,t)|}\mathrm{d}Z(x,t)$, driven by the increment of a spatially-dependent Wiener process such that $\langle \mathrm{d}Z(x,t) \rangle = 0$ and $\langle \mathrm{d}Z(x,t)\mathrm{d}Z(y,s) \rangle = C(x - y)\delta(t - s)\mathrm{d}t\mathrm{d}s$. The spatial correlations $C(x - y)$ are a symmetric function that depends on the distance between two locations in the network. Typical formulations of Langevin equations take the multiplicative noise to be of Stratonovich form (Gardiner 2009), and a related neural field equation can be derived by applying a Kramers-Moyal expansion to a neural master equation (Bressloff 2009).

The impact of multiplicative noise on bump dynamics is analyzed by adapting methods developed for bumps in neural fields with additive noise (Kilpatrick and Ermentrout 2013). Note, we could modify Eq. (2.1) to account for the systematic shift induced by multiplicative noise in the Stratonovich sense. However, this contribution will be $\mathcal{O}(\epsilon)$ in comparison to the $\mathcal{O}(\sqrt{\epsilon})$ amplitude of the noise itself. Thus, we simply truncate the equation to ignore these additional terms, which would only slightly shift the resulting form of the stationary solution we will linearize about, as discussed in Bressloff and Webber (2012). To be explicit, we note that we can compute $\sqrt{\epsilon}\langle u(x,t)$
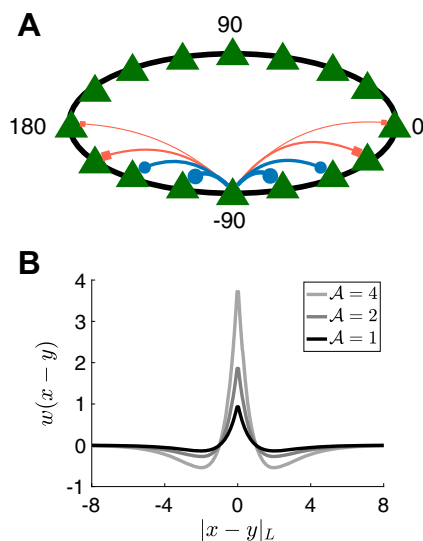
$dZ(x,t)\rangle = \epsilon C(0)\langle \text{sign}[u(x,t)]\rangle dt/2$ (Novikov 1965), smaller than the $\sqrt{\epsilon}$-amplitude noise term we consider. Note, we have run simulations of both the original Eq. (2.1) and the associated mean-corrected equations, and the results are not noticeably different.

2.2 Single bump solutions

Solutions to the noise-free ($Z \equiv 0$) version of Eq. (2.1) can be found explicitly for specific weight functions (Bressloff 2012). In particular, single bump (stationary pulse) solutions exist when $w(x)$ satisfies requirements making it laterally inhibitory (Amari 1977), as Eq. (2.2) is. We construct this solution and demonstrate the use of the interface method for characterizing non-equilibrium dynamics of perturbed bump solutions. This will guide our understanding for applying the interface method to multiple bumps.

In the absence of stochasticity ($Z \equiv 0$), stationary solutions to Eq. (2.1) satisfy $u(x,t) \equiv U(x)$, leading to the implicit equation

$$U(x) = \int_{-L}^{L} w(x-y) H(U(y) - \theta) \mathrm{d}y. \tag{2.4}$$

Unimodal stationary bumps possess a simply-connected active region $\bar{A} = [\bar{x}_1, \bar{x}_2] = \{x | U(x) \geq \theta\}$ (assuming $-L \leq \bar{x}_1 < \bar{x}_2 < L$), which allows us to rewrite Eq. (2.4) as

$$U(x) = \int_{\bar{x}_1}^{\bar{x}_2} w(x-y) \mathrm{d}y. \tag{2.5}$$

For analytical convenience, the translation symmetry of the network Eq. (2.1) can be utilized to shift solutions $U_0(x) = U(x - (\bar{x}_1 + \bar{x}_2)/2)$ so they are centered at zero (Ermentrout 1998; Bressloff 2012):

$$U_0(x) = \int_{-h}^{h} w(x-y) \mathrm{d}y = W(x+h) - W(x-h), \tag{2.6}$$

where $h = (\bar{x}_2 - \bar{x}_1)/2$ and we have defined the antiderivative

$$W(x) = \int_0^x w(y) \mathrm{d}y. \tag{2.7}$$

Linear stability of stationary bumps can be determined by examining the evolution of perturbations $u(x,t) = U_0(x) + \varepsilon \psi(x,t) + \mathcal{O}(\varepsilon^2)$ to the bump profile. Linearizing Eq. (2.1) leads to the following evolution equation for the perturbation

$$\psi_t(x,t) = -\psi(x,t) + w * \left[H'(U_0 - \theta)\psi \right]. \tag{2.8}$$

Separability of solutions $\psi(x,t) = \mathrm{e}^{\lambda t} \bar{\psi}(x)$ can be shown (Carroll et al. 2014), yielding the integral equation for linear stability

$$(\lambda + 1)\bar{\psi}(x) = w(x) * \left[H'(U_0(x) - \theta)\bar{\psi}(x) \right]. \tag{2.9}$$

Note, $(\lambda, \bar{\psi}(x)) = (0, U_0'(x))$ is a solution, since by plugging in we find

$$U_0'(x) = w(x) * \left[H'(u(x,t) - \theta)U_0'(x) \right], \tag{2.10}$$

and Eq. (2.10) arises by differentiating the stationary bump Eq. (2.4). This indicates the bump is marginally stable to perturbations that shift its position (Amari 1977; Kilpatrick and Ermentrout 2013), the main source of error when considering noise and interactions with other bumps.

We can explicitly construct the eigensolutions to Eq. (2.9) by applying the identity $H'(U_0(x) - \theta) = \bar{\alpha} \left[\delta(x-h) + \delta(x+h) \right]$ with $\bar{\alpha} = |U_0'(\pm h)| = w(0) - w(2h)$ (Amari 1977; Bressloff 2012). We find

$$(\lambda + 1)\bar{\psi}(x) = \bar{\alpha} \left[w(x-h)\bar{\psi}(h) + w(x+h)\bar{\psi}(-h) \right],$$

so solutions only depend on the values of $\bar{\psi}(x)$ at $x = \pm h$. Assuming $\bar{\psi}(h) = -\bar{\psi}(-h)$, we find the associated eigenvalue is $\lambda_o = 0$, demonstrating the bump is marginally stable to odd perturbations as mentioned above. For even perturbations $\psi(h) = \psi(-h)$, we find the associated eigenvalue

$$\lambda_e = \frac{2w(2h)}{w(0) - w(2h)}. \tag{2.11}$$

Thus, the stability of the bump will be determined by the sign of $w(2h)$. Typically, the wider bump has $w(2h) < 0$ (Fig. 3a), so it is linearly stable (Amari 1977; Kilpatrick 2016).

We perform the integral in Eq. (2.6) for the case of a weight function of form Eq. (2.2), noting the antiderivative Eq. (2.7) is thus given

$$W(x) = \mathcal{A} \int_0^x (1 - |y|)\mathrm{e}^{-|y|} \mathrm{d}y = \mathcal{A} x \mathrm{e}^{-|x|}. \tag{2.12}$$

Note, for analytical convenience, we approximate $L \to \infty$, obtaining (Coombes and Owen 2005):

$$U_0(x) = \mathcal{A} \left[(x+h)\mathrm{e}^{-|x+h|} - (x-h)\mathrm{e}^{-|x-h|} \right]. \tag{2.13}$$

Self-consistency requires the threshold conditions $U(\pm h) = \theta$ be satisfied, yielding an implicit equation for the bump half-width

$$G(h) := U_0(\pm h) = W(2h) = 2\mathcal{A}h\mathrm{e}^{-2h} = \theta. \tag{2.14}$$

We show $h(\theta)$ for different values of synaptic strength $\mathcal{A}$ in Fig. 3a, and plot stable bump solutions in Fig. 3b, showing they expand in width as the synaptic strength $\mathcal{A}$ is increased.

There are both wide stable and narrow unstable bumps of form Eq. (2.13). A critical value of θ defines the point
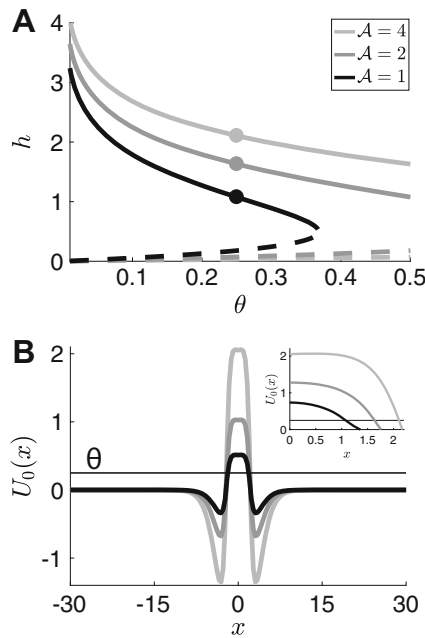
Fig. 3 a Width of bumps, stable (solid) and unstable (dashed), computed using the threshold condition, Eq. (2.14). Stable bump width increases with the synaptic strength $\mathcal{A}$. **b** Examples of bump profiles corresponding to the dots in (**a**), widening as $\mathcal{A}$ is increased for fixed $\theta = 0.25$. Inset shows zoom-in of right threshold crossings

where these branches of Eq. (2.14) annihilate in a saddle-node (SN) bifurcation (Kilpatrick 2016). Differentiating with respect to h, the SN bifurcation occurs where $G'(h) = (1 - 2h)2\mathcal{A}e^{-2h} = 0$ which can be solved for $h_c = 1/2$. Plugging $h_c = 1/2$ into Eq. (2.14), we find $\theta_c = \mathcal{A}e^{-1}$. Thus, we select $\theta < \theta_c$ given synaptic strength $\mathcal{A}$ to ensure solution existence. Furthermore, we can differentiate Eq. (2.14) with respect to the synaptic strength $\mathcal{A}$, yielding

$$\frac{dh}{d\mathcal{A}} = \frac{h}{\mathcal{A}(2h - 1)} > 0,$$

for $h > 1/2$, which occurs for stable bumps as long as $\theta < \theta_c$. Thus, the width of stable bumps will always increase as $\mathcal{A}$ is increased. Before exploring interactions of multiple bumps of the form of Eq. (2.13), we discuss the interface method we will use to obtain low-dimensional approximations for bump dynamics.

2.3 Interface equations for a single bump

Motivated by solutions of the stationary bump type, as in Eq. (2.5), Amari (1977) developed an interface theory for excitation patterns in the noiseless ($Z \equiv 0$) version of the neural field Eq. (2.1). This approach was recently reviewed in Amari (2014), and has been extended to capture the dynamics of other solutions by Coombes and Laing (2011), Coombes et al. (2012), and Avitabile et al. (2017). We extend these techniques further to account for stochastic

perturbations due to the noise term in Eq. (2.1). Interface equations are derived by noting that the output of the Heaviside nonlinearity, Eq. (2.3), is determined by the active region $A(t) = \{x | u(x, t) \geq \theta\}$ of the spatial domain $x \in [-L, L]$. For a single bump, we define the active region $A(t) = [x_1(t), x_2(t)]$, where the interfaces occur at the boundary points $x_1(t)$ and $x_2(t)$,

$$u(x_j(t), t) = \theta, \qquad j = 1, 2. \tag{2.15}$$

We rewrite Eq. (2.1), using our assumed form of the active region $A(t)$ as

$$du(x, t) = \left[-u(x, t) + \int_{x_1(t)}^{x_2(t)} w(x - y) dy \right] dt + \sqrt{\epsilon \cdot |u(x, t)|} dZ(x, t). \tag{2.16}$$

We now derive a stochastic evolution equation for the interfaces, $dx_j = d_j(x_1, x_2, t)dt + g_j(x_1, x_2, t)dz_j$, where d_j is a drift and g_j corresponds to the diffusion term. Differentiating Eq. (2.15) with respect to time, we obtain the following consistency equation for the location of the interfaces $x_j(t)$ and the evolution of the activity variable

$$\alpha_j(t)dx_j(t) + \beta_j(t)(dx_j(t))^2 + du(x_j(t), t) = 0, \tag{2.17}$$

for $j = 1, 2$, where we have defined the spatial gradient at the interface points

$$\alpha_j(t) = \frac{\partial u(x_j(t), t)}{\partial x}, \qquad j = 1, 2,$$

and the second derivative $\beta_j(t) = \frac{1}{2}u_{xx}(x_j(t), t)$ for $j = 1, 2$. The middle term in Eq. (2.17) arises from an application of Itô's lemma (Gardiner 2009). For simplicity, we approximate the spatial gradients using that of the stationary solution for now, $\alpha_1(t) \approx \bar{\alpha} = U_0'(-h)$ and $\alpha_2(t) \approx -\bar{\alpha} = U_0'(h) = -U_0'(-h)$, computed directly from Eq. (2.13). In Coombes et al. (2012) and Gökçe et al. (2017), the dynamic evolution of the gradients $\alpha_j(t)$ is tracked in the case of a deterministic system ($Z \equiv 0$ in Eq. (2.1)). As, discussed, we drop $o(\epsilon)$ terms resulting from multiplicative noise, for simplicity. Thus, the middle term in Eq. (2.17) will vanish, since the noise will have amplitude $\mathcal{O}(\sqrt{\epsilon})$, as we show, and the other terms in $(dx_j)^2$ are vanishingly small. Using the evolution equation for the neural activity, Eq. (2.16), and the interface condition Eq. (2.15), we can describe the evolution of the interfaces by rearranging Eq. (2.17) to find

$$dx_j(t) = \frac{(-1)^j}{\bar{\alpha}} \left(\left[-\theta + \int_{x_1(t)}^{x_2(t)} w(x_j(t) - y) dy \right] dt + \sqrt{\epsilon \cdot \theta} \, dZ(x_j(t), t) \right) \tag{2.18}$$

for $j = 1, 2$. Since the integral in Eq. (2.18) can be evaluated, we employ our definition, Eq. (2.7), of the antiderivative $W(x)$ and write

$$\int_{x_1(t)}^{x_2(t)} w(x_j(t) - y)\mathrm{d}y = W(x_2(t) - x_1(t)),$$

yielding an even simpler form for the interface equations

$$\mathrm{d}x_j(t) = \frac{(-1)^j}{\bar{\alpha}} \Big([-\theta + W(x_2(t) - x_1(t))]\,\mathrm{d}t + \sqrt{\epsilon \cdot \theta}\,\mathrm{d}Z(x_j(t), t) \Big) \quad (2.19)$$

for $j = 1, 2$. We now remark on a number of observations to be made concerning Eq. (2.19). First, in the absence of noise ($Z \equiv 0$), there is a line of fixed points to the resulting equation

