Transient Chaotic Dimensionality Expansion by Recurrent Networks

Christian Keup[ ],[1,2,*,†] Tobias Kühn[ ],[1,2,3,*] David Dahmen[ ],[1] and Moritz Helias[ ][1,4]

[1]*Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA Institut Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany*
[2]*RWTH Aachen University, Aachen, Germany*
[3]*Laboratoire de Physique de l'ENS, Laboratoire MSC de l'Université de Paris, CNRS, Paris, France*
[4]*Department of Physics, Faculty 1, RWTH Aachen University, Aachen, Germany*

Neurons in the brain communicate with spikes, which are discrete events in time and value. Functional network models often employ rate units that are continuously coupled by analog signals. Is there a qualitative difference implied by these two forms of signaling? We develop a unified mean-field theory for large random networks to show that first- and second-order statistics in rate and binary networks are in fact identical if rate neurons receive the right amount of noise. Their response to presented stimuli, however, can be radically different. We quantify these differences by studying how nearby state trajectories evolve over time, asking to what extent the dynamics is chaotic. Chaos in the two models is found to be qualitatively different. In binary networks, we find a network-size-dependent transition to chaos and a chaotic submanifold whose dimensionality expands stereotypically with time, while rate networks with matched statistics are nonchaotic. Dimensionality expansion in chaotic binary networks aids classification in reservoir computing and optimal performance is reached within about a single activation per neuron; a fast mechanism for computation that we demonstrate also in spiking networks. A generalization of this mechanism extends to rate networks in their respective chaotic regimes.

I. INTRODUCTION

While biological neurons communicate by spikes, which are discrete all-or-nothing events, artificial neural networks overwhelmingly use continuous-valued units commonly referred to as "rate neurons." The ramifications of this fundamental distinction between discrete and continuous signaling have been debated concerning learning algorithms [1,2], energy efficiency [3], and information coding [4–11]. Here we study how differences in signaling impact network dynamics underlying classification performance in a reservoir setting [12–14]: Input stimuli influence the dynamical state of a randomly connected network which then acts as the representation, from which the desired output is extracted by a linear readout. For a classification task, the representation thus needs to allow a linear separation of classes. Dynamics promotes this separability by nonlinearly embedding the input into its high-

dimensional state space. This embedding is analogous to the kernel trick used in support vector machines [15]: A generic mapping into a high-dimensional nonlinear feature space tends to improve separability because in N dimensions dichotomies of $2N$ random points can be linearly separated with high probability [16]. Presenting input stimuli as initial conditions to the dynamics of a network, the nonlinear transformation of the representation is determined by the subsequent temporal evolution [Figs. 1(a) and 1(d)]. For example, consider stimuli belonging to different classes, each given by a centroid and local noise [Fig. 1(d)]. Two properties are needed for classification: Differences between stimulus classes must be maintained or amplified to foster discrimination [Fig. 1(c), green]. Similar stimuli, however, should lead to similar representations to support generalization; the distance between trajectories of data points belonging to the same class should have limited growth [Fig. 1(c), dark orange]. This view exposes the tight link to chaos, the sensitivity of the dynamics to initial conditions. For rate networks close to the edge of chaos, separation and generalization are well balanced, leading generally to optimal performance [17–20]. While the theory of deterministic [21,22] and stochastic rate networks [23] is well understood and predicts a clearly defined transition to chaos, its link to chaos in binary networks [24,25] remains elusive. Binary networks are the simplest class of models with discrete signaling between neurons.

---

[*]C. K. and T. K. contributed equally to this work.
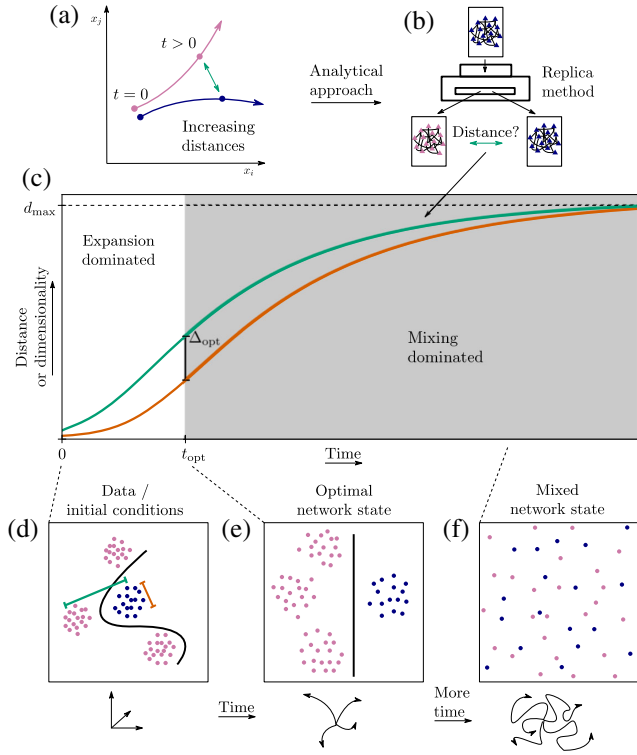[†]Corresponding author.
c.keup@fz-juelich.de

FIG. 1. Transient chaotic dimensionality expansion. The divergence of trajectories with different initial conditions in a chaotic network (a) can be computed by a replica calculation (b), leading to a curve describing the average temporal evolution of the distance (c). Sketch of linearly nonseparable data (d) supplied as initial condition to the dynamics (a). Chaotic dynamics expands the representation into spaces with increasing dimension (c). Interclass separation (green) initially grows quicker than variability within class (dark orange) up to point t_{opt} of optimal separability (e). Subsequently, chaotic mixing causes separability to decline, ultimately leading to a completely mixed state (f).

Here we develop a systematic and model-independent approach to derive mean-field theories for large random networks (Sec. II A). The formalism finds the same set of mean-field equations simultaneously describing binary and rate networks. It shows that a stochastic rate network with properly chosen noise has the same first- and second-order activity statistics as a binary network. The approach allows for replica calculations, the study of ensembles of pairs of networks with identical connectivity in each realization, but different stimuli, as required to assess chaos and computation [Fig. 1(b)]. For stochastic dynamics, one compares two systems with slightly different initial conditions but identical realization of stochasticity [26]. The replica theory exposes that chaos and signal processing in statistically matched rate and binary networks are qualitatively different: Binary networks show a transition to chaos that depends on network size (Secs. II B and II C). In the chaotic regime, distances between states in binary networks increase transiently in a stereotypical manner, confined to a chaotic submanifold whose dimension depends on the

coupling strength and is a fraction of the entire state space [d_{max} in Fig. 1(c), Sec. II D]. Rate networks with statistically matched activity, in contrast, are nonchaotic (Sec. II E). Giving up on the statistical match, rate networks with weak noise in their corresponding chaotic regime show a qualitatively different divergence of state trajectories that sensitively depends on the coupling strength (Sec. II F). Given a distribution of input data whose within-class variability is smaller than the average between-class distances [Fig. 1(d), dark orange and green], the dimensionality expansion of presented stimuli by chaotic binary networks leads to a separation that is optimal for classification after $t_{opt}/\tau = 2\ln 2 \simeq 1.4$ activations per neuron [Figs. 1(c) and 1(e), Sec. II G]. Subsequently, the chaotic mixing leads to a gradual decline of separability [Fig. 1(f)]. Despite the qualitative differences between rate and binary networks, both mechanisms of chaos can be employed to increase classification performance deep in the chaotic regime in a wide range of networks models, including long-short-term-memory (LSTM) and spiking networks (Sec. II H).

II. RESULTS

A. Model-independent field theory of neuronal networks

Here we derive a framework to compute the statistics of neuronal networks in a manner that is largely independent of the employed neuron model. Such a framework is needed to systematically compare different model classes and to assess the generality of results. It must be flexible enough to enable the use of methods such as disorder averages and replica calculations; techniques that are required to systematically derive mean-field equations that allow us to compare networks on a statistical level and to assess how distances between different dynamical states evolve over time and classification of input signals can be achieved (Fig. 1).

We consider a network of N neurons with connectivity matrix $\boldsymbol{J}$, where individual entries are independently and identically distributed as $J_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\bar{g}/N, g^2/N)$; assumptions on the statistics can easily be relaxed as long as higher-order cumulants are suppressed by the large network size. The N neurons have inputs $\boldsymbol{h} = [h_1(t), ..., h_N(t)]$ and outputs $\boldsymbol{x} = [x_1(t), ..., x_N(t)]$. The input-to-output relation of a neuron is often stochastic, so that a conditional probability $\rho[\boldsymbol{x}|\boldsymbol{h}]$ of the output given the input is the most general description of the neural dynamics. The joint statistics of input and output is then

$$\rho[\boldsymbol{x}, \boldsymbol{h}] = \rho[\boldsymbol{x}|\boldsymbol{h}]\rho[\boldsymbol{h}] \tag{1}$$

amounting to a separation of the neurons' input-output functional $\rho[\boldsymbol{x}|\boldsymbol{h}] := \prod_i \rho[x_i|h_i]$ and the input statistics $\rho[\boldsymbol{h}]$.
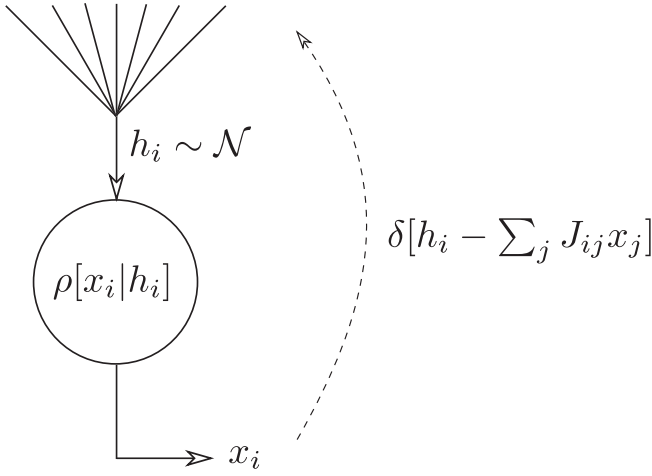
FIG. 2. Summary of model-independent field theory. Conceptual idea to split the network into neuronal dynamics described by conditional probability $\rho[x_i|h_i]$ of neuronal output x_i given its input h_i, and the mapping of output to input by connectivity J_{ij}. The connectivity average affects only the output-to-input mapping and can thus be performed without specifying the neuron model. The formal saddle-point approximation in auxiliary fields $\mathcal{R}(t) \simeq R(t)$ and $\mathcal{Q}(t,s) \simeq Q(t,s)$ [Eqs. (2) and (3)] amounts to a Gaussian approximation of the input $h_i \sim \mathcal{N}(R,Q)$.

Here we denote the functionals by angular brackets and vectors of neuron indices by bold-font symbols.

Any observable O of a neuronal network can be expressed as a functional of the inputs $\boldsymbol{h}$, which have the advantage of being closer to a Gaussian distribution than $\boldsymbol{x}$ due to the convergence of many outputs on one input. Because we do not know the disorder realization (e.g., of the connectivity) in detail, but at most its statistics, we can access only quenched disorder-averaged quantities like

$$\langle O[\boldsymbol{h}]\rangle_{\boldsymbol{J},\boldsymbol{h}} \coloneqq \int \mathcal{D}\boldsymbol{h}\,\langle\rho[\boldsymbol{h}](\boldsymbol{J})\rangle_{\boldsymbol{J}}\,O[\boldsymbol{h}].$$

The description of the network dynamics is self-consistently closed by using a delta distribution $\rho[\boldsymbol{h}] = \delta[\boldsymbol{h} - \boldsymbol{Jx}]$ to enforce that the input to each neuron is composed of a sum of outputs weighted by the synaptic connectivity $\boldsymbol{J}$. The idea of splitting the system into a neuron and a coupling model is illustrated in Fig. 2. Note that Eq. (1) is not a circular definition because $\rho[\boldsymbol{x}|\boldsymbol{h}]$ is a causal functional and $\rho[\boldsymbol{h}] = \delta[\boldsymbol{h} - \boldsymbol{Jx}]$ couples only equal time points of $\boldsymbol{h}$ and $\boldsymbol{x}$, so that the concatenation in Eq. (1) can be understood as a spiral moving forward in time (see also the Appendix 1).

Using the Fourier representation of $\rho[\boldsymbol{h}]$ we obtain, at the expense of introducing the response fields $\hat{\boldsymbol{h}}$, the disorder-averaged input statistics

$$\langle\rho[\boldsymbol{h}](\boldsymbol{J})\rangle_{\boldsymbol{J}} = \left\langle \int \mathcal{D}\boldsymbol{x}\rho[\boldsymbol{x},\boldsymbol{h}]\right\rangle_{\boldsymbol{J}}$$

$$= \int \mathcal{D}\hat{\boldsymbol{h}}\exp{(\hat{\boldsymbol{h}}^T\boldsymbol{h})}\int \mathcal{D}\boldsymbol{x}\langle\exp{(-\hat{\boldsymbol{h}}^T\boldsymbol{Jx})}\rangle_{\boldsymbol{J}}\rho[\boldsymbol{x}|\boldsymbol{h}]$$

obtained by marginalizing over $\boldsymbol{x}$. The connectivity average acts only on the interaction term, which now has the form of a moment-generating function of $\boldsymbol{J}$. In its cumulant expansion, intensive parameters of the system are the first and second cumulant $\bar{g}/N$ and g^2/N, respectively. Higher cumulants would also be suppressed if one assumes the commonly chosen scaling $\propto N^{-\frac{1}{2}}$ of synaptic weights [24,25].

The cumulant expansion suggests to do a Hubbard-Stratonovich transformation with the auxiliary fields $\mathcal{R}(t) \coloneqq (\bar{g}/N)\sum_i x_i(t)$ and $\mathcal{Q}(t,s) \coloneqq (g^2/N)\sum_i x_i(t)x_i(s)$, as outlined in the Appendix 1, so that a saddle-point approximation gives self-consistency relations for the mean inputs and mean time-lagged autocorrelations, a dynamical mean-field theory (DMFT)

$$R(t) = \bar{g}\langle x(t)\rangle_{\Omega(R,Q)}, \tag{2}$$

$$Q(t,s) = g^2\langle x(t)x(s)\rangle_{\Omega(R,Q)}, \tag{3}$$

where the average $\langle\ldots\rangle_{\Omega(R,Q)}$ is defined in Eq. (A9) as an average over $x \sim \langle\rho[x|h]\rangle_h$ and h is a Gaussian process $h \sim \mathcal{N}(R,Q)$. We may think of $x(t)$ as the representative neuron of a homogeneous population, because all neurons with statistically identical connectivity and properties are identical after the disorder average. On the intuitive level, DMFT corresponds to modeling the inputs of all neurons as independent Gaussian processes $h \sim \mathcal{N}(R,Q)$.

Thus, one obtains the DMFT using only the output-to-input relation given by the disordered connectivity while staying agnostic of the neuron model. To instantiate the approximation for the binary model studied here, we must now provide knowledge about the input-to-output relation $\rho[\boldsymbol{x}|\boldsymbol{h}]$.

1. Binary neuron model

We consider the binary neuron model, or kinetic Ising model, with state $x_i \in \{-1,1\}$ [27,28]. The states of all neurons are updated asynchronously by independent Poisson processes with rate τ^{-1} and an activation probability function $T_p : \mathbb{R} \to [0,1]$. It is clear that the form of $\rho[\boldsymbol{x}|\boldsymbol{h}]$ depends on the realization of the update times, which constitute a source of noise, or temporal stochasticity. The update sequence may be thought of as another type of disorder in the sense that it breaks the homogeneity of the time axis by selecting a set of time points where the neuronal state can change. As with the random connectivity, one may study the behavior of the system averaged over this disorder. In this case, the probability of finding a neuron active at time t,

$$p[x_i(t) = 1|h_i] = \int_{-\infty}^t \frac{dt'}{\tau} e^{-(t-t')/\tau} T_p(h_i(t')) \quad (4)$$

is given by the probability $T_p(h_i(t'))$ to be activated at any prior update time point t' and the survivor function $e^{-(t-t')/\tau}$ [29], the probability that no further update happened since. While this knowledge is far from knowing the complete probability functional $\rho[\boldsymbol{x}|\boldsymbol{h}]$ across its infinite time dimension, the information about this single time slice is sufficient to plug into Eq. (2) and obtain, after taking a time derivative, the mean-field equation

$$\tau \frac{d}{dt} R(t) + R(t) = \bar{g} \langle T(h) \rangle_{h \sim \mathcal{N}(R(t), Q(t,t))}, \quad (5)$$

where

$$T(h) = 2T_p(h) - 1. \quad (6)$$

Details are provided in the Appendix 2.

In Eq. (5), only equal-time autocorrelations $Q(t,t)$ appear, because the dynamics is a Markov process; its evolution at time t depends only on the statistics at this very time point, not on the prior history. Closing the equation is thus simple for binary neurons, because, by $x_i \in \{-1,1\}$, their autocorrelation is always 1, so that $Q(t,t) = (g^2/N) \sum_i \langle x_i(t)x_i(t) \rangle = g^2$ when cross-correlations are negligible (see the Appendix 1).

