Robustly encoding certainty in a metastable neural circuit model

Heather L. Cihak [*] and Zachary P. Kilpatrick [†]

Department of Applied Mathematics, University of Colorado, Boulder, Colorado 80309, USA

Localized persistent neural activity can encode delayed estimates of continuous variables. Common experiments require that subjects store and report the feature value (e.g., orientation) of a particular cue (e.g., oriented bar on a screen) after a delay. Visualizing recorded activity of neurons along their feature tuning reveals activity *bumps* whose centers wander stochastically, degrading the estimate over time. Bump position therefore represents the remembered estimate. Recent work suggests bump amplitude may represent estimate certainty reflecting a probabilistic population code for a Bayesian posterior. Idealized models of this type are fragile due to the fine tuning common to constructed continuum attractors in dynamical systems. Here we propose an alternative metastable model for robustly supporting multiple bump amplitudes by extending neural circuit models to include *quantized* nonlinearities. Asymptotic projections of circuit activity produce low-dimensional evolution equations for the amplitude and position of bump solutions in response to external stimuli and noise perturbations. Analysis of reduced equations accurately characterizes phase variance and the dynamics of amplitude transitions between stable discrete values. More salient cues generate bumps of higher amplitude which wander less, consistent with experiments showing certainty correlates with more accurate memories.

I. INTRODUCTION

Working memory involves the essential ability to encode and store information for short periods of time [1]. Since estimation errors can propagate through subsequent computations [2], robust and flexible maintenance of information is key for daily tasks like decision making and planned movement [3–5]. Delayed estimates of a continuous object feature value are encoded by persistent and spatially localized neural activity across multiple brain regions [6] sustained by feature-specific excitation and lateral inhibition [1,7]. Estimate abnormalities can be indicators of neural dysfunctions arising in schizophrenia [8], autism [9], and attention deficit hyperactivity disorder [10]. Thus identifying mechanisms supporting working memory stability may guide diagnostics for predicting neuropsychopathologies [11]. Biologically aligned computational models are useful for identifying how such disorders present and may also act as a testbed for intervention [10,12,13].

We focus here on extending neural circuit models of visuospatial working memory, building on decades of successful interaction between oculomotor delayed response experiments and physiologically inspired models [5,7,14]. In the task, a subject must identify and remember the position of a briefly presented cue and then indicate the remembered location after a few seconds. Neural recordings reveal that the centroid of neural activity bumps encodes the remembered location of the cue during the delay and response [15]. Connections between pyramidal (excitatory) neurons

maintain persistent activity during the delay and interneuron (inhibitory) populations help localize activity to those with similar feature tuning as the cue [1]. Fluctuations in neural and synaptic activity cause the activity bump to wander diffusively, generating response error variance that scales roughly linearly with time [15–18].

Subjects also appear to reliably encode confidence (or certainty) in their delayed estimates [19]. Confidence reports align with accuracy, suggesting delayed estimates are represented probabilistically, possibly by the firing rate level of persistent activity in neurons encoding the estimate [20] which has been observed to increase with training and higher working memory performance [21]. Peak neural activity during retention periods has been shown to increase with strength of evidence [4,22], consistent with Bayesian computation [23–25]. Visual attention [26], stimulus presentation duration [27,28], and cue contrast [26,27] all can increase spike rates and corresponding estimates in neural circuits representing recalled sensory stimuli [29–31]. Overall, these experiments suggest increased (decreased) activity during delay periods generates higher (lower) certainty and more (less) accuracy in estimates [19,22,32].

Our models relate neural activity amplitude and response errors along these lines. Building on physiologically inspired models [7,15] and stochastic methods [33,34] linking neural circuit activity to delayed estimates, we develop a theory of activity-based encoding of confidence and its impact on response accuracy. Larger-amplitude bumps have steeper spatial profiles and wander less in response to fluctuations, better retaining estimates [35]. The theory of bump attractors must be extended to consider how bump amplitude impacts estimate storage and readout [24,36]. Most circuit models support bumps of a single amplitude, generating a bistable

---
[*]Contact author: Heather.Cihak@colorado.edu
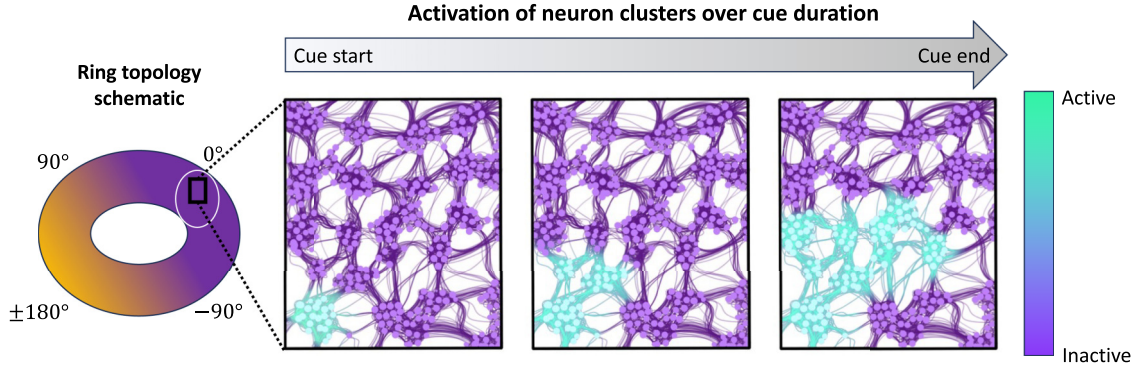[†]Contact author: zpkilpat@colorado.edu

FIG. 1. Schematic of local metastability in a ring attractor network. Consistent with recording [43,44] and modeling [45,46] studies of metastable neural circuits, stronger and/or longer stimuli successively and discretely recruit more active microclusters locally in a neural circuit. Such microclusters may emerge spontaneously in development due to interactions of self-sustained activity with neuronal migration and outgrowth [47,48]. Macroscale connectivity has ring topology akin to that inferred and observed in recordings of neural circuits encoding periodic continuum variables [14,49–51].

amplitude space in which a stable quiescent state and stable wide bump are separated by an unstable narrow bump [37,38]. Bumps are either instantiated or not but cannot encode certainty in bump amplitude since they filter out richer cue information often represented in neural recordings [29]. Here we propose and analyze a mechanism for robust encoding of certainty in activity bumps with graded amplitude values which can be reached in response to variable stimulus features.

Metastability of firing rate states is a common phenomenon in the brain, observed across multiple timescales. Single neurons can occupy multiple possible discrete firing rates states without drive [39], which could arise due to network level phenomena revealed in stochastically switching firing rate sequences [40,41]. Single-neuron models with multiple bistable dendritic compartments exhibit robust and quantized firing rate sets, providing short-term memory of transient inputs represented by quasicontinuous staircaselike firing rate functions [39]. Recent complementary work has suggested strong and sustained oscillatory input from intrinsic cell mechanisms or circuitry may work similarly [42], producing phase-locked states with graded firing rate amplitudes. Alternatively, recorded macroscale neural population activity also exhibits multiple metastable states with local attractor-like dynamics [40]. Furthermore, transitions between metastable state sequences observed in neural data are well captured by clustered population spiking models that globally exhibit discrete firing rate increases [see Fig. 1 and Refs. [40,41,43]].

In contrast to metastable neural circuit models [45,46,52], theory has also pursued finely tuned circuit models to support activity bumps with continuously graded amplitudes [24,25,36]. Fixing the gain of a piecewise linear firing rate function in spatially extended rate models generates activity bumps whose position and amplitude jointly lie on a planar continuum attractor [36]: radial location encodes amplitude and angle encodes position. However, model perturbations destroy the line attractor [53,54] and the bump amplitude wanders in response to noise. Such fragility is alleviated by breaking the symmetry of such continuum attractors, stabilizing a discrete set of attractors separated by saddles [55,56]; more aligned with the discrete and quasicontinuous firing rates sets examined in other studies [39–42].

Thus, we introduce and analyze a neural circuit model supporting metastable dynamics akin to those observed and derived in a number of prior models [39,42,44,45,57] and supported here with staircase shaped input-firing rate relationships. Metastability is conceived as arising from successive activation of neural microclusters with increasing cue salience (Fig. 1). Stable activity bump solutions have multiple graded amplitudes allowing stimulus-dependent encoding of estimate certainty (Sec. II), whose dynamics are characterized by reduced phase-amplitude equations (Sec. III). Our model is more robust to perturbations than prior models with a continuum of amplitudes [36] or an all-or-none (bistable) response [37]. Bumps subjected to fluctuating inputs retain a roughly constant amplitude for long time intervals, and their amplitude dependent wandering dynamics can be determined from reduced equations (Sec. IV).

II. MODEL EQUATIONS

Our network attractor model encodes an angle on the circle $\Delta \in [-\pi, \pi)$, a common requirement of memory and navigational tasks [14,49,51] (see Table I for parameters). Excitatory and inhibitory neural populations are collapsed to a single neural field (integrodifferential) equation organized

TABLE I. Numerical and model parameters for Eq. (1).

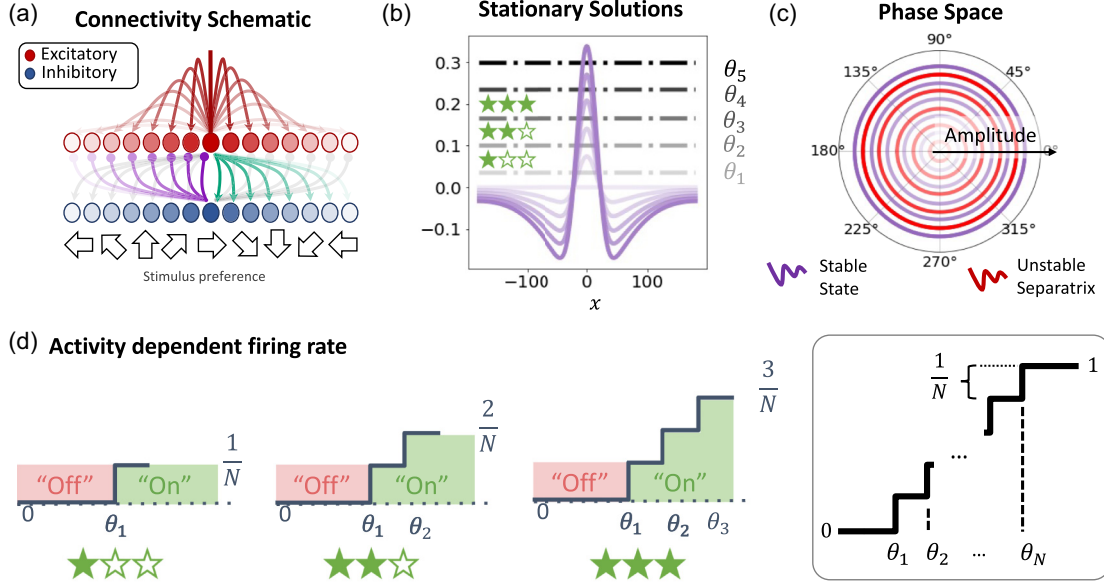| Parameter	Definition	Value
x	Domain	$[-180, 180]°$
dx	Spatial increment	$\frac{360}{n}$ where $n = 2^{12} + 1$
dt	Time step	0.025
A_e	E strength	1.5
A_i	I strength	0.5
κ_e	E synaptic footprint	20
κ_i	I synaptic footprint	1
M	Fourier modes	20
N	Firing rate steps	5
θ	Firing thresholds	[0.035,0.1,0.165,0.234,0.298]
A_c	Cue amplitude	1
a_c	Cue radius	0.02 radians or $\approx 1.15°$

