Precise spatial memory in local random networks

Joseph L. Natale
Department of Physics, Emory University, Atlanta, Georgia 30322, USA

H. George E. Hentschel
Department of Physics, Emory University, Atlanta, Georgia 30322, USA

Ilya Nemenman
Department of Physics, Department of Biology, and Initiative in Theory and Modeling of Living Systems, Emory University, Atlanta, Georgia 30322, USA

(Received 15 November 2019; revised 18 May 2020; accepted 16 June 2020; published 13 August 2020)

Self-sustained, elevated neuronal activity persisting on timescales of 10 s or longer is thought to be vital for aspects of working memory, including brain representations of real space. Continuous-attractor neural networks, one of the most well-known modeling frameworks for persistent activity, have been able to model crucial aspects of such spatial memory. These models tend to require highly structured or regular synaptic architectures. In contrast, we study numerical simulations of a geometrically embedded model with a local, but otherwise random, connectivity profile; imposing a global regulation of our system's mean firing rate produces localized, finely spaced discrete attractors that effectively span a two-dimensional manifold. We demonstrate how the set of attracting states can reliably encode a representation of the spatial locations at which the system receives external input, thereby accomplishing spatial memory via attractor dynamics without synaptic fine-tuning or regular structure. We then measure the network's storage capacity numerically and find that the statistics of retrievable positions are also equivalent to a full tiling of the plane, something hitherto achievable only with (approximately) translationally invariant synapses, and which may be of interest in modeling such biological phenomena as visuospatial working memory in two dimensions.

DOI: 10.1103/PhysRevE.102.022405

I. INTRODUCTION

Biological implementations of working memory bridge the gap between two fundamentally disparate timescales: single neurons process information in $\sim 10^{-3}$ s, whereas organisms interact with their external environments over durations of ~ 1 s or longer. For species from fruit flies to primates, this extension of timescales is reflected at the neural level by elevated spiking activity that persists while a particular memory is being accessed [1].

These excitations tend to be highly localized: For various types of working memory tasks across brain regions, firing rates for only a subset of selectively receptive neurons appear to become elevated [2–5]. Traditionally, these units are considered to be responsible for maintaining the memory, and their so-called *persistent activity*, which can last anywhere from tens of seconds to several minutes, is thought to underlie a multitude of well-studied neural computations [6] (see Ref. [7] for an alternative viewpoint). While the mechanistic drivers of persistent activity are not fully understood—both single-cell and network-level explanations have been proposed over the past several decades, but their relative [illegible] under debate [8]—attractor neural network models have provided phenomenological descriptions [illegible] firing states as fixed points or stable manifolds of the neural dynamics [9–11].

Attractor neural networks were first developed within the context of discrete, long-term associative memory, where each attracting state in a multistable system represented a distinct, stored memory [12]. Continuous-valued variants have since been able to model transient memories, like the firing activity responsible for maintaining an animal's eye position between saccades in one dimension [9] or its heading direction in a two-dimensional (2D) environment [13]. An enduring critique of these architectures is that they typically require highly structured or precisely tuned connection topologies to sustain the desired attractor behavior. For instance, the synaptic connectivity matrices in Ref. [9] satisfy stringent spectral tuning properties that allow certain firing patterns to persist indefinitely. While the need for special structure—including attempts to circumvent the need for recurrent connectivity entirely [14,15]—has been challenged by realizations of persistent activity in the context of (nearly) randomly connected model neurons [16–20], it still seems that some balance between excitation and inhibition is what endows recurrent circuitry with memorylike properties [8].

Recently, a biological instance of continuous attractor dynamics was traced to a circuit in *Drosophila*, consistent with one version of these topological constraints [21]. It has been established that this circuit plays a role in spatial navigation and further suggested that the attractor computation therein derives from high-level network properties—topological configuration, local excitation, and long-range inhibition—rather than "fine-scale" details like the synaptic weight distribution [22]. Although attractor dynamics have

previously been achieved via locally connected network models [23–25], it is yet unclear that networks built from random weights (i.e., unstructured connectivities) can reliably perform spatial memory tasks like those of the fly's internal navigation system. Moreover, the reliability of a randomly weighted network in encoding such memories in larger-dimensional spaces (location on a 2D plane) has yet to be quantified. These possibilities demands investigation, especially since random excitatory-inhibitory networks have been shown to be capable of various other complex computations, including conjunctive encoding for input classification [26] and, with appropriate regulation, emergent selectivity in the context of certain evidence integration tasks [27].

In this article, we ask how well a minimally structured, randomly connected network model [28] can perform a spatial memory task, in which the system must maintain a persistent representation of the geometrical location that corresponds to its most recent stimulation site. To do this, we study the firing-rate dynamics of a system with local, but otherwise random, connections, whose overall activity is regulated by global inhibition. Our network is spatially extended, and we show through numerical simulations that it is able to encode the locations of external stimuli as elevated firing activity in the region proximal to the stimulation site. In other words, it is capable of spatial memory.

We introduce this system in Sec. II and computationally measure its capacity for distinguishing different stimulation locations in Sec. III. We conclude by discussing how the model relates to previous work, and how it might be extended, in Sec. IV. Our intent is not to model any specific biological system, but to demonstrate through simulations how computations similar to those of persistent, continuous attractors are theoretically possible in random networks whose overall firing is controlled globally, rather than through local excitation-inhibition balance [29,30]).

II. MODEL AND METHODS

The network $\mathcal{G} = (\{i\}, \{J_{ij}\})$ consists of excitatory rate neurons $i = 1 \cdots N$ [9,31], embedded on a two-dimensional manifold [32,33]. Specifically, we consider a square plane of side length L, with connections $\{J_{ij}\}$ pointing from neuron j to neighbor i ($i, j = 1 \cdots N$). We choose a set of spatial point coordinates $X = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where each pair $\vec{x}_i = (x_i, y_i)$ is an independent random sample from the bivariate uniform distribution on the interval $[0, L]$. This system has uniform spatial density $\eta = \frac{N}{L^2}$, which is equivalent to $\frac{L}{\sqrt{N}} \equiv \lambda$ as the average separation between neurons.

With matrix elements $\{d_{ij}\}$ representing the Euclidean distances between neurons i and j, we assign a nonzero value to the synapse strength J_{ij} if $d_{ij} < \xi$, where $\xi \ll L$. We prohibit autapses, or self-loops, and invoke periodic boundary conditions in the calculation of d_{ij}. For convenience and to represent all results using the reference plane $[0, L] \times [0, L]$. In all that follows, $L = 1$ and $\xi = 0.06L$ unless otherwise specified. We also choose $N = 2^{12}$, which fixes $\lambda \approx 0.016(L) \approx 0.26\xi$.