To compute $Q(t, t + \Delta t)$ for binary neurons, we need more information about $\rho[\boldsymbol{x}|\boldsymbol{h}]$, namely, the joint probability distribution over two time slices for a neuron:

$$\rho[x(t), x(s)|h] = \underbrace{\rho[x(t)|x(s), h]}_{\overset{t \searrow s}{\longrightarrow} \delta_{x(t),x(s)}} \rho[x(s)|h]. \quad (7)$$

To construct $\rho[x(t)|x(s), h]$ for binary neurons, the basic idea is to iterate the 2×2 states a neuron can assume at the points in time s and t and consider all possible evolutions that match the respective initial and final condition. From such a consideration, we derive $Q(\Delta t) = Q(t, t + \Delta t)$ for stationary dynamics in the Appendix 3 by again taking a time derivative of the saddle-point equation (3) yielding

$$\tau \frac{d}{d\Delta t} Q(\Delta t) + Q(\Delta t)$$
$$= g^2 \int_0^\infty \frac{dt'}{\tau} e^{-t'/\tau} \langle T(h)T(h') \rangle_{(h,h') \sim \mathcal{N}_{R,Q(0),Q(\Delta t + t')}}. \quad (8)$$

This equation is the analogon of the integral equation (5.17) of van Vreeswijk and Sompolinsky [25]. The advantage of the form (8) compared to the classical result is, as detailed in the Appendix 3, that by differentiating once more with respect to Δt and then using Price's theorem [30], it can be cast into a Newtonian form

$$\tau^2 \ddot{Q}(\Delta t) = -V'_{R,Q(0)}(Q(\Delta t)), \quad (9)$$

$$V_{R,Q(0)}(Q) := -\frac{1}{2}Q^2 + g^2 \langle \mathcal{T}(h)\mathcal{T}(h') \rangle_{(h,h') \sim \mathcal{N}_{R,Q(0),Q}}, \quad (10)$$

where $\mathcal{T}$ is a primitive of T, which is $\partial_h \mathcal{T}(h) = T(h)$, and $\mathcal{N}_{R,Q(0),Q}$ is the bivariate Gaussian with stationary mean R and covariance matrix $\begin{pmatrix} Q(0) & Q \\ Q & Q(0) \end{pmatrix}$. We exploit this result in Sec. II E to construct rate models with exactly the same DMFT solution as a binary network.

B. Binary networks are always chaotic in the thermodynamic limit

In the setting of reservoir computing (Fig. 1), a particularly important measure for the classification performance of a network is how the distance between two different dynamical states, each caused by one stimulus, evolves over time. Tracking the evolution of initially small differences between the states amounts to the characterization of chaos [17,18,21,23,31,32]. We assess chaos by studying the time evolution of two systems with infinitesimally different initial conditions but identical connectivity and identical realization of stochasticity, thus, the same sequences of update time points. Technically, this approach amounts to a replica calculation, where one studies the network-averaged correlation between the states of the two systems over time, an approach pioneered by Derrida and Pomeau [33]. Here we do not use the classical annealed approximation of this original work, where the connectivity is redrawn in every time step, but compute the full quenched averages, where the connectivity is constant in time. The calculation leads to a dynamic mean-field theory for the correlation between replicas.

In the Appendix 4, by an approach analogous to the derivation of the ordinary differential equation (ODE) for the autocorrelation (10), we obtain the evolution of the cross-replica equal-time correlation $Q^{(12)}(t)$ in the binary network as

$$\tau \frac{d}{dt} Q^{(12)}(t) = -Q^{(12)}(t)$$
$$+ g^2 (1 - \langle |T(h^{(1)}) - T(h^{(2)})| \rangle_{h^{(1)},h^{(2)}}). \quad (11)$$

Here, $(h^{(1)}, h^{(2)}) \sim \mathcal{N}(\boldsymbol{R}, \boldsymbol{Q})$ is a measure of a pair of Gaussian processes with means $\langle h^{(\alpha)} \rangle = R^{(\alpha)} = \langle T(h) \rangle_{h \sim \mathcal{N}(R^{(\alpha)}, g^2)}$, the stationary solution of Eq. (5), and covariance matrix $\langle\langle h^{(\alpha)} h^{(\beta)} \rangle\rangle = Q^{(\alpha\beta)}(t)$, whose diagonal elements are each $Q^{(\alpha\alpha)}(t) = g^2$.

Since the two replicas are nearly perfectly correlated in the beginning, we know that the correlation between a neuron and its "copy" in the other replica is given by the autocorrelation at first, motivating the ansatz

$$Q^{(12)}(t) = Q(0) - \epsilon(t), \qquad \epsilon(t) \geq 0.$$

As shown in the Appendix 4, an expansion for small ϵ leads to the approximate equation governing the evolution of $\epsilon(t)$,

$$\tau \frac{d}{dt}\epsilon(t) = -\epsilon(t) + \frac{2}{\sqrt{\pi}} g^2 \langle T'(h) \rangle_{h \sim \mathcal{N}(R, Q(0))} \sqrt{\epsilon(t)}, \quad (12)$$

which generalizes the result of van Vreeswijk and Sompolinsky [25] to arbitrary activation functions. As was their conclusion for neurons with hard threshold, we see from Eq. (12) that for any activation function with average positive slope and independent of the parameters, the positive term $\propto \sqrt{\epsilon}$ is always larger than the negative linear term for small ϵ, so an initial deviation between the replicas will grow, indicating chaotic dynamics. Since the calculation becomes exact in the thermodynamic limit, the conclusion is that infinitely large binary networks are always chaotic, with a formally infinite maximum Lyapunov exponent since the slope of the right-hand side of Eq. (12) at $\epsilon = 0^+$ is infinite, leading to an initial growth of ϵ that is faster than exponential. More specifically, $\epsilon(t) \sim t^2$ for $\epsilon \ll 1$, meaning that $\epsilon(0) = 0+$ grows to a finite value in finite time, as opposed to an exponential function. See the Appendix 5 for additional details. Since the slope of the activation function only appears averaged over the input distribution, there is no qualitative difference between different activation functions. In particular, going from a stochastic activation function to the deterministic Heaviside limit changes only the second term in Eq. (12) by a finite factor, and thus does not qualitatively alter the chaotic behavior. This result can also be understood by noting that for the stochastic activation function, the function value at each update is compared to a random number to decide the activity state. The comparison is just like using a Heaviside function but with a randomly drawn threshold at each update.

C. Transition to chaos in finite-size binary networks

In contrast to the theoretical prediction, simulations of binary networks in fact show parameter regimes with regular dynamics (Fig. 4). Since the theory is only exact in the limit of infinite network size, this behavior suggests a finite-size effect. But the result of the replica calculation (12) does not rely on carrying out the $N \to \infty$ limit. Rather, it is expected to be a good approximation for finite, yet large networks $N \gg 1$. How can the theory be reconciled with the simulation?

First, while the square-root term in Eq. (12) is always larger for sufficiently small ϵ, there also exists a point $\partial_t \epsilon^* \overset{!}{=} 0$ where this relationship reverses, and the linear term starts to dominate: The point where the right-hand side of Eq. (12) vanishes,
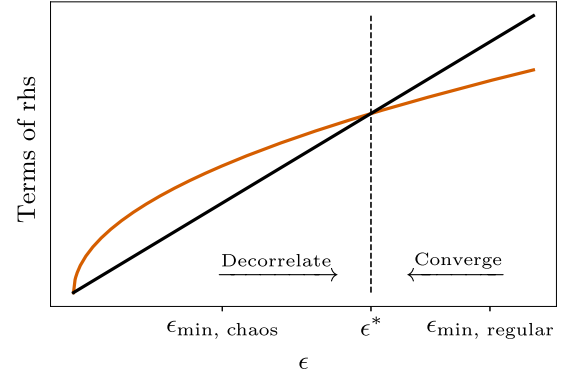

FIG. 3. Fixed-point average distance between replicas implies chaotic subspace. The square-root term $\propto \sqrt{\epsilon}$ (dark orange curve) and the linear term $\propto \epsilon$ in Eq. (12) intersect and produce a fixed point ϵ^* for the covariance $Q(0) - \epsilon$ between the two replicas. The resulting average Hamming distance d^* between states in the two copies of the system is given by ϵ^* and Eq. (16) as $d^* = N\epsilon^*/g^2$. Depending on whether $\epsilon_{\min}$, the minimum decorrelation due to a single flipped spin, is smaller or larger than ϵ^*, the replicas will either decorrelate, or they will converge and forget the perturbation.

$$\sqrt{\epsilon^*} = \frac{2}{\sqrt{\pi}} g^2 \langle T'(h) \rangle_h. \quad (13)$$

This point corresponds to a stable average distance between (partly) decorrelated trajectories, as illustrated in Fig. 3.

Second, in a finite network of N binary neurons, an infinitesimal perturbation cannot be realized, since the smallest possible perturbation is to flip a single spin at index i_{flip}. A single flip implies for the minimally perturbed cross-replica correlation (A21)

$$\begin{aligned} Q(0) - \epsilon_{\min} &= g^2 \langle x^{(1)} x^{(2)} \rangle \\ &= g^2 \frac{1}{N} \left[\sum_{i=1}^{N} (x_i^{(1)} x_i^{(1)}) - \underbrace{x_{i_{\text{flip}}}^{(1)} x_{i_{\text{flip}}}^{(1)}}_{=1} + \underbrace{x_{i_{\text{flip}}}^{(1)} x_{i_{\text{flip}}}^{(2)}}_{=-1} \right] \\ &= Q(0) - 2\frac{g^2}{N}, \end{aligned}$$

so that

$$\epsilon_{\min} = 2\frac{g^2}{N}. \quad (14)$$

Therefore, if $\epsilon_{\min} > \epsilon^*$ the replicas will tend toward more correlation. But as the only possible step below $\epsilon_{\min}$ is having zero different spins and thus perfect correlation, the initial difference should tend to be completely forgotten, resulting in regular dynamics. On the other hand, if $\epsilon_{\min} < \epsilon^*$ an increase of the initial difference is possible.

Thus, the chaos transition criterion in the finite binary network is $\epsilon_{\min} \overset{!}{\leq} \epsilon^*$ resulting in
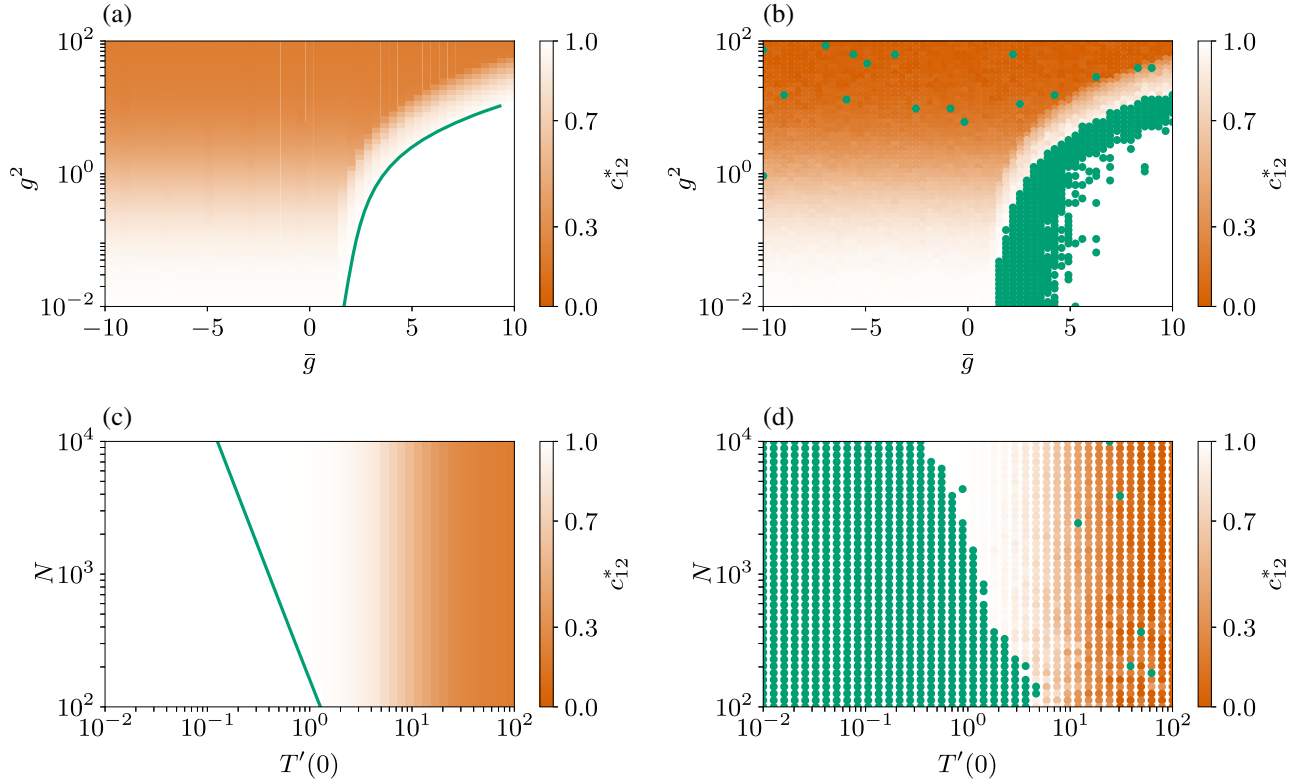
FIG. 4.   Chaos transition and residual correlation in theory (a),(c) and simulations (b),(d). (a) Theoretical prediction of chaos transition [green line, Eq. (15)] and residual correlation coefficient between replicas [dark orange shading, $c_{12}^* = 1 - Q(0)^{-1}\epsilon^*$ and Eq. (13)] for varying mean $\bar{g}$ and variance g^2 of the connectivity. Other parameters are $N = 5000$, $T = \tanh$, $\tau = 10$ ms. (b) As in (a), but each pixel is colored dark orange (chaotic) or green (stable) according to a network simulation. Two identical networks are evolved with identical random numbers, only one being perturbed by flipping four spins, and after $T_{\text{sim}} = 2500$ ms the correlation between the state vectors is computed. The few scattered green dots in the chaotic regime are algorithmic artifacts where the perturbation is unsuccessful; see the Appendix 11 for the perturbation method. (c) Theoretical prediction for varying network size N and slope $T'(0)$ of the activation function $T(h) = \tanh(T'(0)h)$. Other parameters are $\bar{g} = 0$, $g^2 = 0.01$, and $\tau = 10$ ms. (d) As in (c) but from network simulations, with procedure as described for (b).

$$1 \lesssim \sqrt{\frac{2}{\pi}}\, g \langle T'(h) \rangle_h \sqrt{N}. \qquad (15)$$

Because of the scaling with $\sqrt{N}$, it is clear that networks with thousands or even only hundreds of neurons are only nonchaotic if the connectivity is very weak, $g \lesssim N^{-\frac{1}{2}}$, or the dynamics is saturated [which gives a small $\langle T'(h) \rangle_h$]. Also, $N \to \infty$ clearly recovers the limit of strictly chaotic dynamics. For the special case of a Heaviside activation function and vanishing mean connectivity $\bar{g} = 0$, the network is always chaotic, since $\langle H'(h) \rangle_{h \sim \mathcal{N}(0,g^2)} = \sqrt{2/(\pi g^2)}$ results in $\pi/2 \leq \sqrt{N}$, which is certainly true for typical network sizes.

The predicted transition and the residual correlation ϵ^* fit those observed in simulations quite well (Fig. 4). The dependence of the transition on the positive mean connectivity $\bar{g}$ in the upper panels arises because the network settles in a state with nonzero mean activity that depends on $\bar{g}$; it selects one of the two degenerate states in this bistable

"ferromagnetic" regime. The symmetry with respect to a global sign flip of the activity is spontaneously broken. In this state, neurons show a very small average slope $\langle T'(H) \rangle_H$, thus shifting the point of transition to larger g with increasing $\bar{g}$. The predicted residual correlation is independent of N [Fig. 4(c), compare Eq. (13)], while the chaos transition depends on N [compare Eq. (15)].

We obtain the same criterion (15) through a less general, but more intuitive perspective by analyzing the probability that, given a single-spin difference, the difference in inputs is such that during the next updates, another neuron will also be updated to a "wrong" state (see the Appendix 7). This view provides an expression for the average rate of decorrelation caused by an initial single spin flip. Requiring this rate to be unity, we obtain the same chaos transition criterion as Eq. (15). The approach is inspired from and very similar to calculating the divergence rate of flux tubes in spiking networks [34]. Such flux tubes are stable local environments of a phase-space trajectory, while the network is globally unstable. Thus, the phase space can be

partitioned into tubes which diverge from each other, while perturbations within a tube decay. Indeed, the binary network has relatively trivial flux tubes in the input phase space given by those regions that result in the same updated state.

Note that the chaos transition shown in Fig. 4(d) in simulations happens at slightly larger slopes $T'(0)$ than predicted by Eq. (15). Considering the cascade of spin flips evoked by the initial perturbation provides an explanation: If the average proliferation rate of spin flips per time constant is only slightly above 1, the cascade triggered by a single flipped spin still has a large probability of dying out.

