Featured in Physics

Forgetting Leads to Chaos in Attractor Networks

Ulises Pereira-Obilinovic
Center for Neural Science, New York University, New York, New York 10003, USA; Instituto de Ciencias de la Ingeniería, Universidad de O'Higgins, Rancagua, Chile 2841959; and Department of Statistics, The University of Chicago, Chicago, Illinois 60637, USA

Johnatan Aljadeff
Department of Neurobiology, The University of Chicago, Chicago, Illinois 60637, USA and Department of Neurobiology, University of California San Diego, La Jolla, California 92093, USA

Nicolas Brunel*
Departments of Neurobiology and Physics, Duke University, Durham, North Carolina 27710, USA and Departments of Statistics and Neurobiology, The University of Chicago, Chicago, Illinois 60637, USA

(Received 3 December 2021; revised 31 August 2022; accepted 6 December 2022; published 27 January 2023)

Attractor networks are an influential theory for memory storage in brain systems. This theory has recently been challenged by the observation of strong temporal variability in neuronal recordings during memory tasks. In this work, we study a sparsely connected attractor network where memories are learned according to a Hebbian synaptic plasticity rule. After recapitulating known results for the continuous, sparsely connected Hopfield model, we investigate a model in which new memories are learned continuously and old memories are forgotten, using an online synaptic plasticity rule. We show that for a forgetting timescale that optimizes storage capacity, the qualitative features of the network's memory retrieval dynamics are age dependent: most recent memories are retrieved as fixed-point attractors while older memories are retrieved as chaotic attractors characterized by strong heterogeneity and temporal fluctuations. Therefore, fixed-point and chaotic attractors coexist in the network phase space. The network presents a continuum of statistically distinguishable memory states, where chaotic fluctuations appear abruptly above a critical age and then increase gradually until the memory disappears. We develop a dynamical mean field theory to analyze the age-dependent dynamics and compare the theory with simulations of large networks. We compute the optimal forgetting timescale for which the number of stored memories is maximized. We found that the maximum age at which memories can be retrieved is given by an instability at which old memories destabilize and the network converges instead to a more recent one. Our numerical simulations show that a high degree of sparsity is necessary for the dynamical mean field theory to accurately predict the network capacity. To test the robustness and biological plausibility of our results, we study numerically the dynamics of a network with learning rules and transfer function inferred from *in vivo* data in the online learning scenario. We found that all aspects of the network's dynamics characterized analytically in the simpler model also hold in this model. These results are highly robust to noise. Finally, our theory provides specific predictions for delay response tasks with aging memoranda. In particular, it predicts a higher degree of temporal fluctuations in retrieval states associated with older memories, and it also predicts fluctuations should be faster in older memories. Overall, our theory of attractor networks that continuously learn new information at the price of forgetting old memories can account for the observed diversity of retrieval states in the cortex, and in particular, the strong temporal fluctuations of cortical activity.

DOI: 10.1103/PhysRevX.13.011009

Subject Areas: Interdisciplinary Physics, Nonlinear Dynamics, Statistical Physics

*nicolas.brunel@duke.edu

I. INTRODUCTION

Recurrent attractor networks are an influential theoretical model for learning and memory in brain systems [1–4]. In attractor networks, memories correspond to stable patterns of network activity, i.e., fixed-point attractor states of the

network dynamics. To store these memories, an external input associated with each memory modifies the network synaptic connectivity, thanks to a synaptic plasticity rule. When synaptic modifications lead to the creation of a fixed-point attractor of the dynamics, that fixed point is said to be the neural representation of a newly learned memorandum. Memory retrieval occurs when a brief presentation of an external cue correlated with a specific memorandum is followed by an autonomous relaxation to the attractor state associated with it. Decoding the memorandum's identity is possible when the corresponding attractor state is correlated with the learned input.

Theoretical analysis of attractor models shows that they can parsimoniously reproduce key observations of neuronal responses during memory tasks. In particular, they reproduce the phenomenon of selective persistent activity, i.e., the observation that some neurons in multiple cortical areas have persistently elevated firing rates following the presentation of specific items to be maintained in memory during the task, but not others [5–10].

There are two major challenges to classical attractor networks as a mechanism subserving mnemonic representations. First, neuronal activity during delay periods exhibits a high degree of temporal irregularity [11–16], which is at odds with the notion that retrieval states are fixed-point attractors. While extensions of classical attractor models can account for the observed irregularity [17,18], and the nonstationarity of firing rates [18,19], they typically do not account for both irregular and nonstationary behavior, or are not consistent with stable population coding [16] (however, see Ref. [20]). Second, classical attractor models exhibit *catastrophic forgetting*: when the number of learned fixed points exceeds the network's storage capacity, no memory can be retrieved. Catastrophic forgetting occurs because of the statistical symmetry between stored patterns. That symmetry implies that forgetting one memory is statistically equivalent to forgetting all.

Here we propose an attractor network model which overcomes these two limitations, extending work by Tirozzi and Tsodyks on chaotic attractor networks [21] and our recent work on recurrent networks with learning rules inferred from *in vivo* data [22]. In these models, temporal irregularities result from memories retrieved as chaotic attractors. Remarkably, memory retrieval is possible as the network state remains correlated with the stored pattern at all times, despite the ongoing fluctuations. To avoid catastrophic forgetting, we use online learning, a well-known recipe for avoiding catastrophic forgetting [23–31]. In so-called "palimpsest" networks that implement online learning, older memories are overwritten by new ones. Thus, online learning breaks the statistical symmetry between memories, since each memory is characterized by a different age. The effect of such learning rules on network behavior has been studied in networks of binary neurons with fixed-point attractors [24,27,32–34]. To account for the temporal irregularities during retrieval, we study the effects of online learning on memory states in a network of rate neurons that can exhibit firing rate fluctuations.

We use dynamic mean field theory (DMFT) to compute the statistics of network activity in different network states, in the large N, sparse connectivity limit. We find that fixed-point and chaotic attractor memory states coexist in our model. The age of a given stored memory determines the nature of its retrieval state: fixed point for recently learned patterns, and chaotic attractor for older patterns, leading to a continuum of statistically distinguishable memory states. We compare these analytical results to simulations of large networks (up to 10^7 neurons), showing a good agreement between simulations and theory in the limit of extremely sparse networks (i.e., when the average number of connections per neuron scales logarithmically with the network size). The transition to chaos as a function of memory age is well predicted by our theory for more dense networks, i.e., when the average number of connections scales as a power law with the network size. Finally, we show that our theory for the online learning of memories qualitatively holds in more biologically realistic networks with learning rules and transfer functions inferred from *in vivo* data [22], and that these results are highly robust to noise.

II. MODEL

We consider a recurrent neuronal network composed of N neurons whose input currents are described by analog variables h_i, where $i = 1, 2, \ldots, N$. The instantaneous firing rates of neurons are given by the input-output single neuron transfer function (or f-I curve) $\phi(x) = \tanh(x)$. This choice is made to simplify the mathematical analysis while preserving qualitative features of more realistic networks [22] (see Sec. VI). The synaptic input currents obey the standard current-based version of the rate equations,

$$\frac{dh_i(t)}{dt} = -h_i(t) + \sum_{i \neq j}^{N} c_{ij} J_{ij} \phi(h_j(t)) + I_i, \qquad (1)$$

where c_{ij} represents the "structural connectivity" of the network, which is generated as a sparse random Erdős-Rényi graph, i.e., all c_{ij} are i.i.d (independent and identically distributed). Bernoulli random variables, $c_{ij} = 1$ with probability K/N, where K is the average in and out degree; J_{ij} is the strength of the synapse connecting the presynaptic neuron j to the postsynaptic neuron i; and I_i is an external input to neuron i. We assume that $1 \ll K \ll N$, which means that the average number of connections is large but much smaller than the network size. This is a relevant limit for microcircuits in the brain, where each neuron receives a large number of inputs ($K \sim 1000$), and connection probabilities are small, of order $\sim$10% in cortex [35–39] and 1% in hippocampus [40].

The synaptic connectivity matrix is assumed to have been structured through past presentations of external stimuli (patterns) to the network. The stimulus presentation timescale is assumed to be much longer than the timescale of network dynamics, and the synaptic connectivity is assumed to be constant at the scale of retrieval of specific patterns. We use the variable u to refer to the stimulus presentation times, to avoid confusion with t in Eq. (1), where $u \in \mathbb{Z}$. External inputs to the network consist in a stream of time-ordered binary *patterns* $\{\vec{\eta}^u\}$. We assume that η_i^u are i.i.d. binary random variables, such that $\eta_i^u = \pm 1$ with equal probability. We note that while the stored memories are binary, the network dynamics and, therefore, the retrieval states are continuous. The patterns η can be interpreted as the population firing rates when the dynamics in Eq. (1) is driven by a strong random external stimulus I_i symmetrically distributed around zero. The synaptic weights are modified by the patterns in *one shot* using an online version of the covariance rule [41]. According to this rule,

$$J_{ij}^{u+1} = \rho J_{ij}^u + \frac{A}{K}\eta_i^u \eta_j^u. \tag{2}$$

In Eq. (2), the first term on the rhs represents a decay term with a forgetting rate $0 < \rho < 1$ that prevents unbounded growth of synaptic weights. The second term represents a "Hebbian" synaptic modification that is proportional to the covariance of presynaptic and postsynaptic activity, with a strength parameter A. On one extreme, for $\rho = 0$, the learned connectivity from previous patterns is forgotten in a single time step, while in the case $\rho = 1$, there is no forgetting. The case $\rho = 1$ is problematic since it leads to unbounded growth of the weights when the number of presented patterns becomes very large. However, for completeness, we will describe the $\rho = 1$ scenario, in which unbounded growth is avoided by considering a finite number of presented patterns (see Sec. IV).

With such a learning rule, the connectivity matrix at time u is given by

$$J_{ij}^u = \frac{A}{K}\sum_{u'=-\infty}^{u-1} \rho^{u-1-u'}\eta_i^{u'}\eta_j^{u'}. \tag{3}$$

Since the stream of learned patterns is infinite, the properties of the connectivity and the network do not depend on u, so we drop that index and rewrite J_{ij} as

$$J_{ij} = \frac{A}{K}\sum_{\mu=0}^{\infty} e^{-\mu/\tau K}\eta_i^\mu \eta_j^\mu. \tag{4}$$

Here μ is the age of pattern η_i^μ (i.e., how far in the past pattern μ was presented), and τ is a forgetting time constant defined as $\tau = -1/K \log \rho$. We anticipate that to optimize storage capacity, ρ should be close to 1, $\rho \approx 1 - 1/K\tau$, with $\tau \sim \mathcal{O}(1)$, similar to other palimpsest networks [42]. Note that in the sum on the rhs of Eq. (4), we have changed the pattern indices for convenience, such that a pattern of age μ corresponds to the pattern shown at time $u' = u - 1 - \mu$.

III. DYNAMIC MEAN FIELD THEORY

A. Formalism

We start by deriving the DMFT for the network model defined by Eqs. (1) and (4), in the limit of infinitely many neurons $N \to \infty$, synapses per neuron $K \to \infty$, and sparse connectivity $K/N \to 0$ [21,43]. Although the network in Eqs. (1) and (4) stores binary patterns, i.e., $\eta_i^s \in \{-1, 1\}$, here we present a general DMFT for arbitrary $P(\eta)$, with the constraint that $\langle \eta \rangle = 0$. This ensures that the average change in connection strength due to learning a single pattern is zero, and therefore the average synaptic weight does not grow with the number of stored patterns. This could be enforced by a homeostatic mechanism that controls the mean changes in the incoming inputs due to learning [44,45]. We use similar methods as in Refs. [21,46], and anticipate that for some parameters, the dynamics of the network will be chaotic. In the above limit, the instantaneous synaptic inputs to each neuron can then be rewritten of its mean ν_i and random temporal variations around the mean $\zeta(t)$ due to chaotic dynamics,

$$\sum_{j=1, j\neq i}^{N} c_{ij}J_{ij}\phi(h_j(t)) = \nu_i + \zeta_i(t). \tag{5}$$

Therefore, the network dynamics in Eq. (1) becomes statistically equivalent to the following stochastic differential equation:

$$\frac{dh_i(t)}{dt} = -h_i(t) + \nu_i + \zeta_i(t). \tag{6}$$

where the ζ_i's are uncorrelated stochastic processes whose autocorrelation will be computed self-consistently below.

As in classical mean field theories for attractor neuronal network models [2,47], we define the overlaps between network state and the memories stored in the network as

$$m_\mu = \frac{1}{N}\sum_{i=1}^{N}\eta_i^\mu \phi(h_i) = \langle \eta^\mu \phi(h)\rangle_{h,\eta}, \tag{7}$$

where $\langle \cdots \rangle_{h,\eta}$ represents an average over the statistics of the stored patterns η and the input currents h. The network is able to retrieve a pattern stored in memory if it settles in a state in which there is an overlap of order one with the corresponding pattern. This means, in particular, that the stimulus identity can be retrieved from the network activity by a linear decoder. On the other hand, if the overlap is zero

in the large N limit, then the memory is forgotten and cannot be retrieved.