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} = \frac{(-1)^j}{\bar{\alpha}} (-\theta + W(x_2(t) - x_1(t))) \quad (2.20)$$

for $j = 1, 2$, in the space (x_1, x_2) satisfying $W(x_2 - x_1) = \theta$, which is precisely Eq. (2.14) in the case that $x_2 - x_1 = 2h$. Also, note that when the interface locations are symmetric about $x = 0$, then $x_2(t) = -x_1(t) = a(t)$ can be described by a single equation by plugging into Eq. (2.20) to yield

$$\frac{\mathrm{d}a}{\mathrm{d}t} = \frac{1}{\bar{\alpha}} (-\theta + W(2a(t))). \quad (2.21)$$

Lastly, note that we can examine the effects of noise by analyzing the stochastic differential equations

$$\mathrm{d}x_1(t) = \frac{\theta - W(x_2(t) - x_1(t))}{\bar{\alpha}}\mathrm{d}t - \frac{\sqrt{\epsilon \cdot \theta}}{\bar{\alpha}}\mathrm{d}Z(x_1(t), t), \quad (2.22a)$$

$$\mathrm{d}x_2(t) = \frac{-\theta + W(x_2(t) - x_1(t))}{\bar{\alpha}}\mathrm{d}t + \frac{\sqrt{\epsilon \cdot \theta}}{\bar{\alpha}}\mathrm{d}Z(x_2(t), t). \quad (2.22b)$$

As in previous work (Kilpatrick and Ermentrout 2013; Carroll et al. 2014), we can track the stochastic motion of the bump's location by looking at the center of mass $\Delta(t) = (x_1(t) + x_2(t))/2$, evolving as

$$\mathrm{d}\Delta(t) = \frac{\sqrt{\epsilon \cdot \theta}}{2\bar{\alpha}} [\mathrm{d}Z(x_2(t), t) - \mathrm{d}Z(x_1(t), t)]. \quad (2.23)$$

Assuming fluctuations in the active region width are small ($x_2 - x_1 \approx 2h$), we approximate $x_1 \approx \Delta - h$ and $x_2 \approx \Delta + h$, so we can directly compute the mean $\langle \Delta(t) \rangle = \Delta(0)$ and variance $\langle (\Delta(t) - \langle \Delta(t) \rangle)^2 \rangle = Dt$ with diffusion coefficient

$$D = \frac{\epsilon\theta}{4\bar{\alpha}^2} \Big[\langle Z(\Delta - h, t)^2 \rangle + \langle Z(\Delta + h, t)^2 \rangle - 2\langle Z(\Delta - h, t)Z(\Delta + h, t) \rangle \Big],$$

$$= \frac{\epsilon\theta}{2\bar{\alpha}^2} [C(0) - C(2h)], \quad (2.24)$$

where $C(x)$ is the spatial correlation function of the neural field, Eq. (2.1). We derived a related equation in the case of additive noise by directly assuming stochastic motion of the bump's position in Kilpatrick and Ermentrout (2013). Note, Eq. (2.24) provides a formula for the diffusion coefficient of the bump for arbitrary spatial noise correlations $C(x)$, in contrast to the work of Burak and Fiete (2012), which assumes independent fluctuations are generated at each point in the network via a Poisson process.

We demonstrate the accuracy of these approximations by tracking the transient evolution of bumps in numerical simulations and comparing with predictions of Eq. (2.19). First, for a bump unforced by noise that is initiated with a narrower width than its equilibrium width, given by the wide solution to Eq. (2.14), the interfaces relax outward. In fact, these dynamics can be tracked by the interface Eq. (2.21), corresponding to the half-width of the evolving bump. We need only calculate the antiderivative Eq. (2.7) for our specific weight function, Eq. (2.2), given by Eq. (2.12). Thus, we can compute from Eq. (2.13) that

$$\bar{\alpha} = |U_0'(\pm h)| = \mathcal{A} \left[1 - (1 - 2h(\mathcal{A}))\mathrm{e}^{-2h(\mathcal{A})} \right], \quad (2.25)$$

where the half-width depends on the spatial scale $h(\mathcal{A})$. In Fig. 4a, b, we compare the level sets $u(x_j(t), t) = \theta$ of a numerical simulation of Eq. (2.1) initiated with a narrower initial condition to the evolution of the interface Eq. (2.21), showing the expansion is tracked well. We also use Eq. (2.25) to approximate the evolution of noise-driven bumps, described by the stochastic interface Eq. (2.22a, b). We use a cosine spatial correlation function

$$C(x) = \cos(\omega_c x), \quad \omega_c = c\pi/L. \quad (2.26)$$

We demonstrate agreement between the prediction of our interface Eq. (2.22a, b) and the level sets of Eq. (2.1) of a bump perturbed by noise in Fig. 4c, d. Note, for consistency the noise increments $\mathrm{d}Z(x, t)$ generated for the neural field model, Eq. (2.1), are used to generate the noise perturbations $\mathrm{d}Z(x_j, t)$ for the interface Eq. (2.22a, b). A similar approach is used in subsequent single realizations shown in Figs. 7 and 10. Furthermore, we can approximate the diffusion coefficient D corresponding to the rate at which the variance of the bump's centroid grows $\langle (\Delta(t) - \langle \Delta(t) \rangle)^2 \rangle = Dt$, as given by Eq. (2.24), so

$$D = \frac{\epsilon\theta}{2\mathcal{A}^2} \frac{1 - \cos(2\omega_c h)}{\left(1 + (2h(\mathcal{A}) - 1)\mathrm{e}^{-2h(\mathcal{A})}\right)^2}. \quad (2.27)$$

similar to results derived in Bressloff and Webber (2012) and Kilpatrick and Ermentrout (2013). Fixing the noise amplitude ϵ, we find the diffusion coefficient decreases monotonically as the synaptic strength $\mathcal{A}$ is increased (Fig. 4e, f). Our analytical approximation, Eq. (2.27), agrees well with simulations of the full model, Eq. (2.1). Thus,
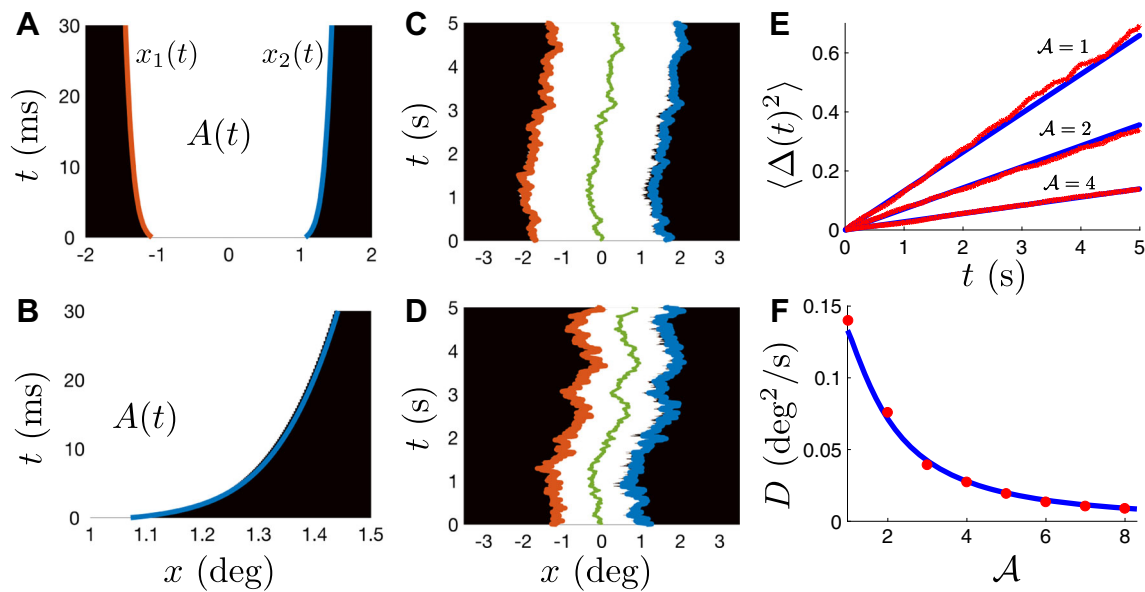
Fig. 4 Low-dimensional interface equations, Eq. (2.19), approximate the dynamics of Eq. (2.1) well for single-bump solutions. **a** Bump of the form $u(x, 0) = 0.25 \cdot U_0(x)$, Eq. (2.13), expands outward to the equilibrium shape $u(x, t) \to U_0(x)$ for $\mathcal{A} = 2$. Interface dynamics $u(x_{1,2}(t), t) = \theta$ are well approximated by Eq. (2.21), computed for the left (red curve) and right (blue curve) interfaces. **b** Zoom-in of (**a**). **c** Bump initially of form $u(x, 0) = U_0(x)$, for $\mathcal{A} = 2$, evolves stochastically in response to noise with correlation structure Eq. (2.26) with $\epsilon = 0.03$. Interface dynamics approximated by Eq. (2.22a, b), plotted for the left (red trajectory) and right (blue trajectory). Centroid (green

trajectory) can be approximated by Eq. (2.23). **d** For $\mathcal{A} = 1$, the bump is narrower and diffuses more in response to the noise. **e** Variance of the bump as a function of time. Theory (blue line) given by Eq. (2.24) is compared with the results of 10^4 numerical simulations (red line) of Eq. (2.1). **f** Diffusion coefficient computed as a function of $\mathcal{A}$. Theory (solid line) and sims (dots) compare favorably, showing a systematic decrease in diffusion as the strength of coupling $\mathcal{A}$ is increased. Threshold $\theta = 0.25$ throughout, and $\omega_c = 25\pi/180$. Numerical simulations are performed using Euler-Maruyama stochastic integration scheme with $dx = 0.005$ and $dt = 0.1$

stronger synaptic inputs (larger $\mathcal{A}$) increase the size of the bumps (larger h), and these wider bumps are more stable to noise perturbations (smaller D).

3 Dynamics of two interacting bumps

Our main interest lies in understanding how multiple bumps interact, as these interactions will contribute to our model's limitations in multi-item WM. Prior to examining WM for an arbitrary number of items, we focus on the case of two interacting bumps, to demonstrate how bump interactions lead to different errors during the delay period of a WM trial. The effective equations for two bumps can then be extended to higher dimensions.

3.1 Interface equations

We begin by extending our interface Eq. (2.19) for one bump to the case of two bumps. Again, since the nonlinearity in Eq. (2.1) is a step function $H(u - \theta)$, its output is determined by the active region $A(t) = \{x | u(x, t) \geq \theta\}$. In the case of a single bump, we defined a simply connected region as in Eq. (2.16). Two bumps would typically be comprised of two disjoint active regions.

However, if the bumps began close enough together, their active regions would overlap and form a single connected domain. In this special case, the dynamics of the system would subsequently be described by the single bump interface Eq. (2.19). Thus, for our analysis here, we assume the active region is comprised of two disjoint subdomains, $A(t) = [x_1(t), x_2(t)] \cup [x_3(t), x_4(t)]$, so Eq. (2.1) becomes

$$
\begin{aligned}
\mathrm{d}u(x, t) = & \left[-u(x, t) + \int_{x_1(t)}^{x_2(t)} w(x - y)\mathrm{d}y \right. \\
& \left. + \int_{x_3(t)}^{x_4(t)} w(x - y)\mathrm{d}y \right] \mathrm{d}t \\
& + \sqrt{\epsilon \cdot |u(x, t)|}\mathrm{d}Z(x, t).
\end{aligned} \tag{3.1}
$$

As mentioned, we assume two bumps have been instantiated far enough apart so that their active regions do not overlap. Assuming continuity of $u(x, t)$, the boundary points of $A(t)$ correspond to the interfaces of the bumps, and satisfy the dynamic threshold equations

$$
u(x_j(t), t) = \theta, \quad j = 1, 2, 3, 4. \tag{3.2}
$$

Differentiating Eq. (3.2) with respect to t, we find the total derivative is again given by Eq. (2.17). Specifying the

integral terms in Eq. (3.1) using Eq. (2.7), we can rewrite integrals $\mathcal{I}$ as

$$
\begin{aligned}
\mathcal{I} :&= \int_{x_1}^{x_2} w(x_j - y)\mathrm{d}y + \int_{x_3}^{x_4} w(x_j - y)\mathrm{d}y \\
&= \int_{x_j - x_2}^{x_j - x_1} w(z)\mathrm{d}z + \int_{x_j - x_4}^{x_j - x_3} w(x_j - y)\mathrm{d}y \\
&= \sum_{k=1}^{4} (-1)^{k-1} W(x_j(t) - x_k(t)). \quad (3.3)
\end{aligned}
$$

We study two cases of the two bump interface equations, which admit different approximations. For a fully deterministic Eq. (2.1), we can derive an integral equation for the time-evolution of the spatial gradients $\alpha_j(t)$ at the interfaces. An alternative approach, which is more straightforward, is to simply approximate the gradients $\alpha_j(t) \equiv \bar{\alpha}_j$ using static quantities derived from stationary solutions of Eq. (2.1). This is easier to employ, especially in the case of stochastic forcing.

Dynamic gradients In the case of a fully deterministic system ($Z \equiv 0$ in Eq. (2.1)), we can follow Coombes et al. (2012) to obtain an analytic formula for $\alpha_j(t)$ by defining $z(x,t) := \frac{\partial u(x,t)}{\partial x}$ and differentiating Eq. (3.1) with respect to x to find

$$
\frac{\partial z(x,t)}{\partial t} = -z(x,t) + \sum_{k=1}^{4} (-1)^{k-1} w(x - x_k(t)),
$$

which we can integrate to yield

$$
\begin{aligned}
z(x,t) = \mathrm{e}^{-t} \int_0^t \mathrm{e}^s \left[\sum_{k=1}^{4} (-1)^{k-1} w(x - x_k(s)) \right] \mathrm{d}s \\
+ z(x,0)\mathrm{e}^{-t}. \quad (3.4)
\end{aligned}
$$

Evaluating Eq. (3.4) at $x_j(t)$, we have:

$$
\begin{aligned}
\alpha_j(t) = \int_0^t \mathrm{e}^{-(t-s)} \left[\sum_{k=1}^{4} (-1)^{k-1} w(x_j(t) - x_k(s)) \right] \mathrm{d}s \\
+ u_0'(x_j(t))\mathrm{e}^{-t}, \quad j = 1,2,3,4, \quad (3.5)
\end{aligned}
$$

where $u(x,0) = u_0(x)$ is the initial condition. Thus, we have a closed system describing the evolution of the interfaces of the two stationary bumps, assuming the active region $A(t)$ remains as two disjoint subdomains

$$
\frac{\mathrm{d}x_j}{\mathrm{d}t} = -\frac{1}{\alpha_j(t)} \left[\sum_{k=1}^{4} (-1)^{k-1} W(x_j(t) - x_k(t)) - \theta \right] \quad (3.6)
$$

for $j = 1,2,3,4$, with $\alpha_j(t)$ defined as in Eq. (3.5). The second order term in Eq. (2.17) will vanish, since there is no noise in this case. As we have performed no truncations, the

pair of Eq. (3.5) and (3.6) exactly characterize the motion of the four bump interfaces (x_1, x_2, x_3, x_4). We compare the evolution of the interfaces given by Eqs. (3.5) and (3.6) to those calculated from the full model Eq. (2.1) in Fig. 5a, b. Bumps can either move away from each other (Fig. 5a) or towards each other (Fig. 5b), depending on the initial distance $2\Delta_x := (x_4(0) + x_3(0) - x_2(0) - x_1(0))/2$ of their centroids from one another. Note, merging occurs extremely rapidly, and repulsion happens otherwise. We can characterize the relaxation rate of merging by linearizing Eq. (2.21) about the equilibrium bump half-width h. As shown in previous stability analyses of stationary bumps in Eq. (2.1), the eigenvalue associated with the decay of width perturbations is $\lambda_e = 2w(2h)/(w(0) - w(2h)) < 0$ (Amari 1977; Kilpatrick and Ermentrout 2013). In our performance calculations, we account for this, and ignore the detailed dynamics of merging.