FIG. 2. Model structure and core dynamics. (a) Excitatory and inhibitory network connectivity depends on difference in stimulus angle preference. (b) Broader and weaker inhibitory connectivity promotes stable and localized activity bumps which can exhibit multiple graded amplitude values due to stairstep firing rate nonlinearities [see panel (d)]. Bump activity $U(x)$ is plotted here as a function of angle x (in degrees, not radians). (c) Phase-amplitude space plots of bumps reveal concentric ring attractors separated by unstable ring repellers, stabilizing amplitude representations. (d) Increasing input successively engages higher firing rate states in the stairstep transfer function.

with a ring topology. Effective input $u(x,t)$ to local clusters at time t is indexed by angular preference along a continuum ($x \in [-\pi, \pi)$ for analysis, but sometimes converted to degrees $180x/\pi$ for plotting), and clusters with similar orientation preferences are strongly coupled by excitation while those with dissimilar preferences effectively inhibit each other [37]. Evolution of network activity is described by the spatially extended Langevin equation:

$$du(x,t) = \left[-u(x,t) + \int_{-\pi}^{\pi} w(x-y)f(u(y,t))dy \right.$$
$$\left. + I_c(x,t) \right]dt + \sqrt{\epsilon}dW(x,t). \quad (1)$$

Recurrent connectivity targeting clusters x from angular position y is described by the effective synaptic kernel, $w(x-y)$ [Fig. 2(a)], which is locally excitatory and laterally inhibitory.

A single stationary bump solution is generated when considering a Heaviside step nonlinearity $f(u) = H(u - \theta)$ with $H(u - \theta) = 1$ if $u \geqslant \theta$ and 0 otherwise [33,34,37,58,59]. To incorporate certainty we examine a sequence of metastable amplitude states; generating stationary bump solutions [Fig. 2(b)] of different amplitudes [Fig. 2(c)] when we consider staircase firing rate functions with N steps [Fig. 2(d)]

$$f(u) = \frac{1}{N} \sum_{k=1}^{N} H(u - \theta_k). \quad (2)$$

Each step on the staircase reflects successive cluster activations (similarly to Ref. [41]), prompting increased population level firing activity associated with distinct metastable states (Fig. 1). Appropriate choices of the thresholds $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_N]$ provide for N stable bump solutions [e.g.,

$N = 5$ in Fig. 2(b)]. Bumps are marginally stable to shifts and so translationally invariant [Fig. 2(c)]. Unstable solutions act as separatrices between the stable bumps. Section III provides details on the stability analysis.

Two limits of Eq. (2) are of interest from previous studies of attractor solutions to Eq. (1). First, taking $N = 1$, we recover a Heaviside nonlinearity, imposing a model with all-or-none responses, either exhibiting stable bumps or no activity, as shown by Amari [37]. This limit has been useful in analyses of the dynamics of bumps as it allows for explicit calculation of solutions, localization of linear stability calculations, and interface methods for determining nonlinear dynamics [35,37,58,60]. Alternatively, fixing $\theta_k = \theta \cdot k/N$ with $k \in \{1, 2, \ldots, N\}$ and taking $N \to \infty$ generates a piecewise linear firing rate function

$$f(u) = \begin{cases} 1, & u \geqslant \theta, \\ u/\theta, & 0 < u < \theta, \\ 0, & u \leqslant 0. \end{cases}$$

Selecting the gain $1/\theta$ fine-tunes the model [54] so it exhibits bumps with a continuum of amplitudes [36]. Other continuous forms of firing rate function $f(u)$ could be obtained in the limit $N \to \infty$ with careful choices of θ_k.

Network connectivity $w(x)$ is assumed to be shaped as the difference $w(x-y) = w_E(x-y) - w_I(x-y)$, collapsing excitation and inhibition into a single population, which can be done rigorously using a separation of timescales analysis [37,61]. Contributions from excitatory and inhibitory populations are given by von Mises distributions $w_k(x) = A_k \exp[\kappa_k[\cos(x) - 1]]$. Approximation of the effective weight function $w(x) = w_E(x) - w_I(x)$ using a finite set of even

Fourier modes allows us to write

$$w(x) = W_0 + \sum_{j=1}^{M} W_j \cos(jx),$$

where the first mode ($j = 1$) is dominant and has positive weight, due to the local excitation and lateral inhibition [1,14,15].

Cue contrast, size, and clarity is parameterized by the convolution:

$$I_c(x,t) = \frac{A_c(t)}{2} w(x) * \left[\text{erf} \frac{x + a_c}{\sigma_c} - \text{erf} \frac{x - a_c}{\sigma_c} \right], \quad (3)$$

where $A_c(t) = A_{\text{cue}} \mathbb{I}_{[t_a^c, t_\omega^c]}(t)$ describes the temporally dependent strength of the cue, $\mathbb{I}_\Omega(t) = 1$ for $t \in \Omega$ (cue on) and 0 otherwise (indicator function), and a_c is the cue halfwidth. Increasing $\sigma_c \geqslant 0$ smooths the input so that in limits $\sigma_c \to \infty$ flattens the profile and $\sigma_c \to 0^+$ yields a top hat convolved with the weight kernel.

Spatially extended Wiener process increments have zero mean, spatial correlations $\langle dW(x,t) dW(y,z) \rangle = C(x - y)\delta(t - s)dt\,ds$ with $\delta(x)$ the Dirac delta distribution, and are scaled to be weak ($0 < \epsilon \ll 1$). Spatial correlations are simulated by spatially convolving white noise increments $d\Upsilon(x,t)$ with an appropriate filter $\mathcal{F}(x)$, so that if $dW(x,t) = \mathcal{F}(x) * d\Upsilon(x,t)$, it can be shown that $C(x - y) = \int_{-\pi}^{\pi} \mathcal{F}(x - z)\mathcal{F}(y - z)dz$ [33].

Numerical simulations (see Appendix A for details) show cues of increasing salience (e.g., strength, time-length, size) generate bumps of increasing amplitude [Fig. 2(b)]. We next derive conditions for bumps, their stability, and their phase-amplitude dynamics in response to perturbations.

III. DETERMINISTIC ANALYSIS

Explicit bump solutions to Eq. (1) can be directly constructed using self-consistency. Stability is determined by an associated linearized operator. An appropriate ansatz inspired by observations from stability calculations then paved the way for low-dimensional reductions of bump dynamics to a set of evolution equations. We conclude this section by identifying how our metastable neural circuit models provide more robust representations of certainty and input angle than past models.

A. Stationary solutions

Time-independent solutions $u(x,t) = U(x)$ to Eq. (1) with $\epsilon \equiv 0$ and $I_c \equiv 0$ satisfy $U(x) = w(x) * f(U(x))$. Decomposing the weight function into M Fourier modes, leveraging trigonometric identities, and examining even solutions, we find stationary solutions take the form:

$$U(x) = \sum_{j=0}^{M} \underbrace{W_j \langle \cos(jx), f(U(x)) \rangle}_{\bar{U}_j} \cos(jx), \quad (4)$$

where $\langle p(x), q(x) \rangle = \int_{-\pi}^{\pi} p(x)q(x)dx$ is an inner product. For any firing rate function, we can form a dense, nonlinear,

implicit system for the coefficients $\bar{U}_j$ [33,38,62]

$$\bar{U}_j = W_j \left\langle \cos(jx), f\left[\sum_{j=0}^{M} \bar{U}_j \cos(jx) \right] \right\rangle. \quad (5)$$

For the staircase firing rate f, Eq. (2), we can find N thresholds ($\theta_1 < \cdots < \theta_N$) such that there are N possible bump solutions. Index bump states as $B = 1, \ldots, N$ (e.g., $B = 1$ and $B = N$ represent the lowest and highest bump-amplitude states), then there are B interfaces (or halfwidths) a_i satisfying the level set conditions, $U(\pm a_i) = \theta_i$. The profile of the Bth bump crosses B levels of the firing rate function, where $1 \leqslant B \leqslant N$, so stationary bumps satisfy

$$U(x) = \frac{2}{N} \sum_{k=1}^{B} \left[W_0 a_k + \sum_{j=1}^{M} \frac{W_j \cos(jx)}{j} \sin(ja_k) \right].$$

Utilizing the threshold-crossing conditions $\theta_i = U(a_i)$ for $i = 1, \ldots, B$ one can implicitly define the half-widths a_i from the system of equations

$$\theta_i = \frac{2}{N} \sum_{k=1}^{B} \left[W_0 a_k + \sum_{j=1}^{M} \frac{W_j \cos(ja_i)}{j} \sin(ja_k) \right] \quad (6)$$

for $i = 1, \ldots, B$. The system Eq. (6) can be numerically solved iteratively across a range of thresholds (see Appendix B). Cascades of saddle node bifurcations for each half-width and threshold pair emerge (see Fig. 9 in Appendix B). Up to $N + 1$ stable solutions (including the quiescent state, $U \equiv 0$) exist for a neural field with an N-step staircase firing rate, separated by N unstable bumps [see Fig. 3(a) for $N = 2$ example]. Alternatively, one can also utilize approximations to stationary bump solutions assuming they are parameterized by a single amplitude A which represents the scaling of the peak [Fig. 3(b)].