We assume that the overlaps do not depend on time [i.e., $m_\mu(t) = m_\mu$], which is trivially true for fixed-point attractors and becomes a good approximation at large N for chaotic attractor memory states. Except when stated otherwise (see Sec. III D), we will assume that only a single memory is retrieved in a given retrieval state. With this assumption, the mean synaptic inputs ν_i to neuron i become

$$\nu_i = A\eta_i^\mu e^{-\mu/\tau K} m_\mu. \tag{8}$$

Note that in Eq. (8), the mean synaptic input depends exponentially on the age of the pattern. For most recent memories ($\mu \ll K$) the mean synaptic current is close to its maximum in magnitude $\nu_\mu \approx A\eta_i^\mu m_\mu$, but it decays exponentially for older memories, $\mu \sim O(K)$.

We next turn to the fluctuations around the mean. The variable $\zeta_i(t)$ is assumed to be a Gaussian random field with mean zero, representing the variability around the mean synaptic inputs to neuron i. Its autocovariance function is given by

$$\mathrm{Cov}(\zeta_i(t), \zeta_i(t+t')) = \gamma A^2 \kappa \langle \phi(h(t))\phi(h(t+t'))\rangle_h, \tag{9}$$

where

$$\kappa = \frac{1}{K}\sum_{\mu=1}^{\infty} e^{-2\mu/\tau K}, \tag{10}$$

and $\gamma = \langle \eta^2 \rangle^2$. In particular, for binary patterns considered here, $\gamma = 1$. In the large K limit, Eq. (10) leads to $\kappa = \tau/2$.

Equations (8) and (10) make it clear that the number of patterns that can be retrieved is of order K. Thus, in the large N limit, we define a continuous age index $s = \mu/K$, such that $s = 0$ for recently learned patterns, while $s \to \infty$ for patterns seen in the distant past.

For a pattern with age s, the dynamics of the currents can be rewritten as

$$\frac{dh_i}{dt} = -h_i + A\eta_i^s e^{-s/\tau} m_s + A\sqrt{\frac{\gamma\tau}{2}} y_i(t), \tag{11}$$

where $y(t)$ is a Gaussian random field with autocovariance function

$$C_s(t') = \langle y(t)y(t+t')\rangle_y = \langle \phi(h(t))\phi(h(t+t'))\rangle_h, \tag{12}$$

where the subscript s in the above equation is here to remind us that the autocovariance function depends on the age of the retreived pattern. By defining the local currents $u_i(t) = h_i(t) - A\eta_i^s e^{-s/\tau} m_s$, Eqs. (11), (7), and (12) can be rewritten as

$$\dot{u} = -u + A\sqrt{\frac{\gamma\tau}{2}} y(t), \tag{13}$$

$$m_s = \langle \eta \phi(u(t) + A\eta e^{-s/\tau} m_s)\rangle_{u,\eta}, \tag{14}$$

while

$$C_s(t') = \langle \phi(u(t) + A\eta e^{-s/\tau} m_s)\phi(u(t+t') + A\eta e^{-s/\tau} m_s)\rangle_{u,\eta}. \tag{15}$$

As in Refs. [46,48], we introduce the synaptic input current autocovariance function,

$$\Delta_s(t') = \langle u(t)u(t+t')\rangle_u, \tag{16}$$

where again s reminds us of the age dependence of this autocovariance function. Analogously to the derivation in Refs. [49,50], we can derive a self-consistent equation for the local-field autocovariance:

$$\frac{d^2\Delta_s(t')}{dt'^2} = \Delta_s(t') - \frac{A^2\gamma\tau}{2} C_s(t'). \tag{17}$$

Using Eq. (6), we can evaluate the age-dependent overlap. Equation (7) becomes

$$m_s = \int D\eta Dx \eta \phi(A[\sqrt{\Delta_{0s}}x + \eta e^{-s/\tau} m_s]). \tag{18}$$

We define Δ_{0s} as the autocovariance at zero time lag $t' = 0$, i.e., $\Delta_s(0) = \Delta_{0s}$, i.e., the variance of the synaptic input currents h_i. The autocovariance in Eq. (15) can be written as

$$C_s(t') = \int D\eta Dz \left[\int Dx \phi(A\{\sqrt{\Delta_{0s} - |\Delta_s(t')|}x + \sqrt{|\Delta_s(t')|}z + \eta e^{-s/\tau} m_s\})\right]^2, \tag{19}$$

where $D\eta = p_\eta(\eta)d\eta$, $Dx = dx e^{-x^2/2}/\sqrt{2\pi}$, and $Dz = dz e^{-z^2/2}/\sqrt{2\pi}$.

We further assume that $\Delta_s(t') \geq 0$, and rescale $\Delta_s(t')$, $\Delta_s(t') \to A^2\Delta_s(t')$. As in Ref. [46], Eq. (17) can be rewritten in terms of a potential,

$$\frac{d^2\Delta_s}{dt'^2} = -\frac{\partial V(\Delta_s, \Delta_{0s})}{\partial \Delta_s}, \tag{20}$$

by defining

$$V(\Delta_s, \Delta_{0s}) = -\frac{\Delta_s^2}{2} + \frac{\tau\gamma}{2A^2}\int D\eta Dz\left[\int Dx\Phi(A\{\sqrt{\Delta_{0s} - |\Delta_s|}x + \sqrt{|\Delta_s|}z + \eta e^{-s/\tau}m_s\})\right]^2, \tag{21}$$

where $\Phi(x) = \int_0^x dr\phi(r)$.

The dynamics in Eqs. (20) and (21) corresponds to equation of particle in a Newtonian potential $V(\Delta_s, \Delta_{0s})$ that depends parametrically on Δ_{0s}. As shown by Ref. [46], the motion of the particle should be such that $[d\Delta_s(0)/dt'] = 0$ and $V(\Delta_{0s}, \Delta_{0s}) = V[\Delta_s(\infty), \Delta_{0s}]$. Therefore, the network dynamics in Eq. (11) is described by three *order parameters*: the overlap m_s, the synaptic input current variance Δ_{0s}, and Δ_{1s}, which is the longtime limit of the autocovariance of the synaptic input currents $\Delta_{1s} = \lim_{t'\to\infty}\Delta_s(t')$. These can be computed self-consistently, as shown below.

As we will see below, there are four qualitatively different types of solutions to the DMF equations. Solutions with $m_s = 0$ correspond to background states (i.e., states that are uncorrelated with all stored memories), while solutions with $m_s > 0$ correspond to memory (or retrieval) states since there exists a finite overlap between the network state and the corresponding stored pattern. Both types of solutions can be either fixed-point attractors of the dynamics [i.e., with autocovariance functions that are constant in time, i.e., $\Delta_s(t) = \Delta_{0s}$ at all times] or chaotic attractors, with an autocovariance function that depends on time. We first focus our attention on the transitions to chaos and then study the system's capacity (largest age at which retrieval states still exist).

B. Transitions to chaos

In fixed-point attractors there are no temporal fluctuations in the input currents. The autocovariance of the local fields in Eq. (16) is then equal to the variance of the local currents at all times [i.e., $\Delta_s(t') = \Delta_{0s}$], which leads to

$$m_s = \int D\eta Dx\eta\phi(A[\sqrt{\Delta_{0s}}x + \eta e^{-s/\tau}m_s]), \tag{22}$$

$$\Delta_{0s} = \frac{\gamma\tau}{2}\int D\eta Dx\phi^2(A[\sqrt{\Delta_{0s}}x + \eta e^{-s/\tau}m_s]). \tag{23}$$

The above equations give the overlap in a retrieval fixed-point attractor state of a memory of age s. However, these fixed points can destabilize and become chaotic, leading to chaotic retrieval states. Importantly, as we will see below in this model, the dynamical properties of the attractors (i.e., fixed point or chaotic) strongly depend on the age of the patterns. Equations (22) and (23) have solutions in parameter space even beyond the transition to chaos, when the assumption of fixed-point attractor memory states with no temporal fluctuations in the input currents is no longer valid. However, as expected, their predictions depart from network simulations [see Fig. 1, dashed line, and Figs. 2(a) and 2(b), red lines]. We refer to the solutions of these equations as the static solutions to DMFT (SMFT).

Analogous to Ref. [46] to find the transition to chaos of memory states, it is necessary to find the point in parameter space where the static solution $\Delta_s(t') = \Delta_{0s}$ becomes unstable. At this point, the autocovariance of the local field $\Delta_s(t')$ transitions from stationary to time dependent. Since the dependence on time of the autocovariance of the local fields is ruled by the Newtonian equation for the position Δ_s at "time" t' of a particle subject to a potential energy [Eq. (17)], finding the transition point is equivalent to finding the critical point $\Delta_{0s}^{\text{chaos}}$ where the potential in Eq. (21) changes its concavity. We also computed the maximum Lyapunov exponent λ analytically using similar methods as Ref. [46] (see Appendix A for details) and found that the transition to chaos, obtained from the condition $\lambda = 0$, coincides with the point where the static solution becomes unstable. As in Ref. [46], from this calculation we conclude that the transition to chaos is determined by the concavity of the classical potential $V(\Delta_s, \Delta_{0s})$ defined in Eq. (20). For a nonchaotic retrieval state the maximum Lyapunov exponent is negative $\lambda < 0$ and the potential $V(\Delta_s, \Delta_{0s})$ is concave; i.e., $[\partial^2 V(\Delta_s, \Delta_{0s})/\partial\Delta_s^2] < 0$. On the other hand, for a chaotic retrieval state the maximum Lyapunov exponent is positive $\lambda > 0$ and therefore the potential $V(\Delta_s, \Delta_{0s})$ is convex; i.e., $[\partial^2 V(\Delta_s, \Delta_{0s})/\partial\Delta_s^2] > 0$. The transition to chaos is given by the change of concavity of the potential $[\partial^2 V(\Delta_s, \Delta_{0s})/\partial\Delta_s^2] = 0$.

Beyond this critical point, solutions for the autocovariance of the local field starting at Δ_{0s} relax to $\lim_{t'\to\infty}\Delta_s(t') \equiv \Delta_{1s}$. At the transition to chaos, $\Delta_{0s} = \Delta_{1s}$. To calculate the transition point in which fixed-point retrieval states become chaotic, we first compute the second derivative of the potential in Eq. (21) as in Appendix C of Ref. [49] and set it to zero, then set $\Delta_{0s} = \Delta_{1s}$, obtaining

$$\frac{A^2\gamma\tau}{2}\int D\eta Dz\{\phi'(A[\sqrt{\Delta_{0s}}z + \eta e^{-s/\tau}m_s])\}^2 = 1. \tag{24}$$

Together Eqs. (22)–(24) describe the curve in the parameter space that separates fixed-point from chaotic memory states.

In the background state, all the overlaps with the stored memories are zero; i.e., $m_s = 0$. The critical line in the

space of parameters for its transition to chaos is thus given by the following set of equations:

$$\frac{A^2\gamma\tau}{2}\int Dz\{\phi'(A\sqrt{\Delta_0}z)\}^2 = 1, \tag{25}$$

$$\Delta_0 = \frac{\gamma\tau}{2}\int Dz\phi^2(A\sqrt{\Delta_0}z). \tag{26}$$

C. Capacity

As there exist two types of memory states, there are two possible ways to compute the capacity: the maximum age at which fixed-point or chaotic retrieval states cease to exist, respectively. However, as we will see later, in a range of parameters of the system, the physically relevant capacity is the one computed from chaotic retrieval states since fixed-point memory states destabilize and become chaotic before reaching capacity for those parameters. In chaotic attractors, the potential defined in Eq. (21) is convex (i.e., $[\partial^2 V(\Delta_s,\Delta_{0s})/\partial\Delta_s^2] > 0$), and the autocovariance of the local currents in Eq. (16) is time dependent. Additionally, as shown by Refs. [46,48], a chaotic solution is characterized by an aperiodic, decreasing autocovariance function. This corresponds to the condition $\lim_{t'\to\infty} V(\Delta_s(t')) = V(\Delta_{0s})$, which is equivalent to

$$\begin{aligned}&-\Delta_{0s}^2 + \frac{\tau\gamma}{A^2}\int D\eta Dx\Phi^2(A[\sqrt{\Delta_{0s}}x + \eta e^{-s/\tau}m_s])\\ &\quad = -\Delta_{1s}^2 + \frac{\tau\gamma}{A^2}\int D\eta Dz\left[\int Dx\Phi(A\{\sqrt{\Delta_{0s}-|\Delta_{1s}|}x + \sqrt{|\Delta_{1s}|}z + \eta e^{-s/\tau}m_s\})\right]^2,\end{aligned} \tag{27}$$

where Δ_{1s} corresponds to the large time limit of $\Delta_s(t)$, $\Delta_s(t)\overset{t\to\infty}{\longrightarrow}\Delta_{1s}$. Therefore, $(\partial V/\partial\Delta)|_{\Delta=\Delta_{1s}} = 0$, which is equivalent to

$$\Delta_{1s} = \frac{\tau\gamma}{2}\int D\eta Dz\left[\int Dx\phi(\{\sqrt{\Delta_{0s}-|\Delta_{1s}|}x + s\sqrt{|\Delta_{1s}|}z + \eta e^{-s/\tau}m_s\})\right]^2. \tag{28}$$

In Eqs. (27) and (28), the overlap m_s is given by Eq. (22). Thus, the three order parameters m_s, Δ_{0s}, and Δ_{1s} can be obtained by solving numerically Eqs. (22), (27), and (28). For fixed-point attractor memory states, there are no temporal fluctuations in the input currents and therefore $\Delta_{1s} = \Delta_{0s}$, recovering from Eqs. (22), (27), and (28) the static solutions in Eqs. (22) and (23). Beyond the transition to chaos, the overlap curve for chaotic attractors is given by Eqs. (22), (27), and (28). The capacity of this network corresponds to the age s^c beyond which older memories cannot be retrieved. At capacity, the corresponding memory state has zero overlap $m_{s^c} = 0$ [see Fig. 2(a)]. Since the transfer function in this model [$\phi(x) = \tanh(x)$] is odd, at capacity we have that $\Delta_{1s^c} = 0$ [see Eq. (28)]. The value of Δ_{0s^c} is then obtained by solving Eq. (27) and is given by

$$\tau = \frac{(A\Delta_{0s^c})^2}{\gamma[\int Dx\Phi^2(A[\sqrt{\Delta_{0s^c}}x]) - (\int Dx\Phi(A[\sqrt{\Delta_{0s^c}}x]))^2]}, \tag{29}$$

with

$$Ae^{-\frac{s^c}{\tau}}\int Dx\phi'(A[\sqrt{\Delta_{0s^c}}x]) = 1. \tag{30}$$

Equations (29) and (30) provide the capacity curve in parameter space (s^c,τ,A) (see the curve that separates the green and the gray regions in Fig. 4). Note that Eq. (30) is obtained by derivating Eq. (22) with respect to the overlap and evaluating the resulting equation at capacity (i.e., $m_{s^c} = 0$), considering the overlap changes continuously with the memory age [see Fig. 5(a)].