Choosing a value for ξ which is small relative to L ensures that connections remain short ranged and that the resulting network is sparse. We argue later that choosing a set of connections $\{J_{ij}\}$ that is too short or too long ranged diminishes the ability of the network to support multiple nontrivial memory states. Quantitatively, since each neuron i interacts with $\sim\pi\xi^2\eta$ downstream neighbors, a typical network realization $\mathcal{G}$ encompasses $\sim\pi N^2(\xi/L)^2$ synapses, or about 1% of all possible connections.

The connection strengths, or synaptic efficiacies, are

$$J_{ij} = \begin{cases} \sim P(\mu, \sigma), & d_{ij} < \xi \quad \text{and} \quad j \neq i, \\ 0, & d_{ij} \geqslant \xi \quad \text{or} \quad j = i, \end{cases} \tag{1}$$

where each J_{ij} is an independent draw from $P(\mu, \sigma)$, representing a lognormal distribution (as argued for in Ref. [34] and elsewhere; we explored other distributions but found no qualitative differences in the results). Since by definition lognormal random variables are positive definite, $J_{ij} > 0$ for all outgoing connections: all neurons are excitatory. In what follows, $\mu = -0.702$ and $\sigma = 0.8752$ (by convention, these parameters refer to the associated normal distribution). These values were taken from fits done during experimental investigations of neural circuit properties in the rat visual cortex [34].

As emphasized above, persistent activity traditionally demands a balance between excitation and inhibition, while our connectivities encompass no explicit inhibition. Therefore, we choose to model inhibition indirectly, imposing its main effect—which we assume is to stabilize the system's total firing activity to a constant value [35,36]—directly. In particular, we insert a term into the usual nonlinear firing-rate equations [9,31] to represent nonlocal inhibitory interactions. In summary, in the absence of synaptic or external inputs, the firing-rate activity $r_i(t)$ decays exponentially over the intrinsic timescale τ. Otherwise, $r_i(t + dt)$ is determined by integrating a nonlinear function of combined input currents $\sum_j J_{ij} r_j(t)$ the from upstream neighbors j and external drive $I_i(t)$ over the short interval $dt \ll \tau$. Thus, for constant $a > 0$,

$$\tau \frac{dr_i}{dt} = -r_i + aN\left(\frac{h_i}{\sum_j h_j}\right), \tag{2}$$

$$h_i = f\left[\sum_j J_{ij} r_j + I_i(t)\right]. \tag{3}$$

This system will ultimately approach a steady state for which $\sum_i r_i(t \gg \tau) = aN$: Global inhibitory interactions, implemented by the second, "activation," term in Eq. (2), create the desired balance. This can be verified by solving for the steady-state conditions $\frac{dr_i}{dt} = 0$. The parameter a in Eq. (2) can be thought of as the system's baseline firing level (the rate at which all neurons would fire if they were to fire at equal rates in the steady state). A complementary interpretation, related to the fraction of active cells in the steady state, will be addressed in detail later. We set $a = 0.02$ and, without loss of generality, choose $\tau = 1$ so that time is measured in unit of τ.

Finally, we adopt for the nonlinearity a version of the firing-rate function introduced by Ref. [13],

$$f(x) = \alpha(\ln\{1 + \ln[1 + e^{\beta(x-\gamma)}]\})^{\delta}, \tag{4}$$

with $\alpha = 18$, $\beta = 0.5$, $\gamma = 16$, and $\delta = 1.5$. We selected these values to place activations $\{h_i\}$ in a biological range (tens or less if measured in Hz) for arguments $x > 0$ spanning two orders of magnitude, with $f(0) \sim 10^{-4} \approx 0$. The reason for the choice given by Eq. (4) is that the gain of this curve increases at a value away from zero and that its behavior in the limit of large inputs is nonsaturating over two orders of magnitude in x. These attributes are intended to better approximate the biological reality [37], as compared with the sigmoidal thresholding functions commonly used in artificial networks (which tend to feature inflection points near values corresponding to zero net input). We note that both of these properties are also satisfied by the rectified linear unit (ReLu) activation function [38], also commonly used in machine learning.

For a realization $\mathcal{G}$ with dynamics given by Eqs. (2) and (3), we quantify how this system performs as a spatial memory architecture. In particular, if a group of neurons local to an arbitrary region of the plane is stimulated externally, will the system be capable of sustaining a persistent representation of the stimulated coordinates? How many distinct stimulation sites can the system reliably encode?

To measure the number of resolvable sites, we perform n_{trials} "external stimulation" computational experiments, sequentially, in Matlab. First, we initialize the system, creating a network realization $\mathcal{G}$ by selecting values for the neuron positions X and connection strengths $\{J_{ij}\}$. We then set the firing rates of all neurons $i = 1 \cdots N$ to $r_i(0) = a$ and evolve Eqs. (2) and (3) from $t = 0$ to $t = 100\tau$, well beyond the point at which the individual firing rates stabilize, using the built-in Runge-Kutta (4,5) solver with $I_i(t) = 0$. The result can be a strong excitation, confined to a local region of the plane, or a fully *delocalized* firing state in which all neurons participate with rates near a. In either case, the rates do not change in time (this holds even if the system is initialized randomly, with rates that sum to the steady-state value aN, instead of uniformly).

To ensure that the system can switch out of this state, we perform a single external stimulation, abitrarily targeting the visual center of the plane, according to the following protocol. With the aforementioned state serving as our initial condition, we locate all neurons contained within an "input" patch of area $\pi\rho^2$ (for now, we choose $\rho = \xi = 0.06L$) centered at $\vec{x}_{\text{stim}} = (0.5, 0.5)$. For this subset of system elements only, we set

$$I_i(t) = A[1 - \Theta(t - \Delta t)] = \begin{cases} A, & t < \Delta t, \\ 0, & t \geqslant \Delta t, \end{cases} \tag{5}$$

where $\Theta(t)$ denotes the Heaviside step function and $\Delta t = 5\tau$. We again solve Eqs. (2) and (3), integrating until $T = 40\tau$, sufficient time for the network to reach a persistent state.

We then repeat this protocol for n_{trials} iterations, each time sampling a random position $\vec{x}_{\text{stim}} = (x_{\text{stim}}, y_{\text{stim}})$ from a uniform grid of 10^4 finely spaced points superimposed on the set of stimulation centers. The resulting state $\{r_i(t = T)\}$ then becomes the new initial condition for the following trial, representing restimulation and new memory formation. We set $n_{\text{trials}} = k \cdot (L/dL)^2$, partitioning stimulations into k successive groups of $(L/dL)^2$ trials that are each composed of independent random permutations of the full list of available grid points $\{x_{\text{stim}}\}$.