Note that for initial conditions that are symmetric about $x = 0$ ($x_4(0) = -x_1(0) = b(0)$ and $x_3(0) = -x_2(0) = a(0)$), the interfaces evolve symmetrically: $x_4(t) = -x_1(t) = b(t) \geq 0$ and $x_3(t) = -x_2(t) = a(t) \geq 0$. In a similar way, the dynamic gradients exhibit odd symmetry: $\alpha_1(t) = -\alpha_4(t) = \beta(t) \geq 0$ and $\alpha_3(t) = -\alpha_2(t) = \alpha(t) \geq 0$. Eqs. (3.5) and (3.6) can be reduced to four equations:

$$
\begin{aligned}
\frac{\mathrm{d}a(t)}{\mathrm{d}t} = \frac{1}{\alpha(t)} [\theta - W(b(t) - a(t)) + W(2a(t)) \\
- W(a(t) + b(t))], \quad (3.7a)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathrm{d}b(t)}{\mathrm{d}t} = \frac{1}{\beta(t)} [W(b(t) - a(t)) - \theta + W(2b(t)) \\
- W(a(t) + b(t))], \quad (3.7b)
\end{aligned}
$$

$$
\begin{aligned}
\alpha(t) = \mathrm{e}^{-t} \int_0^t \mathrm{e}^s [w(a(t) + b(s)) - w(a(t) + a(s)) \\
+ w(a(t) - a(s)) - w(a(t) - b(s))] \mathrm{d}s \\
+ u_0'(a(t))\mathrm{e}^{-t}, \quad (3.7c)
\end{aligned}
$$

$$
\begin{aligned}
\beta(t) = \mathrm{e}^{-t} \int_0^t \mathrm{e}^s [w(b(t) + a(s)) - w(b(t) + b(s)) \\
+ w(b(t) - b(s)) - w(b(t) - a(s))] \mathrm{d}s \\
- u_0'(b(t))\mathrm{e}^{-t}. \quad (3.7d)
\end{aligned}
$$

The system, Eq. (3.7a, b, c and d), is used to calculate the interfaces of the two scenarios shown in Fig. 5a, b.

There is a critical distance $2\Delta_x^c$ between two bumps' initial centroids, which divides solutions that repel ($\Delta_x > \Delta_x^c$) from those that merge ($\Delta_x < \Delta_x^c$). We illustrate this by tracking the centroids of two symmetrically placed bumps for various starting distances Δ_x (Fig. 5c). Similar features of associated spiking network models have been identified in Wei et al. (2012) and Almeida et al. (2015).
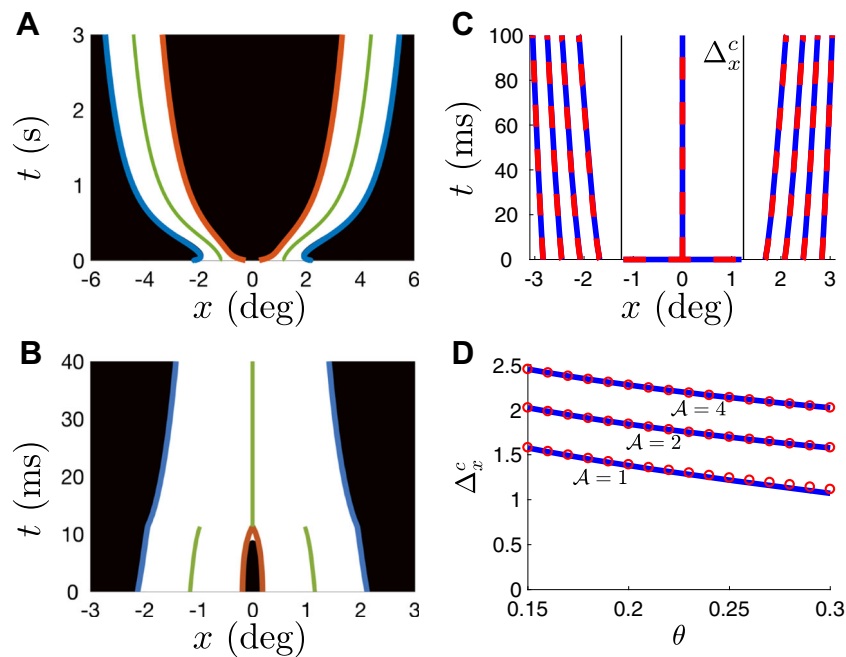
Fig. 5 Two bumps interacting in the deterministic ($Z \equiv 0$) neural field Eq. (2.1). **a** Two bumps repel each other when initiated at $\pm x_0 = \pm 1.25$. The location of the bump interfaces are well-tracked by the curves (solid lines) generated by the low-dimensional Eq. (3.7a, b, c and d). Parameters are $\theta = 0.25$ and $\mathcal{A} = 1$. **b** Two bumps merge when initiated at $\pm x_0 = \pm 1.23$, a short distance for the point of initiation in (**a**). **c** Multiple trajectories of the centroid of the bump for initial conditions $x_0 = \{1.2, 1.6, 2.0, 2.4, 2.8\}$ as predicted by Eq. (3.7a, b, c and d) (solid line) and direct simulation (dashed lines). Bumps repel each other more strongly when they begin close to the critical boundary Δ_x^c (thin line). Within the critical boundary, bumps merge. **d** The critical boundary $\Delta_x^c(\theta)$ is determined by Eq. (3.8) (solid curves), and compared with direct simulations (circles). See Fig. 4 for details on numerical simulations

We can determine an analytical expression that accurately characterizes the critical distance Δ_x^c. Utilizing Eq. (3.7a), we note that if $a'(t) < 0$, bumps will initially move towards one another. Motivated by the findings of our numerical simulations in Fig. 5c, we expect bumps that are initially attracted will continue to move towards one another until they merge. In this case, the critical curve (a^c, b^c), determined by the condition

$$\theta = W(b^c - a^c) - W(2a^c) + W(a^c + b^c) \quad (3.8)$$

divides initial conditions $(a(0), b(0))$ that merge from those that repel each other. Assuming the bumps initially have width $b(0) - a(0) = 2h$, as prescribed by Eq. (2.13), then $\theta = W(b(0) - a(0))$ and Eq. (3.8) simplifies to $W(2a^c) = W(a^c + b^c)$. Thus, defining the right bump's initial centroid $\Delta_x^c = (a^c + b^c)/2$, then $a^c = \Delta_x^c - h$ and $b^c = \Delta_x^c + h$, so $W(2\Delta_x^c - 2h) = W(2\Delta_x^c)$. Applying Eq. (2.7) and simplifying, we find $\Delta_x^c = h/(1 - e^{-2h})$, which is increasing in h for $h > 1/2$. Thus, as expected, wider bumps will always have a wider critical merging distance Δ_x^c. We compare our analytical prediction, Eq. (3.8), to the results of numerical simulations, and find they agree (Fig. 5d). Thus, our interface equations not only track the motion of bumps, but can also predict when they merge with one another.

For the special case in which symmetric bumps initially overlap, merging occurs immediately, and the subsequent interface dynamics are well approximated by Eq. (2.21). In this case, the associated dynamics is given by the decay of the width of the resulting merged bump (inset in Fig. 6a). We can linearly approximate the decay dynamics of the interfaces, where $u(\pm a(t), t) = \theta$, as $a(t) \approx h + (a(0) - h)e^{\lambda_e t}$, with λ_e the eigenvalue in Eq. (2.11). This approximation agrees well with simulations of the full model Eq. (2.1) for a wide range of synaptic strengths $\mathcal{A}$ (Fig. 6a). Interestingly, the rate ($|\lambda_e|$) at which the bump width decays decreases as $\mathcal{A}$ is increased (Fig. 6b). Thus, it takes longer for bumps in networks with strong synapses to reach their equilibrium half-width h.

Static gradient approximation An alternative to computing the integral equations in Eq. (3.5) is to assume a static approximation to the gradients $\alpha_j(t) \equiv \bar{\alpha}_j$. Since our solutions evolve with a profile approximated by sums of the stationary bump solution $U_0(x)$, Eq. (2.5), we consider gradients approximated by $U_0'(x)$. For the weight function Eq. (2.2), these can be computed directly from Eq. (2.25) as

$$\bar{\alpha}_{1,3} = -\bar{\alpha}_{2,4} = \bar{\alpha} = U_0'(-h).$$
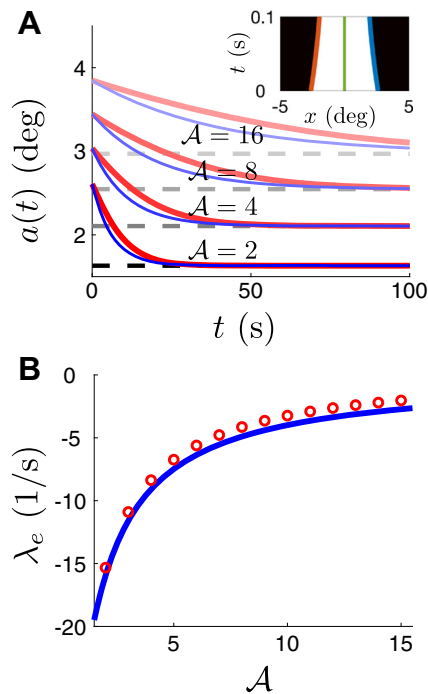
Fig. 6 Decay rate of merged bumps to equilibrium bump shape. **a** Two bumps are initiated at $\pm x_0 = \pm 1.23$, leading to an initial overlap in their active regions (Inset shows simulation of Eq. (2.1) for $\mathcal{A} = 2$). The resulting merged bump has initial half-width $a(0)$ which decays towards the equilibrium half-width $h(\mathcal{A})$ (dashed lines). Interface dynamics $u(\pm a(t), t) = \theta$ are computed directly (red lines) from Eq. (2.1) and by approximation $a(t) \approx a(0) + (a(0) - h)e^{\lambda_e t}$ (blue lines) where λ_e is given by Eq. (2.11). **b**. The rate of decay ($|\lambda_e|$) decreases as $\mathcal{A}$ is increased, meaning networks with stronger synapses have bumps whose widths decay more slowly. Analytical approximation (blue line) using Eq. (2.11) compares well with decay rate of best exponential fit to $a(t)$ in (**a**) (red circles). Threshold $\theta = 0.25$. See Fig. 4 for details of numerical simulations

We can then write the resulting interface equations in a simple form that still captures the interactions between the two bumps, as well as the effects of noise perturbations:

$$dx_1 = \frac{1}{\alpha}\left([\theta - W(x_2 - x_1) + W(x_3 - x_1)\right.$$
$$\left. - W(x_4 - x_1)]\,dt - \sqrt{\epsilon \cdot \theta}\,dZ(x_1, t)\right), \quad (3.9a)$$

$$dx_2 = -\frac{1}{\alpha}\left([\theta - W(x_2 - x_1) + W(x_3 - x_2)\right.$$
$$\left. - W(x_4 - x_2)]\,dt - \sqrt{\epsilon \cdot \theta}\,dZ(x_2, t)\right), \quad (3.9b)$$

$$dx_3 = \frac{1}{\alpha}\left([\theta - W(x_4 - x_3) + W(x_3 - x_2)\right.$$
$$\left. - W(x_3 - x_1)]\,dt - \sqrt{\epsilon \cdot \theta}\,dZ(x_3, t)\right), \quad (3.9c)$$

$$dx_4 = -\frac{1}{\alpha}\left([\theta - W(x_4 - x_3) + W(x_4 - x_2)\right.$$
$$\left. - W(x_4 - x_1)]\,dt - \sqrt{\epsilon \cdot \theta}\,dZ(x_4, t)\right). \quad (3.9d)$$

As in the case of single-bumps, the second order term in Eq. (2.17) is smaller than $\mathcal{O}(\sqrt{\epsilon})$. The system

Eq. (3.9a, b, c and d) accurately approximatess the stochastic dynamics of the two bumps' interfaces (Fig. 7a, b). We can also reduce Eq. (3.9a, b, c and d) to track the centroid of each bump $\Delta_1 = (x_1 + x_2)/2$ and $\Delta_2 = (x_3 + x_4)/2$. To do so, we assume the width of each bump remains approximately constant, so $x_1 \approx \Delta_1 - h$, $x_2 \approx \Delta_1 + h$, $x_3 \approx \Delta_2 - h$, and $x_4 \approx \Delta_2 + h$. In this case, we find two equations for the stochastic dynamics of the centroids:

$$d\Delta_1 = \frac{1}{\alpha}\left(J(\Delta_2 - \Delta_1)dt + \frac{\sqrt{\epsilon \cdot \theta}}{2}d\mathcal{Z}(\Delta_1, t)\right), \quad (3.10a)$$

$$d\Delta_2 = \frac{1}{\alpha}\left(J(\Delta_1 - \Delta_2)dt + \frac{\sqrt{\epsilon \cdot \theta}}{2}d\mathcal{Z}(\Delta_2, t)\right), \quad (3.10b)$$

where odd symmetry of the coupling function $J(\Delta)$ follows from the odd symmetry of $W(x) = -W(-x)$:

$$J(\Delta) = \frac{1}{2}\left(2W(\Delta) - W(\Delta - 2h) - W(\Delta + 2h)\right),$$
$$d\mathcal{Z}(\Delta_j, t) = dZ(\Delta_j + h, t) - dZ(\Delta_j - h, t),$$

for $j = 1, 2$. We use the approximation, Eq. (3.10a, b), to determine the evolution of the centroids in realizations with different initial conditions (Fig. 7c, d). Even though the bump for $\mathcal{A} = 2$ has a lower diffusion coefficient (as shown in Fig. 4e, f), for $\mathcal{A} = 1$, bumps repel each other less. Thus, bumps in networks with weaker synaptic weights can stray less from their initial position, when they are initiated close together. However, when bumps begin far apart, strong connectivity may be more advantageous, since bumps are less perturbed by noise. We examine this tradeoff by determining performance of the network in a two-item WM task, using our approximations and full numerical simulations.