D. Stability of memory states

In this section, we study the stability of a memory state of a given age s with respect to perturbations in neuronal activity in the direction of another, more recent memory state. Thus, we assume that the network state is correlated with two memories: the currently retrieved one, of age s, but also a recent memory ($s = 0$). For simplicity, we focus on the case of binary patterns. This calculation aims to derive dynamical equations for the transient dynamics of the overlaps m_0 and m_s, similarly to a recent calculation for recurrent networks storing sequences of activity [51]. Then, we can use these equations to study numerically the stability of the memory state corresponding to the memory s, i.e., $m_s \sim \mathcal{O}(1)$ and $m_0 \ll 1$.

Similarly to Eq. (11), the dynamics of the currents are approximated by a time-dependent Gaussian random field:

$$\frac{dh_i}{dt} = -h_i + A(\eta_i^0 m_0 + \eta_i^s e^{-s/\tau}m_s) + A\sqrt{\frac{\tau}{2}}y_i(t). \tag{31}$$

At any time, the synaptic input current $h_i(t)$ can be described by a Gaussian random variable:

$$h(t) = A(\eta_i^0 m_0 + \eta_i^s e^{-s/\tau} m_s) + A\sqrt{\frac{\tau\Delta_{0s}(t)}{2}}x, \tag{32}$$

where x is Gaussian random variable with mean 0 and variance 1, and the variance Δ_{0s} is now time dependent.

By applying the scaled inner product $[1/(Nc)]\langle\vec{\eta}^s, \vec{\bullet}\rangle = [1/(Nc)]\sum_{j=1}^{N}\eta_j^N \bullet_j$ to both sides of Eq. (1) we obtain that in the limits $K, N \to \infty$ and $K/N \to 0$ the dynamical equations of both overlaps m_0 and m_1 are given by

$$\frac{dm_0}{dt} = -m_0 + \langle \eta^0 \phi(h)\rangle_{h,\eta}, \tag{33}$$

$$\frac{dm_s}{dt} = -m_s + \langle \eta^s \phi(h)\rangle_{h,\eta}. \tag{34}$$

As previously mentioned, $\langle\cdots\rangle_{h,\eta}$ represents the average over the statistics of the stored patterns η and the input currents h. Plugging Eq. (32) into Eqs. (33) and (34) we obtain

$$\frac{dm_0}{dt} = -m_0 + \int D\eta^0 D\eta^s Dx \eta^0 \phi\left(A\left[\eta_i^0 m_0 + \eta_i^s e^{-s/\tau} m_s + \sqrt{\frac{\tau\Delta_{0s}(t)}{2}}x\right]\right), \tag{35}$$

$$\frac{dm_s}{dt} = -m_s + \int D\eta^0 D\eta^s Dx \eta^s \phi\left(A\left[\eta_i^0 m_0 + \eta_i^s e^{-s/\tau} m_s + \sqrt{\frac{\tau\Delta_{0s}(t)}{2}}x\right]\right). \tag{36}$$

Since stored patterns are binary in our model, the two integrals over the pattern's statistics above are replaced with sums. After some manipulations, we finally obtain the dynamical equations for the overlaps:

$$\frac{dm_0}{dt} = -m_0 + \frac{1}{2}\int Dx\left\{\phi\left(A\left[m_0 + e^{-s/\tau} m_s + \sqrt{\frac{\tau\Delta_{0s}}{2}}x\right]\right) + \phi\left(A\left[m_0 - e^{-s/\tau} m_s + \sqrt{\frac{\tau\Delta_{0s}}{2}}x\right]\right)\right\}, \tag{37}$$

$$\frac{dm_s}{dt} = -m_s + \frac{1}{2}\int Dx\left\{\phi\left(A\left[m_0 + e^{-s/\tau} m_s + \sqrt{\frac{\tau\Delta_{0s}}{2}}x\right]\right) + \phi\left(A\left[e^{-s/\tau} m_s - m_0 + \sqrt{\frac{\tau\Delta_{0s}}{2}}x\right]\right)\right\}. \tag{38}$$

The dynamics of the variance Δ_{0s} and the two-time autocorrelation function can be derived by a system of integro-differential equations (see Sec. II. 2 in the Supplemental Material of Ref. [51] for a derivation in a closely related model). Here, we choose a simpler approach and approximate the dynamics of Δ_{0s} and Δ_{1s} by a system of ordinary differential equations (see Appendix B for more details) whose steady states are given by

$$\begin{aligned}
&-\Delta_{0s}^2 + \frac{\tau}{A^2}\int Dx\{\Phi^2(A[\sqrt{\Delta_{0s}}x + m_0 + e^{-s/\tau}m_s]) + \Phi^2(A[\sqrt{\Delta_{0s}}x + e^{-s/\tau}m_s - m_0])\} = -\Delta_{1s}^2 \\
&+ \frac{\tau}{A^2}\int Dz\left\{\left[\int Dx\Phi(A[\sqrt{\Delta_{0s} - |\Delta_{1s}|}x + \sqrt{|\Delta_1|}z + m_0 + e^{-s/\tau}m_s])\right]^2\right. \\
&\left.+ \left[\int Dx\Phi(A[\sqrt{\Delta_{0s} - |\Delta_{1s}|}x + \sqrt{|\Delta_{1s}|}z + e^{-s/\tau}m_s - m_0])\right]^2\right\}
\end{aligned} \tag{39}$$

and

$$\begin{aligned}
\Delta_{1s} = \frac{\tau}{4}\int Dz\left\{\left[\int Dx\phi(A[\sqrt{\Delta_{0s} - |\Delta_{1s}|}x + \sqrt{|\Delta_{1s}|}z + m_0 + e^{-s/\tau}m_s])\right]^2\right. \\
\left.+ \left[\int Dx\phi(A[\sqrt{\Delta_{0s} - |\Delta_{1s}|}x + \sqrt{|\Delta_{1s}|}z + e^{-s/\tau}m_s - m_0])\right]^2\right\}.
\end{aligned} \tag{40}$$

Equations (37)–(40) are solved using the method described in Appendix B. This is the method used to compute the order parameters m_0, m_s, Δ_{0s}, and Δ_{1s} in Figs. 5(a) and 5(b) (green and blue lines) and in Fig. 6(a) (full lines).

To obtain the instability line in Figs. 4(a) and 6(b) (red line), similar to Ref. [52], we neglect the dynamics of the Δ parameters, and expand the overlap equations [Eqs. (37) and (38)] around the memory state m_s and $m_0 = 0$, leading to

$$\int Dx\phi'\left(A\left[e^{-s/\tau}m_s + \sqrt{\frac{\tau\Delta_{0s}}{2}}x\right]\right) = \frac{1}{A}. \quad (41)$$

Our network simulations show that this approximation matches well the network dynamics (see Figs. 5 and 6) suggesting that the observed memory state instability is mainly governed by the overlap dynamics in Eqs. (37) and (38).

IV. DYNAMICS IN THE ABSENCE OF FORGETTING

To understand the network's different dynamical regimes and how the degree of sparsity c affects the match between the DMFT and numerical simulations, we start by analyzing a simpler scenario, namely, the dynamics of this network in the limit where there is no forgetting ($\tau \to \infty$). In this scenario, we need to consider a finite number of presented patterns p, since the variance of synaptic weights diverges in the limit $p \to \infty$, and the synaptic connectivity can be written as

$$J_{ij} = \frac{Ac_{ij}}{Nc}\sum_{\mu=1}^{p}\eta_i^\mu\eta_j^\mu. \quad (42)$$

Our network model, therefore, coincides with the model studied by Tirozzi and Tsodyks [21] (referred to as the TT model in the following), who provided an analytical description of the network qualitative different behaviors in the sparse coding limit. In the fully connected case $c_{ij} = 1$, it coincides with the model introduced by Hopfield [53] and studied analytically by Kuhn *et al.* [54]. In this limit, the sum on the rhs of Eq. (10) becomes a sum over p patterns, and κ becomes equal to the memory load

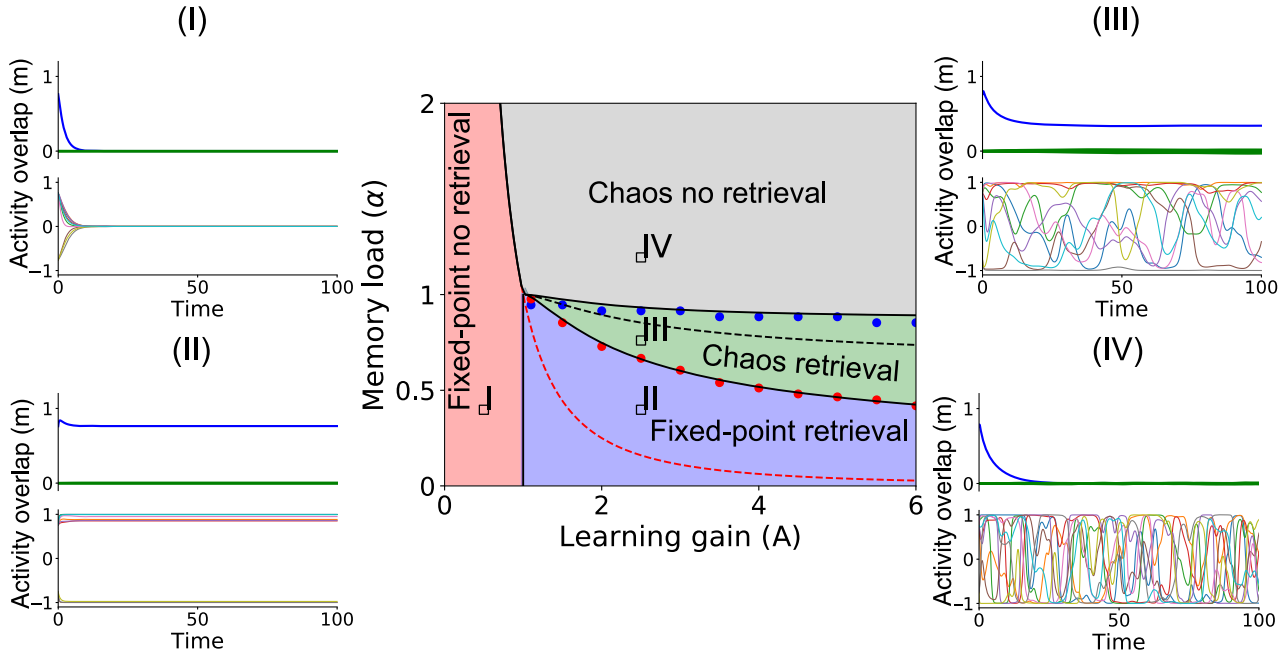

FIG. 1. Center: bifurcation diagram for the TT model in the parameter space spanned by memory load (α) versus the learning gain (A). Full lines: boundaries of the four qualitatively different regions calculated using DMFT. The boundaries between fixed-point and chaotic retrieval states are computed using Eqs. (22)–(24). The capacity curve that separates chaotic nonretrieval states from retrieval states is computed using Eqs. (29) and (30). The vertical line that separates nonretrieval from retrieval states are computed using Eqs. (22) and (23). Red circles: location of the transition to chaos of retrieval states, using simulations of networks of $N = 10^7$ units and $K = 2\log(N)$ connections per neuron. Blue circles: storage capacity, computed using simulations. Dashed red line: transition to chaos of the background state [see Eq. (26)]. Black dashed line: line on which (unstable) fixed-point retrieval states disappear [see Eqs. (22) and (23)]. Surrounding panels, labeled I–IV, show representative numerical simulations in the four qualitatively different regions. In each simulation, the network is initialized close to one of the stored memories. Each plot shows the overlaps of network state with this memory (blue) and other memories (green), and the activity of 10 randomly selected neurons as a function of time. The network parameters A and α are indicated with a square at the bottom left-hand corner of the corresponding roman numbers. The network parameters for the surrounding panels are $N = 10^6$ and $K = 2\log(N)$. The parameter values for A and α for panels I–IV are, respectively, $A = 0.5, 2.5, 2.5, 2.5$ and $\alpha = 11/K, 11/K, 21/K, 33/K$.