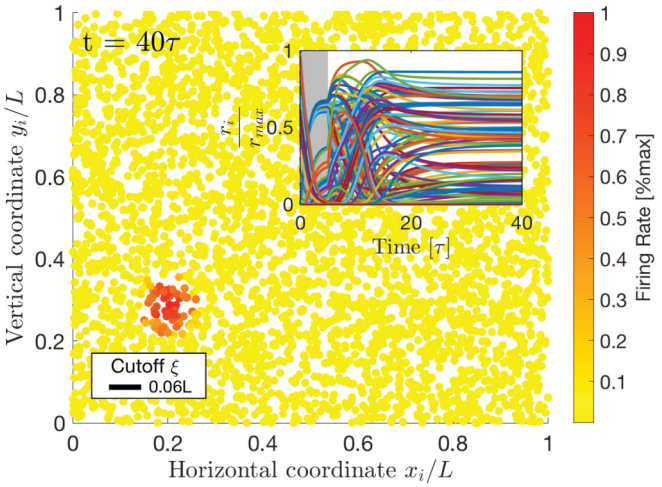

FIG. 1. Sample bump state in a system with $N = 2^{12}$. The scale bar indicates the synaptic cutoff distance ξ, below which $\mathcal{G}$ appears fully connected. Inset: All the neural activities through time. Most of the trajectories remain near zero and cannot be visually distinguished. Stimulation is shown as a gray block of width $\Delta t = 5\tau$.

III. RESULTS

A. Network supports multiple stable attractors

On stimulation, the system initialized as above tends to develop a localized excitation in the vicinity of $\vec{x}_{\text{stim}}$, which quickly coalesces into a roughly circular "bump" of activity [11,39,40]. Figure 1 depicts a representative bump in a system of size $N = 2^{12}$ at $T = 40\tau$. The inset reproduces the firing-rate trajectories for $t \leqslant T$, showing that all rates have stabilized to their final values by T.

While it is free to migrate or spread about $\vec{x}_{\text{stim}}$ during and after stimulation, this activity bump typically assumes a stable shape and location on the plane by the same time T. Analogous behaviors are observed when the system is stimulated from within a previously activated stable state. Then, activities associated with any preexisting bump are rapidly attenuated due to the global inhibition, typically returning to baseline activity values by Δt. Generally, given a sufficiently strong input current amplitude A and adequately long stimulation time Δt, an activity bump will form in any general region of the plane and remain thereafter in the vicinity of $\vec{x}_{\text{stim}}$.

In simulation, our model seems to support only one spatially localized excitation under steady-state conditions, even if stimulated briefly at two locations simultaneously. At least qualitatively, this might be understood by analogy with a simpler system consisting of just two units, representing distant regions of strong firing. If each unit acts according to Eqs. (2) and (3)—loosely, as a self-excitatory, positive-feedback system, with a global inhibition that enters via the normalization $h_i / \sum_j h_j$—then it is easy to imagine that their mutual feedback will lead to a single unit dominating (we ignore oscillations, since the feedback would need to be precisely tuned in order for these to appear). While it is not immediately clear from these equations that simultaneous

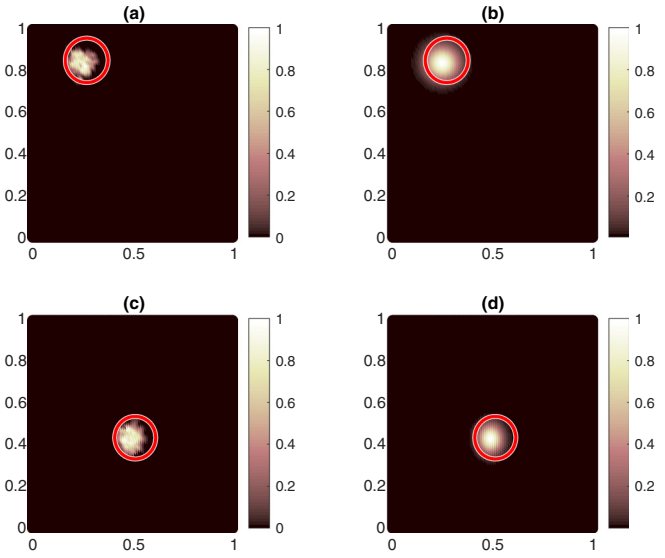

FIG. 2. Spatial activity distributions for sample bumps. Above, two-dimensional Gaussian envelopes were fit to firing-rate activity profiles comprising the final, "bump" states. To do this, horizontal and vertical coordinates of each neuron were binned into 100×100 separate pixels. In panel (a): Example bump, with activities depicted as a percentage of the maximum firing rate. The center of the red circle represents the center of excitation, and its radius is drawn to encompass all units with firing rates observed to exceed some multiple of the baseline firing rate; see Sec. III B for further details. In panel (b): Fit of the (discretized) data depicted in panel (a), with values reported relative to the peak. Panels (c) and (d): Same as top row for a different bump state. Note that active regions for the fits are encompassed by the same circles drawn for their respective firing data. Whether or not the brightly colored pixels for a given bump are, in fact, normally distributed, a main advantage of this Gaussian approximation is that it provides a robust way to track each bump's *width*, as needed in later analyses.

activation at many locations will not lead inevitably to delocalized excitations or multiple small bumps, we are not focused on this here, precisely because we are interested in situations for which there is exactly one driving input at any given moment in time—and only one recent memory, as in the experimental system of Ref. [21]. Thus, as a rule of thumb, we say that the system supports a single bump at any given time [21], in any general spatial region of the plane.

How large are these activity bumps? Although they are not perfectly circular, we observe that excitations do take on a typical size for a fixed cutoff distance ξ. We can therefore speak about an effective bump radius R_{eff}. A simple way to measure R_{eff} would be to choose a firing-rate threshold above which neurons will be considered *active* and compute the radius for the equivalent circular area πR_{eff}^2 occupied by this subset of system elements on the plane. Ideally, we need a criterion that is relatively insensitive to the cutoff distance. In order to track bump sizes without choosing such a threshold in advance, we fit two-dimensional Gaussian curves to the spatial firing-rate distributions associated with each bump (see Fig. 2) and measure $2R_{\text{eff}}$ by the full width at half maximum, as done recently for the experimental system of Ref. [21]. This yields $R_{\text{eff}}(\xi = 0.06L) \approx 0.78\xi \approx 0.05L$. In other words, the bump radius is on the order of the cutoff distance; we expect this to be a generic result.

Taking the ratio $\frac{R_{\text{eff}}}{\lambda} \approx 3$, we see that typical activity bumps are also large in comparison with the average neuron separation λ, as well as the distance $dL = 10^{-2}L$ between adjacent gridpoints. This has an important consequence. If the system is stimulated at a point within (or too near) the area associated with an active bump, it may respond by reverting to the originally active bump state instead of evoking a new memory. This is particularly true if either the input time Δt or amplitude A are insufficiently large but occurs more generally due to the quenched disorder in the neuron positions and connectivity. Certain bumps will emerge as preferred states, which are more strongly favored than others (this limits the network representational capacity, as we determine quantitatively later). Nevertheless, the system does appear to select from a discrete, finite set of constant firing-rate states for the parameter values ($\lambda = N^{-\frac{1}{2}}, \xi \approx 3.84\lambda$) defined above.