3.2 Performance

We study estimation errors of the network encoding locations of two targets, ϕ_1 and ϕ_2. In general, we categorize errors as arising from (a) merging; (b) repulsion; and (c) diffusion of the bumps encoding these targets. The fluctuation-driven random walk of bumps has been characterized in single-item WM models (Compte et al. 2000; Kilpatrick and Ermentrout 2013), and validated in behavioral and electrophysiological experiments (Wimmer et al. 2014; Constantinidis and Klingberg 2016). Bump merging was recently characterized in spiking network models of multi-item WM (Wei et al. 2012; Almeida et al. 2015), motivated by corresponding human psychophysics data (Bays et al. 2009). The merging of item memories has been considered in heuristic models of multi-item WM, and analyses sometimes assume the memory for one of the associated items is then completely lost, so subjects guess to report its location (Zhang and Luck 2008). Similar
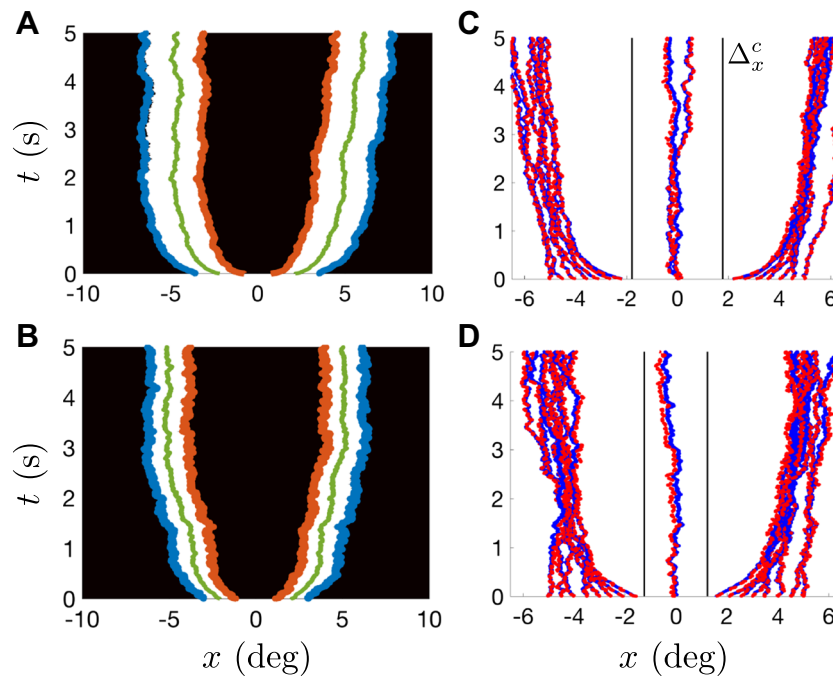
Fig. 7 Stochastic simulations of Eq. (2.1) with noise amplitude $\epsilon = 0.03$ and noise correlations, Eq. (2.26). **a** Two noise-driven bumps repel each other and diffuse when starting at $\pm x_0 = \pm 2$. Our estimate using the static gradient approximation (solid lines), Eq. (3.9a, b, c, and d), tracks the interfaces. Other parameters are $\theta = 0.25$ and $\mathcal{A} = 2$. **b** When $\mathcal{A} = 1$, the bumps repel each other less than in (**a**). **c** Trajectories of the centroid of the noise-driven bumps with centroids initiated at different locations $\pm x_0$ from 1 to 5 deg, spaced 0.5 deg apart. Approximations (solid lines) using the low-dimensional system, Eq. (3.10a, b), agree with direct numerical simulations (dashed lines). When initiated within the region $x_0 < \Delta_x^c$ (thin lines), bumps merge as in the deterministic Fig. 5. Here, $\mathcal{A} = 2$. **d** Same as (**c**), except $\mathcal{A} = 1$, showing bumps repel each other less and also merge in a narrower range. See Fig. 4 for details on numerical simulations

guesses may occur due to attentional lapses, where subjects do not store an item in the first place. We avoid such characterizations in our analysis, and study error solely ascribed to the dynamics of bump attractors in Eq. (2.1). Thus, we assume that when bumps merge, the remaining bump encodes the location of both items corresponding to the original two bumps. Finally, note that repulsion will lead to item memories that diverge from one another when bumps are instantiated close to one another.

Across multiple trials, the task on trial k is to encode both target angles $\phi_1^k, \phi_2^k \in [-180, 180)$ ($\phi_1^k > \phi_2^k$). However, only a single target is probed, for instance by asking the subject to recall the angle corresponding to a particular color (Wilken and Ma 2004; Zhang and Luck 2008; Bays et al. 2009). Due to the symmetry in the system, we compute the mean squared error (MSE) corresponding to the first target angle ϕ_1^k on each trial k

$$\text{MSE} = \langle (\Delta_{1\text{-out}} - \phi_1)^2 \rangle = \frac{1}{K} \sum_{k=1}^{K} (\Delta_{1\text{-out}}^k - \phi_1^k)^2, \quad (3.11)$$

where $\Delta_{j\text{-out}}^1$ is the centroid of the bump encoding target 1 at the end of the trial k (as in Fig. 8a, d). The MSE in Eq. (3.11) can be computed directly from numerical simulations of Eq. (2.1), and we can also approximate the error using our reduced set of centroid equations, Eq. (3.10a, b).

Our approximation of the MSE begins by determining whether or not the bumps merge. To do so, we examine the target distance to see if it is below the critical value, $(\phi_2 - \phi_1) < 2\Delta_x^c$ (Fig. 5d). As shown in Fig. 5b, merging occurs very rapidly, so we do not model the detailed dynamics of merging in our performance calculations. Leveraging Eq. (3.8), which describes the minimal distance at which symmetrically-placed bumps do not merge, we rotate coordinates of $\phi_{1,2}$ so they are symmetric about zero $\tilde{\phi}_{1,2} = (\phi_1 - \phi_2, \phi_2 - \phi_1)/2$. Thus, in Eq. (3.8), assuming the initial bumps are roughly of width $2h$ we have $a^c = (\phi_2 - \phi_1 - 2h)/2$ and $b^c = (\phi_2 - \phi_1 + 2h)/2$, so if $W((\phi_2 - \phi_1 - 2h)/2) < W(\phi_2 - \phi_1)$, the bumps merge, otherwise they repel each other. This boundary is approximated given the weight function, Eq. (2.2), by solving the corresponding equality, $\phi_2 - \phi_1 = \frac{2h}{1-e^{-2h}}$, as the critical distance below which bumps merge. For this subset of cases, the MSE in Eq. (3.11) can be approximated in a straightforward way by noting

$$\begin{aligned} \mathcal{M}_{\text{mg}} &= \langle (\Delta_{1\text{-out}} - \phi_1)^2 \rangle \\ &= \langle (\Delta_{1\text{-out}} - (\phi_1 + \phi_2)/2 - (\phi_1 - \phi_2)/2)^2 \rangle \\ &\approx \langle (\Delta_{1\text{-out}} - (\phi_1 + \phi_2)/2)^2 \rangle + \langle (\phi_1 - \phi_2)^2/4 \rangle \\ &= D \cdot T + \frac{(\phi_2 - \phi_1)^2}{4}, \end{aligned} \quad (3.12)$$
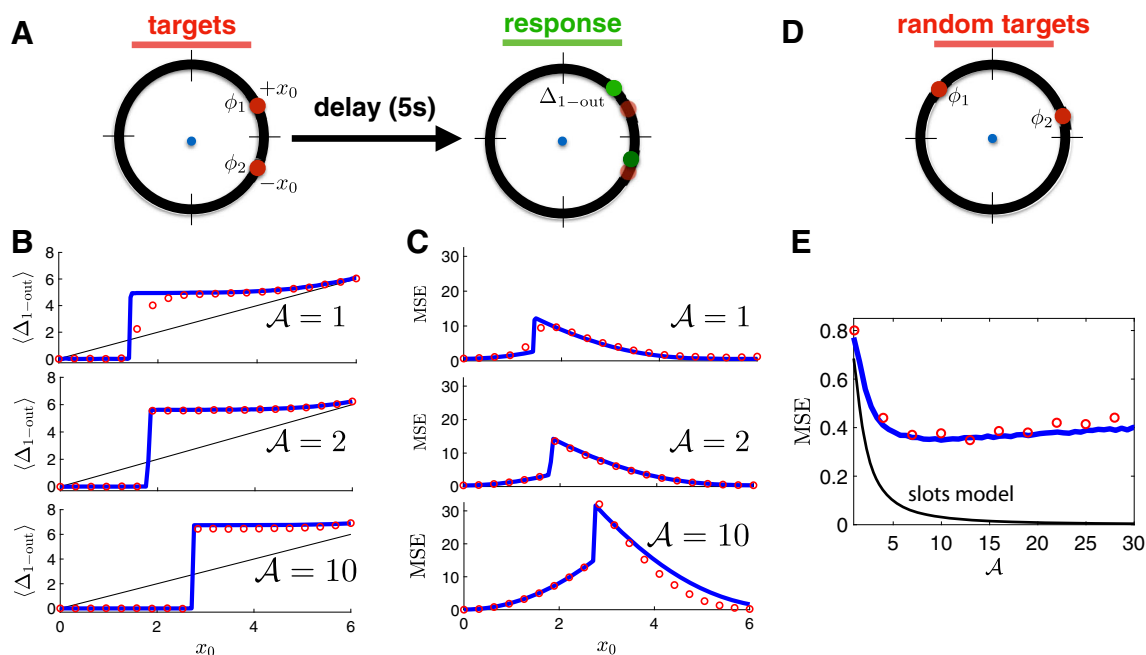
Fig. 8 Recall error in a two-item WM task. **a** Item-distance dependence of errors. Target items $\phi_{1,2}$ are located at $\pm x_0$. Mean squared error (MSE) is computed by comparing output angle $\Delta_{1-\text{out}}$ to input ϕ_1 via Eq. (3.11). **b** Mean location of $\Delta_{1-\text{out}}$ computed as a function of x_0 for low-dimensional Eq. (3.10a, b), (solid lines), and full simulations of the neural field (circles). Merging leads to values close to zero. A transition occurs at the boundary of merging, where repulsion begins. For large enough x_0, $\langle \Delta_{1-\text{out}} \rangle \approx +x_0$. Compare $\mathcal{A} = 1, 2, 10$. **c** MSE as a function of x_0 computed from Eq. (3.11) is maximized near the boundary point, where repulsion is strongest, decreasing as

the bumps are placed further apart. **d** MSE is also computed for uniform random targets $\phi_{1,2}$, averaging across all locations to compute the MSE for each synaptic strength $\mathcal{A}$. **e** MSE is nonmonotonic in $\mathcal{A}$, with fluctuation-driven errors dominating for low $\mathcal{A}$ and merging/repulsion errors dominating for large $\mathcal{A}$. MSE is much larger would be predicted for a slots model, which assumes MSE is unchanged between one and two item memory tasks, so we use $\text{MSE}_{\text{slots}} = D \cdot T$ where D is given by Eq. (2.27). Blue curves and red circles are generated from 10^6 Monte Carlo simulations

since the merged bump rapidly centers at the mean of the two target locations, $(\phi_1 + \phi_2)/2$. We can compute the diffusion coefficient D using Eq. (2.27), our theory for the stochastic dynamics of single bumps, and T is the total delay time, providing us an analytic approximation of the MSE for the case $x_0 < \Delta_x^c$.

If bumps do not merge, we approximate their dynamics using the nonlinear stochastic system, Eq. (3.10a, b). The constituent function $J(\Delta)$ can be computed, given the weight function, Eq. (2.2),

$$J(\Delta) = -2\mathcal{A}e^{-\Delta}\left[\Delta \sinh^2(h) - h \sinh(2h)\right],$$

assuming $\Delta > 0$. Note, the formula for $J(\Delta)$ is more complicated for the case in which $\Delta_2 < \Delta_1$, or their difference is across the periodic boundary at $x = \pm 180$. We consider these other cases in simulations, but do not discuss the formulas in detail here. In the case of noise correlations, Eq. (2.26), we can specify

$$d\mathcal{Z}(\Delta_j, t) = 2\sin(\omega_c h)\left[\sin(\omega_c \Delta_j) \cdot d\xi_1(t) - \cos(\omega_c \Delta_j) \cdot d\xi_2(t)\right],$$

where $d\xi_j(t)$, $j = 1, 2$, are increments of a standard Weiner process. Eq. (3.10a, b) is simulated numerically to estimate $\Delta_{1-\text{out}}^k$ in trial k, which then is plugged into our formula for MSE, Eq. (3.11).

Our approximations are compared to simulations of the full neural field model, Eq. (2.1), in Fig. 8. Targets angles ϕ_1 and ϕ_2 are initially represented by instantiating a bump of form $U_0(x)$, Eq. (2.13), centered at the two locations in the neural field. Distance-dependence is considered first (Fig. 8a), and then trials with random initial angles are considered (Fig. 8d). The system evolves for $T = 5$s (500 time units), and then the centroid of the bumps corresponding to ϕ_1 is read out and compared to the original angle by computing Eq. (3.11). If the bumps merge, both items are represented by the remaining bump. The MSE is similarly computed using the low-dimensional approximation, Eq. (3.10a, b), if the bumps begin sufficiently far apart, otherwise Eq. (3.12) is used to approximate the MSE. These approximations are compared to the full simulations in Fig. 8b, c, for the case in which the initial target angles are $\phi_1 = +x_0$ and $\phi_2 = -x_0$. Merging causes both bumps to have mean position $x = 0$, when the initial targets $\pm x_0$ are sufficiently close (Fig. 8b).

Beyond this boundary, the stored angles repel one another. There is an abrupt transition in the MSE corresponding to this boundary point (Fig. 8c). Importantly, the MSE is limited from below at each x_0 by the variance a single bump ($\langle \Delta^2 \rangle = D \cdot T$), not interacting with another bump. Thus, even though the peak MSE grows significantly for the case $\mathcal{A} = 10$, it is important to note that the MSE will be significantly smaller at large values of x_0, since individual bumps diffuse less for larger values of $\mathcal{A}$ (Fig. 4f).

The lower bound on the MSE produced by a single bump's trajectory is approached when the two bumps are either initiated at the same location ($x_0 = 0$), or when the bumps are initiated sufficiently far from one another. While there will always be vanishingly small repulsive effects that will tend to push bumps farther apart, we see that even for $x_0 \approx 6$, the MSE appears to approach a lower limit. This is because the long-range interactions between bumps are on the order of $e^{-12} \approx 6 \times 10^{-6}$, when using the weight function Eq. (2.2). These effects are smaller than the discretization error produced by the spatial mesh of our numerical integration scheme, so we would expect the strength of repulsion to be weaker than the pinning produced by discretizing, as discussed in Guo and Chow (2005).