D. Dynamics in binary networks is governed by a chaotic submanifold

The chaos in binary networks found in Sec. II B causes nearby trajectories to diverge at first. Because of the fixed point value ϵ_* of the residual correlation found in Sec. II C, however, the network states of the replicas do not decorrelate completely. Instead, any pair of trajectories has an average maximal distance determined by ϵ^*. This limited distance is a result of the two trajectories evolving by the same network connectivity and update sequence. Also, trajectories that are very far apart will converge to this residual correlation. The fixed-point distance ϵ^* is thus a representative of the average distance between any two trajectories in the long-time limit. The corresponding Hamming distance $H^{(12)} = \frac{1}{4}||\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}||^2$, that is, the number of different spins between a pair of binary states, is given by

$$H^{(12)}(t) = \frac{N\epsilon(t)}{2g^2}, \qquad (16)$$

where we use $Q^{(12)} = (g^2/N)(N - 2H^{(12)})$ and the pre-factor of $H^{(12)}$ arises because every flipped spin causes a decrease by 2 (from $+1$ to -1). Even though the $H^{(12)}$ spin flips distinguishing two trajectories can in principle be distributed across any of the N neurons, the subspace spanned by the set of possible trajectories has an approximate dimensionality of

$$d(t) \simeq 2H^{(12)}(t) \simeq \frac{N}{g^2}\epsilon(t). \qquad (17)$$

This relation can be understood by considering two independently drawn binary random vectors of dimension d that have, on average, the distance $H^{(12)} = d/2$, because the average distance between any pair of spins that take the values $x^{(1)}, x^{(2)} \in \{-1, 1\}$ with equal probabilities is $\langle(x^{(1)} - x^{(2)})^2\rangle/4 = 1/2$. Therefore, in the following, we quantify dimensionality via the Hamming distance $H^{(12)}$ using Eq. (17).

Thus, if $\epsilon_{\min} > \epsilon^*$, then $H^{(12)}(\infty) < 1$, and the set of long-term trajectories contains only a single trajectory, thus constituting a limit cycle (although the return time is astronomically large [35]). Irrespective of the initial state, the network is attracted to a stereotypical trajectory; the dynamics is regular. This situation arises for very weak coupling.

If $\epsilon_{\min} < \epsilon^*$, then $H^{(12)}(\infty) > 1$, and there are many trajectories that constitute the attractive subspace. The evolution within the space is chaotic, because for any pair of states with an initial distance $\epsilon < \epsilon^*$, the distance increases; thus, small differences are amplified. A set of trajectories that initially spans a low-dimensional subspace is thus expanded into a higher-dimensional space. For long times, however, any two states differ in only typically $d(\infty)/2$ of their neurons. This limiting dimensionality grows proportional to Ng^2 as seen by inserting Eq. (13) into Eq. (16)

$$d^* = Ng^2\left(\frac{2}{\sqrt{\pi}}\langle T'(h)\rangle_h\right)^2. \qquad (18)$$

The time evolution when starting with a set of trajectories with dimensionality $d(0)$ is given by

$$d(t) = [\sqrt{d_*} - (\sqrt{d_*} - \sqrt{d(0)})e^{-(t/2\tau)}]^2 \qquad (19)$$

obtained by integrating Eq. (12) (see the Appendix 5) and shown in Fig. 6(a). This explicit solution shows that the expansion happens very quickly on a timescale of 2τ, where τ is the average time to have one update per neuron, and then converges to the residual value d^* (18) for long times. This exclusive dependence on τ can intuitively be understood from the right-hand side of Eq. (15), which can be interpreted as the average number of flips n_{spawns} caused by an initial spin flip within one time constant (as obtained in the Appendix 7). Hence, using Eq. (13), one has

$$\frac{d^*}{2} = n_{\text{spawns}}^2, \qquad (20)$$

so that after two time constants have passed, the residual correlation would be reached if the functional form of initial decorrelation would be extrapolated to later times, neglecting saturating terms [see Fig. 6(a) and the Appendix 5]. But because the residual correlation limits the spread of the cascade of flips, in a similar way as the population size limits the growth of an epidemic [36], the growth slows down and asymptotically approaches the residual correlation.

E. Same statistics, different chaotic dynamics in continuously and discretely coupled networks

Having quantified how binary networks with discrete signaling separate different states, as required to understand

classification in reservoir computing (Fig. 1), we now turn to the well-established alternative of units with continuous-valued activity and signaling commonly referred to as "rate models" and typically employed in artificial neuronal networks. Concretely, we consider the coupled set of stochastic differential equations [21,23]

$$\tau \partial_t \boldsymbol{h} = -\boldsymbol{h} + \boldsymbol{J}T(\boldsymbol{h}) + \sqrt{\tau}\boldsymbol{\xi} \qquad (21)$$

with the activation function $T : \mathbb{R} \to [-1, 1]$ given by Eq. (6), timescale τ, and a white noise process ξ with $\langle \xi_i(t)\xi_j(s)\rangle = \sigma_\xi^2 \delta(t-s)\delta_{ij}$. Chaos in such networks has been intensely studied [21,23,32].

We show in Appendix 8 that the model-independent field theory applied to this stochastic rate model yields the same set of self-consistency equations for the first- (5) and second-order statistics (9) as the binary model; also, the conditions on $Q_\infty := \lim_{\tau \to \infty} Q(\tau)$ agree. In contrast to the binary network, however, where the initial value $Q(0) = g^2$ is known, Eq. (9) must be solved with an initial condition for the slope $\dot{Q}(0+)$. This slope is determined by the variance σ_ξ^2 of the noise in Eq. (21). Demanding identical mean-field solutions for the two neuron types, the variance of the noise follows as (see the Appendix 8)

$$\sigma_\xi^2 = \frac{2}{\tau}\sqrt{2[V_{g^2}(Q_\infty) - V_{g^2}(g^2)]}. \qquad (22)$$

Equation (22) tells us that, given a pair of equivalent activation functions (6), moments of connectivity $\bar{g}, g^2$, and timescale τ asynchronously updated binary networks are statistically equivalent in DMFT approximation to rate networks with appropriately chosen Gaussian white noise input. This result is confirmed in simulations by comparing the autocorrelation functions averaged across many neurons in Fig. 5. The good agreement between the autocorrelation that is averaged over all neurons in a network with a single random realization of the coupling matrix and the theoretical curves, which describe ensembles of networks averaged over many realizations of the random couplings, moreover shows that these quantities are self-averaging.

1. Condition for chaos in rate networks

Having established their equivalence on the level of statistics, we now compare the chaotic evolution of binary and rate networks. As its binary counterpart, the rate neuron model can be studied in a replica calculation in dynamical mean-field approximation, which yields the equation of the cross-replica time-lagged covariance of the form [23,32]

$$(\partial_t + 1)(\partial_s + 1)Q^{(12)}(t, s) = g^2 f_T(Q_0, Q^{(12)}), \qquad (23)$$

with $f_T(Q_0, Q^{(12)}) = \langle T(x_1)T(x_2)\rangle$, and the average is taken with respect to $(x_1, x_2) \sim \mathcal{N}(0, \begin{pmatrix} Q_0 & Q^{(12)} \\ Q^{(12)} & Q_0 \end{pmatrix})$. The
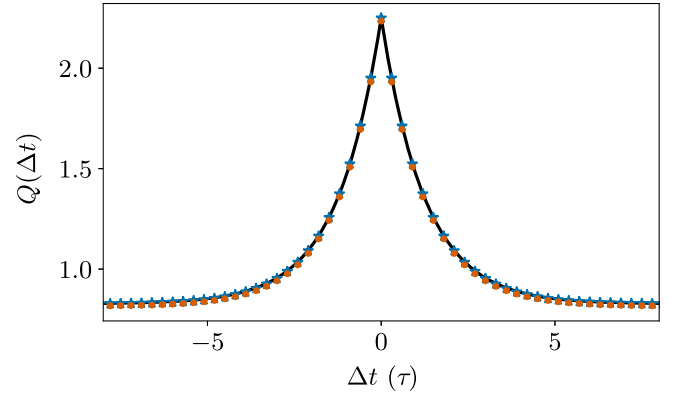

FIG. 5. Matched second-order statistics in binary and rate networks. (a) Autocorrelation functions in simulations of binary (dark orange dots) and rate networks (blue stars). Theoretical curve (black) given by the solution of Eq. (10). Noise amplitude of the rate network chosen by Eq. (22) to obtain matched statistics. Other parameters: $N = 5000$, $g = 1.5$, $\bar{g} = 0$, $\tau = 1$ ms, and $T(h) = \tanh(h - \Theta)$ with $\Theta = 1.173$ such that $\langle x \rangle_{\mathcal{N}(R,Q)} \approx -0.5$ according to the stationary solution of Eq. (5) (see the Appendix 11 for details); simulation duration $T = 5000$ ms; autocorrelation averaged across all neurons. We simulate a single realization of the binary network for 5000 ms and average over five realizations of the rate network running for 1000 ms each. We average over neurons in both cases.

approximation for small differences $Q^{(12)} = Q_0 - \epsilon$ to linear order in ϵ is

$$(\partial_t + 1)(\partial_s + 1)\epsilon(t, s) = g^2 f_{T'}(t, s)\epsilon(t, s), \qquad (24)$$

which is solved by

$$\epsilon(t, t) = \epsilon(0, 0)e^{\lambda_{\max}(g)t}, \qquad (25)$$

where $\lambda_{\max}(g)$ is the largest Lyapunov exponent that follows from an eigenvalue problem, see Sompolinsky *et al.* [21] [their Eqs. (10) and (11)] and Schuecker *et al.* [23] [their Eq. (17)]. The linear stability analysis in Eq. (24) leads to the criterion for the chaos transition [23] [their Eq. (20)]

$$g^2 \langle T(h)T(h)\rangle_{h \sim \mathcal{N}(R^*, Q_0)} - Q_0 \geq 0. \qquad (26)$$

2. No chaos in rate networks with matched statistics

Applying criterion (26) to a network of rate neurons with the noise matched to its binary counterpart via Eq. (22), we obtain $Q_0 = g^2$ by construction. Using that $\langle T(h)T(h)\rangle \leq 1$ because of $|T| \leq 1$, we observe that the condition (26) cannot be fulfilled. The dynamics is therefore always in the regular regime because the frozen noise of amplitude given by Eq. (22) is so large that it drives the dynamics and suppresses chaos. Only asymptotically the chaos transition is approached for an infinite slope of the activation function, $T' \to \infty$, or equivalently $g^2 \to \infty$.

Everything else being identical, the only difference between the two models is the type of signals exchanged

between units, either being discrete or continuous. Therefore chaos in binary networks is intrinsically caused by the discrete signaling. Formally, the difference between the two forms of signaling here shows up in the effective noise ξ: In the rate network, the realization of this noise is identical across the two replicas, because it represents the random realizations of the discrete variables of the binary network whose statistics we want to match. In the binary network, this noise itself changes, because it is intrinsically generated by the discrete switching dynamics, so that the realization is not external and frozen, but depends on the microscopic state.

F. Qualitative differences of chaos between rate and binary networks

In the following, we give up on matching the statistics between rate and binary networks to discuss the qualitative differences of the respective chaotic dynamics.

1. Residual correlation

The first qualitative difference concerns the residual correlation of the replicas. Since there is no term $\propto \sqrt{\epsilon}$ in Eq. (24), there is no residual correlation for small ϵ. Furthermore, Eq. (23) is also valid for small $Q^{(12)}$ and shows that the completely decorrelated state $Q^{(12)} = 0$ is always a fixed point: The expectation value factorizes, and for any point symmetric T, the right-hand side vanishes. In the thermodynamic limit, the residual correlation in rate networks is thus zero for any $g > 1$. A network that has been infinitesimally perturbed eventually has a state that is completely uncorrelated to the unperturbed system.

Trajectories in rate networks of finite size, in fact, show very small residual correlation closely beyond the edge of chaos $g^2 \gtrsim 1$. However, as already noted by Sompolinsky *et al.* [21], the transition at $g = 1$ is not completely sharp in finite-size networks (see also Ref. [37]). For larger networks, however, the residual correlation approaches zero; this is in contrast to binary networks, which in otherwise identical settings have a finite residual correlation (18) even in the large-N limit.

This qualitative difference is shown in Fig. 6: In binary networks, the decorrelation between the original and the perturbed system $d(t)/N = 1 - Q^{(12)}(t,t)/Q_0$ in the long-time limit saturates below unity, on a level that depends on the coupling g by Eq. (18) and is bounded by $8/\pi^2 \simeq 0.81$ in the limit $g \to \infty$, compare the Appendix 5. The quantity $d(t)/N$ can also be interpreted as the relative dimensionality of the explored space. In rate networks with otherwise identical parameters, the relative decorrelation reaches unity independent of g. The decorrelation in rate networks cannot be interpreted in terms of dimensionality, however. Indeed, a recent work demonstrates a structured chaotic attractor in such networks [38].
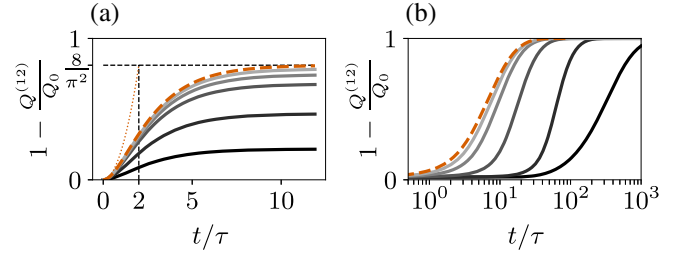

FIG. 6. Discrete chaos versus continuous chaos. (a) Evolution of the decorrelation between replicas in the binary network (19) with $T = \tanh$. Increasing coupling strength from dark to light gray, $g = \{0.5, 1, 2, 3, 5\}$. Dark orange curves: limit $g \to \infty$ or Heaviside T (dashed); quadratic solution for nonsaturated growth derived in the Appendix 5 (dotted). (b) Evolution of decorrelation between replicas in the rate network (21) obtained by numerical solution of Eq. (24) (details in the Appendix 6). Color code as in (a) but using parameters $g = \{1.1, 1.5, 2.5, 5.10\}$ and initial decorrelation of 2% instead of 0%.

2. Transient of decorrelation

The second qualitative difference concerns the transient of decorrelation. The solution (19) shows that the characteristic timescale of decorrelation is 2τ, where τ is the average interval between two state changes of a neuron, exposing that the microscopic state drives the chaotic evolution. Decorrelation slows down only mildly for weaker coupling g, as shown in Fig. 6(a). Moreover, it has a finite slope shortly after the infinitesimal perturbation of the system, reflecting the infinite Lyapunov exponent.

In rate networks, the maximal Lyapunov exponent $\lambda_{\max}(g)$ is finite and depends continuously on the coupling g. Decorrelation therefore starts with a vanishing speed for infinitesimal perturbations, well described by the exponential behavior (25), as shown in Fig. 6(b). The time to reach a given level of decorrelation, moreover, strongly depends on the coupling strength g corresponding to a critical slowing-down at the transition to chaos, which does not occur in binary networks.

G. Computation by transient chaotic dimensionality expansion

We now return to the question how the separation of trajectories by the chaotic dynamics of binary networks affects computation in a setting of reservoir computing (Fig. 1). We investigate the network performance in a pattern classification task: We consider P fixed patterns numbered by the index $1 \leq \alpha \leq P$, each given by a randomly drawn binary vector of length L. Noisy realizations of a pattern are then created by adding Gaussian independent noise of variance σ^2 to each entry of the original pattern, creating P classes of noisy pattern realizations. The network is prepared at $t = 0$ in a fixed initial state consistent with its stationary statistics. A noisy pattern is presented to the network as the initial state of a (fixed)
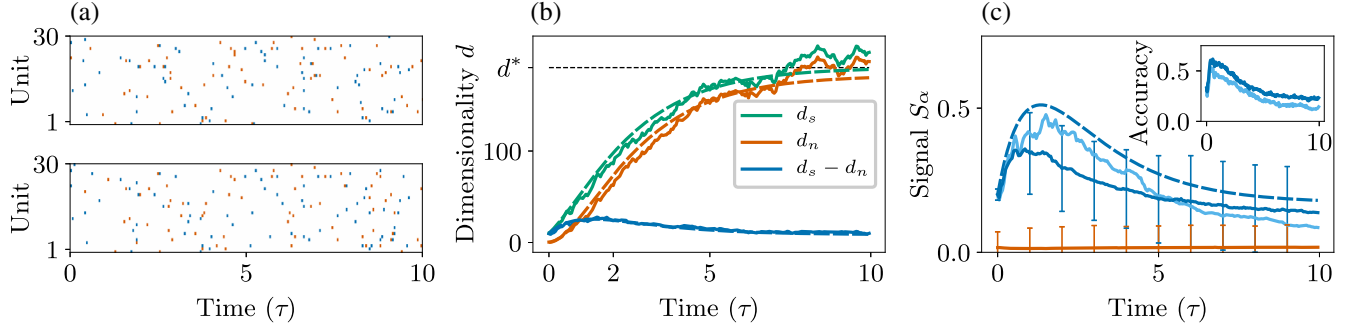
FIG. 7. Transient dimensionality expansion of stimulus representation by a chaotic binary network. (a) Time evolution of the first 50 neurons for two different patterns of initial states; up transitions in red, down in blue. Initial state of first $L = 10$ neurons set to one of the $P = 50$ fixed, random patterns. The initial state of the remaining neurons is kept constant. (b) Evolution of signal and noise subspace dimensionality. Dimensionality $d_s(t)$ given by Eq. (17) explored by the network across different patterns [green solid curve, using averaged simulated distances across all pairs of patterns; dashed curve, theory (19)]. Dimensionality $d_n(t)$ explored across different noisy realizations of a pattern [dark orange solid curve, using averaged simulated distances across all pairs of 20 realizations per pattern; dashed curve, theory (19)]. Noisy realizations of patterns have Gaussian noise with standard deviation $\sigma = 0.3$ added to each of the L entries of the initial pattern state. Difference between signal and noise dimensionality (blue, theory is dashed). (c) Linear readout $S_{\alpha'}(t) = w_{\alpha'}(t)^T[x_\alpha(t) + \xi_{\text{pre}}] + \xi_{\text{post}}$ trained for each time point t to detect stimulus identity by minimizing the quadratic error $\sum_\alpha (S_{\alpha'} - \delta_{\alpha\alpha'})^2$ for the correct stimulus $\alpha = \alpha'$ (blue) and nonmatching stimuli $\alpha \neq \alpha'$ (dark orange). Error bars show the variability across patterns and noise realizations, excluding ξ_{post}. Theoretical prediction (28) (dashed blue). Readout by the theoretical weight vector (A47) using the approximation (A50) (light blue). Inset: classification accuracy of the initial input stimulus based on choosing the readout with largest signal for trained readouts (blue) and approximate readouts (light blue), including ξ_{post}. The approximate readout vectors yield a higher average signal, but also the variance is higher (not shown) resulting in slightly worse classification accuracy. Other parameters: $N = 500$ neurons, coupling strength $g = 0.8$, prereadout noise $\sigma_{\xi,\text{pre}} = 0.1$, and postreadout noise $\sigma_{\xi,\text{post}} = 0.1$ as detailed in the Appendix 10. The training set comprised 100 noisy realizations of each pattern and the test set 20. All theoretical curves are corrected for the probability that a noise realization does not leave the original flux tube; see the Appendix 10 c.

subset of L of the N neurons. The corresponding network state is denoted as $x_\alpha(t)$. At each time t, we train one linear readout $S_{\alpha'}(t) = w_{\alpha'}(t)^T x_\alpha(t)$ per pattern class α' by linear regression to provide the output $S_{\alpha'}(t) = 1$ if the α'th pattern has been presented ($\alpha = \alpha'$) and 0 else ($\alpha \neq \alpha'$; see the Appendix 10 a for details). Thus, we have P readouts, one for detecting each of the presented patterns (one-hot encoding). Classification is performed by selecting the strongest readout signal. Additional noise sources are present at the readout and classification to ensure robustness. The setup, training, and following theory are detailed in the Appendix 10.