B. Stability

Stability of bumps can be determined by examining the linear dynamics of perturbations at the interfaces defined by the level sets $u(x,t) = \theta_i$ analogously to Refs. [33,37,58,59,61]. We study small smooth perturbations of the bump using the ansatz $u(x,t) = U(x) + \psi(x,t)$, where $||\psi|| \ll 1$. For a given bump solution state $1 \leqslant B \leqslant N$, we plug in the ansatz, Taylor expand, and truncate to first order to obtain the linearized dynamics,

$$\partial_t \psi = -\psi + \frac{1}{N} \sum_{k=1}^{B} \sum_{a=\pm a_k} \frac{\psi(a)w(x-a)}{|U'(a_k)|} \equiv \mathcal{L}\psi, \quad (7)$$

where we define the linear operator

$$\mathcal{L}u(x) = -u(x) + w(x) * [f'(U(x))u(x)]. \quad (8)$$

Note, to obtain this result, we have formally Taylor expanded the Heaviside nonlinearities that comprise $f(u)$ and whose discontinuities are shielded by integration against the perturbations $\psi(x,t)$. Formulas for the distributional derivatives are
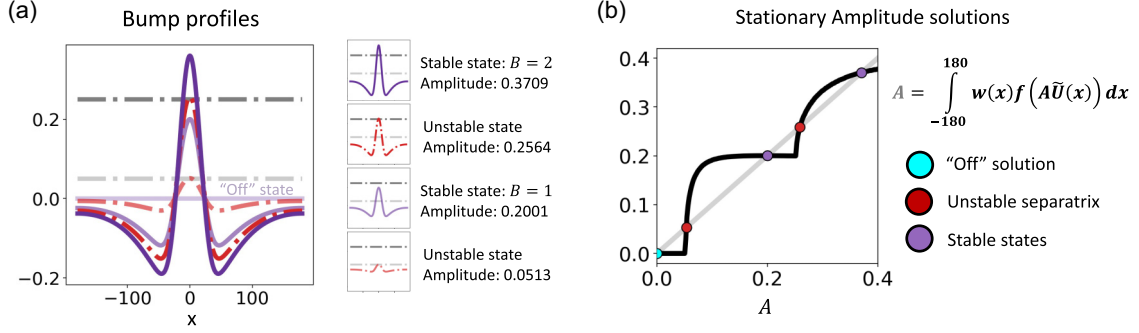
FIG. 3. Stationary bump solutions. (a) All five possible solutions $U(x)$ to Eq. (1) are plotted in degrees x for the case of a staircase firing rate Eq. (2) with $N = 2$: Stable "off" state $U \equiv 0$; stable (purple) and unstable (red) $B = 1$ bump profiles only intersect lower threshold $U(a_1) = \theta_1$ (light dash dot line); stable and unstable $B = 2$ bump profiles intersect both low $U(a_1) = \theta_1$ and high $U(a_2) = \theta_2$ (dark dash dot line) thresholds. (b) Bump solutions $U(x)$ all have roughly the same normalized profile ($\tilde{U}(x) = U(x)/U(0)$) allowing us to represent them by near-exact amplitude solutions to the implicit equation $A = \langle w(x), f(A\tilde{U}(x)) \rangle$, revealing the "off" state (blue); unstable bumps (red); and stable bumps (purple) as points along the line of amplitudes A.

obtained by noting

$$\delta(x + a_k) - \delta(x - a_k) = \frac{d}{dx}[H(x + a_k) - H(x - a_k)]$$
$$= \frac{d}{dx}H(U(x) - \theta_k)$$
$$= H'(U(x) - \theta_k)U'(x)$$

and then dividing by the odd function $U'(x)$. Summing, we then find

$$f'(U(x)) = \frac{1}{N} \sum_{k=1}^{B} \frac{\delta(x - a_k) + \delta(x + a_k)}{|U'(a_k)|}, \quad (9)$$

where we can determine

$$U'(x) = \frac{1}{N} \sum_{k=1}^{B} [w(x + a_k) - w(x - a_k)].$$

Separating solutions $\psi(x, t) = \psi(x)e^{\lambda t}$ and evaluating Eq. (7) at interfaces $x = \pm a_1, \ldots, \pm a_B$ localizes the stability problem to a discrete eigenvalue problem determined by a $2B \times 2B$ matrix. The quiescent solution $u = 0$ is stable [Fig. 2(d)], due to the pure linear decay arising when $\psi(\pm a_k) \equiv 0$ for all k in Eq. (7). For each B where bump solutions exist, we generally find two stationary solutions: a stable wide solution and an unstable narrow solution that is a separatrix between the wide solution and the state below [33,37], finding no more than $N + 1$ stable solutions and N unstable solutions arising due to amplitude quantization of the metastable N step firing rate function.

To illustrate how the stairstep firing rate function impacts the stability problem beyond the standard single step ($N = 1$) case [37], consider $N = 2$, so the linearized and localized eigenproblem becomes

$$(\lambda + 1)\psi(x) = \frac{1}{2} \sum_{k=1}^{2} \sum_{a = \pm a_k} \frac{\psi(a)w(x - a)}{|U'(a_k)|} \quad (10)$$

for $x = \pm a_1, \pm a_2$, a 4×4 system. As expected, the bump is marginally stable to shifts. Assuming $\psi(-a_k) = -\psi(a_k)$ for $k = 1, 2$, plugging in $\lambda = 0$, and enforcing self-consistency,

we obtain a single equation relating perturbations of the inner $\psi(a_2)$ and outer $\psi(a_1)$ interfaces

$$\frac{\psi(a_1)}{\psi(a_2)} = \frac{w(0) - w(2a_1) + w(\Delta a) - w(a_+)}{w(0) - w(2a_2) + w(\Delta a) - w(a_+)},$$

where $\Delta a = a_2 - a_1$ and $a_+ = a_1 + a_2$. This reflects the marginal stability due to translation invariance. We do not expect general bump stability conditions to emerge from examining these perturbations. Stability is often determined by studying width perturbations which are even symmetric $\psi(-a_k) = \psi(a_k)$ for $k = 1, 2$. This generates a 2×2 eigenproblem whose solutions imply stability given positive determinant and negative trace of the associated matrix, providing the conditions

$$2w(2a_1)w(2a_2) > (w(2a_1) + w(2a_2))(w(\Delta a) - w(a_+))$$

and

$$(w(\Delta a) - w(a_+))^2 > -2w(2a_1)w(2a_2)$$
$$+ 2(w(2a_1) + w(2a_2))(w(0) + w(\Delta a) - w(a_+)).$$

Alternative conditions can be constructed for other perturbation types, including cases where interfaces at different levels are shifted in opposite directions.

C. Reduced equation for amplitude evolution

Low-dimensional reductions of neural field dynamics on the ring $x \in [-\pi, \pi]$ can be derived using Fourier decompositions [33,38,62–65]. A complementary approach uses eigenfunctions of the linearized system to partition dynamics into a position variable for a bump (phase) and its amplitude [36,66]. Such an approach starts with the ansatz

$$u(x, t) = A(t)\tilde{U}(x - \Delta(t)) + \sqrt{\epsilon}\psi(x - \Delta(t), t), \quad (11)$$

with $\Delta = \mathcal{O}(\sqrt{\epsilon})$. Changes in amplitude can be large in the event of transitions to neighboring states by external input or accumulated noise; thus we consider $dA > \mathcal{O}(\sqrt{\epsilon})$. In general we assume to leading order that perturbations of the bump shift its amplitude $A(t)$ and/or phase $\Delta(t)$ but otherwise the bump roughly retains its shape defined as $\tilde{U}(x) = U(x)/U(0)$.
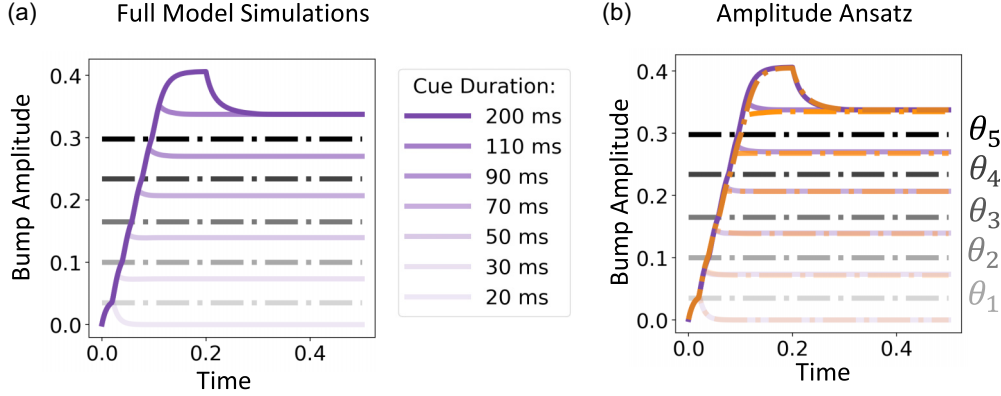
FIG. 4. Stimulus-driven bump-amplitude dynamics. (a) Increasing cue duration in Eq. (1) generates larger activity responses, which settle into larger-amplitude bumps. (b) Amplitude ansatz Eq. (11) and the amplitude evolution Eq. (15) approximate the buildup and relaxation of the bump amplitude in the full Eq. (1) well. Mild inaccuracies arise due to the assumption of fixed shape $u(x, t)/u(0, t) \approx \tilde{U}(x)$.

A low-dimensional description of input-driven bump dynamics can then be obtained by first plugging Eq. (11) into a noise-free version of Eq. (1), expanding and truncating to obtain leading-order terms:

$$A'\tilde{U}(x) - \Delta'A\tilde{U}'(x) + \psi_t - \mathcal{L}(\psi)$$
$$\approx -A\tilde{U}(x) + w(x) * f(A\tilde{U}(x)) + I_c(x + \Delta, t),$$

with the linear operator $\mathcal{L}$ as defined in Eq. (8). To obtain the evolution equation for the amplitude, A, we truncate to leading order and exploit the even symmetry of $\tilde{U}(x)$ to isolate the temporal derivative. To next order we have the following terms:

$$-\Delta'A\tilde{U}'(x) + \psi_t - \mathcal{L}(\psi) \approx I_c(x + \Delta, t). \qquad (12)$$

The eigenvalue associated with shifts ($U(x - \Delta) \approx U(x) - \Delta U'(x)$) is zero, $\mathcal{L}U' \equiv 0$. To ensure bounded solutions we require the right hand side of the equation to be orthogonal to the null space of the adjoint operator

$$\mathcal{L}^*(q(x)) = -q(x) + f'(U)(x)[w(x) * p(x)]. \qquad (13)$$

The null($\mathcal{L}^*$) is spanned by a single odd function defined for a stationary solution in state $1 \leqslant B \leqslant N$, $\phi_o(x) \equiv f'(U(x))U'(x)$, where $f'(U)$ is as defined in Eq. (9):

$$\mathcal{L}^*(f'(U)U') = -f'(U)U' + f'(U)[w * [f'(U)U']] \equiv 0,$$

since $U' = w * f(U) = w * [f'(U)U']$ from integration by parts. Thus taking the inner product of $f'(U(x))U'(x)$ with Eq. (12) yields the evolution equation for Δ. Together we have the system

$$A' = -A + G(A) + J_A(\Delta, t), \qquad (14a)$$

$$\Delta' = -\frac{1}{A}J_\Delta(\Delta, t), \qquad (14b)$$

where $||p||^2 = \langle p(x), p(x) \rangle$ is the squared norm induced by the inner product and

$$G(A) = \frac{\langle \tilde{U}(x), w(x) * f(A\tilde{U}(x)) \rangle}{||\tilde{U}(x)||^2}$$

describes the impact of recurrent connectivity on the bump amplitude, and

$$J_A(\Delta, t) = \frac{\langle \tilde{U}(x), I_c(x + \Delta, t) \rangle}{||\tilde{U}(x)||^2},$$

$$J_\Delta(\Delta, t) = \frac{\langle f'(U(x))U'(x), I_c(x + \Delta, t) \rangle}{\langle f'(U(x))U'(x), \tilde{U}'(x) \rangle},$$

describe how the even and odd parts of the cue input steer the amplitude and phase. The phase is shifted by cues that apply odd perturbations to the bump, though increasing the amplitude A of the bump decreases these shifting responses. Amplitudes relax to a stable steady state once cues are shut off, determined by the basin of attraction demarcated by θ_i where they reside [Fig. 4(a)]. Changing cue contrast, size, and clarity also alters long-term bump amplitudes (see Appendix C and Fig. 10). Assuming separability of the cue $I_c(x, t) = I_A(t)\mathcal{J}(x)$, the phase Δ in Eq. (14) will not shift, so taking $\Delta(0) = 0$ without loss of generality, we can reduce the system to

$$A' = -A + G(A) + \bar{\mathcal{J}}I_A(t), \qquad (15)$$

where $\bar{\mathcal{J}} = \langle \tilde{U}(x), \mathcal{J}(x) \rangle / ||\tilde{U}(x)||^2$. Dynamics of Eq. (15) match the buildup and relaxation to steady-state amplitudes determined from full simulations with low error [Fig. 4(b)]. We can thus use Eq. (15) to approximate the transient dynamics and stable bump profiles expected (freezing $I_A(t) \equiv \bar{I}_A$) as the input amplitude is varied.