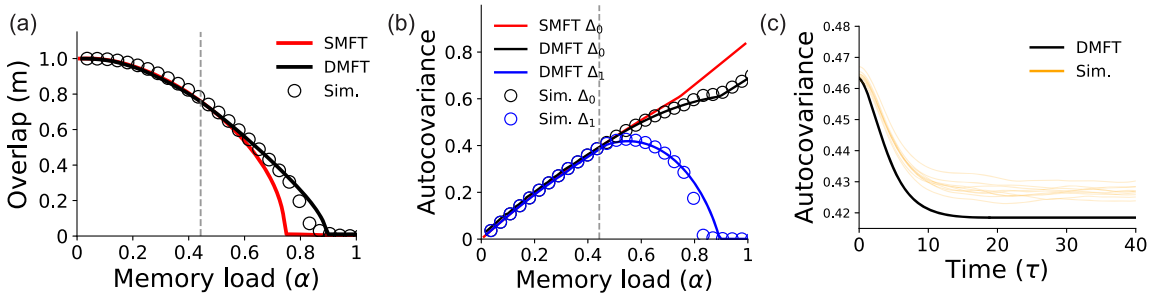

FIG. 2. (a) Overlap versus memory load. Circles: overlaps computed from network simulations (average over 10 realizations). Dashed vertical line: transition to chaos [see Eqs. (22) and (24)]. Black lines: overlaps computed from DMFT [see Eqs. (22), (27), and (28)]. Red lines: overlaps in static solutions to DMFT (SMFT) [see Eqs. (22) and (23)] which are unstable beyond the transition to chaos. (b) Parameters characterizing the autocovariance function Δ_{0s} and Δ_{1s}. As in (a), full lines represent solutions from DMFT [see Eqs. (22), (27), and (28)], while circles represent simulation results. (c) Autocovariance function for $\alpha = 0.5429$ (i.e., $\alpha = 15/K$) [solution of Eq. (20)]. Black, DMFT; yellow, 10 realizations of network simulations. Other parameters: $N = 10^6$, $K = 2\log(N)$, and $A = 5.5$.

$\alpha \coloneqq \kappa = p/K$, i.e., the number of stored patterns divided by the average number of connections. Here we recapitulate the analytical results provided by Tirozzi and Tsodyks and complement them with simulations of large networks to investigate how well the theory matches the results of such simulations at various degrees of sparsity (see Figs. 1–3).

In the TT model, the network dynamics depends on two parameters, the memory load $\alpha = p/K$, i.e., the number of stored patterns scaled by the average number of connections, and A, the strength by which the patterns are imprinted in the connectivity when learned. Figure 1 shows the network bifurcation diagram derived from the DMFT delineating the four different dynamical regimes in parameter space, in which the four qualitatively different types of states described in Sec. III exist: (I) fixed-point background state; (II) fixed-point memory states [i.e., states with a finite overlap $m \sim \mathcal{O}(1)$ with the stored patterns]; (III) chaotic memory states [i.e., with a finite overlap $m \sim \mathcal{O}(1)$ with stored patterns]; (IV) chaotic background state. For small values of the update strength A, the network is weakly coupled, and only the background state, in which the firing rates of all neurons are equal to zero, exists. In this regime, memory retrieval is not possible, and the overlap decays to zero when the network is initialized with any of the stored patterns (see red region I in Fig. 1). For larger values of A and small memory load, any of the stored patterns can be retrieved as fixed-point attractor states. When the network is initialized close to one of the stored patterns, the network goes to a fixed point that is correlated with that pattern (see region II in blue in Fig. 1). Within this parameter region, the red dashed line in Fig. 1 indicates the transition to chaos of the background state. Below the red dashed line, for small memory load, the background state is a fixed point while for larger memory loads it is a chaotic state. Larger memory loads lead to a transition to chaos of the memory states (green region III in Fig 1). In this regime, the network dynamics are chaotic, but it retains a finite overlap with the retrieved memory. Therefore, in spite of chaotic fluctuations, the population activity is stably correlated with the retrieved pattern. Finally, if the memory load further increases, none of the stored patterns can be retrieved, and all memories are forgotten. This is a

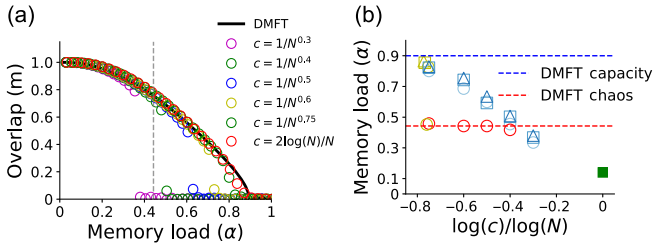

FIG. 3. Effect of connectivity sparseness on the dynamics of the TT model. (a) Overlap versus memory load, for different scalings of connection probability with network size N. Circles: overlap averaged time after transients and over 10 realizations. Horizontal dashed line: transition to chaos [see Eqs. (22)–(24)]. Black line: overlap computed from DMFT [see Eqs. (22), (27), (28)]. Network sizes corresponding to the scalings $c = 1/N^{0.3}, 1/N^{0.4}, 1/N^{0.5}, 1/N^{0.6}, 1/N^{0.75}, 2\log(N)/N$ are $N = 10^5, 2.5\times10^5, 5\times10^5, 10^6, 10^6, 10^6$, respectively. (b) Capacity versus $\log(c)/\log(N)$ [blue symbols for power-law scalings, olive symbols for the $\log(N)/N$ scaling] and transition to chaos versus $\log(c)/\log(N)$ (red circles). The capacity is estimated from network simulations as the minimal memory load for which the overlap is smaller than 0.1. Dashed lines: DMFT blue and red lines correspond to the capacity [see Eqs. (29) and (30)] and transition to chaos [see Eqs. (22)–(24)] predicted by the DMFT, respectively. Green square: capacity in the fully connected case [54] which is equal to the capacity of the Hopfield model (0.14). For each scaling, three symbols correspond to three different networks sizes: $N = \{0.15\times10^5, 0.6\times10^5, 1\times10^5\}$ for $c = 1/N^{0.3}$; $N = \{0.5\times10^5, 1.5\times10^5, 2.5\times10^5\}$ for $1/N^{0.4}$; $N = \{1.5\times10^5, 2\times10^5, 5\times10^5\}$ for $1/N^{0.5}$; $N = \{2\times10^5, 10\times10^5, 20\times10^5\}$ for $1/N^{0.6}$; $N = \{5\times10^5, 10\times10^5, 15\times10^5\}$ for $1/N^{0.75}$; and $N = \{6\times10^5, 20\times10^5, 30\times10^5\}$ for $2\log(N)/N$. For each scaling, circle, square, and triangle correspond to smaller, intermediate, and larger network sizes, respectively. The learning gain is $A = 5.5$.

well-known phenomenon in attractor neuronal networks called catastrophic forgetting. The value of the memory load when this happens is called the memory capacity. Beyond the memory capacity, the only stable attractor is the chaotic background state (the gray region in Fig. 1). The dashed black line in Fig. 1 corresponds to the memory load at which (unstable) fixed-point retrieval states disappear. Interestingly, this line is below the line at which chaotic retrieval states disappear (compare the dashed line with the solid line between the green and gray region in Fig. 1). Thus, chaotic fluctuations allow the network to increase its storage capacity of the network.

Figure 2 compares the results from DMFT with numerical simulations of a very large and sparse network [$N = 10^6$, $K = 2\log(N)$]. For these parameters the DMFT is in excellent agreement with numerical simulations, as shown by comparing the overlaps in memory states [Fig. 2(a)], parameters of the autocovariance [Fig. 2(b)], and the full autocovariance function [Fig. 2(c)]. We next numerically investigate how our theory compares with simulations, for different sparsity scalings with the network size. In networks in which the average number of connections scale logarithmically with the network size [i.e., $K \sim \log(N)$ and $c \sim \log(N)/N$], the storage capacity computed from network simulations is very close to the capacity predicted by the DMFT [see Figs. 3(a) and 3(b)]. However, in the range of network sizes we simulated ($10^4 < N < 10^6$), denser scalings gradually decrease the capacity of the network, and consequently, the DMFT capacity predictions become less accurate [see Figs. 3(a) and 3(b)]. Our simulations show that the capacity depends strongly on the scaling of connection probability with N, but very weakly on N, when N is varied with a fixed scaling between c and N [see Fig. 3(b)]. For all scalings, the increase in network size leads to little changes in the capacity in a large range of network sizes. These results suggest the possibility that in our network, the storage capacity depends purely on the scaling between connection probability and network size in the large N limit. However, our numerical simulations do not allow us to exclude the possibility that finite size effects may cause an extremely slow increase in storage capacity toward the value computed with DMFT with network size, provided connection probability tends to zero as N goes to infinity. The transition to chaos is well predicted by DMFT in networks in which the connection probability $c = 1/N^a$ with $a \geq 0.4$. For less sparse connectivity, retrieval states disappear before the transition to chaos is reached, and therefore chaotic retrieval states no longer exist. For completeness, we also plot in Fig. 3(b) the capacity in fully connected networks obtained analytically by Kuhn *et al.* [54].

V. DYNAMICS OF THE NETWORK WITH FORGETTING

We now turn to a network where connectivity is learned using an online Hebbian synaptic plasticity rule, Eq. (2). In this rule, recent patterns are imprinted in network connectivity using the covariance between presynaptic and postsynaptic activity, while older patterns are forgotten with a decay timescale $\tau = -1/K\log(\rho)$. We assume that an infinite number of patterns has been presented to the network, leading to a connectivity matrix given by Eq. (4).

In this network, the retrieval dynamics of each memory state depends on its age, and there is a continuum of statistically distinguishable memory states parametrized by the memory age (see Sec. III). In a given memory state s, the mean input current depends exponentially on the age of the pattern [see Eq. (11)]. Because of the implicit dependence of the input current autocovariance function on the mean input current [see Eq. (19)], both the mean and the autocovariance of the synaptic input currents h_i vary with the age of the memory. For sufficiently large A, we

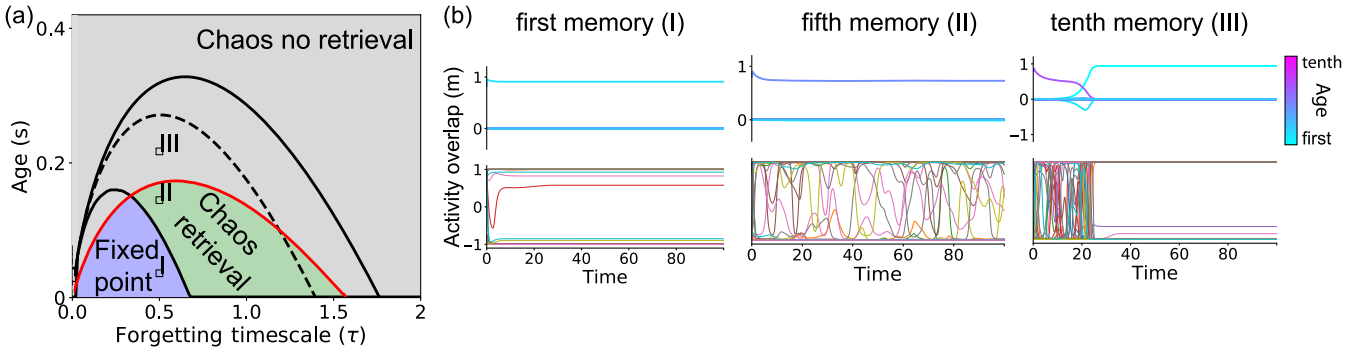

FIG. 4. (a) Bifurcation diagram for the network with forgetting, for $A = 10$. Solid lines: transition to chaos [see Eqs. (22)–(24)] and maximal age at which retrieval states exist, calculated using the DMFT [see Eqs. (29) and (30)]. Dashed line: maximal age at which static retrieval solutions exist (SMFT) [see Eqs. (22) and (23)]. Red line: maximal age at which retrieval states are stable [see Eqs. (39)–(41)]. (b) Retrieval dynamics of memories of age 1, 5, and 10 for the same realization of the connectivity matrix [$N = 10^6$, $K = 2\log(N)$, $A = 10$, $\tau = 0.5$]. These ages correspond to $s = 1/K, 5/K, 10/K$, respectively, indicated with a square at the bottom left-hand corner of the corresponding roman numbers; see I–III in (a).

identify using DMFT three different regions depending on the age s and the memory timescale (Fig. 4): a region in which retrieval states are fixed-point attractors, another region in which retrieval states are chaotic attractors, and finally a region in which no retrieval is possible. Depending on the forgetting timescale, different scenarios are possible as memories age. For short forgetting timescales [$\tau < 0.34$ in Fig. 4(a)], most recent memories are retrieved as fixed-point attractors [blue region in Fig. 4(a) and 4(b)(I)]. As memories age, they become unstable (red line in Fig. 4), and the network typically retrieves one of the most recently learned patterns instead. For larger forgetting timescales ($0.34 < \tau < 0.68$ in Fig. 4), most recent memories are also retrieved as fixed-point attractors. However, unlike for shorter forgetting timescales, fixed-point attractor states become chaotic retrieval states after a certain age [above the line separating blue and green regions in Fig. 4; see an example in Fig. 4(b)(II)]. As memories further age, there is a second transition line where the chaotic attractor becomes unstable, and the network retrieves one of the most recent memories instead [see Fig. 4(b)(III)]. For even larger forgetting timescales ($0.68 < \tau < 1.58$), fixed-point attractor states no longer exist, and even the most recent memories are retrieved as chaotic attractors. Finally, when $\tau > 1.58$, the network is no longer able to retrieve any memory. Note that the maximal capacity (defined as the maximal age at which memories can be retrieved) is given by the red line in Fig. 4. Above this line, retrieval states still exist in a finite region of parameter space, but they are unstable, and the network retrieves more recent patterns instead. Note also that the capacity is optimized at an intermediate value of $\tau \sim 0.6$ for which both types of retrieval states coexist in the network phase space (fixed-point attractors for recent memories, chaotic attractors for older ones). For this value of τ, the capacity is about $s_{\max} = 0.18$, which is significantly lower than the capacity in the TT model.