Clearly, the set of possible trajectories resulting from the different initial-state preparations has dimensionality $d_s(0) = L$ at $t = 0$. The linear separability of pattern classes is thus initially low if $L \ll P \ll N$. From Sec. II D, we know that the chaotic dynamics will quickly increase the dimensionality of the state space that encodes the patterns, eventually approaching that of the chaotic submanifold $d_s(\infty) = d^*$. To explain the effect on the separability of patterns and the classification performance, we must distinguish between the dimensionality of the total set of trajectories (including all patterns and their noisy variations) referred to as the signal dimensionality $d_s(t)$, and the dimensionality of the set of trajectories given by noisy variations of a single pattern referred to as the noise dimensionality $d_n(t)$. Let us at first neglect the noise. With

the increase of $d_s(t)$, the linear separability of patterns also increases. From the property of the linear regression, this means that the average readout signal $S_\alpha(t)$ of the correct pattern class α increases; if network responses were pairwise orthogonal, which to good approximation is satisfied in the high-dimensional signal subspace, the maximal attainable signal would be

$$\hat{S}_\alpha(t) = \frac{d_s(t)}{P},$$

as shown in the Appendix 10 b.

However, also the noise dimensionality $d_n(t)$ spanned by all noisy realizations of the same pattern increases in the same way as $d_s(t)$ due to the chaotic dynamics, as shown in Fig. 7(b). But $d_s(t)$ has a head start because noisy realizations of one pattern are more similar to each other than to other patterns. Now let us assume that different noise realizations cause different responses that lie entirely within and are uniformly distributed across the signal subspace. This means that the noise randomly flips a number of spins that encode the pattern and thus effectively reduces the dimensionality of the space that faithfully encodes the signal. The effective dimension of the space that is available to represent the signal is then

$$\Delta d(t) = d_s(t) - d_n(t). \tag{27}$$

The expected signal is then given by

$$S_a(t) = \frac{d_s(t) - d_n(t)}{P}$$

$$= \frac{d_s(t)}{P}\left(1 - \frac{d_n(t)}{d_s(t)}\right). \qquad (28)$$

This approximate expression overestimates, but captures quite well the overall shape of the average readout signal shown in Fig. 7(c): Initially, the signal rises in relation to the ratio of dimension of representation space and number of patterns, but ultimately, the signal declines, because the dimensionality spanned by the noise approaches that spanned by the signal. As seen in the inset of Fig. 7(c), the classification accuracy mirrors the behavior of the average readout signal. This transient increase of the linear separability of the pattern classes is rooted in the property of the system that the signal dimensionality initially rises faster than the noise dimensionality.

For small initial $d_s(0)$ and in the limit of vanishing noise $d_n(0) \searrow 0$, the maximum of $\Delta d = d_s - d_n$, and therefore S_α, is reached at $\hat{t} \simeq 2\ln 2\tau \simeq 1.39\tau$ [details given in the Appendix 5, Eq. (A29)]; the peak time depends only weakly on g, as shown in Fig. 8(a). The improvement of the separability due to the transient expansion, the maximum $\Delta d(\hat{t})$ compared to its initial value at $t = 0$, is given by Eq. (A30)

$$\frac{\Delta d(\hat{t})}{\Delta d(0)} \simeq \frac{1}{2}\sqrt{\frac{d_*}{d_s(0)} + \frac{1}{4}}$$

$$\overset{g\to\infty}{=} \frac{1}{\pi}\sqrt{\frac{2N}{d_s(0)} + \frac{1}{4}},$$

where the latter expression is the limit of $g \to \infty$ of the former. The improvement of the signal scales with $N^{\frac{1}{2}}$ and $d_s(0)^{-\frac{1}{2}}$, as shown in Fig. 8(b). The theory slightly underestimates the maximum of Δd in simulations, but captures the scaling relation, predicting a slope of approximately $1/2$. So the peak classification accuracy can be improved by larger networks.

Note that in Fig. 7(c), neither in theory nor in the simulation does the average signal drop to zero for $t \to \infty$. In Fig. 7(b), the noise curve saturates to a smaller value than that of the signal; this is because not all noise realizations leave the flux tube of the original pattern, so that the average distance between realizations remains smaller than that between patterns (analyzed in the Appendix 10 c).

H. Generalization to other network models: Transient chaotic SNR amplification

The computational effect described in the last section relies on two factors: first, a high-dimensional space in which trajectories are nonlinearly embedded, and second,
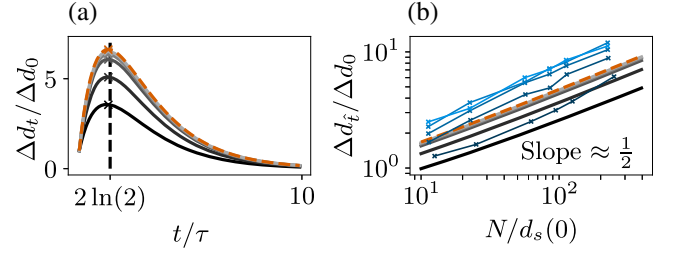

FIG. 8.   Optimal expansion within $2\ln 2$ neuronal activations. (a) Effective dimension $\Delta d(t)$ (27) relative to initial $\Delta d(0)$ for vanishing initial noise amplitude $d_n(0) \searrow 0$. Maxima at $\hat{t} = 2\tau \ln[1 - (1 - \delta)^2/\delta]$, $\delta = \sqrt{d_s(0)/d_*}$ (A28) marked by crosses. The time to maximum for weak initial stimuli $\delta \ll 1$ is well approximated by $\hat{t} \simeq 2\ln 2\tau$ (vertical dashed line). From dark to light gray: $g = 0.5, 1, 2, 3, 5$ and tanh gain function; dashed dark orange curve, Heaviside gain function, limit $g \to \infty$. Initial parameter $N/d_s(0) = 200$. (b) Maximal effective dimension $\Delta d(\hat{t})$ relative to initial $\Delta d(0)$ as a function of $N/d_s(0)$. Same color code as (a). In blue, corresponding simulation results averaged over ten connectivity seeds each and corrected for the probability that a noise realization does not leave the original flux tube; see the Appendix 10 c. Values of g from dark to light as in (a).

the decorrelation dynamics [Eq. (12) and Fig. 7(b)] that causes a small deviation to initially grow slower than a larger deviation. In the setting of a classification task, this mechanism thus enhances the signal-to-noise ratio (SNR).

The first factor is a general feature of all nonlinear neuronal networks. The second factor, we conjecture, should also be a typical property of chaotic networks, because the expansion of distances between three arbitrary trajectories should imply the largest distance to grow faster (in absolute terms) than the two smaller distances, as by a triangle inequality. Therefore, we expect the transient disentanglement of pattern classes to be a general phenomenon in strongly chaotic neuronal networks. To substantiate this claim, we demonstrate this effect in rate networks, and, in a proof-of-principle manner, in a spiking leaky-integrate-and-fire (LIF) network and a LSTM network [39].

1. Rate networks

A network of stochastic rate units with first- and second-order statistics matched to those of a binary network, as in Sec. II E, has a classification performance close to zero, as shown in Fig. 10(a). This is because the matched stochastic rate network is not chaotic; trajectories for different presented patterns converge. A small readout noise thus destroys classification accuracy.

However, how is the performance in the chaotic regime of a rate network? For this, we no longer consider matched statistics and from here on again use rate networks without effective noise and $g^2 > 1$, which are chaotic [21].
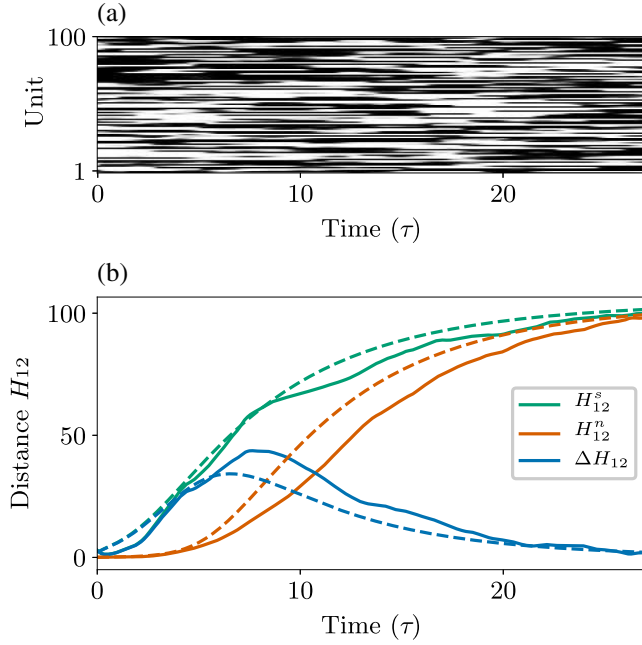
FIG. 9. Evolution of trajectories in a chaotic rate network. (a) Chaotic network activity for $N = 250$, $g = 5.8$, $\sigma_{\text{eff}}^2 = 0$, and $T(h) = \tanh(h)$. Gray scale shows the activity of each neuron between -1 (white) and 1 (black). (b) Time evolution of average distances between patterns (green) and between noisy realizations of a pattern (dark orange). Difference between signal and noise distances (blue). Numerical solutions of Eq. (24) (dashed); for details, see the Appendix 6. Network as in (a) and $P = 50$, $L = 10$, $\sigma = 0.3$.

Similar to binary networks, finite-size chaotic activity [Fig. 9(a)] in rate networks shows a transiently smaller difference between noisy realizations of a pattern than between patterns in the classification task, as shown in Fig. 9(b). This also follows from Eqs. (24) and (25) since the acceleration of the decorrelation is smaller for smaller initial ϵ, making the noise distance grow slower than the signal distance.

Decorrelation in the rate network, however, can take many neuronal time constants if the network is in the mildly chaotic regime [Fig. 9(b)]. The timescale sensitively depends on the recurrent coupling strength, as shown in Fig. 6(b). This can be understood in terms of the decay constant of the time-lagged autocorrelation function: In a noiseless rate network, the timescale of the autocorrelation diverges at the transition to chaos [21].

In the rate network, classification performance jumps to a high value already after the first time step [Fig. 10(c)]. The reason is that all neurons' states are immediately nonlinearly affected by the input pattern. The stimulus is thus immediately projected nonlinearly into an N-dimensional representation that allows linear classification. This becomes apparent by considering that in the time step after stimulus presentation, the input to the network contains a term $\propto (\delta t/\tau) \sum_j J_{ij} T(h_j(0))$.

But even though the dimensionality is immediately N dimensional, the amplitude grows continuously with time and is thus very small at first, $\propto \delta t$. This behavior exposes the qualitative difference in the interpretations of ϵ: In the binary network, there is a direct link between ϵ and the dimensionality of the signal space, while in the rate network, ϵ is a measure of the Euclidean distance between trajectories and is only very indirectly related to the number of dimensions across which this distance is distributed. Hence, the difference between inter- and intrapattern distances in Fig. 9(b) peaks at a later time point than the classification accuracy in Fig. 10(b), because the rate network performs a transient signal amplification rather than a transient dimensionality expansion as in the binary network case.

Even though the dimensionality of the representation immediately after stimulus presentation equals N, distances between stimuli in the new directions are small at first. The classification in the rate network therefore relies on fine-tuned and very large readout weights in the beginning [Fig. 10(d)]. Therefore, the initial classification accuracy is severely impaired by adding even weak noise to the readout [Fig. 10(c)]. This addition of noise, in turn, results in a peak of the accuracy predictable by the theoretical peak in signal amplification [Fig. 9(b)].

In summary, rate networks with continuous signaling perform a transient amplification of the signal-to-noise ratio, rather than a transient dimensionality expansion. Compared to binary networks with discrete signaling, the resulting empirical differences are the strong dependence of the decorrelation timescale on the coupling strength, the existence of a minimal coupling strength required for amplification, the absence of the residual correlation, and the initial high sensitivity to readout noise.

2. Spiking networks

To demonstrate that the same computational effect translates to spiking networks which are furthermore not all-to-all connected, we consider the same task in a purely inhibitory network of LIF neurons with fixed in-degree in the asynchronous-irregular firing state. Trajectories in these networks are known to have a small stable local environment (flux tube) but exhibit chaos for perturbations leaving the flux tube [34,40], just as the binary networks considered in this manuscript. In the binary networks, the flux-tube borders are given by input perturbations that are just sufficient to cause a single spin to flip. By binning the spike trains of the LIF network with a bin width equal to the membrane time constant, we obtain approximately binary vectors if the bin width is small compared to the inverse firing rate. Therefore, we can use the same training and analysis procedure as for the binary networks, and also interpret distances in terms of dimensionality. Further details are given in the Appendix 11. The LIF network dynamics and performance show the same features found in
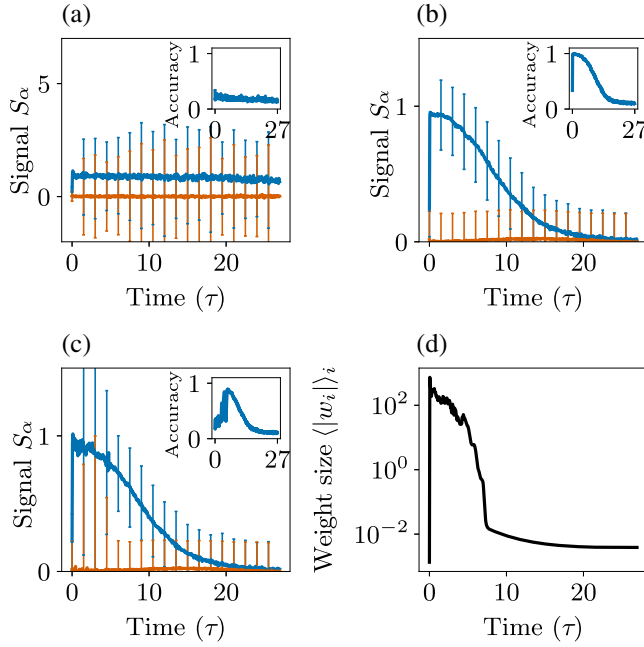
FIG. 10. Stimulus representation in a rate network. (a)–(c) Signal of correct readout (blue), signal of wrong readouts (dark orange), and classification accuracy (inset). (a) Including noise σ_ξ^2 matched to the statistics of a binary network according to Eq. (22). Parameters otherwise as in Fig. 9. Because the frozen effective noise suppresses chaos, trajectories for different stimuli converge toward the same state, and a small readout noise $\sigma_{readout} = 10^{-4}$ results in classification at chance level. (b) Removing the noise in the network, and no readout noise. The network is chaotic as in Fig. 9, and classification accuracy jumps to 1 in the first time step. (c) Including readout noise $\sigma_{readout} = 10^{-4}$ impedes classification accuracy until trajectories are sufficiently separated. (d) Average norm of individual readout weights for the setting in (b). Since trajectories are very close in the initial time period [compare also Fig. 9(b)], initial readout weights are very large, explaining the sensitivity to readout noise in (c).

binary networks: Distances between states quickly grow toward a residual correlation (Fig. 11). Dimensionality expansion takes place on a timescale within which each neuron fires only a single spike or less on average ($\nu^{-1} \simeq 55$ ms). The distance between pattern classes initially grows faster than the noise distance [Fig. 11(b)], and the readout signal and classification performance show a corresponding transient peak [Fig. 11(c)]. As in the binary network, some noise realizations do not leave the flux tube of the unperturbed pattern, causing a reduced long-term average noise distance and a nonzero plateau of the residual classification performance.