In summary, for sufficiently strong input, we observe:

(i) Local stimulation can cause the system to develop stable bumps in essentially any region of the plane;

(ii) The system seems able to transition, smoothly and repeatably, from sustaining one bump state to another (switch between multistable firing patterns);

(iii) Independent stimulations centered at different gridpoints can result in nearly indistinguishable memory bumps.

We take these observations together as the earmarks of dynamical attracting behavior—in particular, the system acts as a discrete approximation to a 2D plane attractor. We identify each achievable bump state with a stored, retrievable memory. By definition, an attracting state persists until stimulation evokes a new bump, so we say that the system stores *spatial memories* encoding the location at which it was most recently stimulated.

If the basins of attraction (from within which stimulation at different $\vec{x}_{\text{stim}}$ values consistently leads to the activation of specific memories) are not infinitely small but of finite size, the system cannot remember arbitrary positions on the plane. It is then natural to ask how many *unique* spatial locations can be distinguished by a given realization of the synaptic matrix. That is, the resolution with which $\vec{x}_{\text{stim}}$ can be decoded requires quantification.

B. Spatial memories effectively span the plane

How many distinct stimulation locations $\vec{x}_{\text{stim}}$ might we anticipate a realization $\mathcal{G}$ to resolve? We expect this *capacity* to depend largely on gross statistics like the average size of the attracting basins rather than on details of the instantial arrangement of neuron positions and synaptic connections associated with a given system configuration.

Since the dynamical equations (2) and (3) are deterministic, the attracting state evoked by stimulation at a given site should be unique, apart from the aforementioned dependencies on the initial state and input-current parameters. This variation can even be minimized: The stronger the external inputs, the more reliably we can anticipate that the system will find an attractor in the vicinity of the stimulation location, independent of where it is currently excited. Thus all that remains to determine the exact set of attractors supported

by a given configuration $\mathcal{G}$ are the the coupling strengths. Accordingly, we expect that the bumps to which excitations attract will be almost exclusively a function of the (quenched) random variable J_{ij}.

We coarsely estimate the system's capacity as follows. Assuming homogeneous basins of attraction and one-to-one retrieval within a basin, the number of reliably stored memories will be equal to the number of basins that fit on the plane. Dividing the $L \times L$ space into equally sized square sections of width $2R_{\text{eff}}^{-2}$ implies, for our parameter values, $\sim 10^2$ distinct, nonoverlapping basins that span the 2D space. Thus our baseline will be $\sim$100 bumps, touching tangentially.

A preliminary step toward more accurately quantifying the number of stimulation locations that the system can reliably encode is simply enumerating all the unique attractors activated during a given series of n_{trials} stimulations. This allows us to conceptualize the capacity in terms of input (stimulation site) to output (bump location) relations. For each stimulation, we track the *center of excitation* $\vec{x}_{\text{COE}}(t) = \sum_{i'} \frac{r_{i'}(t)\vec{x}_{i'}}{aN}$ among cells i' which we identify as actively participating. Instead of accommodating for the uncertainties associated with our Gaussian fits, here we employ simple thresholding to identify active units, for two principal reasons. First, the fixed-threshold criterion $r_i > 10a$ predicts the number of active neurons to within 10 units of the amount given by the more sophisticated *participation number* $p_\nu = (\sum_{i=1}^{N} r_i^\nu)^2 / \sum_{i=1}^{N} r_i^{2\nu}$, with similar qualitative behavior across the surprisingly large range of cutoffs from roughly zero to 10λ. In addition, this fixed-threshold criterion was found to predict centers for bump excitations that coincide well with the measured Gaussian peaks.

For large cutoffs, it is possible that even a fairly nonrestrictive threshold can exclude relatively strongly firing neurons: Our constraint $\sum_i r_i(t) = a$ implies that firing activity within a given bump decreases as bumps increase in size, which is precisely what we observed to happen as we increase ξ. Excitations encompassing zero active neurons were to be assigned a special value of $\vec{x}_{\text{COE}}(t)$, allowing us to count them separately toward the capacity, but this was not observed for the $\xi = 0.06L$ presented below. We enumerate all distinct bumps by counting the unique values of $\vec{x}_{\text{COE}}(T)$ observed to within a specific resolution (we discuss the importance of this resolution below). For n_{trials} large, this number should approach the cardinality of the set of possible memories. The next step will be to quantify how many—or with what fidelity—distinct values of the gridpoint coordinates $\vec{x}_{\text{stim}}$ can be discriminated by these enumerated attractors.

We measure the capacity for a given realization $\mathcal{G}$ as follows. Although each site in the set of $(LdL)^{-2} = 10^4$ available stimulation gridpoints is visited $\frac{n_{\text{trials}}}{L \cdot dL^{-2}} = k$ times each in each series of stimulation events, averaging over all possible initial conditions for each gridpoint would require too much time. Here we choose $k = 10$ to further mitigate finite-sampling errors in the situation described above, in which stimulation near a highly active bump simply reverts the system back to the same attractor after a transient. We also choose to work with an information-theoretic capacity metric, to treat the inherently nonuniform stochasticity associated with the "stimulus-response" records in a natural framework.

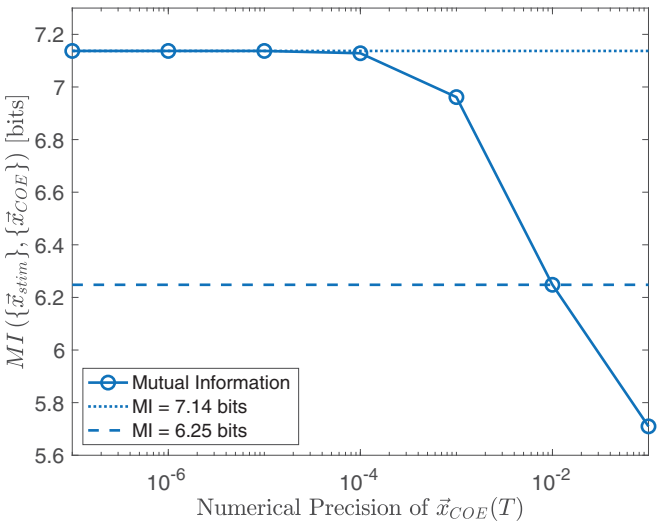

FIG. 3. Mutual information as a function of rounding precision in the center-of-excitation values $\{\vec{x}_{\text{COE}}(T)\}$. Saturation occurs by four decimal places, but in what follows we keep two places to ensure the precision of $\vec{x}_{\text{COE}}$ is not finer than the neuron separation scale λ. This changes the capacity by less than a factor of 2.