Numerical simulations in large, very sparsely connected networks show a good agreement with DMFT, as shown in Fig. 5 in a network with $A = 4$ and $\tau = 0.64$. Figures 5(a) and 5(b) show that both the overlap with the retrieved pattern and the variance Δ_{0s} decay with age. At the age of about $s = 0.18$, retrieval states become unstable, and the network retrieves one of the most recent memories instead. Figures 5(c) and 5(d) show examples of successful retrieval [Fig. 5(c)] and of an unsuccessful retrieval ending in the retrieval of the most recent memory instead [Fig. 5(d)]. Figure 6(a) shows how the overlaps with retrieved memories depend on the forgetting time constant τ. It shows that the overlaps in retrieval states of recent memories decay with τ, since increasing τ means older patterns provide increasing interference with the retrieval of recent patterns. Finally, Fig. 6(b) shows that the location of the two instability lines of retrieval states (in red, the instability toward more recent memories, and in blue, the instability toward chaotic retrieval states) predicted by DMFT are in good agreement with numerical simulations.

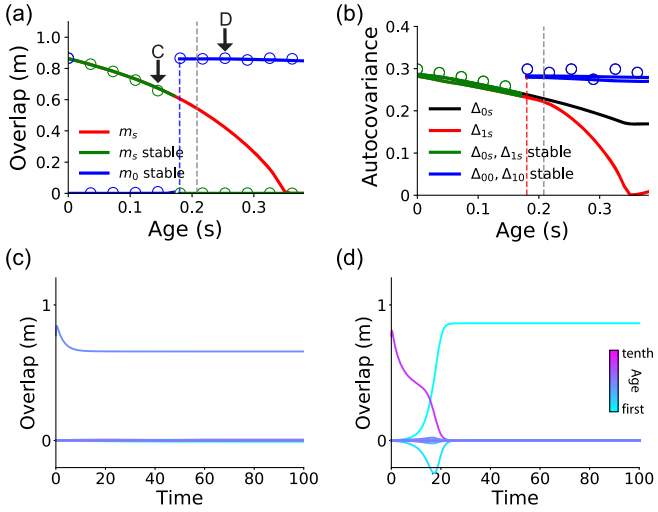

FIG. 5. (a) Overlap versus memory age s in the network with online learning. Green line (green circles): overlap with a memory of age s after the network is initialized close to the corresponding state, calculated from DMFT [see Eqs. (37)–(40)] (network simulations). Blue line and blue circles: overlap with a recent memory that is retrieved instead of the memory of age s. The DMFT blue line corresponds to Eqs. (37)–(40) for $s = 0$. Red line: overlap with a memory of age s in the unstable retrieval state [see Eqs. (22), (27), and (28)]. Vertical dashed gray line: transition to chaos [see Eqs. (22)–(24)]. Vertical dashed blue line: age at which memories become unstable. In simulations, averages are over time and over 10 realizations. (b) Autocovariance parameters versus memory age. Green line (green circles): $\Delta_{0s} = \Delta_{1s}$ computed from DMFT (simulations). Blue line (blue circles): $\Delta_{00} = \Delta_{10}$ computed from DMFT (simulations). Black line and red line: $\Delta_{0s} = \Delta_{1s}$ in unstable retrieval state (DMFT) (c) Dynamics of overlaps for a memory of age $s = 5/K = 0.18$. The memory is retrieved successfully. (d) Dynamics of overlaps for a memory of age $s = 8/K = 0.29$. The memory is not retrieved, as the network goes instead to the attractor state corresponding to the most recent memory. In both (c) and (d), overlaps are color coded by age. Network parameters: $N = 10^6$, $K = 2\log(N)$, $\tau = 0.64$, and $A = 4$.

VI. TOWARD MORE REALISTIC NETWORKS

Do the above results for the dynamics of a network with forgetting hold in a more biologically realistic setting? We numerically investigate this question by analyzing a network whose transfer function and learning rule are both inferred from *in vivo* data, as in Ref. [22]. To this network, we add forgetting and noise. In this model, the synaptic input currents obey standard rate equations,

$$\frac{dh_i(t)}{dt} = -h_i(t) + \sum_{i \neq j}^{N} c_{ij} J_{ij} \phi(h_j(t)) + \sqrt{2\sigma_z^2} z_i(t), \quad (43)$$

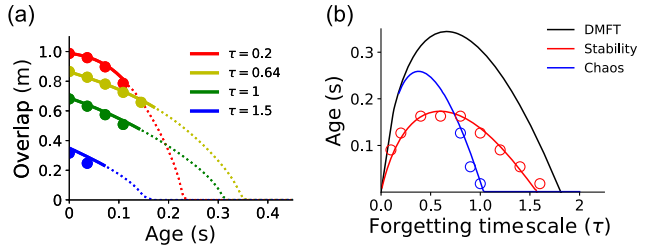

FIG. 6. (a) Overlap versus memory age (s) for different values of the forgetting timescale τ, and $A = 4$. Full lines: overlap computed using DMFT in stable retrieval states [see Eqs. (37)–(40)]. Dashed lines: unstable retrieval states [see Eqs. (22), (27), and (28)]. Circles: simulations (average over time and over 10 realizations). (b) Bifurcation diagram for the network with forgetting, for $A = 4$. Black line: maximal age at which retrieval states exist, calculated using DMFT [see Eqs. (29) and (30)]. Blue line (blue circles): transition to chaos of retrieval states [DMFT [see Eqs. (22)–(24)] (simulations)]. Red line (red circles): maximal age at which retrieval states are stable [DMFT [see Eqs. (39)–(41)] (simulations)]. Simulation parameters: $N = 10^6$, $K = 2\log(N)$.

with two key differences from the previous model described by Eq. (1). First, the input-output transfer function is now a strictly positive sigmoidal function given by

$$\phi(h) = \frac{r_m}{1 + e^{-\beta_m(h-h_m)}}, \tag{44}$$

whose parameters are inferred from data as described in Refs. [22,55]. The maximal firing rate is $r_m = 76.2$ Hz, the slope at the inflection point is $\beta_m = 0.82$, and the input current at the inflection point is $h_0 = 2.46$.

The second difference is that this network is subject to noise, described by the last term in the rhs of Eq. (43), with intensity σ_z.

The network connectivity is now given by

$$J_{ij} = \frac{A}{K}\sum_{\mu=0}^{\infty} e^{-\mu/\tau K} f(\eta_i^\mu) g(\eta_j^\mu), \tag{45}$$

where the functions f and g describe the dependence of the learning rule, respectively, the postsynaptic and presynaptic dependence of a learning rule. In our network, these are given by two step functions:

$$f(\eta) = \begin{cases} q_f & \text{if } x_f \le \eta \\ -(1-q_f) & \text{if } \eta \le x_f, \end{cases} \tag{46}$$

$$g(\eta) = \begin{cases} q_g & \text{if } x_g \le \eta \\ -(1-q_g) & \text{if } \eta \le x_g. \end{cases} \tag{47}$$

The random variables η_i^μ are distributed as $\eta_i^\mu \overset{\text{i.i.d.}}{\sim} \phi(x_i^\mu)$, where the x_i^μ are normally distributed; i.e., $x_i^\mu \overset{\text{i.i.d.}}{\sim} N(0,1)$.

The parameters corresponding to the threshold of the postsynaptic dependence of the learning rules x_f and its

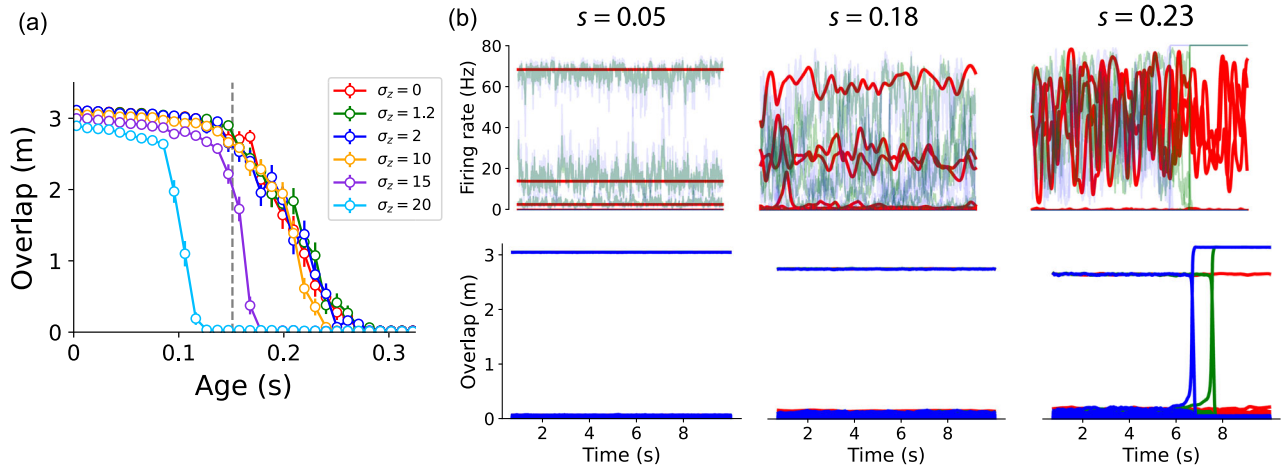

FIG. 7. Dynamics and robustness of a network with transfer function and learning rules inferred from *in vivo* data, with forgetting, and external noise. Dynamics and robustness of a network with transfer functions and learning rules inferred from *in vivo* data, with forgetting, and with external noise. (a) Overlap versus memory age for different values of the standard deviation of external noise σ_z [see Eq. (43)]. The overlap is calculated by averaging over 50 network realizations for each value of σ_z. Error bars: standard error of the mean. The gray vertical dashed line indicates the transition to chaos for $\sigma_z = 0$ estimated from numerical simulations. (b) Dynamics of 10 randomly chosen neurons (top row) and of the overlaps with 200 most recent patterns (bottom row) for three different values of the standard deviation of the noise $\sigma_z = 0$ (red), $\sigma_z = 1.2$ (green), and $\sigma_z = 2$ (blue). Each column corresponds to the retrieval of a memory of age indicated at the top of the panel. The same network connectivity is used in (b) for all network simulations. Note that for $s = 0.23$, the noise destabilizes the corresponding retrieval state, and the network goes to an attractor corresponding to a more recent memory. Parameters: $N = 150\,000$, $c = 1/\sqrt{N}$.

offset q_f are inferred from *in vivo* data as in Refs. [22,55]. The values used here $x_f = 26.6$ and $q_f = 0.83$ correspond to median values of the inferred parameters. As in Ref. [22] we assume that $x_g = x_f$. It is also assumed that the average over the patterns statistics of g is zero, i.e., $\langle g \rangle = 0$, which constrains the parameter q_g. This assumption stems from the requirement of stability when learning many patterns. From a biological standpoint, this assumption could be interpreted as a homeostatic mechanism that prevents the runaway of the average synaptic weights during learning. The value of A is set to be 3 times the median of the inferred values, $A = 10.65$. This choice leads to chaotic dynamics in this model [22]. Lastly, the sparsity is set as $c = 1/\sqrt{N}$.

In the absence of noise ($\sigma_z = 0$), the network exhibits qualitatively similar retrieval dynamics as the one observed in the simpler network with tanh transfer function and binary patterns described in the previous section Sec. V [see Figs. 7(a) and 7(b), red traces]. Newer memories are retrieved as fixed-point attractors, while moderately older memories are retrieved as chaotic attractors. Firing rates are consistent with observations in delay periods in electrophysiological experiments (see Ref. [22]).

In the presence of noise ($\sigma_z > 0$), there is a large range of values of σ_z where the network storage capacity is hardly affected [see Fig. 7(a)], in spite of strong fluctuations at a single neuron level [compare Fig. 7(b) the red ($\sigma_z = 0$) with the green ($\sigma_z = 1.2$) and blue ($\sigma_z = 2.0$) traces].

The instability observed in the previous model, in which moderately old retrieval states become unstable and these memories are overtaken by the most recent ones (see Fig. 5), is also observed in this network. Interestingly, this can happen either in the absence of noise (not shown) or in the presence of noise [see Fig. 7(b), $s = 0.23$].

Overall, these simulation results show that all the aspects of the retrieval dynamics characterized analytically in the previous section hold qualitatively in a more realistic network whose parameters are constrained by *in vivo* data. Additionally, these results are highly robust to noise.