D. Bump robustness to model perturbations

Activity states in metastable neural circuits are robust to dynamic perturbations, and also structural perturbations like changes to connectivity or firing rate relations [39]. Line attractor models which finely encode stimulus differences can be generated by considering piecewise linear firing rate relations [67], as in hand-designed neural circuit models with a continuum of bump attractor amplitudes [24,25,36]. Their low-dimensional dynamics lie on a planar attractor whose angular direction encodes stimulus estimates and radial dimension represents estimate certainty. However, even mild structural perturbations [Fig. 5(a)] destroy the carefully
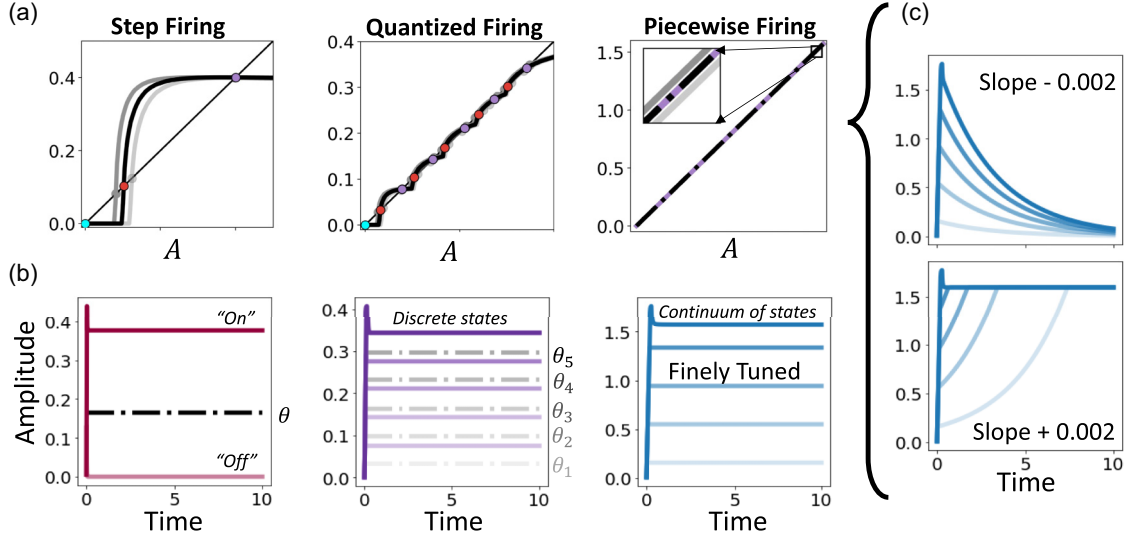
FIG. 5. Metastable neural circuit robustly encodes amplitude. (a) Memory robustness and flexibility tradeoff in circuits with quantized firing rates. Amplitudes of stationary bumps $A = \max_x U(x)$ are represented as intersections (dots) of the inner product $\langle w, f(A\tilde{U}) \rangle$ (thick line) and line of unity (thin line). Single step nonlinearities $f(u) = H(u - \theta)$ support one stable bump, while staircase nonlinearities, Eq. (2), with N steps can support N stable bumps. Equilibria persist in both models even when $f(u)$ is perturbed (gray lines). A finely tuned piecewise linear firing rate supports a continuum of bumps, but perturbations (see inset for slope perturbations) annihilate the line attractor. (b) Inputs of varying time lengths either lead to a single active bump amplitude or no activity for the Heaviside network, a discrete and graded set of bumps amplitudes for the staircase network, and a set of graded amplitudes along a continuum for the piecewise linear network [36]. The Heaviside and staircase networks maintain these solutions under model perturbations. (c) Perturbations to the piecewise linear firing rate function lead to bump collapse or the trivial quiescent solution, breaking the amplitude coding of the finely tuned system.

crafted continuum of amplitudes [Fig. 5(c)], motivating more robust representations. Commonly used single step (Heaviside) nonlinearities in $f(u)$ in Eq. (1) can only support bumps with a single amplitude or a quiescent state, depending on the duration of cues [Fig. 5(b)], but bump solutions are more robust.

Our intermediate solution balances robustness and flexibility by considering staircase firing rate functions, Eq. (2), retaining multiple stable bump-amplitude states even when structurally perturbed [Figs. 5(a) and 5(b)]. Even weak cues can generate bumps, which do not arise in the single step case. On the other hand while the piecewise linear firing rate supports a continuum of possible amplitude states for different cue durations [Fig. 5(b)], infinitesimal structural perturbations (e.g., slope or threshold changes to the firing rate function, connectivity perturbations) annihilate the line attractor [Fig. 5(c)], whereas the quantized firing rate function allows for robustness to structural perturbations while still providing an appreciable resolution of stimulus representations.

IV. STOCHASTIC DYNAMICS OF BUMP PHASE AND AMPLITUDE

Responses from tasks requiring delayed estimates of continuum quantities have been reliably modeled by bump attractor models and their low-dimensional approximations [7,15,56,68]. The phase $\Delta(t)$ (e.g., centroid or peak) of the bump encodes the estimate [15], so the phase variance $\langle (\Delta(t) - \Delta(0))^2 \rangle$ across trials models memory degradation

[33,69] and scales linearly with delay time [17,70] (see, however, Refs. [68,71] for more complex accounts of memory degradation). Strengthening cues in our model increases the salience of bumps and the estimates they encode. Noise in Eq. (1) causes bumps to wander diffusively with larger-amplitude bumps wandering less [33,35,36], and bump amplitudes can transition to neighboring values (Fig. 6). We can derive accurate estimates of the rate of these transitions, providing a new and extended theory of the degradation of delayed estimate accuracy in neural circuits. Our reduced phase-amplitude equations can also be used to estimate phase variance across all possible bump amplitudes.

A. Stochastic phase-amplitude equations

In the analogous deterministic system, we showed the ansatz Eq. (11) decomposes the effects of odd (even) perturbations into shifts (scalings) of the bump. Stochastic perturbations from the spatially extended Wiener process noise in Eq. (1) generate wandering in the phase variable $\Delta(t)$ [33], and occasional transitions in bump amplitude $A(t)$ to neighboring attractors. Plugging in the ansatz Eq. (11) and integrating against the even and odd functions $\tilde{U}(x)$ and $f'(U(x))U'(x)$, we find a coupled system of stochastic differential equations

$$dA = [-A + G(A) + J_A(\Delta, t)]dt + \sqrt{\epsilon}\,dZ_A(\Delta, t), \quad (16a)$$

$$d\Delta = -\frac{1}{A}J_\Delta(\Delta, t)dt - \frac{\sqrt{\epsilon}}{A}dZ_\Delta(\Delta, t), \quad (16b)$$

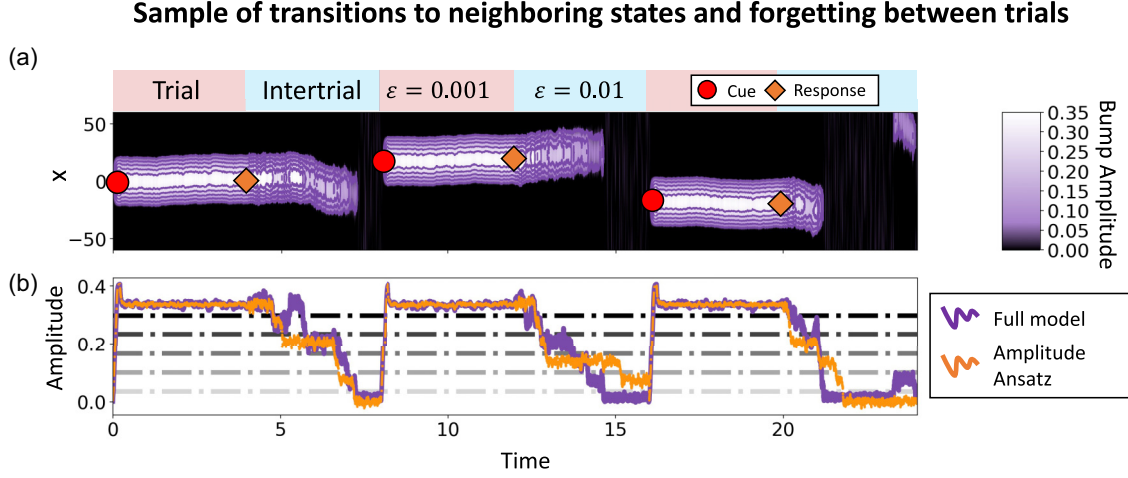### Sample of transitions to neighboring states and forgetting between trials

FIG. 6. Bump-amplitude transitions. (a) Noise ($\epsilon = 0.001$) perturbs neural activity (purple heatmap) so the bump wanders within a trial. Note that level sets (purple lines) $u = \theta_i$. The reduction of attention between trials is modeled by an increase in noise ($\epsilon = 0.01$) which rapidly drives the bump to extinction through transitions in amplitude states. (b) Comparison of amplitude dynamics identified in the full model simulation (purple) Eq. (1) and amplitude ansatz (orange) Eq. (17).

where noise increments are obtained by separating even and odd parts,

$$dZ_A(\Delta, t) = \frac{\langle \tilde{U}(x), dW(x + \Delta, t) \rangle}{||\tilde{U}(x)||^2},$$

$$dZ_\Delta(\Delta, t) = \frac{\langle f'(U(x))U'(x), dW(x + \Delta, t) \rangle}{\langle f'(U(x))U'(x), \tilde{U}'(x) \rangle}.$$

Increasing the bump amplitude dampens the impact of perturbations on the phase. Eq. (16) describes the stochastic dynamics of the bump phase and amplitude, accounting for nonequilibrium dynamics of the amplitude A (see also Ref. [66]). Amplitude dynamics in Eq. (16a) can be further approximated by a quantized chain of Markovian states assuming amplitudes remain near equilibria until fluctuations kick them to neighboring steady states of the deterministic system, which are roots of $A = G(A)$ [33,72]. The phase Δ lies on a continuum ring attractor $[-\pi, \pi)$, wandering with a diffusion coefficient determined by the bump amplitude A. We leverage our phase-amplitude system to estimate the mean time to transition between amplitude states (Fig. 6), which impacts the wandering of bump phase and the estimate retention.