Specifically, we measure the mutual information [41] between random variables $\vec{x}_{\text{stim}}$ and $\vec{x}_{\text{COE}}(T)$ for a realization $\mathcal{G}$. To do this, we obtain the frequencies of occurrence for all observed stimulation locations $\{\vec{x}_{\text{stim}}\}$ and bump centers $\{\vec{x}_{\text{COE}}(T)\}$, over a set of n_{trials} stimulation events. We then use these frequencies as the maximum-likelihood estimates of the corresponding probabilities to form the "plug-in" or naïve estimators for the relevant entropies [42–44], from which we can calculate the mutual information $\text{MI}(\{\vec{x}_{\text{stim}}\}, \{\vec{x}_{\text{COE}})(T)\}$. Since asking how many different attractors were observed for each stimulation position is equivalent to asking how many different stimulation positions lead to the same attractor (i.e., the mutual information is symmetric), we choose the latter. Finally, from the mutual information, we define the capacity

$$C = 2^{\text{MI}(\{\vec{x}_{\text{stim}}\},\{\vec{x}_{\text{COE}}(T)\})}. \tag{6}$$

Since the information is measured over discrete states, we must discretize the values of $\vec{x}_{\text{COE}}(T)$ by rounding them to an appropriate resolution. As seen in Fig. 3, truncating $\vec{x}_{\text{COE}}(T)$ to two decimal places still represents 87.5% of the maximum information or $\approx$6.25 bits. Assuming that the system cannot track bump centers to a precision better than these two decimal places—roughly the theoretical separation between neurons—we arrive at $C \approx 76$ distinct stimulation regions for the values of L, λ, ξ, and ρ used throughout.

In other words, on average, $\mathcal{G}$ is able to store and reliably retrieve a number of memories approximately equal to our naïve, baseline estimate. Unlike in that coarse estimation, we did not require bumps to be nonoverlapping in measuring the capacity—yet the system's recall ability turns out to be nearly as accurate as a fully deterministic discriminator that simply decides in which $R_{\text{eff}} \times R_{\text{eff}}$-sized, homogeneous division of the plane the last stimulation occurred. Thus the

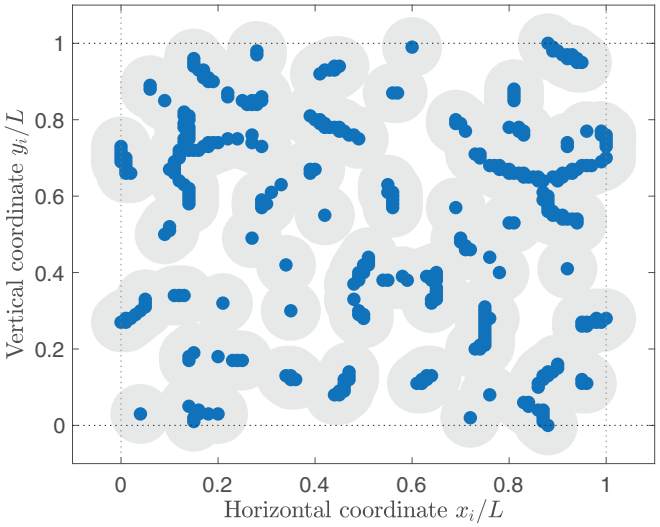

FIG. 4. Spatial distribution of the distinct bumps. Here the different attracting bumps observed over the kn_{trials} computational experiments are distributed in such a way that they span the majority of the 2D plane. Bump centers are shown as blue dots; the radii for their surrounding gray circles are $\approx R_{\text{eff}}$. Dotted lines represent periodic boundaries, included here for clarity.

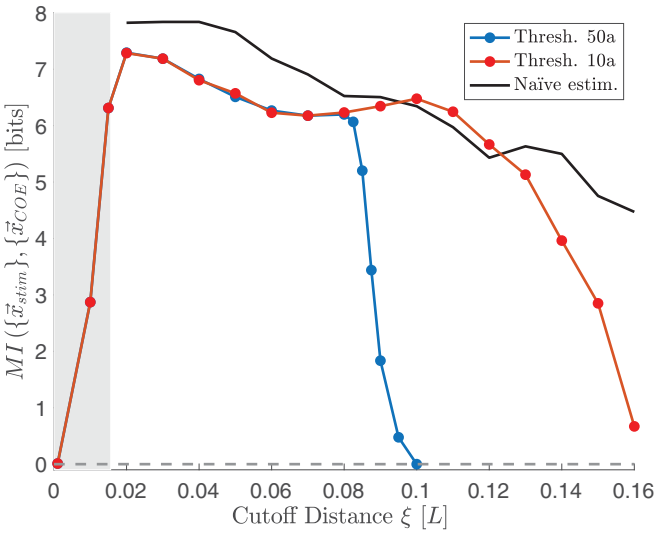

FIG. 5. Information is high for a broad range of cutoffs. That is, varying ξ reveals a broad plateau over which the mutual information remains within a single bit of its maximum. At either extreme of ξ, the information falls to zero as connectivities become too sparse or too dense to support the type of spatial memory discussed throughout. The black curve represents the information $\log_2 \frac{L^2}{\pi R_{\text{eff}}^2}$ corresponding to our original, naïve estimate of C, with $R_{\text{eff}}(\xi)$ adjusted to match the typical values given by Gaussian fits to ∼1000 bumps. Note that the black curve, representing $\xi < \lambda$, exists only outside the shaded gray box, because bumps that did localize for small ξ were too few to quantify accurately. Note also that we could have studied the variation of C with the number of neurons; yet since $\xi \propto \lambda$ and $\lambda = \sqrt{\eta} = \sqrt{N}$ for this system, this measurement would be redundant.

information-theoretic capacity, measured to two decimal digits precision in $\vec{x}_{\text{COE}}(T)$, is also consistent with a typical size for the attracting basins which matches R_{eff} for stable bumps. Furthermore, we observe that the retrievable memories span more or less the entire spatial extent of the $L \times L$ plane. This can be readily observed in Fig. 4, which depicts the set $\{\vec{x}_{\text{COE}}\}$ of unique bumps accounted for over a course of n_{trials} stimulations for one network realization.

C. Mutual information is near optimal for a broad range of parameter values

The cutoff distance is an important length scale in the system. The structure of the network depends crucially on ξ, allowing us to go from completely unconnected neurons in the extreme of $\xi = 0$ to the fully connected network for $\xi = L$. It is important to understand how ξ affects our main findings—in particular, the existence of localized excitations and the number of memories $\mathcal{G}$ can support.

For the unconnected case $\xi = 0$, we have $\{J_{ij}\} = 0$. In the absence of recurrent connections (besides the implicit inhibition), all neurons respond independently to their respective external inputs $I_i(t)$: That is, the $\{r_i\}$ obey a simplified version of Eqs. (2) and (3). In order to write down the dynamics in this case, we first note that neurons outside the stimulation patch have activations $h_i = f(0) \approx 0$ for both $t < \Delta t$ and $t > \Delta t$. These units at first experience an exponential decay in [illegible] and then approach the steady-state value $r_i(t \gg \Delta t) = a$. The $\sim \pi N(\rho/L)^2$ neurons encompassed by the stimulation patch also approach a constant value. To show this, we note that each of the units in this latter subset sees the same input $h_i = f\{A[1 - \Theta(t - \Delta t)]\}$, so that the ratio $h_i(t)/\sum_j h_j(t)$ stays constant. Therefore we can remove the nonlinearities entirely and write

$$\frac{dr_i}{dt} = -r_i + I_i'(t), \tag{7}$$

$$I_i'(t) = \begin{cases} \frac{a}{\pi\rho^2}, & t < \Delta t, \\ a, & t \geqslant \Delta t. \end{cases} \tag{8}$$

Then, in the long-time limit, the unconnected system relaxes to the trivial stable state $\{r_i(t \gg \tau)\} = a$, in which all neurons fire at the same, baseline rate. It cannot sustain any excitations that can be decoded as memories. In the other extreme, $\xi \to L$, it seems unlikely that a fully connected network can support any *localized* excitations.