VII. DISCUSSION

In the present work, we have characterized using dynamical mean field theory the dynamics of an attractor network that learns and forgets through an online Hebbian learning rule, and shown that fixed-point and chaotic memory states coexist. In such a network, recent memories are strongly imprinted in the connectivity, while older memories are exponentially forgotten. In the limit of no forgetting, the network undergoes a global transition to chaos, where all memory states transition to chaos at once when the number of stored patterns surpasses a critical value [see Fig. 8(a)]. When forgetting takes place, memory states are heterogeneous and form a continuum of statistically different memory states. In a broad range of forgetting timescales, recent memories are retrieved in the network as fixed-point attractors, while older memories are retrieved as chaotic attractors [see Fig. 8(b)]. The magnitude of the chaotic fluctuations increases parametrically with the memory age, and the basin of attractions shrinks with age. Interactions between most recent and older memories destabilize older memory states [see Fig. 8(c)]. Our stability analysis explains this effect and accurately predicts the capacity curve, i.e., the curve in parameter space when older memory states destabilize. We contrast our results with simulations of large networks and find that our theory matches well the network dynamics, provided synaptic connectivity is sparse enough. Lastly, we also studied the dynamics of a more biologically realistic network, with learning rules and transfer functions inferred from *in vivo* data, with forgetting and in the presence of noise. Notably, we found that all the features of the retrieval dynamics characterized analytically in our simpler model with forgetting hold qualitatively in this more biologically realistic model. These results are highly robust to noise, and the memory capacity is only weakly affected by levels of

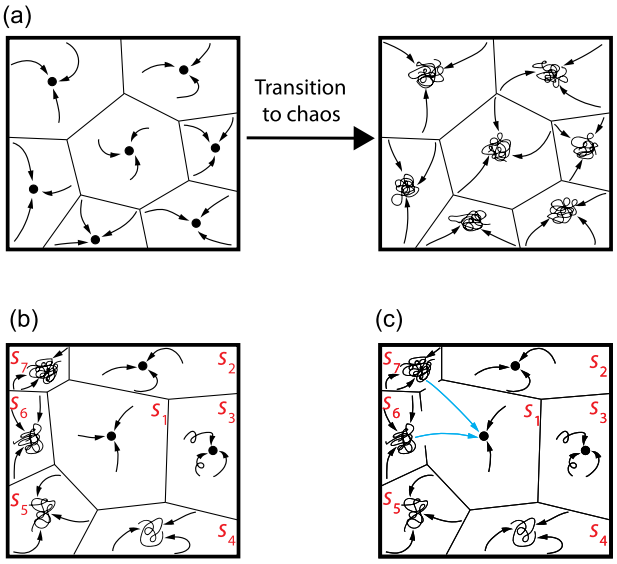

FIG. 8. Schematics of the attractor networks phase space. (a) Extensive transition to chaos of fixed-point attractor memory states in an attractor network with no forgetting. When the memory load α surpasses a critical point all the fixed-point attractors transition to chaos at once. For a given value of the memory load α, all fixed-point or chaotic memory states are statistically equivalent. (b) In an attractor network that forgets, memory states are heterogeneous. Most recent memories are retrieved as fixed-point attractors with a larger basin of attraction, while older memories are chaotic with smaller basins of attraction. The chaotic temporal fluctuations become faster for older memories. (c) As predicted by our stability analysis, older memory states become unstable after a critical memory age, and the network jumps to the most recent memory state (see the light blue arrows).

noise that lead to strong fluctuations at a single neuron level. These results highlight that the theory presented here is robust and holds in biologically realistic scenarios. Overall, our work puts forward a network mechanism that might explain the diversity of neuronal responses in the cortex during memory tasks and provides an analytical framework for dissecting the network mechanisms of such heterogeneity.

A. Experimental predictions

For decades, neuroscientists have used delayed response tasks to investigate the neural mechanisms underlying the maintenance and manipulation of information stored in memory. In these tasks, subjects must keep in memory information about a previously presented stimulus during a delay period, to be able to produce a behavioral response that depends on this information. Neuronal recordings in delay response tasks show that a (typically small proportion) of neurons display elevated tonic activity during the delay period [5–8]. The observed persistent elevated delay activity is consistent with attractor network models in which fixed-point attractors correspond to the memory states [56,57]. However, in many of these recordings, neurons exhibit firing patterns with a high degree of temporal variation. These observations challenge attractor networks as viable models of memory storage [58]. Interestingly, population analysis of neural activity during delay response tasks shows that in spite of the variability observed at a single neuron level, there is a stable population encoding of the memoranda during the delay periods [16]. A possible scenario compatible with the above observations [22,59] is that the single neuron temporal variability and stable population encoding can be explained by attractor networks in the chaotic regime as described here for the TT model in Sec. IV [see Fig. 1, green region, and Fig. 8(a)]. In an attractor network with forgetting, memory states are heterogeneous, and as a memory ages, its activity ranges from fixed point to chaotic. Additionally, for chaotic retrieval states, the temporal fluctuations of their activity get faster with age [see Fig. 6(b)]. Our theory thus predicts that for delay response tasks with aging memoranda (1) retrieval states corresponding to newer memories will present a large proportion of neurons with persistent delay activity consistent with fixed-point attractor dynamics, (2) temporal fluctuations will become more pronounced, and faster, as memories become older, (3) at the behavioral level, the probability of making an error should increase with age, and (4) in error trials with old memoranda, the activity at the population level should be consistent with a transition from a highly irregular state at the beginning of the trial, to a fixed point attractor state corresponding to a more recent memorandum. These predictions can be tested by recording neural activity during delay response tasks performed across the learning and forgetting process of multiple items.

B. Toward more biophysically realistic networks

The TT model has a number of biophysically unrealistic features. The firing rate of neurons ranges from -1 to 1, while in reality the firing rate should be a non-negative quantity. Also, in this model, as well as in the Hopfield model, neurons have the same probability of being active or inactive in any given pattern, while in reality, the probability that neurons are highly active in a given pattern is much smaller than one, and experimentally recorded distributions of firing rates show no evidence of bimodality. We have recently introduced and studied an attractor network similar to the one presented here, but with both transfer function and synaptic plasticity rules inferred from *in vivo* data [22]. We had previously shown that this network displays qualitatively similar dynamics as the TT model. For small learning gain A, and memory loads α, memory states are fixed-point attractors, while for large A and α memory states are chaotic (see Figs. 4 and 6, respectively, in Ref. [22]). Here, we incorporated forgetting to the network in Ref. [22] to evaluate the robustness and biological plausibility of our theory in the online learning case. We found that all aspects characterized analytically in the simpler network with forgetting in Sec. V also hold qualitatively in this more realistic network. Furthermore, these results are highly robust to noise. These results highlight the robustness of our analytical results, showing that they qualitatively hold in more realistic networks with parameters constrained by *in vivo* recordings and are also resilient to noise.

Further analyzing this new network model could help to clarify the network mechanisms underlying the observed changes in neural activity across days during learning novel images [60,61].

C. Relation with other chaotic networks

Random networks of rate units are a popular class of tractable models for exploring the neural mechanisms underlying the temporal fluctuations and heterogeneity of the firing rate patterns observed in the brain. Transition to chaos has been extensively studied in randomly connected networks [46,49,62,63] and the analysis has been extended to multiple scenarios such as networks with multiple populations [48,64,65], networks of clusters of recurrently connected neurons [66], networks driven by external stimuli [67,68], and random networks partially structured with low-rank connectivity [69–71]. The effect of chaos on the computational capabilities of recurrent networks has also been studied extensively [68,72–75]. In the vast majority of the above studies, the focus has been on the transition to chaos of a single fixed point, the background (zero average activity) state. A notable exception is the work in Ref. [21]. In this work by Tirozzi and Tsodyks, they analytically study the highly diluted version of the Hopfield model in a network of continuous "firing rate" units. They derived a DMFT that fully characterizes the

network's dynamics. Their DMFT is a particular case of our model for infinite forgetting timescale $\tau \to \infty$ and number of patterns proportional to K, i.e., $p = \alpha K$. The DMFT for the network in Ref. [21] is recovered in our equations for $\alpha = \kappa = \tau/2$ and $s = 0$. Although Ref. [21] computed the phase diagram of the model analytically, no comparison with network simulations was provided. In Sec. IV of the present work, we complemented the DMFT in Ref. [21] with a systematic comparison between numerical solutions of the DMFT and simulations of large networks. We show that a good agreement between simulations and theory is only obtained for very sparse networks, in which $c \sim 1/N^a$ and $a > 0.8$. (see Sec. VII F). Our work differs from the above studies in that in our network there exists a coexistence between a continuum of statistically distinguishable fixed-point and chaotic memory states [see Figs. 8(b) and 8(c)].

D. Relation with other palimpsest networks

The problem of catastrophic forgetting was recognized soon after the Hopfield model was proposed, and this issue was addressed using various online learning rules [23–25] that lead to gradual forgetting of old memories, thereby successfully avoiding catastrophic forgetting. This class of models is sometimes referred to as palimpsest models. In all these models, the price to pay for avoiding catastrophic forgetting is a substantial decrease in storage capacity. This phenomenon is also present in our model, where the capacity is only about $s_{\max} = 0.18$ for $A = 10$, to be compared with a capacity of order 0.9 in the TT model.

Online learning rules have also been widely studied in networks with binary synapses [26–28,30–34,76,77]. In these models, synapses have two states, one with strong and the other with weak (or zero) efficacy. Transitions between states are induced stochastically by the activity of presynaptic and postsynaptic neurons. Under general conditions, these networks also behave as palimpsests, maintaining a memory of recently shown patterns but gradually forgetting patterns shown in the past. Most studies of such networks have used an ideal observer approach to estimate the maximal age at which a pattern can still be retrieved from synaptic connectivity [27,28,30,31,77]. Storage capacity has also been investigated in attractor networks with binary synapses, using either binary neurons [33,34,77] or firing rate units [32]. We expect that the results presented here about the diversity of retrieval states should hold in sparsely connected networks with such synapses, but the capacity should depend drastically on the implementation of the Markov process describing the transition between states.

E. Relation to random networks with low-rank structure

Recently, low-rank networks have been used as a framework for modeling the dynamics and computations in neuronal circuits in the brain [69–71,78,79]. In these networks, the connectivity is composed of a low-rank matrix plus a random matrix with i.i.d. Gaussian entries, i.e., $J_{ij} = \sum_{s=1}^{p} (A/N)\eta_i^s \xi_j^s + X_{ij}$, where $X_{ij} \overset{\text{i.i.d.}}{\sim} N(0, g^2/N)$ and $p \sim \mathcal{O}(1)$. Attractor networks and low-rank networks are very similar from a mean field perspective. In the attractor network analyzed here, for a given memory state s, in the large N and K limit, the result of our DMFT is equivalent to the following low-rank network: $(A/N)\eta_i^s \eta_j^s e^{-s/\tau} + (A\tau/2\sqrt{N})X_{ij}$, where $X_{ij} \overset{\text{i.i.d.}}{\sim} N(0,1)$. Therefore, the mean input current due to the low-rank component in low-rank networks is equivalent to the mean input current of the retrieved memory in attractor networks, while the random component is equivalent to the quenched noise that arises from the stored memories (see Ref. [79]). Because these networks are similar in the mean field limit, we hypothesize that by introducing heterogeneity in the strength of low-rank components, the coexistence between chaotic and fixed-point attractors displayed in our network can also be obtained in low-rank networks.

F. Network sparsity

The DMFT derived here, as well as the ones in Refs. [21,80], can be obtained using a path integral formalism, expanding the generating functional in powers of $c = K/N \ll 1$, and then neglecting the $\mathcal{O}(K^2/N^2)$ corrections [43]. However, the conditions on c and N for which these higher order corrections can be neglected remained unclear. Our numerical simulations show that our DMFT accurately predicts the network dynamics in very sparse networks with $c \sim \log(N)/N$ (Fig. 3). For denser connectivity, the agreement between capacity predicted by theory and simulations degrades progressively [see Figs. 3(a) and 3(b)]. On the other hand, the transition to chaos is well predicted by theory up to connection probabilities $c \sim 1/N^{0.4}$, at which point the chaotic region disappears altogether. These results are consistent with results on sparse networks of binary neurons where the disruptive effect of correlations between neurons on storage capacity can be avoided when $K \ll \log(N)$ [81]. However, our numerical simulations do not allow us to rule out the possibility that finite size effects may cause an extremely slow increase in capacity with network size for denser scalings.

This code is available at Ref. [82].

ACKNOWLEDGMENTS

U. P.-O. thanks the Swartz Foundation for its support and the Champaign Public Library for providing the physical space where an early version of this work was developed. N. B. was supported by R01MH115555, R01NS112917, ONR N00014-17-1-3004, and NSF IIS-1430296. J. A. was supported by DARPA Award No. D21AP10162. We thank anonymous reviewers whose comments have helped to improve the paper significantly.

APPENDIX A: MAXIMUM LYAPUNOV EXPONENT

Here we calculate the maximum Lyapunov exponent analytically similarly as in Ref. [48]. For a rigorous approach using the path integral formalism, see Ref. [49].