B. Mean time for amplitude transitions

Defining stable ($\bar{A}_i^s$) and unstable ($\bar{A}_i^u$) bump amplitudes of the noise-free system ($A = G(A)$), we have observed (Fig. 6) that the full system tends to dwell near stable amplitudes ($\bar{A}_i^s$) on short timescales, eventually hopping to neighboring values ($\bar{A}_{i\pm1}^s$). In Eq. (16), the amplitude A must pass through unstable bump amplitudes ($\bar{A}_i^u$ or $\bar{A}_{i+1}^u$) when transitioning. Between transitions and in the absence of inputs, the translation symmetry of the spatially extended Wiener process statistics ensures Eq. (16a) behaves as a one-dimensional stochastic differential equation

$$dA = [-A + G(A)]dt + \sqrt{\epsilon}d\bar{Z}_A(t). \tag{17}$$

Fluctuation-induced transitions in amplitude A are determined by analyzing the associated Fokker-Planck equation of

Eq. (17). We can then formulate the mean exit time problem for $A(t)$ to depart the interval $[\bar{A}_i^u, \bar{A}_{i+1}^u]$ when starting at $\bar{A}_i^s$ ($i = 0, 1, \ldots, N$). On the boundaries, $\bar{A}_0^u \to -\infty$ and $\bar{A}_{N+1}^u \to \infty$. The variance and diffusion coefficient of the noise in Eq. (17) can be determined as $\langle \bar{Z}_A(t)^2 \rangle = D_A t$, where

$$D_A = \frac{\langle \tilde{U}(x), \tilde{U}(x) * C(x) \rangle}{||\tilde{U}(x)||^4}.$$

The probability density $p(A, t)$ evolves according to the Fokker-Planck equation

$$p_t = -\frac{\partial}{\partial A}[(-A + G(A))p] + \frac{D_A}{2}p_{AA}, \tag{18}$$

and $p(A, 0) = \delta(A - \bar{A}_0)$, the amplitude starts at some value $A_0 \in [\bar{A}_i^u, \bar{A}_{i+1}^u]$. We expect $A_0 = \bar{A}_i^s$, but to determine first passage time statistics, we determine quantities across the interval. Since it determines the timescale on which a stationary approximation of A in Eq. (16b) is valid, as well as the higher-order dynamics of A, we are interested in the random time $T(A_0)$ the amplitude in Eq. (17) escapes the interval $[\bar{A}_i^u, \bar{A}_{i+1}^u]$. The mean time $\mathcal{T}(A_0) = \langle T(A_0) \rangle$ is determined by leveraging the backward Fokker-Planck (FP) equation [73], describing the evolution of the probability $q \equiv p(A, t|A_0, 0)$ we find the amplitude at A at time t given it started at A_0 at $t = 0$. The state variable in the backward FP equation is the initial condition A_0 and we use the adjoint linear operator of Eq. (18) to define the flux

$$q_t = [-A_0 + G(A_0)]q_{A_0} + \frac{D_A}{2}q_{A_0A_0}, \tag{19}$$

$$= -\mathcal{J}(A, t|A_0, 0).$$

The probability we find the amplitude within $[\bar{A}_i^u, \bar{A}_{i+1}^u]$ at time t is given by integrating the density

$$\int_{\bar{A}_i^u}^{\bar{A}_{i+1}^u} p(A, t|A_0, 0)dA = \mathcal{G}(A_0, t) = P(T(A_0) > t),$$

where the last equality follows from the fact that the amplitude leaves the interval after t if it has not left by then. Integrating

the backward FP Eq. (19), we obtain a related equation for $\mathcal{G}(A_0, t)$

$$\mathcal{G}_t = [-A_0 + G(A_0)]\mathcal{G}_{A_0} + \frac{D_A}{2}\mathcal{G}_{A_0 A_0}, \qquad (20)$$

with boundary conditions $\mathcal{G}(A_0, 0) = 1$ if $A_0 \in [\bar{A}_i^u, \bar{A}_{i+1}^u]$ and 0 otherwise, while $\mathcal{G}(\bar{A}_i^u, t) = \mathcal{G}(\bar{A}_{i+1}^u, t) = 0$. The mean first passage time can then be computed

$$\mathcal{T}(A_0) = -\int_0^\infty t\mathcal{G}_t(A_0, t)dt = \int_0^\infty \mathcal{G}(A_0, t)dt.$$

A differential equation for $\mathcal{T}(A_0)$ can be derived by integrating Eq. (20) over $t \in (0, \infty)$, finding

$$[-A_0 + G(A_0)]\mathcal{T}' + \frac{D_A}{2}\mathcal{T}'' = -1, \qquad (21)$$

along with boundary conditions $\mathcal{T}(\bar{A}_i^u) = \mathcal{T}(\bar{A}_{i+1}^u) = 0$. Equation (21) can be solved by integrating to find

$$\mathcal{T}(A) = \frac{2}{D_A \Lambda(\bar{A}_i^u, \bar{A}_{i+1}^u)}\left[\Lambda(\bar{A}_i^u, A_0)\int_{A_0}^{\bar{A}_{i+1}^u}\frac{\mathcal{V}(\bar{A}_i^u, y')}{v(y')}dy' \right.$$
$$\left. -\Lambda(A_0, \bar{A}_{i+1}^u)\int_{\bar{A}_i^u}^{A_0}\frac{\mathcal{V}(\bar{A}_i^u, y')}{v(y')}\right],$$

where $\mathcal{V}(a, b) = \int_a^b v(x)dx$ and $\Lambda(a, b) = \int_a^b \frac{dx}{v(x)}$ and

$$v(A_0) = \exp\left[\frac{2}{D_A}(V(\bar{A}_i^u) - V(A_0))\right]. \qquad (22)$$

Stochastic amplitude dynamics are strongly determined by the potential function [Fig. 7(a)], formed by integrating $V(A) = \int_{-\infty}^A [A' - G(A')]dA'$, which biases transitions to lower-amplitude states over time. The energy barrier the stochastic particle must surmount is lower on the left side [Fig. 7(a)], so increases in neural variability following task-relevant epochs (when variability is lower, due perhaps to attention [74,75]) could serve to annihilate persistent activity [see Fig. 6 and Refs. [33,66]].

To approximate the rate of transition over either barrier, we assume that the mean time of escape over either boundary will be roughly the same $\mathcal{T}^{\pm}(\bar{A}_i^s) \approx \mathcal{T}(\bar{A}_i^s)$ as is often the case even with asymmetric potentials [72,73]. We then approximate the rate of transition over either barrier $r_i^{\pm} \approx \pi_i^{\pm}(\bar{A}_i^s)/\mathcal{T}(\bar{A}_i^s)$ as the escape probability scaled by the mean time. Exit probabilities $\pi_i^{\pm}(\bar{A}_i^s)$ are determined by deriving the appropriate differential equation. First, integrate the probability current through the boundary of interest $\mathcal{J}(\bar{A}_{i+1}^u, t|A_0, 0)$ or $-\mathcal{J}(\bar{A}_i^u, t|A_0, 0)$. For instance, the probability the particle exits via $A = \bar{A}_{i+1}^u$ after time with t is

$$g^+(A_0, t) \equiv \int_t^\infty \mathcal{J}(\bar{A}_{i+1}^u, t'|A_0, 0)dt'$$
$$= \int_t^\infty \left[(A_0 - G(A_0))q - \frac{D_A}{2}q_{A_0}\right]dt'.$$

Using the fact that $q = p(\bar{A}_{i+1}^u, t|A_0, 0)$ satisfies Eq. (19), we find that $g_+(A_0, t)$ satisfies

$$g_t^+ = (-A_0 + G(A_0))g_{A_0}^+ + \frac{D_A}{2}\partial_{A_0}^2 g_{A_0 A_0}^+.$$

(a) **Mean Exit Time (Escape Potential Well)**

Potential landscape   A(t)   Separatrices

(b) **Transition rates between states ($\varepsilon = 0.01$)**

High rate   Low rate

0.002   0.0056   0.1048   0.0818   0.0495

0 1 2 3 4 5

1.4391   1.5258   1.0979   3.9940   3.1061

(c) **Mean Transition Time (seconds)**

Low Amplitude   Middle Amplitude   High Amplitude
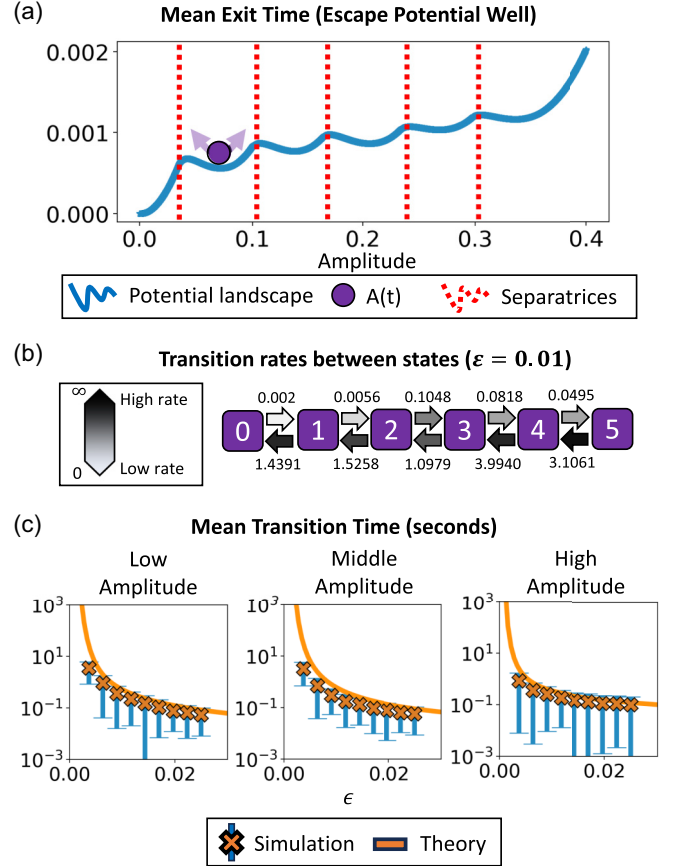
ϵ

Simulation   Theory

FIG. 7. Amplitude potential landscape and transition dynamics. (a) Amplitude potential well landscape $V(A)$ (blue line) determines drift via the descent of its gradient $-V'(A) = -A + G(A)$. Potential peaks (red dashed) separate stable minima. Stochastic fluctuations drive amplitude (purple particle) to escape minima, usually *downhill* towards the off state ($A \equiv 0$). (b) Markov chain approximation of well-hopping dynamics. Transition rates from a state $\bar{A}_i^s$ to its neighbor $\bar{A}_{i\pm1}^s$ are approximated $r_i^{\pm} \approx \pi_i^{\pm}(\bar{A}_i^s)/\mathcal{T}(\bar{A}_i^s)$ by the ratio of the escape probability and mean first passage time. Stable bump amplitudes are enumerated from 0 (the off state) and 5 (the highest-amplitude state). Noise $\epsilon = 0.01$. (c) Mean transition times were estimated by averaging over 1000 simulations (Mean: orange "x"; blue lines: standard deviation) comparing well with theory (orange line). See Appendix B for simulation details.