3. LSTM networks

Finally, to demonstrate the existence of the computational effect in a powerful specialized machine-learning architecture, we consider the same task in a recurrent LSTM network [39]. The architecture is similar to the rate networks we consider but contains a large number of additional dynamical "gating" variables which control when, where, and by how much the cell states interact with each other, resulting in considerably more complex dynamics than in a rate network with fixed coupling matrix. We use a vanilla pytorch implementation, choosing the initialization parameters such that the network exhibits spontaneous chaotic fluctuations over a moderate range of timescales, as seen in Fig. 12(a). Details about the parameters and task implementation are given in the Appendix 11. In Figs. 12(b) and 12(c), we see a behavior in close analogy to that shown in Figs. 9(b) and 10 for rate networks. Signal and noise distances increase differentially fast, and there is a pronounced transient peak in the classification accuracy. Other than in the binary and LIF networks, which use discrete signaling, the rate and LSTM networks do not have locally stable flux tubes and no residual plateau in the classification performance.
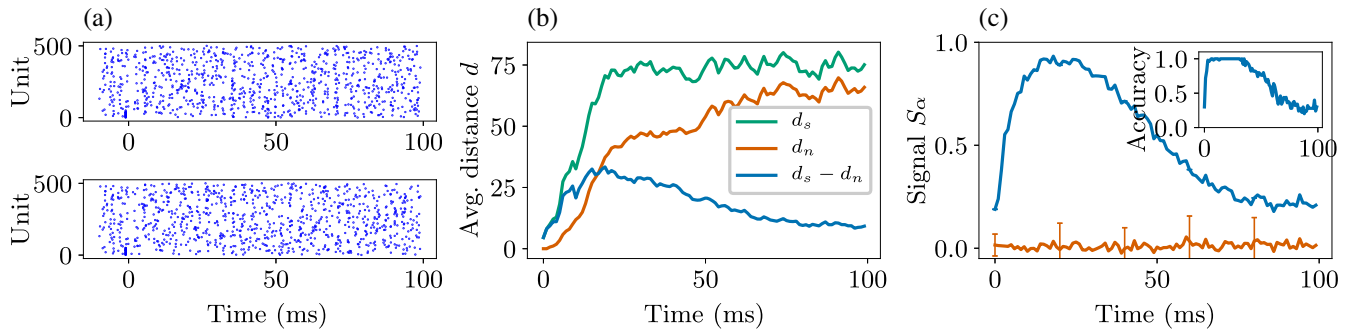

FIG. 11. Transient chaotic dimensionality expansion in a spiking network. (a) Raster plots of spiking activity in the network for two different patterns; average firing rate per neuron $\nu \simeq 19$ Hz. Parameters $N = 500$, in-degree $K = 125$, weights $J_{ii} = -1.0$ mV, and membrane time constant $\tau_m = 10$ ms; additional detail given in the Appendix 11. (b) Evolution of average Euclidean distance between different patterns (green), noisy realizations of a pattern (dark orange), and the difference of the two (blue). (c) Signal of matching readout (blue), nonmatching readouts (dark orange), and classification performance (inset).
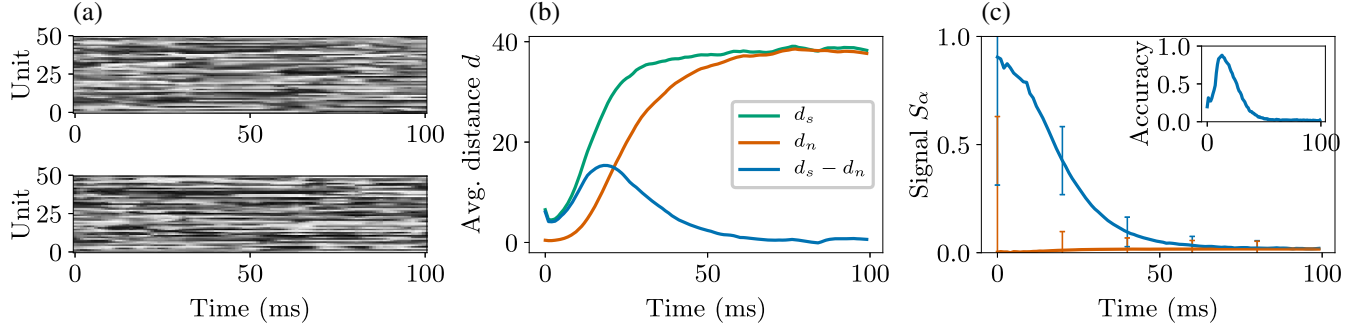
FIG. 12. Transient chaotic SNR amplification in a LSTM network. (a) Chaotic network activity of the first 50 units for two different patterns. Parameters $N = 200$ and $g = 5.8$ for the hidden to input-, forget- and cell-gates weights and all other weights using a standard $\mathcal{U}(-N^{-\frac{1}{2}}, N^{-\frac{1}{2}})$ distribution, with additional detail given in the Appendix 11. Gray scale shows the activity of each hidden unit between -1 (white) and 1 (black). (b) Evolution of average Euclidean distance between different patterns (green), noisy realizations of a pattern (dark orange), and the difference of the two (blue). The short initial decay in pattern distances is likely due to a short networkwide contraction of the dynamics after pattern presentation caused by the interaction of the gating variable states and the L changed hidden states. (c) Signal of correct readout (blue), incorrect readouts (dark orange), and classification performance (inset).

III. DISCUSSION

This manuscript compares the effect of discrete and continuous signaling on the dynamics and function of neuronal networks. Focusing on binary classification as a fundamental computation, it addresses the question how the temporal dynamics can be used to represent stimuli. Separating representations translates into asking how state trajectories diverge or converge if different stimuli are presented. Technically, this amounts to quantifying chaos in such networks.

A model-independent path-integral approach enables comparisons across models. We find that the dynamic mean-field theory is of identical structure for networks of binary units and for continuous-rate networks. In binary networks, we discover a network-size-dependent transition to chaos and the existence of a chaotic submanifold. We elucidate the qualitative differences to chaos in rate networks in terms of the mechanism causing chaos, timescales, and parameter regimes.

Applied to classification, chaotic dynamics causing a relative dimensionality expansion of representations leads to a mechanism of fast and transient computation in binary networks with discrete signaling. We describe a generalization of this effect as a transient signal-to-noise amplification in chaotic rate networks with continuous signaling.

The remainder of the discussion puts these results into context of the literature, mentions limitations, and provides an outlook.

A. Differences and similarities across neuron models

1. Transition to chaos in binary networks at finite size

We demonstrate that there is a transition to chaos in finite-size binary networks described by a field-theoretical replica calculation. Our results are consistent with works on sparse random boolean networks with synchronous update

showing a chaos transition for in-degree $K = 2$ [33,41], which relates here to the transition at $N \approx 2$ for the Heaviside activation function. Derrida and Pomeau [33] approximate the disorder by annealed averages and use synchronous update, while we compute the quenched averages and employ asynchronous update. In mean-field theory, the in-degree in sparse networks plays a similar role as the network size in dense networks. Other works that investigated the edge of chaos numerically in discretely coupled networks have also found small in-degrees as critical coupling [18,20,42].

2. Correspondence of DMFT in binary and rate networks

The model-independent field theory presented here exposes a one-to-one match of the stationary activity statistics in dynamical mean-field approximation of binary and rate networks. Exposing identities between neuron models is useful to see if and how the results generalize. Steps in this direction where already taken in Grytskyy *et al.* [43], who showed that weak pairwise correlations can be explained by linearizing LIF neurons, Hawkes processes, and binary neurons, mapping them to noisy linear rate models. The results presented here are more general since they apply not only to the linearization of the models but hold for the nonlinear behavior as well. The equivalence of time-lagged autocorrelations is shown here for stationary statistics; for nonstationary dynamics also the effective noise strength should vary as a function of the time ([44], Chap. 6.4).

3. Assumptions on connectivity

The assumption of Gaussian connectivity $J_{ij} \sim \mathcal{N}(\bar{g}/N, g^2/N)$ straightforwardly generalizes to other connectivities, as long as higher than second cumulants are suppressed by powers of N^{-1}. Scaling the mean connectivity as $\bar{g}/N$ yields a consistent approximation in $1/N$.

Sparse connectivity, however, typically leads to a scaling $\bar{g}/\sqrt{N}$. Formally, a consistent treatment therefore requires Gaussian fluctuations of the mean activity field $\langle \mathcal{R}^2 \rangle \sim 1/N$, which is possible in the presented framework. Such fluctuations are, however, suppressed by negative feedback [45] in the inhibition-dominated (balanced) regime $\bar{g} < 0$. For multiple populations, the DMFT equations acquire population indices, but stay structurally the same (cf. Ref. [25] for binary neurons and Refs. [32,46] for rate neurons). Scale-free distributions of weights can violate the assumptions and require a different approach [47].

4. Relation of the model-independent path-integral formulation to earlier work

The seminal work by Sompolinsky *et al.* [21] on rate neurons used statistical field theory [22], and the work by van Vreeswijk and Sompolinsky [25] on binary neurons relied on a disorder average of the master equation [27,28]. The statistical field theory that we develop here captures both model classes and is similar to the Martin-Siggia-Rose–de Dominicis–Janssen formalism [48,49] for rate neurons (reviewed, e.g., in Refs. [50–52]). In particular, this formulation exposes the identical structure of the mean-field approximations.

Binary networks with asymmetric connectivity show nonequilibrium dynamics, so that the Ising Hamiltonian cannot be used. Instead, complete information about the system dynamics needs to be captured. Full information is supplied by the master equation, for which an established approach is the Doi-Peliti formalism [53,54]. The fields in the latter approach, however, have no intuitive physical interpretation, even though they allow the construction of mean-field equations and fluctuation corrections [55]. Closer to our method are the approaches by Sommers [56], Andreanov *et al.* [57], and Lefèvre and Biroli [58], which can be obtained as special cases from our formulation.

5. Different chaotic dynamics in binary and rate networks

Qualitative differences between chaos in binary and rate networks can be summarized as follows: (i) Rate networks with activity statistics matched to that of binary networks are nonchaotic. This shows that discrete signaling provides a different mechanism that drives chaos in binary networks, and that the mechanism causing chaos in rate networks is not effective in binary networks. A marginally chaotic solution is approached in matched rate networks when sending the activation function to the Heaviside limit. This is consistent with the finding that rate networks are always chaotic if the activation function has an infinite slope [32]. (ii) In the limit of large numbers of neurons (now unmatched) rate networks have a critical coupling strength beyond which they transition to chaos, while binary networks are always chaotic in this limit. At finite network sizes, binary networks have a size-dependent, critical coupling strength, which is typically very low. (iii) Decorrelation of trajectories in binary networks is generally faster than in rate models. In binary networks, it takes place on a timescale given by the interval between state changes of individual neurons and is only mildly affected by the network coupling. In rate models, it strongly depends on the coupling strength showing a critical slowing-down at the transition to chaos. Stochasticity gradually smooths out this divergence [23]. (iv) Trajectories in binary networks decorrelate only up to a residual correlation, while those in rate networks completely decorrelate. (v) Binary networks have an infinite Lyapunov exponent, so that decorrelation starts off with a finite slope even for infinitesimal initial perturbations. In rate networks, the initial decorrelation is an exponential function, whose slope therefore vanishes for infinitesimal perturbations.

6. Origin of the difference

In the rate network, the noise is external and frozen, but a binary network's noise realization depends acutely on the initial value of the system. Thus, perturbing the initial value also changes the noise realization. In particular, due to the thresholding operation that produces the discrete signal, a tiny perturbation in the input can cause a flip of the neuron, and consequently, a macroscopic change of the network state; the probability that this change happens increases with the number of targets that receive this perturbation, thus, with network size. This increase is due to the strong synapses $|J_{ij}| \propto N^{-\frac{1}{2}}$ (for $\langle\langle J_{ij}^2 \rangle\rangle \sim gN^{-1}$). For large networks, the growth of the perturbation corresponds to a macroscopic change in the noise realization.

The presented replica calculation provides a complementary explanation for the qualitative difference between binary and rate neurons. The question of a chaos transition is reduced to studying how correlations are transferred from the inputs of a pair of neurons to their outputs: The change of the correlation between replicas is proportional to the mismatch between the correlation at time t, the second line on the right of Eq. (11), and the correlation transmitted through a pair of neurons and connectivity given by the third line, a function $c_{\text{out}}(c_{\text{in}})$. Such a transmission curve $c_{\text{out}}(c_{\text{in}})$ is shown for discrete signaling and for continuous signaling in Fig. 13. Clearly, if $c_{\text{out}}(c_{\text{in}}) < c_{\text{in}}$ for a c_{in} close to perfect correlation just below unity, the correlation decreases over time, the dynamics is chaotic; if $c_{\text{out}}(c_{\text{in}}) > c_{\text{in}}$, the correlation regenerates, the dynamics is regular. While the slope $c'_{\text{out}}(c_{\text{in}} \rightarrow 1)$ for discrete signaling diverges as $\propto (1 - c_{\text{in}})^{-\frac{1}{2}}$, it stays finite for continuous signaling. The infinite slope for networks with discrete signaling leads to an infinite Lyapunov exponent for $N \rightarrow \infty$. For continuous signaling, the slope is finite as long as the slope of the activation function is bounded, resulting in finite Lyapunov exponents. This analytical view of the different transitions to chaos in continuous and
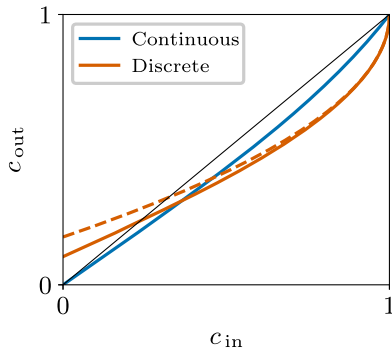
FIG. 13.   Correlation transmission by a pair of neurons. Correlation coefficient c_{out} between outputs as a function of the correlation coefficient c_{in} between inputs. Discrete signaling [dark orange, Eq. (A44)]; continuous signaling [blue, Eq. (A41)]. Approximation for discrete signaling in the limit $c_{in} \to 1$ (dashed dark orange) shows a behavior of the form $\propto -(1 - c_{in})^{\frac{1}{2}}$, so the slope diverges as $\propto (1 - c_{in})^{-\frac{1}{2}}$. For comparison: identity mapping with unit slope (black line).

discrete networks is also consistent with numerical findings [59].

Correlation transmission by pairs of neurons is well studied experimentally and theoretically; e.g., Refs. [60–63]. A diverging slope of the correlation-transmission curve, shown here for binary neurons, has also been demonstrated for spiking neurons without reset [62] and for the LIF model [61,64–66], suggesting that these model classes behave similarly with regard to the transition to chaos.

7. Flux tubes in binary and spiking networks

There is a tight link between the replica calculation and chaos in spiking networks examined in terms of the divergence rate between flux tubes [40]. Flux tubes are neighboring portions of the phase space within which perturbations of the state do not cause a global change of subsequent activity [34]. Binary neurons are formally simpler than spiking models, because one can investigate changes of network states directly instead of analyzing spike patterns (see the Appendix 7), binary neurons do not have additional internal degrees of freedom, such as the membrane potential, and their activation times are given by predetermined update times. Therefore, perturbations inside a flux tube are not forgotten exponentially as in LIF neurons, but instantly. Just like in LIF networks, the distance to a flux-tube boundary shrinks with network size as $1/N$. The divergence rate between flux tubes scales as $\sqrt{N}$, as opposed to N in the LIF network [40]. Both are consistent with an infinite Lyapunov exponent for $N \to \infty$. Chaotic spiking activity has also been investigated in Lajoie *et al.* [67,68], who showed that chaotic quadratic-integrate-and-fire networks exhibit a reduced spike pattern entropy, indicating that they explore only a lower-dimensional manifold in phase space.

B. Computation in the chaotic regime

1. Transient chaotic SNR amplification

We find that strongly chaotic neuronal networks of different types invariably exhibit a transiently improved separability of low-dimensional inputs. This is at first sight surprising, because chaotic dynamics amplifies noise as well as informative differences. However, the variability within a class (noise) is typically smaller than the variability across classes. We find the latter to be amplified more strongly than the former, thus improving the linear separability of the classes. We argue that this relative amplification is a general effect in nonlinear, high-dimensional chaotic systems. There is an analogy to astrophysics: Space in the Universe is locally expanding everywhere. Thus, all points are drawn apart, forming diverging trajectories. As a result of the ubiquitous expansion, galaxies move apart ever faster the greater the distance between them, as described by the Hubble constant. In the same way, trajectories in the network that are farther apart (different classes) separate faster than trajectories that are initially closer (noisy realizations within the same pattern class). Unlike the Universe, however, the network state space is higher dimensional and inherently expands along highly curved directions. Therefore, the faster expanding, larger differences are also more strongly affected by the nonlinearity and are more quickly embedded into the surrounding higher dimensions. Also, unlike the Universe, the state space volume of the network is finite and constant, so some directions must shrink to conserve the total volume. Trajectories therefore do not diverge indefinitely but reach a stable average distance determined by the volume of the limiting chaotic attractor. They continue to be mixed by the expanding and shrinking dynamics, such that information about their initial distance relations is eventually forgotten and classification performance subsides.