We quantify the precise dependence of our findings on the value of the cutoff distance in Fig. 5. We generated this plot by progressively decreasing ξ for an initial, fully connected realization $\mathcal{G}$. Here we chose $k = 1$, stimulating at the first 10^4 of the 10^5 sites used to generate Fig. 3, and rounded the measured information values to a precision of two decimal places in $\vec{x}_{\text{COE}}$ as decided above. Clearly, the mutual information quickly drops to zero below the average neuron separation λ. This means that the system attains only states that are delocalized—effectively all neurons contribute to the excitation, but none exceed the threshold $r_i > 10a$ to be considered "active"—which we identify as the single, trivial state.

At the other extreme, the mutual information returns to zero for large values of ξ. This can be explained in terms of the circumstances discussed in Sec. III A, in which it becomes difficult for the network to switch out of its preferred states. As the cutoff distance increases above $\xi \approx 7\lambda$ (or $\approx 5\lambda$ for the stricter threshold of $r_i = 50a$), more neurons are directly involved in sustaining a given excitation, and the structure of the basins of attractions changes so as to accommodate fewer feasible memories. As in the case of insufficient stimulation time or amplitude, the success or failure of a given stimulation in evoking a nearby bump is somewhat history dependent (in the sense that some memories may be retrievable from some initial states, but not from certain preferred states), yet invariably the system comes to favor a single state in the limit that the network becomes fully connected. For the $10a$ threshold, the network cannot reliably store any spatial memories for roughly $\xi > 0.16L \approx 10\lambda$.

Between these two extremes, there is an optimal value $\xi^* \approx 0.02L$, for which the greatest number of stimulation gridpoints can be distinguished. Moreover, starting at this value, there is a plateau in the system's accuracy from roughly $\xi = 0.02L \cdots 0.11L \approx \lambda \cdots 7\lambda$, across which the mutual information varies by only ~ 1 bit. More precisely, the gap between the highest and lowest points on the $10a$-threshold curve of Fig. 5 corresponds to the difference between resolving $C \approx 156$ and $C \approx 72$ distinct stimulation sites. These values are of the same rough order of magnitude, and their average is nearly equal to our very first baseline estimate of 100 distinct, homogeneous basins. We note in particular that the cutoff distance $\xi = 0.06L$ used throughout the rest of the paper is nominally three times larger than ξ^*, but different by less than the aforementioned bit in terms of the information.

In principle, the capacity should also depend on how reliably the system accesses its attractors for (or indeed, whether the set of accessible attractors changes with) different values of the size of the input patch, ρ. Figure 6 records the dependence of the mutual information on ρ. Outside this range, the system will attract to (possibly different) preferred states, but between roughly 2λ and 6λ we observe that the system attracts to the same bump state regardless of the specific value of ρ (not explicitly depicted). This gives the appearance that the system really is tracking the stimulation centers in computing its final states, at least for input patch sizes in this range.

To the extent that different proxies for $\vec{x}_{\mathrm{COE}}$ agree, this suggests that the system does in fact encode a coarse representation of the stimulation location—the bump centers of excitations—rather than tracking high-dimensional quantities like the real-valued firing rates. That is, although an experimental system wired according to our prescription for $\{J_{ij}\}$ could indeed store information in individual firing rates for other purposes, we are not merely imposing but discovering [illegible]ensional summary variable $\vec{x}_{\mathrm{COE}}$ is sufficient to predict the stimul[illegible]ation region to a considerable accuracy. [illegible] toward testing this hypothesis would be to systematically map [illegible]he basins of attraction for a given realization $\mathcal{G}$, and check whether the steep decrease shown in Fig. 6 occurs when the stimulation patch grows large enough to extend into multiple basins besides that of the targeted memory.

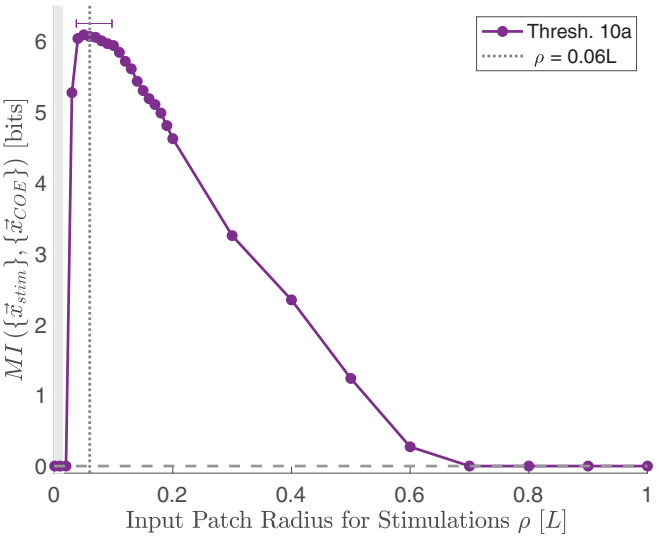

FIG. 6. Mutual information peaks for a range of small input patch sizes. In particular, in the neighborhood of $\rho = \xi = 0.06L$, mutual information values do not vary significantly. We verified that the system tends to fall into the same attractor regardless of the specific value of ρ until a large percentage of neurons are stimulated, thereby activating the aforementioned "preferred," or global, states. At roughly the same value after which comes a decrease in information with ξ, we observe a similar decrease in information with ρ. This continues monotonically, until $\rho > 50\lambda$, after which stimulation leads only to excitations below the activity threshold.

Together, the above results suggest that our randomly weighted network can sustain local excitations for a range of parameter values. In general, these excitations can serve reliably as spatial memories encoding the system's most recent stimulation location if the number of neurons activated via stimulation and local synaptic input is small relative to the system size N. This can be achieved by choosing ξ less than approximately $\mathcal{O}(10\lambda)$, which ensures that a given neuron synapses with anywhere from roughly $\pi(\lambda)^2\eta \cdots \pi(10\lambda)^2\eta \approx 10^0 \cdots 10^2$ neighbors.