Taking $t \to 0^+$ and defining $\pi_i^+(A_0) := g^+(A_0, 0)$, we see that $\mathcal{J}(\bar{A}_{i+1}^u, 0|A_0, 0)$ vanishes if $A_0 \neq \bar{A}_{i+1}^u$, since $p(\bar{A}_{i+1}^u, 0|A_0, 0) = \delta(A_0 - \bar{A}_{i+1}^u)$, so $g_t^+(A_0, 0) \to 0$ and

$$(-A_0 + G(A_0))\partial_{A_0}\pi_i^+(A_0) + \frac{D_A}{2}\partial_{A_0}^2\pi_i^+(A_0) = 0,$$

where $\pi_i^+(\bar{A}_{i+1}^u) = 1$, $\pi_i^+(\bar{A}_i^u) = 0$, and $\pi_i^+(A_0) + \pi_i^-(A_0) = 1$. We solve and $\pi_i^+(A_0) = \mathcal{N}(A_0)/\mathcal{N}(\bar{A}_{i+1}^u)$ and $\pi_i^-(A_0) = 1 - \pi_i^+(A_0)$, where $\mathcal{N}(A) = \int_{\bar{A}_i^u}^A \frac{dy}{v(y)}$. The escape probability and exit rate associated with the left boundary of each well is larger than for the right boundary [Fig. 7(b)], so $A(t)$ will tend towards 0, and all bumps are eventually extinguished given a long delay time, as suggested by behavior [76].

Mean transition time estimates align well with full system simulations [Fig. 7(c)], making two key predictions. First,
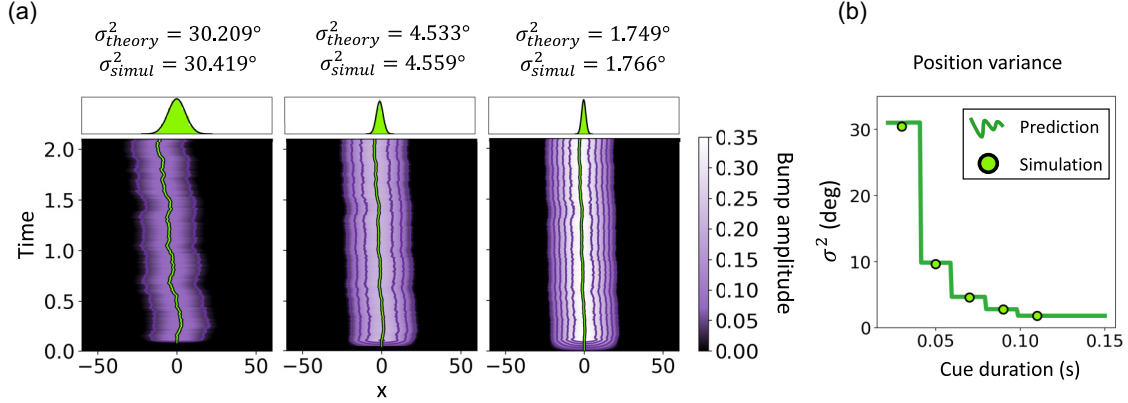
FIG. 8. Increased cue duration decreases wandering. (a) Top: Estimate distributions computed from bump peak after delay for cue durations $t_c = 30, 70, 110$ ms. Bottom: Wandering bumps $u(x, t)$ of low-, medium-, and high-amplitude following cues of different lengths. (b) Response variance decreases with cue duration due to increasing bump amplitude. Noise $\epsilon = 0.01$. See Table I for other parameters and Appendix A for numerical methods.

higher-amplitude bumps transition to neighboring amplitudes (usually lower) quicker than low-amplitude bumps. This does not negate our overall claim that high-amplitude bumps are more robust, since transitions from high amplitudes still generate medium to high-amplitude bumps. Second, transition frequency increases with the fluctuation strength. Asymptotic errors in approximating the potential exponentially impact passage time estimates, which is obvious at low noise levels.

C. Phase variance estimates

We now study how bump amplitude shapes the wandering of the bump phase $\Delta(t)$. As has been found previously, higher-amplitude bumps wander less [33,36]. Since the phase encodes the remembered stimulus value Δ_0, the variance $\langle(\Delta(t) - \Delta_0)^2\rangle$ measures recall error. Variance is determined by analyzing the reduced and forced equation for the phase, Eq. (16b), which we write out in terms of integrals without external inputs

$$d\Delta = -\frac{\sqrt{\epsilon}}{A} \frac{\langle f'(U(x))U'(x), dW(x + \Delta, t)\rangle}{\langle f'(U(x))U'(x), \tilde{U}'(x)\rangle}. \quad (23)$$

Fixing the amplitude $A \approx \bar{A} \in \bar{A}_{0:N}^s$ in Eq. (23) in the case of rare transitions due to weak noise and/or short delays, we compute variance $\langle(\Delta - \Delta_0)^2\rangle = D(\epsilon)t$, where

$$D(\epsilon) = \frac{\epsilon}{\bar{A}^2} \frac{\langle f'(U(x))U'(x), f'(U(x))U'(x) * C(x)\rangle}{[\langle f'(U(x))U'(x), \tilde{U}'(x)\rangle]^2}.$$

Larger amplitudes $\bar{A}$ reduce the variance (Fig. 8), well predicted by our theory, as in findings showing higher certainty reduces response errors [19,22]. Increased neural responses (bumps of higher amplitude) occur in response to longer, brighter, clearer, and larger cues generating more accurate responses (i.e., there is less wandering).

V. DISCUSSION

Metastability is a powerful mechanism for supporting robust representation of information in neural circuits [43–45]. Recent works have begun to explore possible roles of multistability in neuron firing arising from cell mechanisms such

as nonlinear denditric compartmentalization [39] and intrinsic subthreshold oscillation bands [42] which link steplike firing rate sets of neurons and graded amplitudes of neuron activity. On the macroscale, we have proposed a neural circuit model inspired by microclustered architecture [45,46] of such multistable neurons with population firing dependent on cluster activation [41] which sustains neural population activity bumps with multiple amplitudes. Rather than employing a fragile model with a fine-tuned transfer function [24,25,36], we considered a quantized firing rate function generating a robust model reminiscent of metastable single-neuron models with bistable dendritic compartments [39]. Our neural circuit model's dynamics can be reduced to evolution equations that clearly account for how stochasticity and perturbations impact delay encoding and confidence. Our analysis provides a simple and understandable theory for increased accuracy of delayed estimates made from more salient cues [22,28,29].

Our neural field model can support up to N pairs of bumps, each pair including one stable and one unstable, when its stairstep firing rate function possesses N steps. Active solutions of the neural field asymptotically relax to similarly shaped bump profiles, strongly suggesting an ansatz for low-dimensionalizing system dynamics. Both external inputs and noise can drive neural activity bumps between neighboring amplitude values, as described by our reduced system. Our reduced equations not only accurately predict the wandering of bumps in response to noise but also the timing and preference of amplitude transitions. We find bumps formed from more salient cues are more resilient to fluctuations and generate more accurate response estimates.

Our analysis could be extended in several ways. Short-term plasticity can further stabilize bumps during delay periods [77], effects that could be analyzed using an interface based analysis [59]. We could also consider models with separate excitatory and inhibitory populations and develop theory separately tracking each bump's phase and amplitude dynamics [34]. Our phase-amplitude ansatz makes near-equilibrium assumptions about the shape of the bump, but perturbations may warp the bump profile in ways not well described by multiplicative scalings. Consideration of such additive changes to neural population responses could more fully characterize the
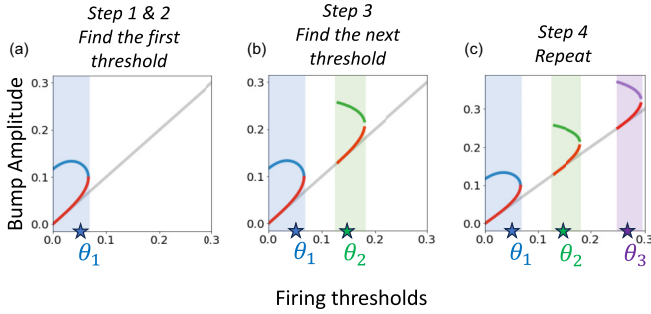
FIG. 9. Iteratively identifying bumps and stairstep firing thresholds. We briefly illustrate the sequential procedure of determining stable or unstable bump branches and stairstep firing thresholds. (a) Select θ_1 and find the $B = 1$ set of bump solution branches (blue branch is stable); (b) After choosing θ_1 find the second set of branches (green or red lines in green region). (c) Iterate for thresholds $\theta_{3:N}$. Gaps between solution regions arise since $B = k + 1$ bumps must all have higher amplitudes than $B = k$ bumps (lowest point on neighboring red branch sits above blue branch).

continuum of spontaneous modulations to neural tuning [78]. Accounting for dynamic perturbations of bump profiles may also improve the accuracy of our amplitude transition theory. Also, our analysis is limited to firing rate functions with a few steps, $N = \mathcal{O}(1)$, but could be extended to examine the limit of many steps ($N \gg 1$) with frequent transitions between or the impact of more irregular step spacings.

Quantized representation of inputs and even behavior variables is supported by a number of other computational and experimental studies. Clustered spiking networks can also generate staircase firing rate relations [41,46], suggesting that a mean-field analysis of such models could be a fruitful direction of future study. Complementarily, modular cortical networks exhibiting clustering at larger scales have also demonstrated improved working memory performance [6,79–81], further supporting the hypothesis that architectures that engender metastability could underlie more robust delay encoding. More recently, an analysis of neural activity in hypothalamus has revealed that general behavioral states like aggression may be encoded along a discrete and approximate line attractor [82]. It is important to note that the destruction of a line attractor may not be catastrophic for coding delayed estimates or other states as long as the resulting evolution of the dynamics is slow compared to the needed encoding time [83].