2. Elucidation by dimensionality in binary networks

This qualitative picture is made concrete in binary networks: Their discrete state space allows an interpretation of the growing average distances in terms of dimensionality. This allows us to express the improvement of classification accuracy in terms of the difference between the dimensionality of the representation of the signal d_s and the number of dimensions corrupted by noise d_n. Their temporal evolution follows stereotypic decorrelation curves obtained from a replica calculation. Dimensionality and separability are linked, because each dimension allows the linear separation of two additional random features [16,69]. The effect can also be viewed as a dynamical version of the usually statically applied kernel trick [15].

3. Relation of chaos and computational power

It has been argued that close to the edge of chaos, networks show an optimal trade-off between separation of

stimuli and generalization [17,18,20,31]. Our analysis provides a time-dependent perspective on this hypothesis. Indeed, the signal and noise dimensionalities that approximate the classification performance can be compared to the kernel quality and generalization rank (Vapnik-Chervonenkis dimension [15]) measures introduced by Legenstein and Maass [20]. Our results show that binary networks generally have a short memory lifetime. But in contrast to rate networks, it also does not reduce strongly when moving deeper into the chaotic regime. For tasks that require only short memory, performance can benefit from the increased separation even deep in the chaotic regime, in particular because the peak of the informative dimensionality $\max(d_s - d_n)$ increases with network size. This is in line with the results of Snyder *et al.* [42], where for small readout delay, performance stays high when increasing the in-degrees [e.g., their Figs. 3(a) and 3(b)]. Overall, these observations raise the question whether the effect could be combined with a prolonged memory lifetime, for example, by heterogeneous time constants of neurons or synapses, clustered connectivity [70], or by feeding the readouts back into the network.

4. Other related works

Recurrent networks can be transformed to deep feed-forward networks with weight sharing by "unrolling" them in time. Successive layers then correspond to adjacent time steps in the recurrent network. Chaotic iterative maps were shown to yield an increase of the dimensionality of representations toward deeper layers [71]. Training specifically on low-dimensional representations yields facilitated feature generation by dimensionality expansion in early layers and feature selection and generalization by dimensionality suppression in later layers [72]. Farrell *et al.* [73] investigated (continuous) recurrent networks in a classification task similar to ours, also considering the strongly chaotic regime. They focused on late-time compression of the representation by training and found that chaos benefits learning of the task, for which our results provide a principled explanation. Their results provide clues on how training interacts with the random connectivity, an interesting avenue for future work.

The mechanism of transient computation we investigate coexists with non-normal amplification [74–77], which is caused by effective feed-forward structures embedded in nonorthogonalizable coupling matrices. Non-normal amplification is especially strong in the chaotic regime [74]. The here described mechanism, however, also applies to normal matrices. Finally, recent developments in statistical mechanics of computation in neuronal networks are reviewed by Bahri *et al.* [78].

5. A mechanism of fast computation in spiking networks

The peak classification performance in binary reservoirs is reached on the scale of a single neuronal time constant after stimulus onset. This scale also holds approximately in the LIF network. However, the peak time can likely be even shorter for higher in-degrees or firing rates, as these influence the divergence rate of flux tubes in such networks [40]. The mechanism may explain how computation can spread rapidly through the hierarchical networks of the brain: The time window corresponds to one activation per contributing neuron on average, allowing the fast feed-forward processing latencies of $\lesssim 50$ ms per stage measured experimentally in cortical areas [79,80]. This perspective suggests how networks that employ discrete communication may compute rapidly on the basis of a few spikes rather than requiring a prolonged averaging over time. A possible impediment to computation in a chaotic system is its sensitivity to initial conditions, which, apart from the input patterns, are kept fixed in this manuscript. One solution would be a mechanism which quenches variability at appropriate times. Another possibility was exposed by Lajoie *et al.* [81], who showed that chaotic spiking networks can reliably encode inputs despite changing initial conditions, because each input confines the chaotic activity to a different manifold.

6. Experimental evidence and predictions

The olfactory system is a potential candidate to rely on the transient computational mechanism we describe, because it is specialized on classification of patterns without a temporal component. In the vertebrate olfactory bulb, an odor activates a comparably low-dimensional pattern of glomeruli, the input layer to a higher-dimensional recurrent network that needs to separate representations to enable classification of odor identity by subsequent processing stages. The insect antennal lobe shares this basic organization. Recordings in the olfactory systems in zebrafish [82], locust [83], and rats [84] show a representation of stimuli that is consistent with the here found mechanism of transient dimensionality expansion.

In zebrafish, the activities of mitral cells in the olfactory bulb show a high correlation for similar odors shortly after stimulus presentation. Subsequently, they decorrelate on a timescale of approximately 800 ms, reaching a residual correlation of about 40% [see Ref. [82], Fig. 2(e)]. The discriminability of these similar odors by a linear readout from the mitral cells improves within the same time span to nearly error-free classification [see Ref. [82], Fig. 2(i)]. In the process, the population statistics stays approximately constant. These features are in line with transient dimensionality expansion, except that the classification accuracy does not decline again after the improvement. However, this missing decline could be caused by feedback stabilizing the representation after recognition or be related to the sustained presentation of the odor stimulus. Recordings in the locust antennal lobe show qualitatively similar behavior [83]. In particular, these experiments report a decoding accuracy that is highest within the

transient phase. In the rat olfactory bulb, inhalation also triggers a fast decorrelation transient of approximately 100 ms, during which, odor identity is encoded in the instantaneous spike pattern, and decoding accuracy rapidly peaks after approximately 70 ms before declining to a lower level [84]. Future work should systematically investigate if the dynamics in these biological systems is in fact chaotic, for which, suitable analysis methods are available [85,86]. Also, the analysis of inter- and intraclass distances and classification by linear readouts is applicable to experimental spiking data. The computational mechanism we describe needs a reliable initial state from which trajectories diverge. In the rat olfactory system, this reset could be tied to inhalation onset. In cortex, stimuli seem to quench the variability of spontaneous activity to evoke relatively low-dimensional responses [87–90]. To check for chaotic dynamics, one could therefore analyze the growth of intertrial variability during and after stimulation offset. We finally list concrete testable predictions for neural systems that implement classification by transient chaotic dimensionality expansion:

  (i) Variability is small or quenched at stimulus onset, then transiently increases and reaches a stable value.
  (ii) Not only the interclass distances but also the intraclass (noise) distances increase, although initially slower.
  (iii) Decoding accuracy based on linear readouts trained at each time point shows a peak, and this peak occurs before the distances saturate.

ACKNOWLEDGMENTS

APPENDIX

1. Model-independent mean-field theory for random networks

This section presents a self-contained derivation of the model-independent mean-field theory for networks with

Gaussian random connectivity $J_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\bar{g}/N, g^2/N)$. The N neurons have inputs $\boldsymbol{h}(t) = [h_1(t), ..., h_N(t)]$ and outputs $\boldsymbol{x}(t) = [x_1(t), ..., x_N(t)]$, and the neuronal dynamics is described by the conditional probability functional $\rho[x_i|h_i]$. For deterministic neurons, where $x_i = f[h_i]$ is some causal functional of the input, one may set $\rho[x_i|h_i] = \delta(x_i - f[h_i])$. We use vectorial notation to denote

$$\rho[\boldsymbol{x}|\boldsymbol{h}] = \prod_{i=1}^{N} \rho[x_i|h_i], \tag{A1}$$

because, given their inputs $\{h_i\}$, neurons are otherwise pairwise independent. The probability functional $\rho[x_i|h_i]$ is assumed to be strictly causal, which is, $x_i(t)$ is independent of $h_i(s > t)$; a more explicit notation would be $\rho[x_i(\circ+)|h_i(\circ)]$ denoting that the time argument $x_i(t + \epsilon)$ must be infinitesimally advanced by $\epsilon > 0$ compared to the argument of $h_i(t)$ for ρ to depend on h.

The joint statistics of input and output is then

$$\rho[\boldsymbol{x}, \boldsymbol{h}] = \rho[\boldsymbol{x}|\boldsymbol{h}]\rho[\boldsymbol{h}]. \tag{A2}$$

The distribution of the inputs $\boldsymbol{h}$ is given as the marginalization over $\boldsymbol{x}$ as

$$\rho[\boldsymbol{h}] = \int \mathcal{D}\boldsymbol{x}\rho[\boldsymbol{x}, \boldsymbol{h}]$$
$$= \int \mathcal{D}\boldsymbol{x}\rho[\boldsymbol{x}|\boldsymbol{h}]\rho[\boldsymbol{h}]. \tag{A3}$$

The connectivity $\boldsymbol{J}$ couples the outputs $\boldsymbol{x}$ of the neurons to the input $\boldsymbol{h}$ as

$$\boldsymbol{h}(t) = \boldsymbol{J}\boldsymbol{x}(t).$$

So, in the marginalization (A3) over $\boldsymbol{x}$, we need to set

$$\rho[\boldsymbol{h}(\circ)|\boldsymbol{x}(\circ)] = \delta[\boldsymbol{h} - \boldsymbol{J}\boldsymbol{x}]$$
$$= \int \mathcal{D}\hat{\boldsymbol{h}} \exp(\hat{\boldsymbol{h}}^T \boldsymbol{h}) \exp(-\hat{\boldsymbol{h}}^T \boldsymbol{J}\boldsymbol{x}), \tag{A4}$$

where the path-integral measure is $\int \mathcal{D}\hat{\boldsymbol{h}} = \prod_t \int_{-i\infty}^{i\infty} [d\hat{h}(t)/2\pi i]$, and the inner product is meant as $\hat{\boldsymbol{h}}^T \boldsymbol{h} = \sum_{i=1}^{N} \int_{-\infty}^{\infty} dt \hat{h}_i(t) h_i(t)$. By connecting the outputs back to the inputs, Eq. (A2) may seem to take a circular structure like $\rho[\boldsymbol{x}, \boldsymbol{h}] = \rho[\boldsymbol{x}|\boldsymbol{h}]\rho[\boldsymbol{h}|\boldsymbol{x}]$. But since the first conditional probability is causal, and the second couples only equal time points, Eq. (A2) is more accurately represented as

$$\rho[\boldsymbol{x}, \boldsymbol{h}] = \rho[\boldsymbol{x}(\circ+)|\boldsymbol{h}(\circ)]\rho[\boldsymbol{h}(\circ)|\boldsymbol{x}(\circ)],$$

which is ordered in time, resulting in a spiraling structure.

Performing the disorder average $\langle\ldots\rangle_{\boldsymbol{J}}$ of Eq. (A3), the only term affected is the last exponential factor in the second line of Eq. (A4), which yields

$\langle\exp(-\hat{\boldsymbol{h}}^T\boldsymbol{J}\boldsymbol{x})\rangle_{\boldsymbol{J}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(\bar{g}/N,g^2/N)}$

$$= \exp\left(-\frac{\bar{g}}{N}\sum_{i=1}^{N}\hat{h}_i^T\sum_{j=1}^{N}x_j + \frac{g^2}{2N}\sum_{i,j=1}^{N}(\hat{h}_i^T x_j)^2\right). \quad (A5)$$

Here, the scalar product in the last term in the exponent rewrites explicitly as

$$(\hat{h}_i^T x_j)^2 = \iint dt ds\ \hat{h}_i(t)\hat{h}_i(s)x_j(t)x_j(s).$$

The terms suggest the introduction of the auxiliary fields $\mathcal{R}(t) := (\bar{g}/N)\sum_j x_j(t)$ and $\mathcal{Q}(t,s) := (g^2/N)\sum_j x_j(t)x_j(s)$ to rewrite Eq. (A5) as

$$\prod_i \exp\left(-\hat{h}_i^T\mathcal{R} + \frac{1}{2}\hat{h}_i^T\mathcal{Q}\hat{h}_i\right), \quad (A6)$$

where the bilinear form is to be read as

$$\hat{h}_i^T\mathcal{Q}\hat{h}_i = \iint dt ds\ \hat{h}_i^T(t)\mathcal{Q}(t,s)\hat{h}_i(s).$$

The appearance of the product sign and the neuron-independent fields $\mathcal{R}$ and $\mathcal{Q}$ signifies that the problem becomes completely symmetric with regard to neurons. Enforcing the definitions of the auxiliary fields by Dirac distributions represented in the Fourier domain analogous to Eq. (A4) yields another pair of fields $\hat{\mathcal{R}}$ and $\hat{\mathcal{Q}}$ and brings Eq. (A3) into the form

$$\langle\rho[\boldsymbol{h}]\rangle_{\boldsymbol{J}} \overset{(A3),(A4)}{=} \int \mathcal{D}\boldsymbol{x}\rho[\boldsymbol{x}|\boldsymbol{h}]\langle\delta[\boldsymbol{h}-\boldsymbol{J}\boldsymbol{x}]\rangle_{\boldsymbol{J}} \quad (A7)$$

$$\overset{(A5),(A6)}{=} \int \mathcal{D}\{\mathcal{Q},\mathcal{R},\hat{\mathcal{Q}},\hat{\mathcal{R}}\}\exp\left(-\frac{N}{\bar{g}}\hat{\mathcal{R}}^T\mathcal{R} - \frac{N}{g^2}\hat{\mathcal{Q}}^T\mathcal{Q}\right)$$

$$\times \prod_i \int \mathcal{D}\{x_i,\hat{h}_i\}\rho[x_i|h_i]$$

$$\times \exp\left(\hat{h}_i^T h_i - \hat{h}_i^T\mathcal{R} + \frac{1}{2}\hat{h}_i^T\mathcal{Q}\hat{h}_i + \hat{\mathcal{R}}^T x_i + x_i^T\hat{\mathcal{Q}}x_i\right).$$
$$(A8)$$

Adding the normalization condition by integrating over $\boldsymbol{h}$,

$$1 \equiv \int \mathcal{D}\boldsymbol{h}\langle\rho[\boldsymbol{h}]\rangle_{\boldsymbol{J}},$$

we note that this integral affects only the last two lines in Eq. (A8). The exponent in the second line can be considered an action of a field theory for the auxiliary fields $\{\mathcal{Q},\mathcal{R},\hat{\mathcal{Q}},\hat{\mathcal{R}}\}$. The integral in the last two lines

appears to the power of N, so that one may rewrite the full expression as

$$\int \mathcal{D}\{\mathcal{Q},\mathcal{R},\hat{\mathcal{Q}},\hat{\mathcal{R}}\}\exp(N\Omega[\mathcal{R},\mathcal{Q},\hat{\mathcal{R}},\hat{\mathcal{Q}}])$$

with

$\Omega[\mathcal{R},\mathcal{Q},\hat{\mathcal{R}},\hat{\mathcal{Q}}]$

$$:= -\frac{\mathcal{R}^T\hat{\mathcal{R}}}{\bar{g}} - \frac{\mathcal{Q}^T\hat{\mathcal{Q}}}{g^2} + \ln \int \mathcal{D}\{x,h,\hat{h}\}\rho[x|h]$$

$$\times \exp\left(\hat{h}^T h - \hat{h}^T\mathcal{R} + \frac{1}{2}\hat{h}^T\mathcal{Q}\hat{h} + \hat{\mathcal{R}}^T x + x^T\hat{\mathcal{Q}}x\right).$$

We now compute the values of the auxiliary fields that provide the dominant contribution to the probability mass. The appearance of N in the exponent $N\Omega[\mathcal{R},\mathcal{Q}]$ suggests to perform the integration over the fields $\{\mathcal{Q},\mathcal{R},\hat{\mathcal{Q}},\hat{\mathcal{R}}\}$ in saddle-point approximation, demanding $(\delta\Omega/\delta\{\mathcal{Q},\mathcal{R},\hat{\mathcal{Q}},\hat{\mathcal{R}}\}) \overset{!}{=} 0$, which yields four conditions for the saddle-point values $R,Q,\hat{R},\hat{Q}$ of the fields

$$R(t) = \bar{g}\langle x(t)\rangle_{\Omega(R,Q)},$$

$$\hat{R}(t) = \bar{g}\langle\hat{h}(t)\rangle_{\Omega(R,Q)} \equiv 0,$$

$$Q(t,s) = g^2\langle x(t)x(s)\rangle_{\Omega(R,Q)},$$

$$\hat{Q}(t,s) = \frac{g^2}{2}\langle\hat{h}(t)\hat{h}(s)\rangle_{\Omega(R,Q)} \equiv 0.$$

Here the expectation value is $\langle\ldots\rangle_{\Omega(R,Q)} = \int\mathcal{D}x\ldots\int\mathcal{D}\{h,\hat{h}\}\rho[x|h]\exp(\hat{h}^T h - \hat{h}^T R + \frac{1}{2}\hat{h}^T Q\hat{h})$. The denominator coming from the outer derivative of the logarithm appearing in the expression for Ω does not contribute, because the normalization condition of the latter distribution is unity, since the exponential term is the moment-generating functional of a Gaussian process $h \sim \mathcal{N}(R,Q)$ and $\rho[x|h]$ is normalized, allowing us to rewrite

$$\langle\ldots\rangle_{\Omega(R,Q)} = \int \mathcal{D}x\ldots\langle\rho[x|h]\rangle_{h\sim\mathcal{N}(R,Q)}. \quad (A9)$$

The auxiliary fields $\hat{h}$ and all their powers are zero on expectation, which is a consequence of the normalization (see Refs. [52,91], Sec. X).