IV. DISCUSSION

We have showed through numerical simulations that short-range, but otherwise unstructured, connectivities can support spatial memory via persistent firing if the overall activity of the network is constrained by global regulation of firing activies. The spatial regions that can be remembered (discriminated) with fidelity effectively tile an $L \times L$ planar section, with a resolution of $\mathcal{O}(\lambda^{-1})$ distinct sites. Essentially equivalent to our naïve count of nonoverlapping memories spanning the manifold, this performance corresponds to an information-theoretic capacity that scales as $C \propto \sqrt{N}/L$, or $C \propto \sqrt{\eta}$ in terms of the neuron density. Although this dependence can be verified by testing larger system sizes N and L, the capacity measure studied here is ultimately determined by the typical bump size R_{eff}, which comes in turn from the chosen cutoff distance ξ. Given that we can always write

the cutoff as a multiple of the neuron separation scale, $\xi = \xi(\lambda) = \xi(L/\sqrt{N})$, we expect that N and L enter only as the ratio λ, and that varying the cutoff distance is equivalent to varying the number of neurons up to statistical fluctuations that are larger at smaller N.

Since the average neuron separation sets the scale of the problem at the outset, it is not necessarily surprising that the optimal cutoff distance $\xi^* \approx \lambda$. What is unexpected in our results is the fact that a spatial memory spanning a two-dimensional manifold can be achieved without explicit tuning of synaptic connections. This is reinforced by the fact that we observe not just an isolated peak at ξ^*, but a broad plateau of near-optimal cutoff distances.

While it was originally maintained [8] that only tuned connectivity profiles can produce continuous attractors, the idea that random networks support memory on short timescales is not entirely new [16,19,28,45]. Indeed, recent work argues that quasi-random topologies, refined via some non-linear Hebbian learning rule, can give rise to attractor dynamics in the specific context of persistent neural activity as a substrate for working memory [46]. Here, we are interested in *using* such random networks to store spatial memories that effectively span a continuous manifold [33]. Crucially, we accomplish this using a network structure which emphatically requires no learning.

Similarly, distance-dependent topologies [47] have been implemented in previous models, including the seminal work on continuous neural attractors [39]. While other studies have realized attractor dynamics with connectivities that are relatively unstructured (i.e., except for local profiles) [23–25], we are aware of only two related studies that attempt to quantify the contribution of sparse, short-range (1D nearest-neighbor) connections formally to the localization of firing-rate excitations [48,49]. As we do, both of these respect Dale's Principle [50] for the signs of synaptic connections only indirectly [51] and explore random weights. While it may be interesting to explore the spectra of our $\{J_{ij}\}$ in the context of Anderson localization, or the notion of "spatially structured" disorder developed in Ref. [49], a more obvious generalization of our model would be to relax the hard-threshold cutoff condition to a connection probability. For example, we could set $J_{ij} \propto e^{-|\vec{x}_i - \vec{x}_j|/\xi}$, or another function of $d_{ij} = \|\vec{x}_i - \vec{x}_j\|$ (as in, for example, Refs. [23–25,52]).

A drawback to our model, in the form presented here, is that the system of Eqs. (1)–(5) incorporates no explicit noise terms. Fundamental to our results is the firing-rate constraint $\sum_i r_i(t) = aN$, an imposition which corresponds only approximately to the biological reality for real circuits (as in Ref. [21]). In our future work, we propose to replace the constant parameter a by a Gaussian process $\alpha(t) = a + \eta(t)$. We expect that, for small amounts of noise, the system will retain its qualitative behavior but with a reduced capacity. On the other hand, for $\eta(t)$ with large variance, it is possible that [illegible] store memories with high fidelity due to longer bump "excursions" or full delocalization, as in Figs. 5

If these assumptions regarding the inclusion of noise are found to hold, it would be interesting to explore noise parameters that place the firing-rate variability in a regime consistent with previous experiments [8,37] while respecting our sparsity constraints. Yet we reiterate that our goal is not to model any known experimental system. Indeed, whether or not our model relates to specific, observable experimental systems remains to be seen. In anticipation of such *in vivo* analogs, we offer the following predictions regarding which features of our model might be used to infer whether short-range, randomly weighted connections drive a given instance of persistent activity.

First, in the best case scenario, novel technologies may allow researchers to probe structural properties directly. This promises a trivial way of checking whether synaptic matrices are untuned, as in Eq. (1), and is already underway for the fly [53–55]. While the emerging picture for *Drosophila* is one of decidedly nonrandom connectivity, this may not hold for significantly larger organisms. Indeed, the number of possible synapses in a neural system scales as $\mathcal{O}(N^2)$. Thus genetic encoding of precise values for some billions of pairwise connections even in modestly sized vertebrates is simply not feasible. On the other hand, it is plausible that regularity appears at the level of local rules superimposed on essentially random connectivities, as in canonical microcircuit models [56], which would be consistent with our setup.

In the absence of structural information, the firing-rate activities themselves can also help support or reject our model. Since most classic continuous-attractor architectures have translationally invariant connections, they are able to host bumps at virtually any location [57]. Our $\{J_{ij}\}$, on the other hand, lack such a symmetry. This leads to discrete attractors [58] with variable spacing and portions of the plane that cannot be reliably encoded. Such "discrete approximations" to attracting manifolds have even been touted as more robust than their continuous counterparts, for example, to perturbations in the synaptic weights [59]. It would be interesting to quantify the fraction or extent of the plane that the system can remember in the presence of the aforementioned noise.

In addition, while continuous attractor models accommodate a degree of drift or diffusion for activity bumps following their settlement on the manifold [60], tracking $\vec{x}_{\mathrm{COE}}(t)$ reveals that excursions in our random networks occur predominantly *before* $t = T$; see the inset of Fig. 1. Thus, comparing the observed distribution of displacements, between the tested $\vec{x}_{\mathrm{stim}}$ values and the corresponding $\vec{x}_{\mathrm{COE}}(t)$ could also distinguish our model.

Finally, the raw activity measurements $\{r_i(t)\}$ are also subject to what is known as network reverse-engineering, or automated inference methods that operate directly on data to reconstruct network interaction structures [61]. Although we do not advocate applying out-of-the-box algorithms to glean structural information in general, there do exist certain signatures and gross statistics which can be used to differentiate truly random graphs from more complex or subtle architectures at a coarse level [62].

Our model is one of many that attempt to capture the ability of different neural systems to support localized excitations that encode real-valued quantities. Here, we eschew structured topographic mappings [21] in favor of a random

connectivity that we find to be capable of storing similar neural representations. Whether or not *in vivo* circuits conforming to the specifications of our model are found experimentally to underlie one of these interesting systems, in our view such random, activity-constrained networks should still be taken seriously as null models for recurrent neural computation [27].

ACKNOWLEDGMENTS

This work was supported in part by the following grants: James S. McDonnell Foundation Grant No. 220020321, NSF Grants No. 1208126 and No. 1822677, and NIH Grant No. 1 R01 EB022872. We thank Itai Pinkovezky, K. Michael Martini, and Vivek Jayaraman for insightful discussions.