Our neuromechanistic model provides several links between circuit features and behavioral response trends, providing a testbed for physiological theories of increased errors and impaired processing for continuum estimates in schizophrenia [8,28], autism spectrum disorders [9], or attention deficit hyperactivity disorder [84]. Early detection of such abnormalities using noninvasive psychophysics could speed diagnoses and the implementations of behavioral interventions to help manage executive function in neurodivergent populations [85,86]. Models that can connect aberrant response statistics to underlying neurophysiology require carefully balancing mathematical tractabilty and the inclusion

of hitherto unexplored features of the underlying biological circuits.

Python code for simulating and analyzing our neural field models are available [87].

APPENDIX A: MODEL EQUATION SIMULATIONS

Convolutions and spatially filtered noise were computed using fast Fourier transforms. Euler-Maruyama is used to time step Eq. (1) with initial conditions and inputs centering bumps at $x = 0$. Numerical quadrature is performed using Riemann sums. Table I gives simulation parameters unless otherwise indicated in figure captions.

Amplitude transition times are found by (1) initializing simulations starting with a bump having an amplitude corresponding to a stable stationary bump, (2) running a stochastic simulation until the estimated amplitude crosses through a neighboring unstable bump value (or until a maximum time is reached), and (3) recording the time of transition for 1000 trials with transitions detected within the delay or terminating when 25 successive or 100 cumulative trials have failed to transition, which we take to indicate that the mean transition time is too close to or far beyond the cutoff time for our parameterized method to make an accurate estimate. Bump amplitudes are estimated as the peak activity value $\max_x u(x, t)$.

APPENDIX B: ITERATIVE CONSTRUCTION OF BUMPS

Bumps are constructed by identifying threshold intervals in which solutions of successively higher amplitude exist. Starting by solving the threshold condition Eq. (6) at the first level ($i = 1$) and constraining $\theta_1 < U(0)$, we can find θ_1 values that admit stable or unstable branches of $B = 1$ bumps. Then, finding the peak of the maximum $U(0)$ for the stable $B = 1$ bumps, we constrain an interval of possible θ_2 values and use the Fourier coefficient equations Eq. (5) to determine the next larger family of bumps of sufficiently high amplitude so that $\theta_2 < U(0)$. For a satisfactory θ_2, we can continue branches of stable or unstable bumps. This process is then repeated by choosing an appropriate $\theta_{k+1} > U(0)$ for all $B = k$ bumps calculated from the Fourier decomposition given by Eq. (5), using Eq. (5) to compute the next branches such that $U(0) > \theta_{k+1}$ until $k + 1 = N$. See Fig. 9 for an illustration.

APPENDIX C: AMPLITUDE DEPENDENCE ON CUE CHARACTERISTICS

Here we qualitatively compare bump amplitudes to other cue properties. Experiments show longer encoding periods lead to higher accuracy and increased neural responses [27,28], consistent with longer cue durations generating bumps of higher amplitude [Fig. 4(a)]. Increasing cue contrast (stimulus amplitude) also increases neural responses [see Refs. [26,27] and Fig. 10(a)]. Larger cues can also elicit higher neural responses [see Ref. [88] and Fig. 10(b)]. Cue blurriness can also impact detection and encoding [see Refs. [30,89] and Fig. 10(c)].
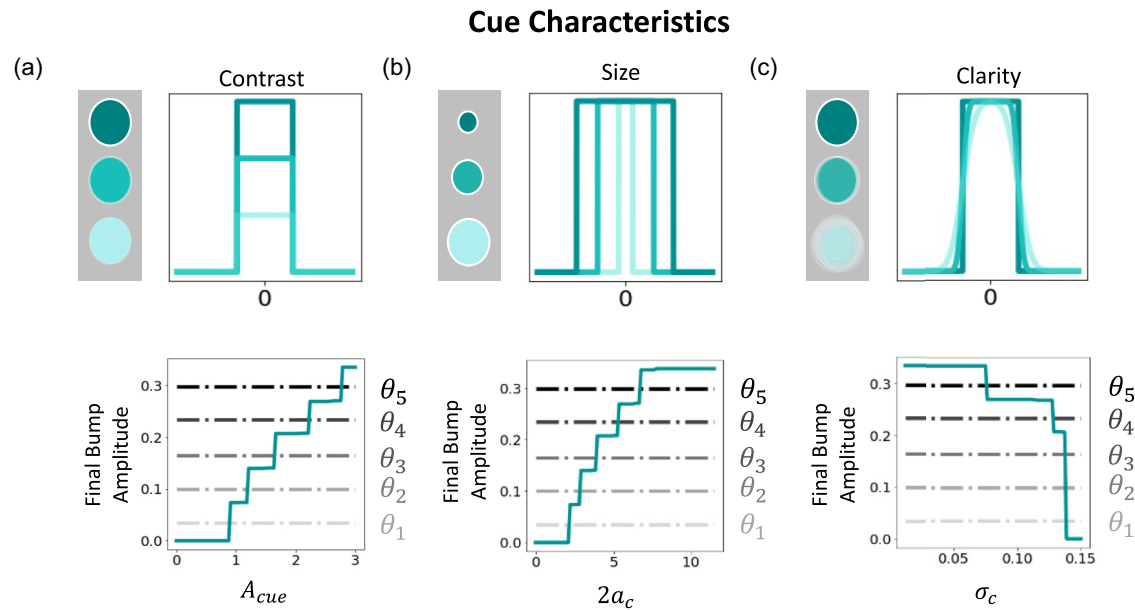
Cue Characteristics

FIG. 10. Cue characteristics and certainty. Cue profile ($A_c(t)[\text{erf} \frac{x+a_c}{\sigma_c} - \text{erf} \frac{x-a_c}{\sigma_c}]$ portion of $I_c(x,t)$) features are varied and resulting bump amplitude determined. (a) Contrast controls cue strength A_{cue} (bottom panel), whose increase generates higher-amplitude bumps. (b) Wider cues generated by increasing the diameter $2a_c$ increase bump amplitude. (c) Cue clarity (sharpness) is reduced by decreasing σ_c in Eq. (3), decreasing encoded bump amplitude.

[1] P. S. Goldman-Rakic, Cellular basis of working memory, Neuron **14**, 477 (1995).

[2] K. P. Raghubar, M. A. Barnes, and S. A. Hecht, Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches, Learn. Indiv. Differ. **20**, 110 (2010).

[3] B. K. Hulse and V. Jayaraman, Mechanisms underlying the neural computation of head direction, Annu. Rev. Neurosci. **43**, 31 (2020).

[4] R. Romo, C. D. Brody, A. Hernández, and L. Lemus, Neuronal correlates of parametric working memory in the prefrontal cortex, Nature (London) **399**, 470 (1999).

[5] S. Funahashi, C. Bruce, and P. Goldman-Rakic, Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex, J. Neurophysiol. **61**, 331 (1989).

[6] C. E. Curtis, Prefrontal and parietal contributions to spatial working memory, Neuroscience **139**, 173 (2006).

[7] C. Constantinidis and T. Klingberg, The neuroscience of working memory capacity and training, Nat. Rev. Neurosci. **17**, 438 (2016).

[8] H. Stein, J. Barbosa, and A. Compte, Towards biologically constrained attractor models of schizophrenia, Curr. Opin. Neurobiol. **70**, 171 (2021).

[9] S. Vattikuti and C. C. Chow, A computational model for cerebral cortical dysfunction in autism spectrum disorders, Biol. Psychiatr. **67**, 672 (2010).

[10] T. Klingberg, H. Forssberg, and H. Westerberg, Increased brain activity in frontal and parietal cortex underlies the development of visuospatial working memory capacity during childhood, J. Cogn. Neurosci. **14**, 1 (2002).

[11] P. R. Montague, R. J. Dolan, K. J. Friston, and P. Dayan, Computational psychiatry, Trends Cogn. Sci. **16**, 72 (2012).

[12] X.-J. Wang and J. H. Krystal, Computational psychiatry, Neuron **84**, 638 (2014).

[13] T. L. Hodgson, G. Ezard, and F. Hermens, Eye movements in neuropsychological tasks, in *Current Topics in Behavioral Neurosciences, Vol. 41*, edited by T. Hodgson (Springer, Cham, 2019), pp. 393–418.

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

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

[16] J. M. Fuster, Unit activity in prefrontal cortex during delayed-response performance: Neuronal correlates of transient memory, J. Neurophysiol. **36**, 61 (1973).

[17] J. M. White, D. L. Sparks, and T. R. Stanford, Saccades to remembered target locations: An analysis of systematic and variable errors, Vision Res. **34**, 79 (1994).

[18] A. A. Faisal, L. P. Selen, and D. M. Wolpert, Noise in the nervous system, Nat. Rev. Neurosci. **9**, 292 (2008).

[19] M. Honig, W. J. Ma, and D. Fougnie, Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions, Proc. Natl. Acad. Sci. USA **117**, 8391 (2020).

[20] P. M. Bays, S. Schneegans, W. J. Ma, and T. F. Brady, Representation and computation in visual working memory, Nat. Hum. Behav. **8**, 1016 (2024).

[21] H. Tang, X.-L. Qi, M. R. Riley, and C. Constantinidis, Working memory capacity is enhanced by distributed prefrontal activation and invariant temporal dynamics, Proc. Natl. Acad. Sci. USA **116**, 7095 (2019).

[22] M. A. Basso and R. H. Wurtz, Modulation of neuronal activity by target uncertainty, Nature (London) **389**, 66 (1997).

[23] J. M. Beck, W. J. Ma, R. Kiani, T. Hanks, A. K. Churchland, J. Roitman, M. N. Shadlen, P. E. Latham, and A. Pouget, Probabilistic population codes for Bayesian decision making, Neuron **60**, 1142 (2008).

[24] A. Kutschireiter, M. A. Basnak, R. I. Wilson, and J. Drugowitsch, Bayesian inference in ring attractor networks, Proc. Natl. Acad. Sci. USA **120**, e2210622120 (2023).

[25] J. M. Esnaola-Acebes, A. Roxin, and K. Wimmer, Flexible integration of continuous sensory evidence in perceptual estimation tasks, Proc. Natl. Acad. Sci. USA **119**, e2214441119 (2022).

[26] C. J. McAdams and J. H. Maunsell, Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4, J. Neurosci. **19**, 431 (1999).

[27] I. I. Groen, G. Piantoni, S. Montenegro, A. Flinker, S. Devore, O. Devinsky, W. Doyle, P. Dugan, D. Friedman, N. F. Ramsey *et al.*, Temporal dynamics of neural responses in human visual cortex, J. Neurosci. **42**, 7562 (2022).

[28] J. C. Badcock, D. R. Badcock, C. Read, and A. Jablensky, Examining encoding imprecision in spatial working memory in schizophrenia, Schizophr. Res. **100**, 144 (2008).