We consider that a small and infinitesimally brief perturbation is given to our network at time t':

$$\frac{dh_i(t)}{dt} = -h_i(t) + \sum_{i\neq j}^{N} c_{ij}J_{ij}\phi[h_j(t)] + h_i^0(t'). \tag{A1}$$

One can define the susceptibility of the network to this perturbation as

$$\chi_{il}(t,t') = \frac{\partial h_i(t)}{\partial h_l^0(t')}. \tag{A2}$$

From Eq. (A1), we obtain the dynamics for the susceptibility:

$$\frac{d\chi_{il}(t,t')}{dt} = -\chi_{il}(t,t') + \sum_{i\neq j}^{N} c_{ij}J_{ij}\phi'[h_j(t)]\chi_{jl}(t,t') + \delta_{il}\delta(t-t'). \tag{A3}$$

Let us now consider two replicas of the same system,

$$\left(1+\frac{d}{dt_a}\right)\chi_{il}(t_a,t_c) = \sum_{i\neq j}^{N} c_{ij}J_{ij}\phi'[h_j(t_a)]\chi_{jl}(t_a,t_c) + \delta_{il}\delta(t_a-t_c), \tag{A4}$$

$$\left(1+\frac{d}{dt_b}\right)\chi_{il}(t_b,t_d) = \sum_{i\neq j}^{N} c_{ij}J_{ij}\phi'[h_j(t_b)]\chi_{jl}(t_b,t_d) + \delta_{il}\delta(t_b-t_d), \tag{A5}$$

and define

$$G(t_a,t_b,t_c,t_d) \equiv \frac{1}{N}\sum_{i,j}\langle \chi_{ij}(t_a,t_c)\chi_{ij}(t_b,t_d)\rangle, \tag{A6}$$

where $\langle\cdots\rangle$ is the average over the quenched disorder of the stored patterns and the structural connectivity of the network c_{ij}.

Combining Eqs. (A4) and (A5) we obtain

$$\left(1+\frac{\partial}{\partial t_a}\right)\left(1+\frac{\partial}{\partial t_b}\right)G(t_a,t_b,t_c,t_d) = \frac{\tau A^2}{2}\langle\phi'(h_s(t_a))\phi'(h_s(t_b))\rangle G(t_a,t_b,t_c,t_d) + \delta(t_a-t_c)\delta(t_b-t_d). \tag{A7}$$

In this calculation, we have neglected the $\mathcal{O}(1/N)$ cross terms, and also assumed the independence of J_{il}^2 with $\phi'(h_s(t_a))$ and $\phi'(h_s(t_b))$.

From the definition of the potential in Eq. (21), it can be shown that the second derivative of the potential, which provides its concavity, is given by

$$\frac{\partial^2 V(\Delta_s,\Delta_{0s})}{\partial\Delta_s^2} = \frac{\tau A^2}{2}\langle\phi'(h_s(t_a))\phi'(h_s(t_b))\rangle - 1. \tag{A8}$$

Defining

$$\Gamma = t_a - t_b, \tag{A9}$$

$$\Gamma' = t_c - t_d, \tag{A10}$$

$$T = t_a + t_b, \tag{A11}$$

$$T' = t_c + t_d, \tag{A12}$$

Eq. (A7) becomes

$$\left(1+\frac{\partial}{\partial\Gamma}+\frac{\partial}{\partial T}\right)\left(1+\frac{\partial}{\partial T}-\frac{\partial}{\partial\Gamma}\right)G(T,T',\Gamma,\Gamma')=\left(1+\frac{\partial^2 V(\Delta_s,\Delta_{0s})}{\partial\Delta_s^2}\right)G(T,T',\Gamma,\Gamma')+\delta(T-T')\delta(\Gamma-\Gamma'). \quad \text{(A13)}$$

Further manipulating the above equation, we obtain

$$\left(\left[1+\frac{\partial}{\partial T}\right]^2-1-\frac{\partial^2}{\partial\Gamma^2}-\frac{\partial^2 V(\Delta_s,\Delta_{0s})}{\partial\Delta_s^2}\right)G(T,T',\Gamma,\Gamma')=\delta(T-T')\delta(\Gamma-\Gamma'). \quad \text{(A14)}$$

At this point, we notice that the following term,

$$H(\Gamma)\equiv-\frac{\partial^2}{\partial\Gamma^2}-\frac{\partial^2 V(\Delta_s,\Delta_{0s})}{\partial\Delta_s^2}, \quad \text{(A15)}$$

has the form of a quantum mechanical Hamiltonian where $-(\partial^2/\partial\Gamma^2)$ is analog to the kinetic energy and $-[\partial^2 V(\Delta_s,\Delta_{0s})/\partial\Delta_s^2]$ to the quantum mechanical potential. Then, a complete base must exist in Hilbert space such that

$$H(\Gamma)\psi_k(\Gamma)=E_k\psi_k(\Gamma). \quad \text{(A16)}$$

Since the $\psi_k(\Gamma)$ are a complete basis of the Hilbert space, we use the ansatz

$$G(T,T',\Gamma,\Gamma')=\sum_{l=0}^{\infty}C_l e^{\lambda_l(T-T')}\psi_l(\Gamma)\psi_l^*(\Gamma'). \quad \text{(A17)}$$

Plugging the above expression in Eq. (A14), we obtain

$$\sum_{l=0}^{\infty}[\lambda_l(\lambda_l+2)+E_l]C_l e^{\lambda_l(T-T')}\psi_l(\Gamma)\psi_l^*(\Gamma')=0. \quad \text{(A18)}$$

Therefore, we conclude that

$$\lambda_l=-1\pm\sqrt{1-E_l}. \quad \text{(A19)}$$

The minimum energy solution of the quantum mechanical problem or ground state of the Hamiltonian, which corresponds to the maximum λ_l, is given by

$$E_0=\frac{\partial^2 V(\Delta_s,\Delta_{0s})}{\partial\Delta_s^2}. \quad \text{(A20)}$$

The maximum Lyapunov exponent can be computed as [49]

$$\lambda=\lim_{t-t'\to\infty}\frac{1}{2(t-t')}\log\left[\frac{1}{N}\sum_{ij}\chi_{ij}^2(t,t')\right]. \quad \text{(A21)}$$

For large networks, the quantity $(1/N)\sum_{ij}\chi_{ij}^2(t,t')$ is self-averaging, leading to

$$\lambda=\lim_{t-t'\to\infty}\frac{1}{2(t-t')}\log\left[\frac{1}{N}\sum_{ij}\langle\chi_{ij}^2(t,t')\rangle\right]. \quad \text{(A22)}$$

Note that

$$G(2t,0,0,0)=\frac{1}{N}\sum_{ij}\langle\chi_{ij}^2(t,t')\rangle. \quad \text{(A23)}$$

Therefore, the maximum Lyapunov exponent is given by

$$\lambda=\lim_{t\to\infty}\frac{1}{2t}\log G(2t,0,0,0)\rangle=1\pm\sqrt{1+\frac{\partial^2 V(\Delta_s,\Delta_{0s})}{\partial\Delta_s^2}}. \quad \text{(A24)}$$

Thus, the transition to chaos is determined by the concavity of the classical potential $V(\Delta_s,\Delta_{0s})$ defined in Eq. (20). For a nonchaotic retrieval state we have that $\lambda<0$ and the potential $V(\Delta_s,\Delta_{0s})$ is concave; i.e., $[\partial^2 V(\Delta_s,\Delta_{0s})/\partial\Delta_s^2]<0$. For a chaotic retrieval state we have that $\lambda>0$ and the potential $V(\Delta_s,\Delta_{0s})$ is convex; i.e., $[\partial^2 V(\Delta_s,\Delta_{0s})/\partial\Delta_s^2]>0$. The transition to chaos is given by the critical point $[\partial^2 V(\Delta_s,\Delta_{0s})/\partial\Delta_s^2]=0$. This critical point is calculated as explained in Sec. III B.

APPENDIX B: RELAXATION DYNAMICS FOR Δ_{0s} AND Δ_{1s}

For computing the order parameters m_0, m_s, Δ_{0s}, and Δ_{1s} in Figs. 5(a) and 5(b) (green and blue lines), we replace the full integro-differential dynamics of the variance Δ_{0s} and the two-time autocorrelation function (see Sec. II.2 in the Supplemental Material of Ref. [51] for the corresponding equations in a closely related model) by a relaxational dynamics for Δ_{0s} and Δ_{1s} given by

$$\frac{d\Delta_{1s}}{dt} = -\Delta_{1s} + \frac{\tau}{4}\int Dz\left\{\left[\int Dx\phi(A[\sqrt{\Delta_{0s} - |\Delta_{1s}|}x + \sqrt{|\Delta_{1s}|}z + m_0 + e^{-s/\tau}m_s])\right]^2 + \left[\int Dx\phi(A[\sqrt{\Delta_{0s} - |\Delta_{1s}|}x + \sqrt{|\Delta_{1s}|}z + e^{-s/\tau}m_s - m_0])\right]^2\right\} \tag{B1}$$

and

$$\frac{d\Delta_{0s}}{dt} = -\Delta_{0s} + \sqrt{\Delta_{1s}^2 + \psi_1(m_0, m_s, \Delta_{0s}, \Delta_{1s}) - \psi_2(m_0, m_s, \Delta_{0s}, \Delta_{1s})}, \tag{B2}$$

with

$$\psi_1(m_0, m_s, \Delta_{0s}, \Delta_{1s}) = \frac{\tau}{2A^2}\int Dx\{\Phi^2(A[\sqrt{\Delta_{0s}}x + m_0 + e^{-s/\tau}m_s]) + \Phi^2(A[\sqrt{\Delta_{0s}}x + e^{-s/\tau}m_s - m_0])\}, \tag{B3}$$

$$\psi_2(m_0, m_s, \Delta_{0s}, \Delta_{1s}) = \frac{\tau}{2A^2}\int Dz\left\{\left[\int Dx\Phi(A[\sqrt{\Delta_{0s} - |\Delta_{1s}|}x + \sqrt{|\Delta_1|}z + m_0 + e^{-s/\tau}m_s])\right]^2 + \left[\int Dx\Phi(A[\sqrt{\Delta_{0s} - |\Delta_{1s}|}x + \sqrt{|\Delta_{1s}|}z + e^{-s/\tau}m_s - m_0])\right]^2\right\}. \tag{B4}$$

Note that the dynamics for the overlaps is given by Eqs. (37) and (38). Therefore, the full dynamics for the order parameters $m_0, m_s, \Delta_{0s}, \Delta_{1s}$ is ruled by a system of integro-differential equations given by Eqs. (B1), (B2), (37), and (38). We initialize the dynamics using several initial conditions of the form

$$m_0 = \delta_{m_0}, \tag{B5}$$

$$m_s = 1, \tag{B6}$$

$$\Delta_{0s} = d_0, \tag{B7}$$

$$\Delta_{1s} = d_1, \tag{B8}$$

where d_0 and d_1 are fixed and δ_{m_0} is small and variable. The steady state solutions are shown in Figs. 5(a) and 5(b). This method makes an accurate prediction for the memory age s^* above which networks jump from an older to the most recent memory state [see Figs. 5(a) and 5(b)].

APPENDIX C: NETWORK SIMULATIONS AND NUMERICAL SOLUTIONS TO THE DMFT

The network simulations and numerical solutions to the DMFT were generated using custom PYTHON scripts.

[1] J. J. Hopfield, *Neural Networks and Physical Systems with Emergent Collective Computational Abilities*, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).

[2] D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Spin-Glass Models of Neural Networks*, Phys. Rev. A **32**, 1007 (1985).

[3] D. J. Amit, *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge University Press, Cambridge, England, 1992).

[4] N. Brunel, *Network Models of Memory*, in *Methods and Models in Neurophysics: Lecture Notes of the Les Houches Summer School 2003, Session LXXX*, edited by C. Chow, B. Gutkin, D. Hansel, C. Meunier, and J. Dalibard (Elsevier, New York, 2005), pp. 407–476.

[5] J. M. Fuster and G. E. Alexander, *Neuron Activity Related to Short-Term Memory*, Science **173**, 652 (1971).

[6] Y. Miyashita, *Neuronal Correlate of Visual Associative Long-Term Memory in the Primate Temporal Cortex*, Nature (London) **335**, 817 (1988).

[7] S. Funahashi, C. J. Bruce, and P. S. Goldman-Rakic, *Mnemonic Coding of Visual Space in the Monkey's Dorsolateral Prefrontal Cortex*, J. Neurophysiol. **61**, 331 (1989).

[8] P. S. Goldman-Rakic, *Cellular Basis of Working Memory*, Neuron **14**, 477 (1995).

[9] D. Liu, X. Gu, J. Zhu, X. Zhang, Z. Han, W. Yan, Q. Cheng, J. Hao, H. Fan, R. Hou *et al.*, *Medial Prefrontal Activity during Delay Period Contributes to Learning of a Working Memory Task*, Science **346**, 458 (2014).

[10] Z. V. Guo, N. Li, D. Huber, E. Ophir, D. Gutnisky, J. T. Ting, G. Feng, and K. Svoboda, *Flow of Cortical Activity Underlying a Tactile Decision in Mice*, Neuron **81**, 179 (2014).

[11] A. Compte, C. Constantinidis, J. Tegnér, S. Raghavachari, M. Chafee, P. S. Goldman-Rakic, and X.-J. Wang, *Temporally Irregular Mnemonic Persistent Activity in Prefrontal Neurons of Monkeys during a Delayed Response Task*, J. Neurophysiol. **90**, 3441 (2003).