Performance on the two-item WM task with random initial targets ϕ_1 and ϕ_2 is considered in Fig. 8e. Recall variability, represented by the MSE, is greater than what would be predicted by a model that allows distinct slots for each item. Note, there have been efforts to revise the slot model (Zhang and Luck 2008; Cowan 2010), so that error increases when considering two items versus one item. However, the increases in error arising from neural activity dynamics we observe are much more nuanced than would be possible for previous phenomenological slots or resources models (Zhang and Luck 2008; Bays and Husain 2008). Both items (bumps) are stored in a single network, producing interactions between bumps when items are initially close, which contributes an additional source of variability to the recall. Merging produces a systematic shift in the remembered location of items, as does repelling. The frequency of these interactions grows as the synaptic strength parameter $\mathcal{A}$ is increased, counteracting the reduction in diffusion also produced by increasing $\mathcal{A}$. This tradeoff produces a non-monotonic dependence of the MSE on $\mathcal{A}$ (Fig. 8e), so there is an optimal $\mathcal{A}$ for two-item storage with low-diffusion of bumps and low-probability of bump interaction. This optimum occurs when $\mathcal{A} \approx 10$, so even though the peak MSE is much larger than for the cases $\mathcal{A} = 1, 2$ (Fig. 8c), the average MSE is smaller since bumps are less susceptible to stochastic perturbations. Note, the MSE in the interacting bumps model is larger than would be predicted by a slots model that assumes MSE is unchanged as the number of items is increased up to some fixed capacity (Fig. 8e).

Our interacting bumps model can account for the item-dependent increase in the variability of recall in two-item WM tasks (Wilken and Ma 2004; Bays et al. 2009). This arises due to the nonlinear interactions between the bumps, which add to the variability already present due to the dynamic fluctuations in the network. We now examine item-dependent changes in recall variability for tasks with more than two items, showing our analysis extends to the case of multiple interacting bumps.

4 Multiple interacting bumps

Recent models of multi-item WM focus on uncovering the nature of item-number limitations, as they impact response variability (Ma et al. 2014). Phenomenological models can be altered to capture errors that either reflect a finite capacity or the distribution of resources (Zhang and Luck 2008), but physiologically-inspired models account for the architecture and dynamics of neural circuits underlying WM storage (Bays 2015). The work of Edin et al. (2009), Wei et al. (2012), and Almeida et al. (2015) has shown that a recurrent spiking network can support multiple bumps that each individually encode a different item. Our model is a tractable version of these previous studies, allowing us to derive explicit expressions describing limitations of the network.

Prior to developing effective equations for bump interfaces, we consider the problem of network capacity. This is one way in which our model differs from the standard resource model of WM. Only a finite number of bumps can be stored in the recurrent network, and this upper limit is determined by the choice of the synaptic strength parameter $\mathcal{A}$. However, we note this upper limit is quite large. We can approximate this limit by again examining a stationary solution problem.

4.1 Network capacity

We frame the problem of identifying network capacity by attempting to identify multi-bump stationary solutions to Eq. (2.1) in the absence of noise ($Z \equiv 0$). Finite multi-bump solutions are not stable in the limit $L \to \infty$ (Laing and Troy 2003b), since multiple active regions exert a repulsive drift on one another. If bumps are spaced evenly around the domain, the conformation is stable since the repulsive forces acting on each bump from either direction balance. Thus, stable multi-bump solutions constitute a periodic pattern that wraps around the domain. One question is just how the minimal period of this pattern changes as the synaptic strength $\mathcal{A}$ is changed. Since $\mathcal{A}$ increases the width of single bump solutions, one might expect the capacity to decrease as $\mathcal{A}$ is increased. We demonstrate in fact that the capacity of the network grows as the synaptic strength $\mathcal{A}$ is increased (Fig. 9a).
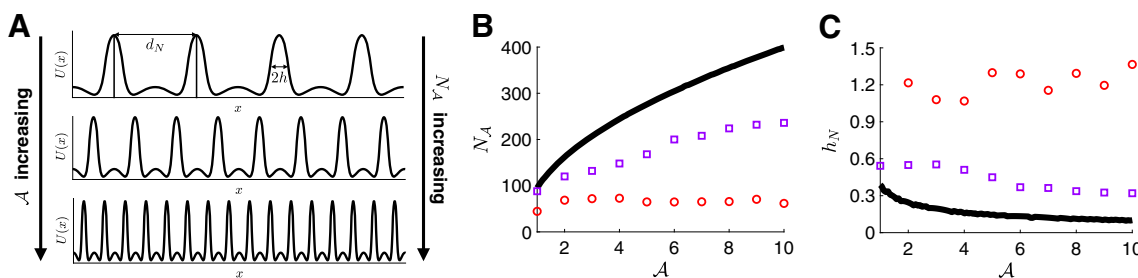
Fig. 9 Capacity of the network, Eq. (2.1), is bounded by the number of stable bumps $N_\mathcal{A}$ that can be instantiated in the periodic domain $x \in [-180, 180)$. **a** The capacity tends to increase with synaptic strength $\mathcal{A}$, since stronger synapses lead to narrower bumps in multibump solutions. **b** The maximal number $N_\mathcal{A}$ of bumps that can be packed into the domain increases with $\mathcal{A}$. Periodic solutions have spacing $d_N = 360/N$ between bump centroids, and the same half-width h_N for each bump. Solid line shows result from theory Eq. (4.4), squares are results from deterministic simulations, and circles are from stochastic simulations (See main text). Note for stochastic networks, the capacity is fairly flat across $\mathcal{A}$. **c** The bump half-width h_N associated with the maximal capacity solution narrows for increasing synaptic strength $\mathcal{A}$ owing to the strong recurrent inhibition for networks with large $\mathcal{A}$. Threshold $\theta = 0.25$

Network capacity can be bounded by examining the upper limit on the number of possible bumps in a periodic solution to Eq. (2.1). These numbers will tend to be much larger than those imposed by a slots model of WM capacity (Zhang and Luck 2008), so our model will still behave approximately as a resource model since the capacity $N_\mathcal{A}$ is quite high. The capacity can be estimated by studying the existence of multi-bump solutions, comprised of multiple stationary active regions of the same width, spaced an even distance apart. For example, a two-bump solution with centroids at $x = \pm 90$ has the form

$$U(x) = \int_{-90-h_2}^{-90+h_2} w(x-y)\mathrm{d}y + \int_{90-h_2}^{90+h_2} w(x-y)\mathrm{d}y.$$

As in the case of single bumps, there is one unknown, which is the half-width of each bump h_2. Self-consistency of the threshold conditions $U(-90 \pm h_2) = U(90 \pm h_2) = \theta$ yields an implicit equation

$$\theta = \int_0^{2h_2} w(y)\mathrm{d}y + \int_{180-2h_2}^{180} w(y)\mathrm{d}y, \quad (4.1)$$

which follows from the periodicity of the weight function ($w(180 + \alpha) = w(180 - \alpha)$). Computing integrals in Eq. (4.1) for the weight function Eq. (2.2), we find the implicit equation for the half width h_2 is

$$\theta = 2\mathcal{A}\left(h_2 \mathrm{e}^{-2h_2} + \mathrm{e}^{-180}\left[90 - (90 - h_2)\mathrm{e}^{2h_2}\right]\right), \quad (4.2)$$

so if Eq. (4.2) has a solution we expect the network with synaptic strength $\mathcal{A}$ to have capacity of at least two items. Note $\mathrm{e}^{-180} \approx 6.17 \times 10^{-79}$ is extremely small, so networks in which single bumps exist will likely also possess two bump solutions, since Eq. (4.2) is a very mild perturbation of Eq. (2.14). This approach can be generalized to the case of more than two bumps ($N > 2$), yielding analogous equations to Eq. (4.2) for the associated bump half-width.

Solutions with N distinct bumps have a regular spacing between each of the bump, forming a periodic pattern that tiles the domain. The spacing between the centroid of each bump is computed by partitioning the domain by N, $d_N = 360/N$. We can write an N-bump solution in the form

$$U(x) = \sum_{j=1}^{N} \int_{c_j-h_N}^{c_j+h_N} w(x-y)\mathrm{d}y$$

where $c_j = -180 + (j + 1/2)d_N$ is the location of the centroid of the jth bump. Threshold conditions $U(c_j \pm h_N) = \theta$ for $j = 1, ..., N$ yield the implicit equation for the half-width

$$\theta = \sum_{j=0}^{N-1} \int_{jd_N}^{jd_N+2h_N} w(y)\mathrm{d}y. \quad (4.3)$$

For N large, Eq. (4.3) will contain many very small terms corresponding to the interactions between distant bumps. It is easier to express the sum of integrals in Eq. (4.3) by approximating the N-bump solution on $x \in [-180, 180]$ with an infinite-bump solution on $x \in (-\infty, \infty)$. Placing the core bump at $x = 0$, and other bumps with centroids at jd_N for j a nonzero integer, yields the implicit equation

$$\theta = \sum_{j=-\infty}^{\infty} \int_{jd_N}^{jd_N+2h_N} w(y)\mathrm{d}y,$$

which we can integrate directly using the weight function Eq. (2.2), resulting in the following equation that must be solvable for a neural field on $x \in (-\infty, \infty)$ to have periodic bump solutions spaced d_N apart:

$$\theta = \mathcal{A}\left(2h_N \mathrm{e}^{-2h_N} + \mathcal{S}(h_N, d_N)\right), \quad (4.4)$$

where

$$\mathcal{S}(h, d) = \sum_{j=1}^{\infty} \mathrm{e}^{-jd}\left[4h\cosh(2h) - 2jd\sinh(2h)\right],$$

$$= \frac{4h\cosh(2h)}{\mathrm{e}^d - 1} - \frac{2d\mathrm{e}^d\sinh(2h)}{(\mathrm{e}^d - 1)^2}.$$

We also require $h_N \in [0, d_N/2)$, since the active region of each bump cannot overlap with another. Thus, the capacity $N_{\mathcal{A}}$ is approximated by the maximum value N for which Eq. (4.4) possesses a solution. We plot $N_{\mathcal{A}}$ as a function of $\mathcal{A}$ in Fig. 9b (solid line), showing our analytically determined capacity is an increasing function of synaptic strength. This may seem surprising, since one might expect that bumps will be wider as $\mathcal{A}$ grows. However, this is only true when bump interactions are not considered. Many bumps will tend to interact through strong inhibition, which narrows them, decreasing h_N (Fig. 9c), but they will still be sustained by the strong recurrent excitation generated by increasing $\mathcal{A}$.

Due to both the infinite domain approximation in Eq. (4.4) and inevitable truncation errors in our numerical root-finding scheme, we find that the maximal number of bumps predicted by our theory overestimates what we find in deterministic simulations (circles in Fig. 9b, c). Recall, we are identifying roots of a transcendental equation, which can be quite sensitive to truncation errors. Thus, in practice we would certainly not expect the capacity predicted by our theory to be obtained in simulations of Eq. (2.1) with the addition of stochastic forcing. While calculations using Eq. (4.4) provide a clean method for estimating the upper bound on bump number in the network, in practice, we expect these solutions to be sensitive to perturbations arising in a numerical integration scheme (See for example discussion in Guo and Chow 2005).

Thus, we performed a coarser estimate of the network capacity by using a simple numerical simulation method, to find how many bumps can be packed into the domain $x \in [-180, 180)$ for each $\mathcal{A}$. We ran deterministic numerical simulations of Eq. (2.1) in the absence of noise with the initial condition $u(x, 0) = \sin(N_{\mathcal{A}} \pi x / 180)$, allowing them to equilibrate after long time ($t \to 5$s). At this point, we counted the number of bumps remaining, and computed their half-width, plotting the result in comparison to our analysis in Fig. 9b, c (squares). In fact, this method provides a considerably lower bound on the estimate, but the trend of the maximum bump number increasing with $\mathcal{A}$ is still present.

To estimate capacity from stochastic simulations, we used the steady state of deterministic simulations as an initial condition for 10^4 Monte Carlo simulations of Eq. (2.1) for $t = 5$s, using parameters from Fig. 4. We then averaged the number of bumps at the end of all simulations. In this case, we found the capacity is fairly insensitive to $\mathcal{A}$ (Fig. 9b), but typically in the range of 60–70 bumps, well above the single digit range predicted in Edin et al. (2009) and Wei et al. (2012). A major reason for this is that these previous models assumed a wider footprint of inhibitory connectivity, whereas our model assumes inhibition with a similar spatial scale to excitation (See also Lim and Goldman 2014; Rosenbaum et al. 2017).

4.2 Interface equations

We extend the interface equations derived for the interactions of two bumps to account for interactions between an arbitrary number of bumps in the network. As in the case of two bumps, if multiple bumps' initial active regions overlap, they will merge. This sets an upper bound on the capacity of the network (as in Fig. 9). Our analysis proceeds now by projecting the dynamics of bumps in the network to equations that simply track bump interfaces.

The active region in this case is given by the union of N finite intervals, which we assume to be disjoint, $A(t) = \cup_{j=1}^{N} [a_j(t), b_j(t)]$. Note, in the case that bumps' active regions overlap, we simply redefine a set of fewer than N finite intervals, with some corresponding to the merged bumps. In the case of N distinct active regions, Eq. (2.1) becomes

$$\mathrm{d}u(x, t) = \left[-u(x, t) + \sum_{j=1}^{N} \int_{a_j(t)}^{b_j(t)} w(x - y)\mathrm{d}y \right] \mathrm{d}t + \sqrt{\epsilon \cdot |u(x, t)|}\mathrm{d}Z(x, t).$$

The dynamic threshold equations are then given

$$u(a_j(t), t) = u(b_j(t), t) = \theta, \qquad j = 1, ..., N. \quad (4.5)$$

As before, we differentiate Eq. (4.5), rewrite the corresponding integrals using Eq. (2.7), and assume the gradients at the threshold are static and approximated by the stationary bump gradients ($\bar{\alpha} = U_0'(-h)$). We can approximate the interface dynamics by the following stochastic system:

$$\mathrm{d}a_j = \frac{1}{\bar{\alpha}} \left(\left[\theta - \sum_{k=1}^{N} (W(a_j - a_k) - W(a_j - b_k)) \right] \mathrm{d}t - \sqrt{\epsilon \cdot \theta}\, \mathrm{d}Z(a_j, t) \right), \quad (4.6a)$$

$$\mathrm{d}b_j = -\frac{1}{\bar{\alpha}} \left(\left[\theta - \sum_{k=1}^{N} (W(b_j - a_k) - W(b_j - b_k)) \right] \mathrm{d}t - \sqrt{\epsilon \cdot \theta}\, \mathrm{d}Z(b_j, t) \right), \quad (4.6b)$$

for $j = 1, ..., N$. Furthermore, the dynamics of the centroid of each bump j can be tracked by using the change of variables, $\Delta_j = (a_j + b_j)/2$, along with the approximations $a_j \approx \Delta_j - h$ and $b_j \approx \Delta_j + h$, yielding the system of N stochastic differential equations

$$\mathrm{d}\Delta_j = \frac{1}{\bar{\alpha}} \left(\sum_{k \neq j} J(\Delta_k - \Delta_j)\mathrm{d}t + \frac{\sqrt{\epsilon \cdot \theta}}{2}\mathrm{d}\mathcal{Z}(\Delta_j, t) \right), (4.7)$$

for $j = 1, ..., N$, where $J(\Delta)$ and $\mathrm{d}\mathcal{Z}(\Delta_j, t)$ are defined as in Eq. (3.10a, b). While we could employ the low-dimensional Eq. (4.7) to approximate how the network,

Eq. (2.1), performs on multi-item WM tasks, we opt to preserve the interface information in Eq. (4.6a, b). Interactions between multiple bumps are much better captured when the width of bumps is also considered along with their position. Truncating to only consider the centroid ignores width perturbations, which for example ignores bump annihilation events.