[29] J. R. Cohen, K. K. Sreenivasan, and M. D'Esposito, Correspondence between stimulus encoding-and maintenance-related neural processes underlies successful working memory, Cereb. Cortex **24**, 593 (2014).

[30] C. McGettigan, A. Faulkner, I. Altarelli, J. Obleser, H. Baverstock, and S. K. Scott, Speech comprehension aided by multiple modalities: Behavioural and neural interactions, Neuropsychologia **50**, 762 (2012).

[31] F. Cazettes, B. J. Fischer, and J. L. Pena, Cue reliability represented in the shape of tuning curves in the owl's sound localization system, J. Neurosci. **36**, 2101 (2016).

[32] T. Meyer, X.-L. Qi, T. R. Stanford, and C. Constantinidis, Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks, J. Neurosci. **31**, 6266 (2011).

[33] Z. P. Kilpatrick and B. Ermentrout, Wandering bumps in stochastic neural fields, SIAM J. Appl. Dyn. Syst. **12**, 61 (2013).

[34] H. L. Cihak, T. L. Eissa, and Z. P. Kilpatrick, Distinct excitatory and inhibitory bump wandering in a stochastic neural field, SIAM J. Appl. Dynam. Syst. **21**, 2579 (2022).

[35] N. Krishnan, D. B. Poll, and Z. P. Kilpatrick, Synaptic efficacy shapes resource limitations in working memory, J. Comput. Neurosci. **44**, 273 (2018).

[36] S. Carroll, K. Josić, and Z. P. Kilpatrick, Encoding certainty in bump attractors, J. Comput. Neurosci. **37**, 29 (2014).

[37] S.-i. Amari, Dynamics of pattern formation in lateral-inhibition type neural fields, Biol. Cybern. **27**, 77 (1977).

[38] P. C. Bressloff and S. R. Carroll, Stochastic neural fields as gradient dynamical systems, Phys. Rev. E **100**, 012402 (2019).

[39] M. S. Goldman, J. H. Levine, G. Major, and D. W. Tank, Robust persistent neural activity in a model integrator with multiple hysteretic dendrites per neuron, Cereb. Cortex **13**, 1185 (2003).

[40] T. Sasaki, N. Matsuki, and Y. Ikegaya, Metastability of active CA3 networks, J. Neurosci. **27**, 517 (2007).

[41] L. Mazzucato, A. Fontanini, and G. La Camera, Dynamics of multistable states during ongoing and evoked cortical activity, J. Neurosci. **35**, 8214 (2015).

[42] K. P. Champion, O. Gozel, B. S. Lankow, G. B. Ermentrout, and M. S. Goldman, An oscillatory mechanism for multi-level storage in short-term memory, Commun. Biol. **6**, 829 (2023).

[43] G. La Camera, A. Fontanini, and L. Mazzucato, Cortical computations via metastable activity, Curr. Opin. Neurobiol. **58**, 37 (2019).

[44] B. A. Brinkman, H. Yan, A. Maffei, I. M. Park, A. Fontanini, J. Wang, and G. La Camera, Metastable dynamics of neural circuits and networks, Appl. Phys. Rev. **9**, 011313 (2022).

[45] P. C. Bressloff, Metastable states and quasicycles in a stochastic Wilson-Cowan model of neuronal population dynamics, Phys. Rev. E **82**, 051903 (2010).

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

[47] S. Okujeni and U. Egert, Self-organization of modular network architecture by activity-dependent neuronal migration and outgrowth, eLife **8**, e47996 (2019).

[48] L. Beaulieu-Laroche, E. H. Toloza, M.-S. Van der Goes, M. Lafourcade, D. Barnagian, Z. M. Williams, E. N. Eskandar, M. P. Frosch, S. S. Cash, and M. T. Harnett, Enhanced dendritic compartmentalization in human cortical neurons, Cell **175**, 643 (2018).

[49] K. Zhang, Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory, J. Neurosci. **16**, 2112 (1996).

[50] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser, Microstructure of a spatial map in the entorhinal cortex, Nature (London) **436**, 801 (2005).

[51] S. S. Kim, H. Rouault, S. Druckmann, and V. Jayaraman, Ring attractor dynamics in the *Drosophila* central brain, Science **356**, 849 (2017).

[52] A. Thompson, A. Litwin-Kumar, and B. Doiron, Relating spontaneous dynamics and stimulus coding in competitive networks (unpublished).

[53] H. S. Seung, How the brain keeps the eyes still, Proc. Natl. Acad. Sci. USA **93**, 13339 (1996).

[54] S. Lim and M. S. Goldman, Balanced cortical microcircuitry for maintaining information in working memory, Nat. Neurosci. **16**, 1306 (2013).

[55] C. D. Brody, R. Romo, and A. Kepecs, Basic mechanisms for graded persistent activity: Discrete attractors, continuous attractors, and dynamic representations, Curr. Opin. Neurobiol. **13**, 204 (2003).

[56] Z. P. Kilpatrick, B. Ermentrout, and B. Doiron, Optimizing working memory with heterogeneity of recurrent cortical excitation, J. Neurosci. **33**, 18999 (2013).

[57] M. T. Schaub, Y. N. Billeh, C. A. Anastassiou, C. Koch, and M. Barahona, Emergence of slow-switching assemblies in structured neuronal networks, PLoS Comput. Biol. **11**, e1004196 (2015).

[58] S. Coombes and M. R. Owen, Evans functions for integral neural field equations with Heaviside firing rate function, SIAM J. Appl. Dynam. Syst. **3**, 574 (2004).

[59] H. L. Cihak and Z. P. Kilpatrick, Multiscale motion and deformation of bumps in stochastic neural fields with dynamic connectivity, Multiscale Model. Simul. **22**, 178 (2024).

[60] S. Coombes, H. Schmidt, and I. Bojak, Interface dynamics in planar neural field models, J. Math. Neurosci. **2**, 9 (2012).

[61] D. J. Pinto and G. B. Ermentrout, Spatially structured activity in synaptically coupled neuronal networks: II. lateral inhibition and standing pulses, SIAM J. Appl. Math. **62**, 226 (2001).

[62] R. Veltz and O. Faugeras, Local/global analysis of the stationary solutions of some neural field equations, SIAM J. Appl. Dynam. Syst. **9**, 954 (2010).

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

[64] B. Ermentrout, Neural networks as spatio-temporal pattern-forming systems, Rep. Prog. Phys. **61**, 353 (1998).

[65] C. R. Laing and C. C. Chow, Stationary bumps in networks of spiking neurons, Neural Comput. **13**, 1473 (2001).

[66] Z. P. Kilpatrick, Ghosts of bump attractors in stochastic neural fields: Bottlenecks and extinction, Discr. Contin. Dynam. Syst. B **21**, 2211 (2016).

[67] C. K. Machens, R. Romo, and C. D. Brody, Flexible control of mutual inhibition: A neural model of two-interval discrimination, Science **307**, 1121 (2005).

[68] M. F. Panichello, B. DePasquale, J. W. Pillow, and T. J. Buschman, Error-correcting dynamics in visual working memory, Nat. Commun. **10**, 3366 (2019).

[69] Y. Burak and I. R. Fiete, Fundamental limits on persistent activity in networks of noisy neurons, Proc. Natl. Acad. Sci. USA **109**, 17645 (2012).

[70] C. J. Ploner, B. Gaymard, S. Rivaud, Y. Agid, and C. Pierrot-Deseilligny, Temporal limits of spatial working memory in humans, Eur. J. Neurosci. **10**, 794 (1998).

[71] D. P. Bliss, J. J. Sun, and M. D'Esposito, Serial dependence is absent at the time of perception but increases in visual working memory, Sci. Rep. **7**, 14739 (2017).

[72] B. Lindner, M. Kostur, and L. Schimansky-Geier, Optimal diffusive transport in a tilted periodic potential, Fluct. Noise Lett. **01**, R25 (2001).

[73] C. W. Gardiner, *Handbook of Stochastic Methods: For Physics, Chemistry, and the Natural Sciences* (Springer, Berlin, 2009).

[74] J. F. Mitchell, K. A. Sundberg, and J. H. Reynolds, Differential attention-dependent response modulation across cell classes in macaque visual area V4, Neuron **55**, 131 (2007).

[75] M. R. Cohen and J. H. Maunsell, Attention improves performance primarily by reducing interneuronal correlations, Nat. Neurosci. **12**, 1594 (2009).

[76] H. Shin, Q. Zou, and W. J. Ma, The effects of delay duration on visual working memory for orientation, J. Vision **17**, 10 (2017).

[77] Z. P. Kilpatrick, Synaptic mechanisms of interference in working memory, Nature Sci. Rep. **8**, 7879 (2018).

[78] I. Arandia-Romero, S. Tanabe, J. Drugowitsch, A. Kohn, and R. Moreno-Bote, Multiplicative and additive modulation of neuronal tuning with population activity affects encoded information, Neuron **89**, 1305 (2016).

[79] C. Constantinidis and X.-J. Wang, A neural circuit basis for spatial working memory, Neuroscientist **10**, 553 (2004).

[80] Z. P. Kilpatrick, Delay stabilizes stochastic motion of bumps in layered neural fields, Physica D **295-296**, 30 (2015).

[81] Z. P. Kilpatrick, Interareal coupling reduces encoding variability in multi-area models of spatial working memory, Front. Comput. Neurosci. **7**, 82 (2013).

[82] A. Nair, T. Karigo, B. Yang, S. Ganguli, M. J. Schnitzer, S. W. Linderman, D. J. Anderson, and A. Kennedy, An approximate line attractor in the hypothalamus encodes an aggressive state, Cell **186**, 178 (2023).

[83] D. Sussillo and O. Barak, Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks, Neural Comput. **25**, 626 (2013).

[84] F. X. Castellanos, F. F. Marvasti, J. L. Ducharme, J. M. Walter, M. E. Israel, A. Krain, C. Pavlovsky, and D. W. Hommer, Executive function oculomotor tasks in girls with ADHD, J. Am. Acad. Child Adolesc. Psychiatr. **39**, 644 (2000).

[85] T. Klingberg, Training and plasticity of working memory, Trends Cognit. Sci. **14**, 317 (2010).

[86] B. Kluwe-Schiavon, B. Sanvicente-Vieira, C. H. Kristensen, and R. Grassi-Oliveira, Executive functions rehabilitation for schizophrenia: A critical systematic review, J. Psychiatr. Res. **47**, 91 (2013).

[87] H. L. Cihak, Robust certainty encoding, https://github.com/cihakh/RobustCertaintyEncoding (2024).

[88] A. C. Dobbins, R. M. Jeo, J. Fiser, and J. M. Allman, Distance modulation of neural activity in the visual cortex, Science **281**, 552 (1998).

[89] J. Cunningham, T. Nicol, C. King, S. G. Zecker, and N. Kraus, Effects of noise and cue enhancement on neural responses to speech in auditory midbrain, thalamus and cortex, Hearing Res. **169**, 97 (2002).