[12] M. Shafi, Y. Zhou, J. Quintana, C. Chow, J. Fuster, and M. Bodner, *Variability in Neuronal Activity in Primate Cortex*

during Working Memory Tasks, Neuroscience **146**, 1082 (2007).
[13] O. Barak, M. Tsodyks, and R. Romo, *Neuronal Population Coding of Parametric Working Memory*, J. Neurosci. **30**, 9424 (2010).
[14] O. Barak and M. Tsodyks, *Working Models of Working Memory*, Curr. Opin. Neurobiol. **25**, 20 (2014).
[15] D. Kobak, W. Brendel, C. Constantinidis, C. E. Feierstein, A. Kepecs, Z. F. Mainen, X. L. Qi, R. Romo, N. Uchida, and C. K. Machens, *Demixed Principal Component Analysis of Neural Population data*, eLife **5** (2016).
[16] J. D. Murray, A. Bernacchia, N. A. Roy, C. Constantinidis, R. Romo, and X.-J. Wang, *Stable Population Coding for Working Memory Coexists with Heterogeneous Neural Dynamics in Prefrontal Cortex*, Proc. Natl. Acad. Sci. U.S.A. **114**, 394 (2017).
[17] F. Barbieri and N. Brunel, *Irregular Persistent Activity Induced by Synaptic Excitatory Feedback*, Front. Comput. Neurosci. **1**, 5 (2007).
[18] G. Mongillo, O. Barak, and M. Tsodyks, *Synaptic Theory of Working Memory*, Science **319**, 1543 (2008).
[19] M. Lundqvist, A. Compte, and A. Lansner, *Bistable, Irregular Firing and Population Oscillations in a Modular Attractor Memory Network*, PLoS Comput. Biol. **6**, e1000803 (2010).
[20] S. Druckmann and D. B. Chklovskii, *Neuronal Circuits Underlying Persistent Representations Despite Time Varying Activity*, Curr. Biol. **22**, 2095 (2012).
[21] B. Tirozzi and M. Tsodyks, *Chaos in Highly Diluted Neural Networks*, Europhys. Lett. **14**, 727 (1991).
[22] U. Pereira and N. Brunel, *Attractor Dynamics in Networks with Learning Rules Inferred from In Vivo Data*, Neuron **99**, 227 (2018).
[23] J. Nadal, G. Toulouse, J. Changeux, and S. Dehaene, *Networks of Formal Neurons and Memory Palimpsests*, Europhys. Lett. **1**, 535 (1986).
[24] G. Parisi, *A Memory which Forgets*, J. Phys. A **19**, L617 (1986).
[25] M. Mézard, J. Nadal, and G. Toulouse, *Solvable Models of Working Memories*, J. Phys. (Paris) **47**, 1457 (1986).
[26] M. Tsodyks, *Associative Memory in Neural Networks with Binary Synapses*, Mod. Phys. Lett. B **04**, 713 (1990).
[27] D. J. Amit and S. Fusi, *Learning in Neural Networks with Material Synapses*, Neural Comput. **6**, 957 (1994).
[28] S. Fusi, P. J. Drew, and L. F. Abbott, *Cascade Models of Synaptically Stored Memories*, Neuron **45**, 599 (2005).
[29] S. Fusi and L. Abbott, *Limits on the Memory Storage Capacity of Bounded Synapses*, Nat. Neurosci. **10**, 485 (2007).
[30] S. Lahiri and S. Ganguli, *A Memory Frontier for Complex Synapses*, in *Advances in Neural Information Processing Systems* (2013), pp. 1034–1042, https://papers.nips.cc/paper/2013/hash/7f24d240521d99071c93af3917215ef7-Abstract.html.
[31] M. K. Benna and S. Fusi, *Computational Principles of Synaptic Memory Consolidation*, Nat. Neurosci. **19**, 1697 (2016).
[32] S. Romani, D. J. Amit, and Y. Amit, *Optimizing One-Shot Learning with Binary Synapses*, Neural Comput. **20**, 1928 (2008).
[33] A. M. Dubreuil, Y. Amit, and N. Brunel, *Memory Capacity of Networks with Stochastic Binary Synapses*, PLoS Comput. Biol. **10**, e1003727 (2014).
[34] Y. Huang and Y. Amit, *Capacity Analysis in Multi-State Synaptic Models: A Retrieval Probability Perspective*, J. Comput. Neurosci. **30**, 699 (2011).
[35] A. Mason, A. Nicoll, and K. Stratford, *Synaptic Transmission between Individual Pyramidal Neurons of the Rat Visual Cortex In Vitro*, J. Neurosci. **11**, 72 (1991).
[36] H. Markram, J. Lübke, M. Frotscher, A. Roth, and B. Sakmann, *Physiology and Anatomy of Synaptic Connections between Thick Tufted Pyramidal Neurones in the Developing Rat Neocortex.*, J. Physiol. **500**, 409 (1997).
[37] C. Holmgren, T. Harkany, B. Svennenfors, and Y. Zilberter, *Pyramidal Cell Communication within Local Networks in Layer 2/3 of Rat Neocortex*, J. Physiol. **551**, 139 (2003).
[38] A. M. Thomson and C. Lamy, *Functional Maps of Neocortical Local Circuitry*, Front. Neurosci. **1**, 19 (2007).
[39] S. Lefort, C. Tomm, J.-C. F. Sarria, and C. C. Petersen, *The Excitatory Neuronal Network of the C2 Barrel Column in Mouse Primary Somatosensory Cortex*, Neuron **61**, 301 (2009).
[40] S. J. Guzman, A. Schlögl, M. Frotscher, and P. Jonas, *Synaptic Mechanisms of Pattern Completion in the Hippocampal CA3 Network*, Science **353**, 1117 (2016).
[41] T. J. Sejnowski, *Storing Covariance with Nonlinearly Interacting Neurons*, J. Math. Biol. **4**, 303 (1977).
[42] M. Mézard, J.-P. Nadal, and G. Toulouse, *Solvable Models of Working Memories*, J. Phys. (Paris) **47**, 1457 (1986).
[43] R. Kree and A. Zippelius, *Continuous-Time Dynamics of Asymmetrically Diluted Neural Networks*, Phys. Rev. A **36**, 4421 (1987).
[44] T. Toyoizumi, M. Kaneko, M. P. Stryker, and K. D. Miller, *Modeling the Dynamic Interaction of Hebbian and Homeostatic Plasticity*, Neuron **84**, 497 (2014).
[45] T. Vogels, H. Sprekeler, F. Zenke, C. Clopath, and W. Gerstner, *Inhibitory Plasticity Balances Excitation and Inhibition in Sensory Pathways and Memory Networks*, Science **334**, 1569 (2011).
[46] H. Sompolinsky, A. Crisanti, and H.-J. Sommers, *Chaos in Random Neural Networks*, Phys. Rev. Lett. **61**, 259 (1988).
[47] M. Tsodyks and M. Feigel'Man, *The Enhanced Storage Capacity in Neural Networks with Low Activity Level*, Europhys. Lett. **6**, 101 (1988).
[48] J. Kadmon and H. Sompolinsky, *Transition to Chaos in Random Neuronal Networks*, Phys. Rev. X **5**, 041030 (2015).
[49] A. Crisanti and H. Sompolinsky, *Path Integral Approach to Random Neural Networks*, Phys. Rev. E **98**, 062120 (2018).
[50] J. Schücker, S. Goedeke, D. Dahmen, and M. Helias, *Functional Methods for Disordered Neural Networks*, arXiv:1605.06758.
[51] M. Gillett, U. Pereira, and N. Brunel, *Characteristics of Sequential Activity in Networks with Temporally Asymmetric Hebbian Learning*, Proc. Natl. Acad. Sci. U.S.A. **117**, 29948 (2020).
[52] C. Gastaldi, T. Schwalger, E. De Falco, R. Q. Quiroga, and W. Gerstner, *When Shared Concept Cells Support Associations: Theory of Overlapping Memory Engrams*, PLoS Comput. Biol. **17**, e1009691 (2021).

[53] J. J. Hopfield, *Neurons with Graded Response Have Collective Computational Properties Like Those of Two-State Neurons*, Proc. Natl. Acad. Sci. U.S.A. **81**, 3088 (1984).
[54] R. Kühn, S. Bös, and J. L. van Hemmen, *Statistical Mechanics for Networks of Graded-Response Neurons*, Phys. Rev. A **43**, 2084 (1991).
[55] S. Lim, J. L. McKee, L. Woloszyn, Y. Amit, D. J. Freedman, D. L. Sheinberg, and N. Brunel, *Inferring Learning Rules from Distributions of Firing Rates in Cortical Neurons*, Nat. Neurosci. **18**, 1804 (2015).
[56] D. J. Amit and N. Brunel, *Model of Global Spontaneous Activity and Local Structured Activity during Delay Periods in the Cerebral Cortex*, Cereb. Cortex **7**, 237 (1997).
[57] X.-J. Wang, *Synaptic Basis of Cortical Persistent Activity: The Importance of NMDA Receptors to Working Memory*, J. Neurosci. **19**, 9587 (1999).
[58] M. Lundqvist, P. Herman, and E. K. Miller, *Working Memory: Delay Activity, Yes! Persistent Activity? Maybe Not*, J. Neurosci. **38**, 7013 (2018).
[59] J. Aljadeff, M. Gillett, U. Pereira-Obilinovic, and N. Brunel, *From Synapse to Network: Models of Information Storage and Retrieval in Neural Circuits*, Curr. Opin. Neurobiol. **70**, 24 (2021).
[60] G. Huang, S. Ramachandran, T. S. Lee, and C. R. Olson, *Neural Correlate of Visual Familiarity in Macaque Area V2*, J. Neurosci. **38**, 8967 (2018).
[61] M. Garrett, S. Manavi, K. Roll, D. R. Ollerenshaw, P. A. Groblewski, N. D. Ponvert, J. T. Kiggins, L. Casal, K. Mace, A. Williford *et al.*, *Experience Shapes Activity Dynamics and Stimulus Coding of VIP Inhibitory Cells*, eLife **9**, e50340 (2020).
[62] G. Wainrib and J. Touboul, *Topological and Dynamical Complexity of Random Neural Networks*, Phys. Rev. Lett. **110**, 118101 (2013).
[63] R. Engelken, F. Wolf, and L. Abbott, *Lyapunov Spectra of Chaotic Recurrent Neural Networks*, arXiv:2006.02427.
[64] J. Aljadeff, M. Stern, and T. Sharpee, *Transition to Chaos in Random Networks with Cell-Type-Specific Connectivity*, Phys. Rev. Lett. **114**, 088101 (2015).
[65] O. Harish and D. Hansel, *Asynchronous Rate Chaos in Spiking Neuronal Circuits*, PLoS Comput. Biol. **11**, e1004266 (2015).
[66] M. Stern, H. Sompolinsky, and L. Abbott, *Dynamics of Random Neural Networks with Bistable Units*, Phys. Rev. E **90**, 062710 (2014).
[67] K. Rajan, L. F. Abbott, and H. Sompolinsky, *Stimulus-Dependent Suppression of Chaos in Recurrent Neural Networks*, Phys. Rev. E **82**, 011903 (2010).
[68] J. Schuecker, S. Goedeke, and M. Helias, *Optimal Sequence Memory in Driven Random Networks*, Phys. Rev. X **8**, 041029 (2018).
[69] F. Mastrogiuseppe and S. Ostojic, *Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks*, Neuron **99**, 609 (2018).
[70] I. D. Landau and H. Sompolinsky, *Coherent Chaos in a Recurrent Neural Network with Structured Connectivity*, PLoS Comput. Biol. **14**, e1006309 (2018).
[71] I. D. Landau and H. Sompolinsky, *Macroscopic Fluctuations Emerge in Balanced Networks with Incomplete Recurrent Alignment*, Phys. Rev. Res. **3**, 023171 (2021).
[72] T. Toyoizumi and L. F. Abbott, *Beyond the Edge of Chaos: Amplification and Temporal Integration by Recurrent Networks in the Chaotic Regime*, Phys. Rev. E **84**, 051908 (2011).
[73] D. Sussillo and L. F. Abbott, *Generating Coherent Patterns of Activity from Chaotic Neural Networks*, Neuron **63**, 544 (2009).
[74] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, *Exponential Expressivity in Deep Neural Networks through Transient Chaos*, Adv. Neural Inf. Process. Syst. **29**, 3360 (2016), https://papers.nips.cc/paper/2016/hash/148510031349642de5ca0c544f31b2ef-Abstract.html.
[75] C. Keup, T. Kühn, D. Dahmen, and M. Helias, *Transient Chaotic Dimensionality Expansion by Recurrent Networks*, Phys. Rev. X **11**, 021064 (2021).
[76] D. J. Amit and S. Fusi, *Constraints on Learning in Dynamic Synapses*, Network **3**, 443 (1992).
[77] Y. Amit and Y. Huang, *Precise Capacity Analysis in Binary Networks with Multiple Coding Level Inputs*, Neural Comput. **22**, 660 (2010).
[78] F. Schuessler, A. Dubreuil, F. Mastrogiuseppe, S. Ostojic, and O. Barak, *Dynamics of Random Recurrent Networks with Correlated Low-Rank Structure*, Phys. Rev. Res. **2**, 013111 (2020).
[79] M. Beiran, A. Dubreuil, A. Valente, F. Mastrogiuseppe, and S. Ostojic, *Shaping Dynamics with Multiple Populations in Low-Rank Recurrent Networks*, Neural Comput. **33**, 1572 (2021).
[80] M. Tsodyks, *Associative Memory in Asymmetric Diluted Network with Low Level of Activity*, Europhys. Lett. **7**, 203 (1988).
[81] B. Derrida, E. Gardner, and A. Zippelius, *An Exactly Solvable Asymmetric Neural Network Model*, Europhys. Lett. **4**, 167 (1987).
[82] https://github.com/ulisespereira/chaos-forgetting-palimpsest.