We demonstrate the efficacy of the low-dimensional description, Eq. (4.6a, b) in capturing the dynamics of the bump interfaces in Fig. 10. One category of dynamics that is much more common in the case of multiple bumps is annihilation of a bump by two neighboring bumps on either side (Fig. 10a). Notice, the middle bump does not merge with the bump on the left or right, but is extinguished by their combined recurrent inhibition. This is captured reasonably well by our interface Eq. (4.6a, b). Note, annihilation events can also occur in the case of two bumps being instantiated close to one another. The combination of noise and lateral inhibition can lead to one bump's extinction without it merging with the neighboring bump. Bump edges fluctuate in response to noise, and multiple bumps can collectively repel each other when they are not too close (Fig. 10b). Merging occurs when two bumps begin close to one another, but are not in the vicinity of other bumps (Fig. 10c).

4.3 Performance

Estimation errors for WM tasks involving an arbitrary number of targets N, $\{\phi_1, \phi_2, ..., \phi_N\}$, still mostly arise from merging, repulsion, or diffusion. However, recurrent inhibition from multiple active regions can also result in annihilation of bumps, so the activity is extinguished separately from a neighboring bump (as shown in Fig. 10a). These errors combine to shape the response variability, measured by the MSE in Eq. (3.11), as a function of the item number N. Our results demonstrate much more

consistency with a resource model of multi-item WM. In particular, we see that the response variability for individual items increases with the number of items stored in memory, starting with the difference between $N = 1$ and $N = 2$ discussed in the previous section (Fig. 8e).

We have to make specific choices about how multiple bumps encode the multiple target items that initially instantiate them. As in the case of two-items, we ignore probabilistic effects that could further contribute to the error beyond that captured by the stochastic dynamics of the bumps (Bays et al. 2009). Lapses and swaps are not considered in our model. When two bumps merge with one another, or one bump is annihilated, the items associated with vanishing bumps are then associated to the closest remaining bump. Handling bump annihilation events parsimoniously singles out the contributions to the error engendered solely by the neural circuit dynamics described by our recurrent network.

The MSE of a single item is computed using the formula, Eq. (3.11), as before. Recall, we assume that a subject is probed about a single item in WM, as this is the protocol typical used to measure behavioral errors (Ma et al. 2014). Furthermore, this allows us to significantly reduce the space of dynamics that we must track in each numerical simulation. Rather than having to track the locations of all bumps, it is sufficient to only follow the dynamics of bumps in the vicinity of the bump originating from the item of interest. We examine performance when target angles $\{\phi_1, \phi_2, ..., \phi_N\}$ are randomly placed about the domain $x \in [-180, 180)$ (Fig. 11a). First, we study the scaling of recall variability, measured by MSE, as a function of the number of items stored N. As expected by our two-bump performance results, we find that the MSE grows steadily with N (Fig. 11b). This is suggestive of the resource model of WM, since the increase occurs across all N. Note, the simulations of the interface equations, Eq. (4.6a, b), (solid line) agree with with simulations of the full model, Eq. (2.1),
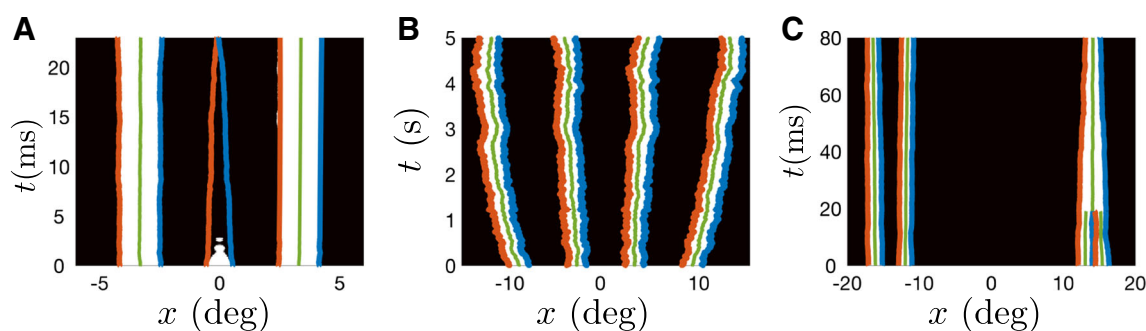

Fig. 10 Categories of bump interactions that induce WM errors. **a** Annihilation occurs when a bump is closely flanked by two others, leading to extinction of the inner bump. **b** Repulsion drives bumps away from one another, due to the recurrent inhibition acting at long distances. **c**. Bumps will merge if two begin close and are not in the vicinity of other bumps. Interface equations, Eq. (4.6a, b), track the dynamics of full numerical simulations of Eq. (2.1). Parameters are $\epsilon = 0.03$ and $\theta = 0.25$
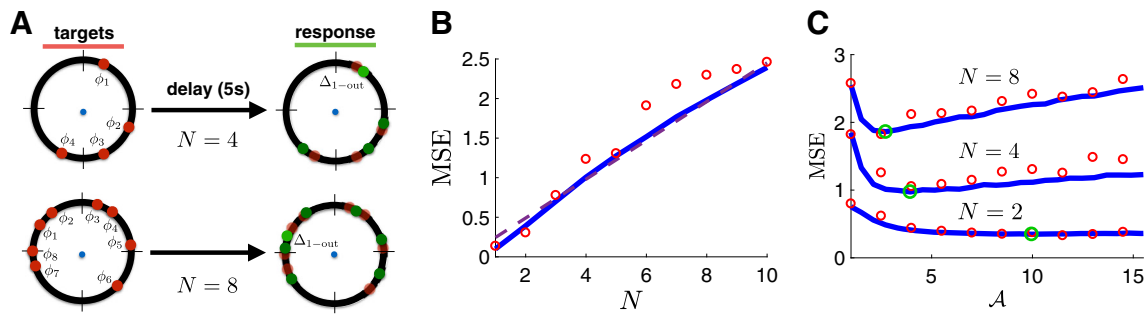
Fig. 11 Recall error in multi-item WM task. **a** There are N target items $\{\phi_1, \phi_2, ..., \phi_N\}$ whose locations are chosen uniformly randomly around $x \in [-180, 180]$. Mean squared error (MSE) is computed by comparing the output angle Δ_{1-out} of a randomly chosen input ϕ_1 via Eq. (3.11). **b**. MSE increases with the number of items N, computed from Eq. (4.6a, b), (solid line), agreeing with numerical simulations of Eq. (2.1), (circles). Trend is fit by a linear function (dashed line). We fix $\mathcal{A} = 5$. **C**. MSE is nonmonotonic in synaptic strength $\mathcal{A}$, owing to the tradeoff between fluctuation-driven errors (at small $\mathcal{A}$), and bump interaction errors (at large $\mathcal{A}$). Note, the optimal $\mathcal{A}$ (large circles) decreases as N is increased

even though the interface equations are considerably faster to numerically solve. In addition, we find the MSE trend is well fit by a linear function of N. Next, we study the effect of varying the synaptic strength $\mathcal{A}$ for different item number counts. Extending our findings for two-item memory, we see there is an optimal $\mathcal{A}$ that minimizes the MSE, and this optimum decreases as N is increased (Fig. 11c). Thus, as more items must be stored in the network, it is advantageous to have networks with narrower bumps to decrease the probability of bump interactions.

5 Extensions to multiple stimulus features

Thus far, we have studied a network that represents multiple items in WM as bumps within a single-layer neural field. However, WM tasks typically employ stimuli with multiple features (e.g., color and orientation), which likely are represented by populations of neurons with tuning that varies along multiple feature dimensions (Mante et al. 2013; Schneegans and Bays 2017). In light of this, we now consider an extension of Eq. (2.1) to a multilayer neural field model with spatially-organized interlaminar connectivity. Each distinct layer represents a different value of a quantized stimulus feature. For example, the discrete stimulus feature may be color, and the continuous variable, represented by position within each layer, could encode orientation (as represented in Figs. 8a, d and 11a). When multi-item WM is represented by the multilayer neural field, interaction-based errors decrease since stimuli of different colors interact less.

The multilayer neural field model we analyze builds on a model introduced in Kilpatrick (2013), which represents WM for a single-item using multiple bumps in distinct but coupled network layers. Here, we consider a variation of this model that allows for the representation of multiple items,

each with their own bump, possibly in distinct layers of the network (Fig. 12a). Activity in this neural field model is now described by the following set of evolution equations:

$$\mathrm{d}u_j(x,t) = \left[-u_j(x,t) + \sum_{k=1}^{M} w_{jk} * H(u_k - \theta) \right] \mathrm{d}t \\ + \sqrt{\epsilon \cdot |u_j|} \mathrm{d}Z_j, \quad (5.1)$$

where $j = 1, ..., M$ are the layers of the network. Each distinct layer represents a different color value, whereas spatial locations x correspond to the preferred stimulus orientation of neurons there. Synaptic connectivity within and between layers is described by layer-dependent weight function

$$w_{jk}(x) = \mathcal{A}_{jk} \left(1 - \frac{|x|_L}{\sigma_{jk}} \right) \mathrm{e}^{-|x|_L/\sigma_{jk}}, \quad (5.2)$$

so $\mathcal{A}_{jk}$ and σ_{jk} scale the amplitude and width of synaptic connectivity. We chose to normalize connectivity within layers, $\mathcal{A}_{jj} := \mathcal{A}$, and take weaker connectivity between layers $0 \leq \mathcal{A}_{jk} < \mathcal{A}$ ($j \neq k$). In addition, we normalize the width of connectivity within layers $\sigma_{jj} := 1$ and take narrow connectivity between layers, $0 \leq \sigma_{jk} < 1$. The antiderivative of Eq. (5.2) is then given

$$W_{jk}(x) = \int_0^x w_{jk}(y)\mathrm{d}y = \mathcal{A}_{jk}x\mathrm{e}^{-|x|/\sigma_{jk}}, \quad (5.3)$$

and $W_{jj}(x) = W(x)$ from Eq. (2.7). Noise to each layer is weak and multiplicative $\sqrt{\epsilon \cdot |u_j(x,t)|} \mathrm{d}Z_j(x,t)$, where we define spatially-dependent Wiener processes such that $\langle \mathrm{d}Z_j(x,t) \rangle = 0$ and $\langle \mathrm{d}Z_j(x,t)\mathrm{d}Z_k(y,s) \rangle = C_{jk}(x - y)\delta(t-s)\mathrm{d}t\mathrm{d}s$ for $j, k = 1, ..., M$. In numerical simulations, we consider independent cosine correlated noise to each layer $C_{jj}(x) = \cos(\omega_c x)$, where $\omega_c = c\pi/L$, so that $C_{jk}(x) \equiv 0$ ($j \neq k$), but see Kilpatrick (2013) for more details on the impact of noise correlations between layers.

We can extend our analysis of Eq. (2.1) to consider interactions of multiple bumps in Eq. (5.1). Our reduction
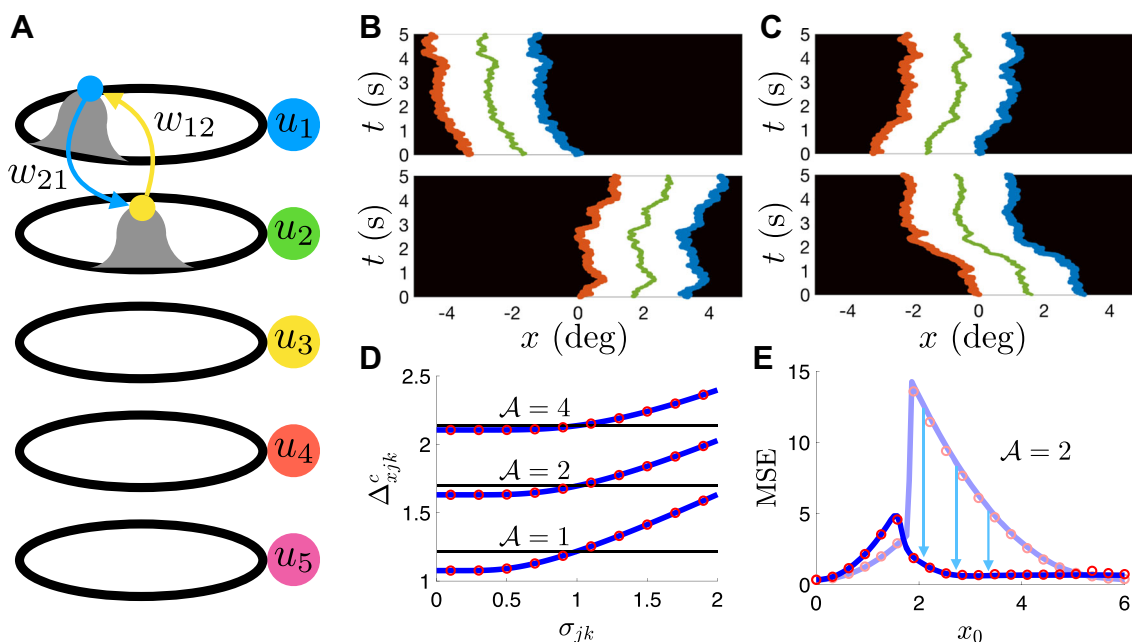
Fig. 12 Interaction-related errors are reduced in a multilayer network model, Eq. (5.1). **a** Recurrent network is now comprised of multiple rings with both within layer connectivity $w_{jj}(x - y)$ and interlayer connectivity $w_{jk}(x - y)$, as described by Eq. (5.2). The centroid location x of a bump within a layer represents an encoded orientation, whereas the layer in which it resides encodes color. Bumps in distinct layers communicate via weak interlayer connectivity. **b** Bumps initiated in two distinct layers (1 and 2) repel one another when initiated far enough apart in the x-dimension (at $\pm x_0 = \pm 1.8$ here). **c** Bumps initiated in two distinct layers (1 and 2) are attracted to one another when initiated close enough together in the x-dimension (at $\pm x_0 = \pm 1.6$ here). **d** The critical boundary Δ^c_{xjk} between collocation and repulsion

is determined by Eq. (5.8), (solid curves), and agrees with direct simulations (circles). Note, the minimal distance for bumps to repel one another increases as the width of interlaminar connections $\sigma_{jk} = \sigma_{kj}$ is increased. **e** MSE as a function of x_0 computed from Eq. (3.11) is maximized near Δ^c_{xjk}, where repulsion is strongest, in the multilayer neural field, Eq. (5.1), for bumps initiated in different layers (1 and 2). Low-dimensional approximation, Eq. (5.9a, b), (dark blue line) agrees well with simulations of full model Eq. (5.1) (dark red circles). Compare with the MSE in the single-layer neural field, Eq. (2.1), which is appreciably larger (light blue line and circles). Unless stated otherwise, parameters are $\theta = 0.25$, $\mathcal{A} = 2$, $\mathcal{A}_{12} = \mathcal{A}_{21} = 0.1$, $\sigma_{12} = \sigma_{21} = 0.5$, $\epsilon = 0.03$, and $w_c = 25\pi/180$

proceeds by projecting the dynamics of bumps in multiple layers to equations tracking interfaces of each bump in its respective layer. Note that now bumps can only merge if they begin in the same layer j. If bump centroids are initially close in the x-dimension but reside in different layers, the bumps will colocate to roughly the same x location in their respective layers (Fig. 12b). Repulsion between layers can occur as well (Fig. 12c). For N bumps, the active region is given by N finite intervals, assumed to be disjoint:

$$A(t) = \cup_{n=1}^{N} [a_n(t), b_n(t)] \times \{l(n)\},$$

where $\{l(n)\}$ denotes the layer in which bump n resides. As before, if bumps' active regions overlap, we redefine a smaller set of intervals. In the case of N active regions, Eq. (5.1) becomes

$$du_j = \left[-u_j + \sum_{n=1}^{N} \int_{a_n(t)}^{b_n(t)} w_{jl(n)}(x - y) dy \right] dt$$
$$+ \sqrt{\epsilon \cdot |u_j(x, t)|} dZ_j(x, t).$$