The equations thus lead to the result in the main text, Eqs. (2) and (3).

2. Derivation of the mean-field equation for binary networks

Having obtained the saddle-point solution to the path integral developed in the previous section, the first result is the time evolution of the mean input activity $R(t)$. For

binary networks, we need to insert information specific to the neuron model in order to compute $\langle x(t) \rangle_{h \sim \mathcal{N}(R(t), Q(t,t))}$ in Eq. (2). This means we need $\rho[x_i(t)|h_i]$ for a binary neuron. Note that only the probability distribution of the activity at a single time point is needed, which is much simpler to obtain than a distribution across all time points, which would include not only the dependence on the input history, but also on the neuron's own activity state. For the most compact presentation, here we use the bitlike $n \in \{0, 1\}$ representation instead of the Ising $x \in \{-1, 1\}$ representation used in the main text. The results between the two can be easily related by the mapping

$$x = 2n - 1. \qquad \text{(A10)}$$

The bitlike $n \in \{0, 1\}$ representation has the advantage that we need to consider only the active state in averages, since the inactive $n = 0$ state does not contribute. In this case, the probability of finding a neuron active at some time t,

$$p[n_i(t) = 1|h_i] = \int_{-\infty}^{t} \frac{dt'}{\tau} e^{-(t-t')/\tau} T_p(h_i(t')), \qquad \text{(A11)}$$

is given by the probability $T_p(h_i(t'))$ to be activated at any prior update time point t' and the survivor function $e^{-(t-t')/\tau}$ [29], the probability that no further update happened since. Therefore, plugging into Eq. (2)

$$R(t) = \bar{g} \langle x(t) \rangle_{\Omega(R,Q)}$$
$$\stackrel{(A10)}{=} \bar{g}(2\langle n(t) \rangle_{\Omega(R,Q)} - 1)$$
$$= \bar{g} \int_{-\infty}^{t} \frac{dt'}{\tau} e^{-(t-t')/\tau} \langle 2T_p(h) - 1 \rangle_{h \sim \mathcal{N}(R(t'), Q(t',t'))},$$

where we use $\int_{-\infty}^{t} (dt'/\tau) e^{-(t-t')/\tau} = 1$ from the second to the third line. Taking a time derivative and using Eq. (6), we obtain the mean-field equation

$$\tau \frac{d}{dt} R(t) + R(t) = \bar{g} \langle T(h) \rangle_{h \sim \mathcal{N}(R(t), Q(t,t))}. \qquad \text{(A12)}$$

Note that here we need only the input variance $Q(t,t)$, which is trivially given by the mean activity and, potentially, zero-time-lag cross-correlations of the outputs. However, the mean-field equation does not depend on the autocorrelation at nonzero time lag, which is derived in the next section. Therefore, at least as far as cross-correlations are negligible, Eq. (A12) is closed.

3. Derivation of the ODE for autocorrelations in binary networks

Here we derive the form (9) for the evolution of the autocorrelation.

The correlation functions in the Ising and bitlike representation are, according to Eq. (A10), related as

$$q_I(t, s) := \langle x(t) x(s) \rangle$$
$$= \langle [2n(t) - 1][2n(s) - 1] \rangle$$
$$= 4\langle n(t)n(s) \rangle - 2\langle n(t) \rangle - 2\langle n(s) \rangle + 1, \qquad \text{(A13)}$$

where by $Q(t, s) = g^2 q_I(t, s)$ we obtain the quantity considered in the main text in Eq. (9). Defining $q(t, s) := \langle n(t)n(s) \rangle$, we have

$$q(t, s) = \int dh \rho(h) \sum_{n(t)=0}^{1} \sum_{n(s)=0}^{1} n(t)n(s)\rho[n(t), n(s)|h]$$
$$= \langle \rho[n(t) = 1, n(s) = 1|h] \rangle_{h \sim \rho}, \qquad \text{(A14)}$$

where we write $\langle \ldots \rangle_{h \sim \rho}$ as a short form of $\int dh \rho(h) \ldots$. The latter joint probability is decomposed, analogous to Eq. (1), as

$$\rho[n(t) = 1, n(s) = 1|h]$$
$$= \rho[x(t) = 1|x(s) = 1, h]\rho[x(s) = 1|h]. \qquad \text{(A15)}$$

We obtain the first conditional probability on the right by considering the possibilities to reach the final state $x(t) = 1$ given that $x(s) = 1$,

$$\rho[x(t) = 1|x(s) = 1, h]$$
$$= P(\text{no updates in } [s, t])$$
$$\quad + P(\text{last update in } [s, t] \text{ to up state})$$
$$= e^{-\frac{t-s}{\tau}} + \int_{s}^{t} e^{-\frac{t-t'}{\tau}} T_p(h(t')) \frac{dt'}{\tau}. \qquad \text{(A16)}$$

Likewise, we obtain the latter conditional probability on the right of Eq. (A15) as

$$\rho[x(s) = 1|h] = \int_{-\infty}^{s} e^{-(s-t')/\tau} T_p(h(t')) \frac{dt'}{\tau}. \qquad \text{(A17)}$$

Combining Eqs. (A14), (A16), and (A17), we get

$$q(t,s) = \left\langle \int_{-\infty}^{s} e^{-(s-t')/\tau} T_p(h(t')) \frac{dt'}{\tau} \left(e^{-(s-t')/\tau} + \int_{s}^{t} e^{-(t-t'')/\tau} T_p(h(t'')) \frac{dt''}{\tau} \right) \right\rangle_{h\sim\rho}$$

$$= \int_{-\infty}^{s} e^{-(t-t')/\tau} \langle T_p(h(t')) \rangle_{h\sim\rho} \frac{dt'}{\tau} + \int_{-\infty}^{s} e^{-(s-t')/\tau} \int_{s}^{t} e^{-(t-t'')/\tau} \langle T_p(h(t')) T_p(h(t'')) \rangle_{h\sim\rho} \frac{dt''}{\tau} \frac{dt'}{\tau}.$$

In the stationary state, the first integral in the last line reduces to $\langle T_p(h) \rangle_{h\sim\rho} e^{-(t-s)/\tau}$. Also, $q(t,s) =: q(t-s)$ is then a function of the time lag $\Delta t := t - s$ alone

$$q(\Delta t) = \langle T_p(h) \rangle_{h\sim\rho} e^{-\Delta t/\tau} + \int_{-\infty}^{s} e^{-(s-t')/\tau} \int_{s}^{s+\Delta t} e^{-(s+\Delta t - t'')/\tau} \langle T_p(h(t')) T_p(h(t'')) \rangle_{h\sim\rho} \frac{dt''}{\tau} \frac{dt'}{\tau}.$$

Differentiating by Δt, we get

$$\tau \frac{d}{d\Delta t} q(\Delta t) = -q(\Delta t) + \int_{-\infty}^{s} e^{-(s-t')/\tau} \langle T_p(h(t')) T_p(h(s + \Delta t)) \rangle_{h\sim\rho} \frac{dt'}{\tau}$$

$$= -q(\Delta t) + \int_{0}^{\infty} e^{-t/\tau} \langle T_p(h(0)) T_p(h(t + \Delta t)) \rangle_{h\sim\rho} \frac{dt}{\tau}, \tag{A18}$$

where we substitute $s - t' \to t$ in the last step and use the stationarity to shift the time arguments of the h by $t - s$. Using Eqs. (A10), (A13), and (6) and assuming stationarity, we get the result (8) in the main text, where the gain function T instead of T_p appears. Shifting the integration variable t by Δt, we obtain

$$\tau \frac{d}{d\Delta t} q(\Delta t) = -q(\Delta t) + \int_{\Delta t}^{\infty} e^{-(t-\Delta t)/\tau} \langle T_p(h(0)) T_p(h(t)) \rangle_{h\sim\rho} \frac{dt}{\tau},$$

and performing another derivative $\tau d/d\Delta t$ yields

$$\tau^2 \frac{d^2}{d\Delta t^2} q(\Delta t) = -\tau \frac{d}{d\Delta t} q(\Delta t) + \int_{\Delta t}^{\infty} e^{-(t-\Delta t)/\tau} \langle T_p(h(0)) T_p(h(t)) \rangle_{h\sim\rho} \frac{dt}{\tau} - \langle T_p(h(0)) T_p(h(\Delta t)) \rangle_{h\sim\rho}$$

$$= q(\Delta t) - \langle T_p(h(0)) T_p(h(\Delta t)) \rangle_{h\sim\rho}, \tag{A19}$$

where for the second equality, we use the first-order differential equation (A18). Closing the equation in the mean-field approximation amounts to setting the measure of $h \sim \rho \equiv \mathcal{N}(R, Q)$ to the Gaussian process with mean R and variance Q, as determined by the saddle-point equations (2) and (3). This approximation neglects fluctuations of $\mathcal{R}$ and $\mathcal{Q}$, which is justified if the system is not close to the critical point and the average connectivity scales at most like $1/N$ [28], but even if the latter condition is relaxed to a $1\sqrt{N}$ scaling, this merely leads to an additional term in the input fluctuations taking into account pairwise correlations [92].

Moving to the $[-1, 1]$ representation by using Eqs. (A10), (A13) and (6) and multiplying Eq. (A19) by g^2 changes $T_p \to T$ and $g^2 q \to Q$ so that we obtain Eq. (9). Here, in addition, we introduce $\mathcal{N}_{R,Q(0),Q}$ as the bivariate Gaussian with stationary mean R and covariance matrix $\begin{pmatrix} Q(0) & Q \\ Q & Q(0) \end{pmatrix}$.

This Gaussian expectation value allows us to employ Price's theorem [30], which states that

$$\frac{d}{dQ} \langle \mathcal{T}(h) \mathcal{T}(h') \rangle_{(h,h')\sim\mathcal{N}_{R,Q(0),Q}}$$

$$\overset{\text{(Price's theorem)}}{\equiv} \langle T(h) T(h') \rangle_{(h,h')\sim\mathcal{N}_{R,Q(0),Q}},$$

where $\mathcal{T}(x) := \int^x T(x') dx'$ is the primitive of T.

4. Replica calculation for chaos

a. Model-independent replica calculation

To assess the transition to chaos, we perform a replica calculation that considers a pair of networks with identical connectivity but slightly different initial conditions for the neurons. We use superscripts (1) and (2) to distinguish the two systems. The correlation between the two replicas is a

measure of the distance between their respective states in terms of the squared Euclidean distance

$$d^{(12)}(t) \coloneqq ||\boldsymbol{x}^{(1)}(t) - \boldsymbol{x}^{(2)}(t)||^2$$

$$= \sum_{\alpha=1}^{2} \sum_{i=1}^{N} x_i^{(\alpha)} x_i^{(\alpha)} - 2 \sum_{i=1}^{N} x_i^{(1)} x_i^{(2)}. \quad \text{(A20)}$$

The first term, on expectation over realizations of the activity, approaches the average autocorrelation in the two replicas and the latter term the inter-replica correlation. For Ising spins, the expression simplifies to $2N - 2 \sum_{i=1}^{N} x_i^{(1)} x_i^{(2)}$.

The formal derivation of mean-field equations that approximate these quantities proceeds analogous to the Appendix 1: The analog expression to Eqs. (A3) and (A4) reads

$$\rho[\boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}] = \int \mathcal{D}\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\} \rho[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} | \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}]$$

$$\times \prod_{\alpha=1}^{2} \delta[\boldsymbol{h}^{(\alpha)} - \boldsymbol{J}\boldsymbol{x}^{(\alpha)}].$$

Here the conditional density $\rho[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} | \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}]$ is a joint distribution across the two replicas, because it must allow the representation of update processes or stochastic activations of corresponding neurons that have identical realizations between the two replicas.

The important point is the identical matrix $\boldsymbol{J}$ appearing in the product of the latter two Dirac distributions, which, after introducing Fourier representations as in Eq. (A4) and taking the disorder average over $\boldsymbol{J}$ analogous to Eq. (A5), yields

$$\langle \exp(-\hat{\boldsymbol{h}}^{(1)T} \boldsymbol{J} \boldsymbol{x}^{(1)} - \hat{\boldsymbol{h}}^{(2)T} \boldsymbol{J} \boldsymbol{x}^{(2)}) \rangle_{\boldsymbol{J} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\frac{\bar{g}}{N}, \frac{g^2}{N})}$$

$$= \prod_{i=1}^{N} \prod_{\alpha=1}^{2} \exp\left(-\frac{\bar{g}}{N} \hat{h}_i^{(\alpha)T} \sum_{j=1}^{N} x_j^{(\alpha)} + \frac{g^2}{2N} \sum_{j=1}^{N} (\hat{h}_i^{(\alpha)T} x_j^{(\alpha)})^2\right)$$

$$\times \exp\left(\frac{g^2}{N} \sum_{j=1}^{N} \hat{h}_i^{(1)T} x_j^{(1)} \hat{h}_i^{(2)T} x_j^{(2)}\right).$$

The penultimate line is the same contribution for each replica as in the single system; it is treated in the same manner by introducing pairs of auxiliary fields $\{\mathcal{R}^{(\alpha)}, \hat{\mathcal{R}}^{(\alpha)}, \mathcal{Q}^{(\alpha\alpha)}, \hat{\mathcal{Q}}^{(\alpha\alpha)}\}_{\alpha \in \{1,2\}}$. The last line couples the two replicas and can be decoupled similarly by defining

$$\mathcal{Q}^{(12)}(t, s) \coloneqq \frac{g^2}{N} \sum_{j=1}^{N} x_j^{(1)}(t) x_j^{(2)}(s).$$

This definition is enforced by inserting a δ constraint represented as a Fourier integral with the corresponding

conjugate field $\hat{\mathcal{Q}}^{(12)}(s, t)$. The integral over $\{\mathcal{R}^{(\alpha)}, \hat{\mathcal{R}}^{(\alpha)}, \mathcal{Q}^{(\alpha\beta)}, \hat{\mathcal{Q}}^{(\alpha\beta)}\}_{\alpha,\beta \in \{1,2\}}$ is then taken in saddle-point approximation with the resulting nontrivial saddle-point equations

$$R^{(\alpha)}(t) = \bar{g} \langle x^{(\alpha)}(t) \rangle_{\Omega(\{R^{(\alpha)}, Q^{(\alpha\beta)}\})},$$

$$Q^{(\alpha\beta)}(t, s) = g^2 \langle x^{(\alpha)}(t) x^{(\beta)}(s) \rangle_{\Omega(\{R^{(\alpha)}, Q^{(\alpha\beta)}\})}. \quad \text{(A21)}$$

The remaining response fields vanish $\hat{R}^{(\alpha)} = \hat{Q}^{(\alpha\beta)} \equiv 0$. The expectation value in Eq. (A21) is taken with the measure

$$\langle \ldots \rangle_{\Omega(\{R^{(\alpha)}, Q^{(\alpha\beta)}\})} \quad \text{(A22)}$$

$$= \int \mathcal{D}\{x^{(1)}, x^{(2)}\} \ldots \langle \rho[x^{(1)}, x^{(2)} | h^{(1)}, h^{(2)}] \rangle_{(h^{(1)}, h^{(2)})}, \quad \text{(A23)}$$

where $(h^{(1)}, h^{(2)}) \sim \mathcal{N}(\{R^{(\alpha)}, Q^{(\alpha\beta)}\})$ is a pair of Gaussian processes with cumulants

$$\langle\!\langle h^{(\alpha)}(t) \rangle\!\rangle = R^{(\alpha)}(t),$$

$$\langle\!\langle h^{(\alpha)}(t) h^{(\beta)}(s) \rangle\!\rangle = Q^{(\alpha\beta)}(t, s).$$

The distance (A20) between the replica in mean-field approximation can then be written as

$$d^{(12)}(t) = N g^{-2} \left(\sum_{\alpha=1}^{2} Q^{(\alpha\alpha)}(t, t) - 2 Q^{(12)}(t, t) \right).$$

b. Application to binary networks

The zero-lag cross-replica correlation is then given with Eqs. (A21) and (A22) as

$$Q^{(12)}(t, t) \overset{(A21, A22)}{=} g^2 \sum_{x^{(1)}(t), x^{(2)}(t) = -1}^{1} x^{(1)} x^{(2)}$$

$$\times \langle \rho(x^{(1)}, x^{(2)}, t | h^{(1)}, h^{(2)}) \rangle_{(h^{(1)}, h^{(2)}) \sim \mathcal{N}(\{R^{(\alpha)}, Q^{(\alpha\beta)}\})}.$$

To construct $\rho(x^{(1)}, x^{(2)}, t | h^{(1)}, h^{(2)})$, first note that both neurons are updated by the same stochastic realizations of the update process. This process has two random components: The drawing of the update time point t', which, for the Poisson updates, has a distribution of $e^{-(t-t')/\tau}(dt'/\tau)$ for the last event to have appeared in $[t', t' + dt]$, and the stochastic activation depending on the gain function $T_p \in [0, 1]$, whose value for both replicas is compared to the same realization of a uniformly distributed random number $r \in [0, 1]$.