[1] P. S. Goldman-Rakic, Neuron **14**, 477 (1995).
[2] J. M. Fuster and G. E. Alexander, Science **173**, 652 (1971).
[3] S. Funahashi, C. J. Bruce, and P. S. Goldman-Rakic, J. Neurophysiol **61**, 331 (1989).
[4] J. M. Fuster and J. P. Jervey, Science **212**, 952 (1981).
[5] J. W. Gnadt and R. A. Andersen, Experim. Brain Res. **70**, 216 (1988).
[6] D. MacNeil and C. Eliasmith, PLoS ONE **6**, e22885 (2011).
[7] M. Lundqvist, P. Herman, and E. K. Miller, J. Neurosci. **38**, 7013 (2018).
[8] J. Zylberberg and B. W. Strowbridge, Annu. Rev. Neurosci. **40**, 603 (2017).
[9] H. S. Seung, Proc. Natl. Acad. Sci. USA **93**, 13339 (1996).
[10] D. J. Amit and N. Brunel, Cereb. Cortex **7**, 237 (1997).
[11] A. Compte, N. Brunel, P. S. Goldman-Rakic, and X.-J. Wang, Cereb. Cortex **10**, 910 (2000).
[12] J. J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982).
[13] K. Zhang, J. Neurosci. **16**, 2112 (1996).
[14] S. Ganguli and P. Latham, Neuron **61**, 499 (2009).
[15] M. S. Goldman, Neuron **61**, 621 (2009).
[16] W. Maass, T. Natschläger, and H. Markram, Neural. Comput. **14**, 2531 (2002).
[17] N. Bertschinger and T. Natschläger, Neural. Comput. **16**, 1413 (2004).
[18] R. Singh and C. Eliasmith, J. Neurosci. **26**, 3667 (2006).
[19] D. V. Buonomano and W. Maass, Nat. Rev. Neurosci. **10**, 113 (2009).
[20] O. Barak, M. Rigotti, and S. Fusi, J. Neurosci. **33**, 3844 (2013).
[21] S. S. Kim, H. Rouault, S. Druckmann, and V. Jayaraman, Science **356**, 849 (2017).
[22] K. S. Kakaria and B. L. de Bivort, Front. Behav. Neurosci. **11**, 8 (2017).
[23] A. Renart, P. Song, and X.-J. Wang, Neuron **38**, 473 (2003).
[24] J. A. Goldberg, U. Rokni, and H. Sompolinsky, Neuron **42**, 489 (2004).
[25] R. Rosenbaum and B. Doiron, Phys. Rev. X **4**, 021039 (2014).
[26] J. B. George, G. M. Abraham, Z. Rashid, B. Amrutur, and S. K. Sikdar, Sci. Rep. **8**, 1403 (2018).
[27] A. J. Sederberg and I. Nemenman, PLoS Comp. Biol. **16**, e1007875 (2020).
[28] B. Kriener, H. Enger, T. Tetzlaff, H. E. Plesser, M.-O. Gewaltig, and G. T. Einevoll, Front. Comput. Neurosc. **8**, 136 (2014).
[29] C. van Vreeswijk and H. Sompolinsky, Science **274**, 1724 (1996).
[30] C. van Vreeswijk and H. Sompolinsky, Neural. Comput. **10**, (1998).
[31] S. Druckmann and D. B. Chklovskii, Curr. Biol. **22**, 2095 (2012).
[32] F. P. Battaglia and A. Treves, Phys. Rev. E **58**, 7738 (1998).
[33] R. Monasson and S. Rosay, Phys. Rev. Lett. **115**, 098101 (2015).
[34] S. Song, P. J. Sjöström, M. Reigl, S. Nelson, and D. B. Chklovskii, PLoS Biol. **3**, e68 (2005).
[35] Y. Roudi and A. Treves, PLoS Comp. Biol. **4**, e1000012 (2008).
[36] R. Monasson and S. Rosay, Phys. Rev. E **87**, 062813 (2013).
[37] F. Barbieri and N. Brunel, Front. Neurosci. **2**, 3 (2008).
[38] R. H. Hahnloser and H. S. Seung, in *Advances in Neural Information Processing Systems (NIPS 2000)* (MIT Press, Cambridge, MA, 2001), pp. 217–223.
[39] S.-i. Amari, Biol. Cybern. **27**, 77 (1977).
[40] X.-J. Wang, Trends Neurosci. **24**, 455 (2001).
[41] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[42] A. Antos and I. Kontoyiannis, Rand. Struct. Algor. **19**, 163 (2001).
[43] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek, Phys. Rev. Lett. **80**, 197 (1998).
[44] L. Paninski, Neural Comput. **15**, 1191 (2003).
[45] J. Griffith, Biophys. J. **3**, 299 (1963).
[46] U. Pereira and N. Brunel, Neuron **99**, 227 (2018).
[47] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition* (Cambridge University Press, Cambridge, 2014).
[48] A. Amir, N. Hatano, and D. R. Nelson, Phys. Rev. E **93**, 042310 (2016).
[49] H. Tanaka and D. R. Nelson, Phys. Rev. E **99**, 062406 (2019).
[50] E. Catsigeras, Appl. Math. **4**, 15 (2013).
[51] M. Zhu and C. J. Rozell, PLoS Comp. Biol. **11**, e1004353 (2015).
[52] C. Huang, D. A. Ruff, R. Pyle, R. Rosenbaum, M. R. Cohen, and B. Doiron, Neuron **101**, 337 (2019).
[53] D. Turner-Evans, S. Wegener, H. Rouault, R. Franconville, T. Wolff, J. D. Seelig, S. Druckmann, and V. Jayaraman, eLife **6**, e23496 (2017).
[54] R. Franconville, C. Beron, and V. Jayaraman, eLife **7**, e37017 (2018).
[55] P. H. Li, L. F. Lindsey, M. Januszewski, M. Tyka, J. Maitin-Shepard, T. Blakely, and V. Jain, Microscopy Microanalysis **25**(S2), 1364 (2019).
[56] O. Sporns, Front. Comput. Neurosc. **5**, 5 (2011).
[57] S. Wu, K. M. Wong, C. A. Fung, Y. Mi, and W. Zhang, F1000 **5** (2016).
[58] H. K. Inagaki, L. Fontolan, S. Romani, and K. Svoboda, Nature **566**, 212 (2019).

[59] Z. P. Kilpatrick, B. Ermentrout, and B. Doiron, J. Neurosci. **33**, 18999 (2013).

[60] K. Wimmer, D. Q. Nykamp, C. Constantinidis, and A. Compte, Nat. Neurosci. **17**, 431 (2014).

[61] J. L. Natale, D. Hofmann, D. G. Hernández, and I. Nemenman, in *Quantitative Biology: Theory, Computational Methods and Examples of Models*, edited by B. Munsky, W. Hlavacek, and L. Tsimring (MIT Press, Cambridge, MA, 2018).

[62] E. Estrada, *The Structure of Complex Networks: Theory and Applications* (Oxford University Press, Oxford, UK, 2011).