The dynamic threshold equations are then given by

$$u_{l(n)}(a_n(t), t) = u_{l(n)}(b_n(t), t) = \theta, \tag{5.4}$$

for $n = 1, ..., N$. Differentiating Eq. (5.4), rewriting the integrals using Eq. (2.7), and using the static gradient approximation, we can write the interface dynamics as

$$da_n = \frac{1}{\bar{\alpha}} \left(\left[\theta - \sum_{m=1}^{N} (W_{l(n)l(m)}(a_n - a_m) \right. \right.$$
$$\left. \left. - W_{l(n)l(m)}(a_n - b_m)) \right] dt - \sqrt{\epsilon \cdot \theta} dZ_n(a_n, t) \right), \tag{5.5a}$$

$$db_n = -\frac{1}{\bar{\alpha}} \left(\left[\theta - \sum_{m=1}^{N} (W_{l(n)l(m)}(b_n - a_m) \right. \right.$$
$$\left. \left. - W_{l(n)l(m)}(b_n - b_m)) \right] dt - \sqrt{\epsilon \cdot \theta} dZ_n(b_n, t) \right), \tag{5.5b}$$

for $j = 1, ..., N$, where the layer-dependence of the bump interactions is expressed via the modulation of the interlaminar weight antiderivatives $W_{l(n)l(m)}$ given by Eq. (5.3).

We now demonstrate how this multilayer network structure can lead to an effective reduction in bump interactions, even in the noise-free system ($Z_j \equiv 0$, $\forall j =$

1, ..., M). For simplicity, we focus on the case of two symmetrically-placed bumps in two distinct layers

$$A(0) = [-b_0, -a_0] \times \{j\} \cup [a_0, b_0] \times \{k\},$$

of a symmetric network ($W_{jk}(x) \equiv W_{kj}(x), \; \forall j, k$). Therefore, we can write Eq. (5.5a, b) as

$$a_1'(t) = \frac{1}{\alpha} \big[\theta - W(b_1 - a_1) - W_{jk}(b_2 - a_1) \\ + W_{jk}(a_2 - a_1) \big], \tag{5.6a}$$

$$b_1'(t) = -\frac{1}{\alpha} \big[\theta - W(b_1 - a_1) - W_{jk}(b_2 - b_1) \\ + W_{jk}(a_2 - b_1) \big], \tag{5.6b}$$

$$a_2'(t) = \frac{1}{\alpha} \big[\theta - W(b_2 - a_2) - W_{kj}(a_2 - a_1) \\ + W_{kj}(b_2 - a_1) \big], \tag{5.6c}$$

$$b_2'(t) = -\frac{1}{\alpha} \big[\theta - W(b_2 - a_2) - W_{kj}(b_2 - a_1) \\ + W_{kj}(b_2 - b_1) \big], \tag{5.6d}$$

so by symmetry, $a(t) := a_2(t) = -b_1(t)$ and $b(t) := b_2(t) = -a_1(t)$, and Eq. (5.6a, b, c and d) reduces to

$$a'(t) = \frac{1}{\alpha} [\theta - W(b(t) - a(t)) + W_c(2a(t)) \\ - W_c(a(t) + b(t))], \tag{5.7a}$$

$$b'(t) = \frac{1}{\alpha} [W(b(t) - a(t)) - \theta + W_c(2b(t)) \\ - W_c(a(t) + b(t))], \tag{5.7b}$$

where $W_c(x) := W_{jk}(x) \equiv W_{kj}(x)$. Note, Eq. (5.7a, b) is an approximation of the evolution of the interfaces, since we use a static approximation for the interface gradients. However, the fixed points of Eq. (5.7a, b) and the exact system should be the same. As in the single-layer case, there is a critical distance Δ_{xjk}^c which divides solutions that drift apart ($\Delta_x > \Delta_{xjk}^c$) from those tend towards collocation ($\Delta_x < \Delta_{xjk}^c$), given that one bump is in layer j and the other in k. An implicit analytical expression for Δ_{xjk}^c can be determined by looking for the equilibria of Eq. (5.7a, b). As before, the critical curve (a^c, b^c) is determined by the zeros of the right hand side of the $a'(t)$ equation, yielding

$$\theta = W(b^c - a^c) + W_c(a^c + b^c) - W_c(2a^c), \tag{5.8}$$

a curve dividing initial conditions ($a(0), b(0)$) into those that lead to collocation and those that repel. As before, we can assume $b^c - a^c = 2h$, and find the critical curve in terms of the stationary bump half-width h and centroid $\Delta_{xjk}^c = (a^c + b^c)/2$, yielding $W_c(2(\Delta_{xjk}^c - h)) = W_c(2\Delta_{xjk}^c)$. Defining $\mathcal{A}_c := \mathcal{A}_{jk} = \mathcal{A}_{kj}$ and $\sigma_c := \sigma_{jk} = \sigma_{kj}$, we can thus apply Eq. (5.3) and solve for $\Delta_{xjk}^c = h/(1 - e^{-2h/\sigma_{jk}})$. Clearly, Δ_{xjk}^c decreases with σ_{jk}. Thus, since $0 \leq \sigma_{jk} < 1$ ($j \neq k$), we expect the critical distance Δ_{xjk}^c to be narrower

here than in the single layer case (Fig. 12d). Interestingly, there is no dependence of Δ_{xjk}^c on the amplitude $\mathcal{A}_{jk}$ of interlaminar connectivity, although this will affect the rate at which the bump collocate. Since bumps are in different layers ($j \neq k$), both of them will still remain after a collocation event and will be then centered at the same location in their respective layers (See also (Folias and Ermentrout 2011)).

To consider the combination of bump interactions and stochastic forcing on the dynamics of bumps in the multilayer network, Eq. (5.1), we also derive a reduction of Eq. (5.5a, b), which tracks the centroid of two bumps. For the purposes of comparison with the single-layer model, we gain significant insight by comparing performance on two-item WM tasks. Starting with Eq. (5.5a, b) assuming $N = 2$, we assume the width of each bump remains constant, so $a_j = \Delta_j - h$ and $b_j = \Delta_j + h$ for $j = 1, 2$. As before, we assume symmetric interlaminar connectivity, $W_c(x) := W_{12}(x) \equiv W_{21}(x)$, and we also assume noise is symmetric and independent. The resulting equations for the bump centroids are

$$d\Delta_1 = \frac{1}{\alpha} \left(J_c(\Delta_2 - \Delta_1)dt + \frac{\sqrt{\epsilon \cdot \theta}}{2} d\mathcal{Z}_1(\Delta_1, t) \right), \tag{5.9a}$$

$$d\Delta_2 = \frac{1}{\alpha} \left(J_c(\Delta_1 - \Delta_2)dt + \frac{\sqrt{\epsilon \cdot \theta}}{2} d\mathcal{Z}_2(\Delta_2, t) \right), \tag{5.9b}$$

where

$$J_c(\Delta) = \frac{1}{2} [2W_c(\Delta) - W_c(\Delta - 2h) - W_c(\Delta + 2h)],$$
$$d\mathcal{Z}_j(\Delta_j, t) = dZ_j(\Delta_j + h, t) - dZ_j(\Delta_j - h, t).$$

Our low-dimensional approximation, Eq. (5.9a, b), for the evolution of the centroids can be used to compute the estimation errors of the network Eq. (5.1) in the case of two target locations, ϕ_1 and ϕ_2. Similar to the single-layer network, Eq. (2.1), errors arise due to merging, repulsion, and diffusion, but the effects of interactions between bumps will be weaker when bumps are initiated in two different layers. Also, note in the collocation events discussed above, bumps are attracted to the same relative position within a layer, but two bumps still remain. As before, we consider a task wherein the target ϕ_1 is probed at the end of a delay-period and compute the MSE according to Eq. (3.11).

One distinction in comparison to the single layer network case is that bumps that begin close enough, $\phi_2 - \phi_1 < 2\Delta_{x12}^c = h/(1 - e^{-2h/\sigma_{jk}})$ (Fig. 12c), will collocate and limit each others' diffusion via the attraction of interlaminar connections. The stochastic dynamics of this effect has been analyzed in detail in Kilpatrick (2013) and Bressloff and Kilpatrick (2015) assuming the relative distance between bumps is small. In essence, a linear approximation is applied to Eq. (5.9a, b) and the variance of (Δ_1, Δ_2) can be

computed directly for the resulting multivariate Ornstein-Uhlenback process

$$d\Delta_1 = \kappa\,[-\Delta_1 + \Delta_2]\,dt + d\mathbf{Z}_1, \qquad (5.10a)$$

$$d\Delta_2 = \kappa\,[+\Delta_1 - \Delta_2]\,dt + d\mathbf{Z}_2, \qquad (5.10b)$$

where $\kappa = (w_c(0) - w_c(2h))/\bar{\alpha}$ $(w_c(x) := w_{12}(x) = w_{21}(x))$ and

$$d\mathbf{Z} = \frac{\sqrt{\epsilon \cdot \theta}}{2\bar{\alpha}} \begin{pmatrix} d\mathcal{Z}_1(\Delta_1, t) \\ d\mathcal{Z}_2(\Delta_2, t) \end{pmatrix},$$

so that for symmetric and independent noise, $\langle \mathbf{Z}_1(t)\mathbf{Z}_2(t)\rangle \equiv 0$ and $\langle \mathbf{Z}_1^2(t)\rangle = \langle \mathbf{Z}_2^2(t)\rangle = Dt$ with D defined as in Eq. (2.24). We can in fact directly compute the first and second moments of Δ_1 from Eq. (5.10a, b) as in Kilpatrick (2013). For simplicity, we present our results for the case of symmetric initial conditions $\phi_1 = -\phi_2 = x_0$:

$$\langle \Delta_1(t) \rangle = x_0 e^{-2\kappa t},$$

$$\langle \Delta_1^2(t) \rangle = x_0^2 e^{-4\kappa t} + \frac{D}{2}t + \frac{D}{8\kappa}\left[1 - e^{-4\kappa t}\right].$$

Thus, for collocating bumps, we can approximate Eq. (3.11), similar to Eq. (3.12) as

$$\begin{aligned}
\mathcal{M}_{\text{mg}} &= \langle (\Delta_{1-\text{out}} - \phi_1)^2 \rangle \\
&= \langle \Delta_{1-\text{out}}^2 \rangle - 2x_0\langle \Delta_{1-\text{out}}\rangle + x_0^2 \\
&= \frac{D}{2}T + \frac{D}{8\kappa}\left[1 - e^{-4\kappa T}\right] + x_0^2\left(1 - e^{-2\kappa T}\right)^2,
\end{aligned}$$

so notice that the interlaminar coupling dampens the long-term fluctuations in the bumps' positions so the diffusion is half that of the single-layer network (Kilpatrick 2013). Additional contributions arise from the competition between stochastic fluctuations driving the bumps apart and the coupling pulling them together. This gives us an analytic approximation of the MSE for collocating bumps, $x_0 < \Delta_{x12}^c$.

When bumps do not attract one another, then we can use our nonlinear approximation Eq. (5.9a, b) to estimate the error in recall via direct simulation. We compare this approximation to simulations of the multilayered neural field, Eq. (5.1) in Fig. 12e. The distance-dependence of MSE is demonstrated in comparison to the MSE in the single layer neural field, Eq. (2.1). As expected, because the connectivity between layers is weaker and narrower than connectivity within layers, bumps that are initiated in the same layer interact more strongly than bumps in different layers. For example, items of the same color will tend to interact in WM more strongly than items of different colors. We expect this result will extend to the case of randomized item locations, and will explore this more in subsequent work.

We conclude that an interacting bumps model of multi-item WM can capture several key features of error. Synaptic fluctuations lead to the time-dependent scaling typically observed in parametric WM tasks (White et al. 1994; Ploner et al. 1998; Wimmer et al. 2014). More important for multi-item WM tasks is the impact of item number on the reliability of storage (Ma et al. 2014). Here, we have shown that interactions in item memory can be described by the dynamics of multiple bumps in a common network. Specifically, items that are closer to one another in one or more feature dimensions will have associated bumps that interact more strongly, potentially leading to bump annihilations or repulsions. Notably, our mechanism is much more suggestive of the resource model of WM (Bays and Husain 2008) than of a slots model, which would assume that the first few items stored do not have associated memories that interact in any way (Zhang and Luck 2008).

6 Discussion

Working memory is a central feature of cognition, which plays an important role in attention (Gazzaley and Nobre 2012) and motor planning (Ikkai and Curtis 2011). Limitations on the fidelity of working memory can therefore limit other cognitive functions. We have proposed a simplified model to account for item and temporal limitations in multi-item WM, based on recent studies of spiking networks (Wei et al. 2012; Almeida et al. 2015). Similar to these previous studies, we associate the memory of an item in space with the location of a bump attractor, subject to fluctuations and interactions with other bumps.

The advantage of our model is that we were able to analyze the dynamics of the network, and reduce the dynamics to a low-dimensional system describing the interfaces of the bumps. Errors in recall occur due to merging, repulsion, and annihilation events resulting from bump interactions, which can also be captured by the corresponding interface equations. This is in addition to the typical fluctuation-induced errors known to arise during the delay-period of visual WM tasks (Wimmer et al. 2014; Constantinidis and Klingberg 2016). Importantly, we have shown that the strength of synaptic coupling in a recurrent network shapes the mean squared error (MSE) in WM tasks. More weakly coupled networks support narrower bump attractors that interact less, but are more subject to noise fluctuations. Strongly coupled networks possess wider bumps, which are less subject to noise fluctuations. There is an optimal scaling of synaptic strength that minimizes the MSE, trading off reduced effects of noise with bumps that do not interact too strongly. This optimal scaling strength decreases for tasks requiring memory of more items N.

Interface methods were used to reduce the neural field model to a system of a few differential equations, corresponding to the threshold crossing points of dynamically evolving bumps. An exact description of the evolution of

the interfaces can be obtained by evolving an integral equation describing the dynamically evolving gradient at the interfaces (Coombes et al. 2012). However, these integral equations are much less straightforward to derive for the case in which stochastic forcing is incorporated into the evolution equations, so we employed a static gradient approximation, ignoring the perturbations of the gradient near the interfaces. We expect that interface equations that capture these fluctuations in the gradient would provide a more accurate approximation of the dynamics of the full neural field model, Eq. (2.1). Nonetheless, we were able to derive a reasonably accurate approximation for the dynamics of multiple remembered items during the delay period of a WM task. This low-dimensional description could be leveraged in future work to explore the effects of bump interactions in two- and three-dimensional feature space.

Controversy still remains as to whether errors in multiple item WM are best explained using a slots (Zhang and Luck 2008) or resource (Bays and Husain 2008) model of item storage. Our model mostly supports the latter hypothesis, since recall errors depend on the item number across all item counts. Furthermore, we expect that the practical capacity of the network will be relatively high, since many bumps can be stored in a single recurrent network, and there will not tend to be an abrupt drop in accuracy at any particular item count. Thus, we see no strong evidence of a small and fixed capacity in our model. Rather, the capacity of the network is much larger, and would likely not be revealed by the typical single digit item counts used in WM experiments (Ma et al. 2014). Furthermore, our model predicts that if two angles are initially placed close to one another, there will be more error in their recalled locations. This is due to the nonlinear interactions in the network. This is consistent with a recent experimental study of multi-item working memory tasks analyzed in Almeida et al. (2015).

Computational models of multi-item WM that use multiple activity bumps have been explored previously in the works of Macoveanu et al. (2006), Edin et al. (2009), and Wei et al. (2012). However, none of these works provided a concise description of the neural and synaptic mechanics underlying specific forms of recall error, as we have done using our interface equations approach. Primarily, these previous works explored trends in large-scale simulations. One exception is Edin et al. (2009), who were able to derive a steady-state estimate of the capacity of their network, showing it to be somewhere between two and seven items due to inhibition decreasing the excitability of inactive regions of the network. Our network does not exhibit this low capacity, likely due to the fact that excitation and inhibition are balanced so the total excitatory and inhibitory input to the network sums to zero. This approach is motivated by recent work suggesting the common features of cortical variability arise from balanced and spatially-organized excitation and inhibition (Lim and Goldman 2014; Rosenbaum et al. 2017). In bump attractor models of WM, we expect the mean network activity will tend to increase with item number in networks imbalanced towards excitation (Edin et al. 2009), and stay relatively constant in networks with roughly balanced excitation/inhibition (Wei et al. 2012). In this regard, our results are more in line with Wei et al. (2012), especially because Edin et al. (2009) consider the incorporation of external inputs which further modifies the network's mean excitability.

We could also extend our analyses to consider WM for objects residing in a higher-dimensional feature space. For instance, we could carry out a more thorough analysis of the mulitlayer neural field model introduced in Section 5, or analyze the problem of a stimulus space covering two continuous dimensions. Note that working memory tasks usually require subjects to remember both the color and position of remembered items, so that, for example, the color can be used to indicate which item the subject should recall (Bays and Husain 2008). Studying the interactions of neural activity along two stimulus dimensions would also aid in providing a mechanistic explanation for swap errors. Swap errors can occur when a subject uses the stored location of a different item to report the remembered location of a cued item (Bays et al. 2009). Thus, it may be that both item interactions and fluctuations that produce errors in the stored position may affect the stored color in a similar way. Recently, Schneegans and Bays (2017) considered a related neural population model that could simultaneously represent color, orientation, and location of a stimulus, an extension of a model initially discussed in Bays (2014). Indeed, stochasticity in the neural code could account for swap errors and response variability, although they did not consider the detailed dynamics of the neural populations during a delay period.

Our mechanistic model of multiple item WM limitations has a distinct advantage over heuristic parameterized models, in that it is linked to constraints of physiology. However, we expect there are many extensions of the framework, which would likely more reliably reflect actual physiology. In particular, while there is evidence for persistent neural activity in visual WM, we expect that the strongly bistable nature of bumps in Eq. (2.1) may be inconsistent with the various accumulating and decaying activity traces observed in some cortical areas during WM tasks (Murray et al. 2017; Zylberberg and Strowbridge 2017). Such temporally heterogeneous activity, distinct from the relatively stable trace of a bump attractor, may still provide a stable population code. Thus, it would be interesting to explore how dynamic neural activity traces representing different items would interact in a computational model, and whether the principles of a resource-type model of WM would arise there as well.

Compliance with Ethical Standards

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

References

Almeida, R., Barbosa, J., Compte, A. (2015). Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *Journal of Neurophysiology*, *114*(3), 1806–1818.

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

Amari, S. (2014). Heaviside world: excitation and self-organization of neural fields. In *Neural fields* (pp. 97–118). Springer.

Avitabile, D., Desroches, M., Knobloch, E. (2017). Spatiotemporal canards in neural field equations. *Physical Review E*, *95*(4), 042,205.

Barak, O., & Tsodyks, M. (2014). Working models of working memory. *Current Opinion in Neurobiology*, *25*, 20–24.

Bays, P.M. (2014). Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience*, *34*(10), 3632–3645.

Bays, P.M. (2015). Spikes not slots: noise in neural populations limits working memory. *Trends in Cognitive Sciences*, *19*(8), 431–438.

Bays, P.M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851–854.

Bays, P.M., Catalao, R.F., Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7–7.

Bressloff, P.C. (2005). Weakly interacting pulses in synaptically coupled neural media. *SIAM Journal on Applied Mathematics*, *66*(1), 57–81.

Bressloff, P.C. (2009). Stochastic neural field theory and the system-size expansion. *SIAM Journal on Applied Mathematics*, *70*(5), 1488–1521.

Bressloff, P.C. (2012). Spatiotemporal dynamics of continuum neural fields. *Journal of Physics A: Mathematical and Theoretical*, *45*(3), 33,001–33,109.

Bressloff, P.C., & Kilpatrick, Z.P. (2015). Nonlinear langevin equations for wandering patterns in stochastic neural fields. *SIAM Journal on Applied Dynamical Systems*, *14*(1), 305–334.

Bressloff, P.C., & Webber, M.A. (2012). Front propagation in stochastic neural fields. *SIAM Journal on Applied Dynamical Systems*, *11*(2), 708–740.

Burak, Y., & Fiete, I.R. (2012). Fundamental limits on persistent activity in networks of noisy neurons. *Proceedings of the National Academy of Sciences*, *109*(43), 17,645–17,650.

Buschman, T.J., Siegel, M., Roy, J.E., Miller, E.K. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences*, *108*(27), 11,252–11,255.

Carroll, S., Josić, K., Kilpatrick, Z. (2014). Encoding certainty in bump attractors. *Journal of Computational Neuroscience*, *37*(1), 29–48.

Compte, A., Brunel, N., Goldman-Rakic, P.S., Wang, X.J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, *10*(9), 910–923.

Constantinidis, C., & Klingberg, T. (2016). The neuroscience of working memory capacity and training. *Nature Reviews Neuroscience*, *17*(7), 438–449.

Coombes, S., & Laing, C. (2011). Pulsating fronts in periodically modulated neural field models. *Physical Review E*, *83*(1), 011,912.

Coombes, S., & Owen, M.R. (2005). Bumps, breathers, and waves in a neural network with spike frequency adaptation. *Physical Review Letters*, *94*(14), 148,102.

Coombes, S., & Schmidt, H. (2010). Neural fields with sigmoidal firing rates: approximate solutions. *Discrete and Continuous Dynamical Systems Series 28*, 1369–1379.

Coombes, S., Schmidt, H., Bojak, I. (2012). Interface dynamics in planar neural field models. *The Journal of Mathematical Neuroscience*, *2*(1), 9.

Cowan, N. (2010). The magical mystery four: how is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51–57.

Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegnér, J., Compte, A. (2009). Mechanism for top-down control of working memory capacity. *Proceedings of the National Academy of Sciences*, *106*(16), 6802–6807.

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

Folias, S., & Ermentrout, G. (2011). New patterns of activity in a pair of interacting excitatory-inhibitory neural fields. *Physical Review Letters*, *107*(22), 228,103.

Fougnie, D., Suchow, J.W., Alvarez, G.A. (2012). Variability in the quality of visual working memory. *Nature Communications*, *3*, 1229.

Funahashi, S., Bruce, C.J., Goldman-Rakic, P.S. (1989). Mnemonic coding of visual space in the monkey9s dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*(2), 331–349.

Gardiner, C. (2009). Handbook of stochastic methods, Vol. 4, Springer, Berlin.

Gazzaley, A., & Nobre, A.C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*(2), 129–135.

Gökçe, A., Avitabile, D., Coombes, S. (2017). The dynamics of neural fields on bounded domains: an interface approach for dirichlet boundary conditions. *Journal of Mathematical Neuroscience*, *7*(1), 1–12.

Goldman-Rakic, P.S. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477–485.

Gorgoraptis, N., Catalao, R.F., Bays, P.M., Husain, M. (2011). Dynamic updating of working memory resources for visual objects. *Journal of Neuroscience*, *31*(23), 8502–8511.

Guo, Y., & Chow, C.C. (2005). Existence and stability of standing pulses in neural networks: II. Stability. *SIAM Journal on Applied Dynamical Systems*, *4*(2), 249–281.

Ikkai, A., & Curtis, C.E. (2011). Common neural mechanisms supporting spatial working memory, attention and motor intention. *Neuropsychologia*, *49*(6), 1428–1434.

Keshvari, S., Van den Berg, R., Ma, W.J. (2013). No evidence for an item limit in change detection. *PLoS Computational Biology*, *9*(2), e1002,927.

Kilpatrick, Z.P. (2013). Interareal coupling reduces encoding variability in multi-area models of spatial working memory. *Frontiers in Computational Neuroscience*, *7*(82), 1–14.

Kilpatrick, Z.P. (2016). Ghosts of bump attractors in stochastic neural fields: Bottlenecks and extinction. *Discrete Contin Dyn Syst Ser B*, *21*, 2211–2231.

Kilpatrick, Z.P. (2017). Synaptic mechanisms of interference in working memory. bioRxiv 149435.

Kilpatrick, Z.P., & Ermentrout, B. (2013). Wandering bumps in stochastic neural fields. *SIAM Journal on Applied Dynamical Systems*, *12*(1), 61–94.

Kilpatrick, Z.P., Ermentrout, B., Doiron, B. (2013). Optimizing working memory with heterogeneity of recurrent cortical excitation. *The Journal of Neuroscience*, *33*(48), 18,999–19,011.

Laing, C.R., & Troy, W.C. (2003a). Pde methods for nonlocal models. *SIAM Journal on Applied Dynamical Systems*, *2*(3), 487–516.

Laing, C.R., & Troy, W.C. (2003b). Two-bump solutions of amari-type models of neuronal pattern formation. *Physica D: Nonlinear Phenomena*, *178*(3), 190–218.

Laing, C.R., Troy, W.C., Gutkin, B., Ermentrout, G.B. (2002). Multiple bumps in a neuronal model of working memory. *SIAM Journal on Applied Mathematics*, *63*(1), 62–97.

Lara, A.H., & Wallis, J.D. (2012). Capacity and precision in an animal model of visual short-term memory. *Journal of Vision*, *12*(3), 13–13.

Lim, S., & Goldman, M.S. (2014). Balanced cortical microcircuitry for spatial working memory based on corrective feedback control. *Journal of Neuroscience*, *34*(20), 6790–6806.

Lu, Y., Sato, Y., Si, A.mari. (2011). Traveling bumps and their collisions in a two-dimensional neural field. *Neural Computation*, *23*(5), 1248–1260.

Luck, S.J., & Vogel, E.K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279.

Luck, S.J., & Vogel, E.K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400.

Ma, W.J., Husain, M., Bays, P.M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356.

Macoveanu, J., Klingberg, T., Tegnér, J. (2006). A biophysical model of multiple-item working memory: a computational and neuroimaging study. *Neuroscience*, *141*(3), 1611–1618.

Mante, V., Sussillo, D., Shenoy, K.V., Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84.

Murray, J.D., Bernacchia, A., Roy, N.A., Constantinidis, C., Romo, R., Wang, X.J. (2017). Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. In *Proceedings of the National Academy of Sciences, (Vol 114, No. 2, pp. 394–399)*.

Novikov, E.A. (1965). Functionals and the random-force method in turbulence theory. *Soviet Physics – JETP*, *20*(5), 1290–1294.

Pinto, D.J., & Ermentrout, G.B. (2001). Spatially structured activity in synaptically coupled neuronal networks: II. Lateral inhibition and standing pulses. *SIAM Journal on Applied Mathematics*, *62*(1), 226–243.

Ploner, C.J., Gaymard, B., Rivaud, S., Agid, Y., Pierrot-Deseilligny, C. (1998). Temporal limits of spatial working memory in humans. *European Journal of Neuroscience*, *10*(2), 794–797.

Rosenbaum, R., Smith, M.A., Kohn, A., Rubin, J.E., Doiron, B. (2017). The spatial structure of correlated neuronal variability. *Nature Neuroscience*, *20*(1), 107.

Schneegans, S., & Bays, P.M. (2017). Neural architecture for feature binding in visual working memory. *Journal of Neuroscience*, *37*(14), 3913–3925.

van den Berg, R., Shin, H., Chou, W.C., George, R., Ma, W.J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785.

Wei, Z., Wang, X.J., Wang, D.H. (2012). From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *Journal of Neuroscience*, *32*(33), 11,228–11,240.

White, J.M., Sparks, D.L., Stanford, T.R. (1994). Saccades to remembered target locations: an analysis of systematic and variable errors. *Vision Research*, *34*(1), 79–92.

Wilken, P., & Ma, W.J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 11–11.

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

Zhang, W., & Luck, S.J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233.

Zylberberg, J., & Strowbridge, B.W. (2017). Mechanisms of persistent activity in cortical circuits: possible neural substrates for working memory. *Annual Review of Neuroscience*, *40*, 603–627.