The four possible outcomes of this update of states $(x^{(1)}, x^{(2)})$ are $(-1, -1)$, $(1, 1)$, both of which lead to $x^{(1)} x^{(2)} = +1$ and $(-1, 1)$, $(1, -1)$, both of which lead to $x^{(1)} x^{(2)} = -1$. One thus needs only to distinguish two

outcomes: The event $x^{(1)}x^{(2)} = -1$ takes place if the random variable r is in between the values of the two gain functions, $T_p(h^{(1)}) < r < T_p(h^{(2)})$, which happens with probability $p_{\text{diff}} = |T_p(h^{(1)}) - T_p(h^{(2)})|$ and the other event $x^{(1)} \cdot x^{(2)} = +1$ with $1 - p_{\text{diff}}$. So in total, we get at the time t' of update

$$\langle x^{(1)}(t')x^{(2)}(t')\rangle_r = (-1) \times p_{\text{diff}} + (+1) \times (1 - p_{\text{diff}}) \tag{A24}$$

$$= 1 - 2p_{\text{diff}}$$
$$= 1 - 2|T_p(h^{(1)}(t')) - T_p(h^{(2)}(t'))|. \tag{A25}$$

Taken together with the asynchronous update time point, we thus have

$$Q^{(12)}(t,t) = g^2 \int_{-\infty}^{t} \frac{dt'}{\tau} e^{-(t-t')/\tau}(1 - 2\langle |T_p(h^{(1)}(t')) - T_p(h^{(2)}(t'))|\rangle_{(h^{(1)},h^{(2)})\sim\mathcal{N}(\{R^{(\alpha)},Q^{(\alpha\beta)}\})}).$$

Taking a derivative with respect to t, we obtain an ODE governing the time evolution of the cross-replica correlation

$$\tau\frac{d}{dt}Q^{(12)}(t,t) = -Q^{(12)}(t,t) + g^2(1 - 2\langle |T_p(h^{(1)}(t)) - T_p(h^{(2)}(t))|\rangle_{(h^{(1)},h^{(2)})\sim\mathcal{N}(\{R^{(\alpha)},Q^{(\alpha\beta)}\})}). \tag{A26}$$

Note that $Q^{(12)}(t)$ also appears implicitly in the distribution of $h^{(1)}, h^{(2)}$, rendering the equation nonlinear. This implies the result (11) in the main text, which follows by replacing $2[T_p(h^{(1)}) - T_p(h^{(2)})] = T(h^{(1)}) - T(h^{(2)})$ due to Eq. (6).

So far, we have proceeded without approximation apart from the saddle-point approximation. Perfect correlation of the replicas $Q^{(12)}(t,t) = Q_0 = g^2$ is clearly a fixed point, since then $h^{(1)}(t) = h^{(2)}(t)$ and the right-hand side

vanishes. We now wish to assess the stability of this solution, that is, whether a perturbation of one replica results in the recovery of perfect correlation (regular dynamics) or in a decorrelation of the replicas (chaos). Making the ansatz $Q^{(12)}(t,t) = Q_0 - \epsilon(t)$ and using the mean-field approximation of the input distribution, the last term of the ODE (A26) becomes, by substituting $H := (h^{(1)} + h^{(2)})/2, h := (h^{(1)} - h^{(2)})/2$ and then expanding in h and $\epsilon/(2Q_0)$:

$$(|T_p(h^{(1)}) - T_p(h^{(2)})|)_{\binom{h^{(1)}}{h^{(2)}}\sim\mathcal{N}\left(\binom{R}{R},\binom{Q_0 \quad Q_0-\epsilon}{Q_0-\epsilon \quad Q_0}\right)}$$

$$= (|T_p(H+h) - T_p(H-h)|)_{\binom{H}{h}\sim\mathcal{N}\left(\binom{R}{0},\binom{Q_0-\frac{\epsilon}{2} \quad 0}{0 \quad \frac{\epsilon}{2}}\right)}$$

$$= \frac{1}{2\pi\sqrt{\frac{\epsilon}{2}(Q_0-\frac{\epsilon}{2})}} \iint \exp\left(-\frac{H^2}{2(Q_0-\frac{\epsilon}{2})} - \frac{h^2}{2\frac{\epsilon}{2}}\right)|T_p(R+H+h) - T_p(R+H-h)|dHdh$$

$$= \frac{1}{2\pi\sqrt{\frac{\epsilon}{2}(Q_0-\frac{\epsilon}{2})}} \iint e^{-\frac{H^2}{2(Q_0-\frac{\epsilon}{2})}} e^{-\frac{h^2}{\epsilon}}|2T'_p(R+H)h + \mathcal{O}(h^3)|dHdh$$

$$= 2\frac{1}{\sqrt{2\pi Q_0}} \int e^{-\frac{H^2}{2Q_0}}|T'_p(R+H)|dH \frac{1}{\sqrt{2\pi\frac{\epsilon}{2}}} \int |h|e^{-\frac{h^2}{\epsilon}}dh[1 + \mathcal{O}(\epsilon)]$$

$$= 2\langle T'_p(H)\rangle_{H\sim\mathcal{N}(R,Q_0)} \sqrt{\frac{\epsilon}{\pi}} + \mathcal{O}\left(\epsilon^{\frac{3}{2}}\right). \tag{A27}$$

Plugging this result and the ansatz $Q^{(12)}(t) = Q_0 - \epsilon(t)$ back into Eq. (A26) then yields Eq. (12).

5. Growth of perturbations in binary networks

Starting from Eq, (12)

$$\tau \frac{d}{dt} \epsilon(t) = -\epsilon(t) + c\sqrt{\epsilon(t)},$$

where $c = (2/\sqrt{\pi})g^2 \langle T'(h) \rangle_{h \sim \mathcal{N}(R,g^2)}$, we integrate the differential equation

$$\tau^{-1} \int_0^t dt = -\int_{\epsilon(0)}^{\epsilon(t)} \frac{d\epsilon}{\epsilon - c\sqrt{\epsilon}}$$
$$= -2\ln(c - \sqrt{\epsilon})|_{\epsilon(0)}^{\epsilon(t)},$$

so the solution is

$$\epsilon(t) = [c - (c - \sqrt{\epsilon(0)})e^{-(t/2\tau)}]^2.$$

Expressed in terms of the dimensionality $d(t) = (N/g^2)\epsilon(t)$ and $d_* = (N/g^2)c^2$ (18),

$$d(t) = [\sqrt{d_*} - (\sqrt{d_*} - \sqrt{d(0)})e^{-(t/2\tau)}]^2.$$

In the long-time limit, the solution reaches the fixed point

$$d_* = d(t \to \infty).$$

The fastest increase happens for a Heaviside gain function $T = -1 + 2\theta \in \{-1, 1\}$ for which $\langle T'(h) \rangle_{h \sim \mathcal{N}(0,g^2)} = 2/(\sqrt{2\pi}g)$ so $d^* = Ng^2(4/\pi)(4/2\pi g^2) = N(8/\pi^2)$. This shows that the binary network decorrelates only to a dimensionality of $d^*/N = 8/\pi^2 \simeq 0.81$. The distance as a function of the time is then

$$\frac{d(t)}{N} = \frac{8}{\pi^2} \left[1 - \left(1 - \frac{\pi}{\sqrt{8}} \sqrt{\frac{d(0)}{N}} \right) e^{-(t/2\tau)} \right]^2.$$

The maximal signal-to-noise ratio is obtained by using d_s with a nonzero initial $d_s(0)$ and a noise distance d_n^0 that initially vanishes $d_n^0(0) = 0$, leading to

$$\frac{d_s(t) - d_n^0(t)}{d_*} = [1 - (1 - \delta)e^{-(t/2\tau)}]^2 - (1 - e^{-(t/2\tau)})^2,$$

where we define $\delta := \sqrt{d_s(0)/d_*}$. The maximum of this function is at

$$0 = -(1 - \delta)[1 - (1 - \delta)e^{-(\hat{t}/2\tau)}] + (1 - e^{-(\hat{t}/2\tau)})$$

$$\Rightarrow e^{-\frac{\hat{t}}{2\tau}} = \frac{\delta}{1 - (1 - \delta)^2}. \tag{A28}$$

For small $\delta \ll 1$, this expression yields

$$e^{-(\hat{t}/2\tau)} \simeq \frac{\delta}{1 - (1 - 2\delta)} = \frac{1}{2},$$

so the time of the maximum becomes approximately independent of the initial value δ and thus independent of the signal-to-noise ratio

$$\hat{t} \simeq 2\tau \ln 2 \simeq 1.39\tau. \tag{A29}$$

The maximum is with $e^{-(\hat{t}/2\tau)} \simeq \frac{1}{2}$,

$$\frac{d_s(\hat{t}) - d_n^0(\hat{t})}{d_*} \simeq \left[1 - (1 - \delta)\frac{1}{2} \right]^2 - \left(1 - \frac{1}{2} \right)^2$$
$$= \frac{\delta}{2} + \frac{\delta^2}{4}.$$

The latter expression shows that this maximum relative to the initial signal is

$$\frac{d_s(\hat{t}) - d_n^0(\hat{t})}{d_s(0) - d_n^0(0)} = \frac{[d_s(\hat{t}) - d_n^0(\hat{t})]/d_*}{d_s(0)/d_*}$$
$$\simeq \frac{\frac{\delta}{2} + \frac{\delta^2}{4}}{\delta^2} = \frac{1}{2\delta} + \frac{1}{4}$$
$$= \frac{1}{2}\sqrt{\frac{d_*}{d_s(0)}} + \frac{1}{4}. \tag{A30}$$

For the Heaviside nonlinearity, the result becomes with $d^*/N = 8/\pi^2$,

$$\frac{d_s(\hat{t}) - d_n^0(\hat{t})}{d_s(0) - d_n^0(0)} = \frac{1}{\pi}\sqrt{2\frac{N}{d_s(0)}} + \frac{1}{4}.$$

6. Growth of perturbations in rate networks

On small timescales, the distance evolves in proportion to the Lyapunov exponent

$$d(t) = d(0)\exp(\lambda_{\max}t).$$

So, $d_s(t) - d_n(t) = [d_s(0) - d_n(0)]\exp(\lambda_{\max}t)$. In particular, there is no maximum expected on a timescale of the neuronal dynamics. The typical timescale instead is determined by the maximal Lyapunov exponent.

To derive an equation for the time evolution of the correlation between replicas in the rate network, we use the pair of equations obtained from the mean-field description

$$\tau\partial_t x^\alpha(t) = -x^\alpha(t) + h^\alpha(t)$$
$$\langle h^\alpha(t)h^\beta(s) \rangle = Q^{\alpha\beta}(t,s) \equiv g^2 \langle T(x^\alpha(t))T(x^\beta(s)) \rangle. \tag{A31}$$

We may thus write

$$x^\alpha(t) = \frac{1}{\tau} \int_{-\infty}^{t} e^{-\frac{t-t'}{\tau}} h^\alpha(t') dt'.$$

So the correlation function $c^{\alpha\beta}(t, s) := \langle x^\alpha(t) x^\beta(s) \rangle$ obeys

$$c^{\alpha\beta}(t, s) = \frac{1}{\tau} \int_{-\infty}^{t} e^{-\frac{t-t'}{\tau}} \frac{1}{\tau} \int_{-\infty}^{s} e^{-\frac{s-s'}{\tau}} \langle h^\alpha(t') h^\beta(s') \rangle dt' ds',$$

which becomes in differential form

$$\tau \partial_t c^{\alpha\beta}(t, s) = -c^{\alpha\beta}(t, s) + \frac{1}{\tau} \int_{-\infty}^{s} e^{-\frac{s-s'}{\tau}} \langle h^\alpha(t) h^\beta(s') \rangle ds'.$$

This differential equation allows the integration along the t direction by one time step δ,

$$c^{\alpha\beta}(t + \delta, s) = \left(1 - \frac{\delta}{\tau}\right) c^{\alpha\beta}(t, s)$$
$$+ \frac{\delta}{\tau} \int_{-\infty}^{s} e^{-\frac{s-s'}{\tau}} \langle h^\alpha(t) h^\beta(s') \rangle ds',$$

which requires only $c^{\alpha\beta}$ and $\langle hh \rangle$ in the t past and in the s past. The integration can be done for $t \in [0, T]$, where T is a desired final point.

As a result, one has $c^{\alpha\beta}(t, s)$ for $t \in [0, T]$. In the next update step, we move into the s direction by

$$c^{\alpha\beta}(t, s + \delta) = \left(1 - \frac{\delta}{\tau}\right) c^{\alpha\beta}(t, s)$$
$$+ \frac{\delta}{\tau} \int_{-\infty}^{t} e^{-(t-t')/\tau} \langle h^\alpha(t') h^\beta(s) \rangle dt',$$

where the latter integral can be computed because it requires only

$$\langle h^\alpha(t) h^\beta(s) \rangle = g^2 \langle T(x_1) T(x_2) \rangle_{(x_1, x_2) \sim \mathcal{N}(0, c^{\alpha\beta}(t, s))} \qquad \text{(A32)}$$

computed in the previous step.

It makes sense to introduce as an auxiliary variable

$$q^{\alpha\beta}(t, s) := \frac{1}{\tau} \int_{-\infty}^{t} e^{-\frac{t-t'}{\tau}} \langle h^\alpha(t') h^\beta(s) \rangle dt',$$

for which we can assume the symmetry $q^{\alpha\beta}(t, s) = q^{\beta\alpha}(t, s)$ to write the updates

$$c^{\alpha\beta}(t + \delta, s) = \left(1 - \frac{\delta}{\tau}\right) c^{\alpha\beta}(t, s) + \frac{\delta}{\tau} q^{\alpha\beta}(s, t), \qquad \text{(A33)}$$

$$c^{\alpha\beta}(t, s + \delta) = \left(1 - \frac{\delta}{\tau}\right) c^{\alpha\beta}(t, s) + \frac{\delta}{\tau} q^{\alpha\beta}(t, s). \qquad \text{(A34)}$$

The auxiliary variable $q^{\alpha\beta}(t, s)$ obeys the differential equation $(\tau \partial_t + 1) q^{\alpha\beta}(t, s) = \langle h^\alpha(t) h^\beta(s) \rangle$, which yields the update equation

$$q^{\alpha\beta}(t + \delta, s) = \left(1 - \frac{\delta}{\tau}\right) q^{\alpha\beta}(t, s) + \frac{\delta}{\tau} \langle h^\alpha(t) h^\beta(s) \rangle. \qquad \text{(A35)}$$

So the required sequence of updates is
  (i) Start at $t = 0$.
 (ii) Assume we compute $c^{\alpha\beta}(t', s')$ and $q^{\alpha\beta}(t', s')$ until this point t for all $(t' < t, s' < t)$.
(iii) Compute $c^{\alpha\beta}(t + \delta, \forall s' \le t)$ using Eq. (A33).
 (iv) Compute $c^{\alpha\beta}(\forall t' \le t, t + \delta)$ using Eq. (A34).
  (v) Compute $\langle h^\alpha(t + \delta) h^\beta(s') \rangle \forall s' \le t$ and $\langle h^\alpha(t') h^\beta(t + \delta) \rangle \forall t' \le t$ using Eq. (A32) and the result from the previous step.
 (vi) Compute $q^{\alpha\beta}(t' \le t + \delta, t + \delta)$ by iterating Eq. (A35) with zero initial condition and starting with t' sufficiently far back in the past.
(vii) Compute $c^{\alpha\beta}(t + \delta, t + \delta)$ by average of Eqs. (A33) and (A34) using new value $q^{\alpha\beta}(t + \delta, t)$ and $q^{\alpha\beta}(t, t + \delta)$, respectively.
(viii) Go to the next time slice $t \to t + \delta$, return to step iii.

7. Flux tubes in binary networks

It has been shown by Puelma Touzel and Wolf [34] that the borders of flux tubes in spiking networks of inhibitory LIF neurons are related to changes in the global order of spikes. In particular, if a perturbation creates an additional spike or causes the omission of an expected one, the mean firing rate will stay constant, but the order of future spikes is very likely to be irrevocably changed. The divergence rate of two trajectories can be assessed by calculating the mean number of unexpected spike order changes caused by a single such perturbation, resulting in a branching process. In the context of binary networks, we can ask the equivalent question: Given a flip of a single neuron's activity variable, how many wrong update results will occur on average in the following time τ?

The flip of one neuron $x_j \to -x_j$ causes a change $\Delta h_i = -2 J_{ij} x_j$ in the input of neurons it is connected to. Across different target neurons, Δh is therefore distributed as

$$\rho(\Delta h) = \mathcal{N}\left(\pm 2 \frac{\bar{g}}{N}, 4 \frac{g^2}{N}\right),$$

and the probability of a neuron to be updated into the wrong state due to the perturbation in the input is

$$p(\text{flip}(x) | \Delta h) = \langle |T_p(h + \Delta h) - T_p(h)| \rangle_{h \sim \mathcal{N}(R, Q_0)}$$
$$\overset{|\Delta h| \ll 1, T_p' \ge 0}{\approx} |\Delta h| \langle T_p'(h) \rangle_{h \sim \mathcal{N}(R, Q_0)},$$

where the absolute value enters because both directions of perturbation cause a positive probability of wrong updating, and we assume $T' \ge 0$ for simplicity. Now we ask the following question: How many downstream flips n_{spawns} will, on average, be triggered in the network during one

time constant, given a single original flip? This quantity controls whether the decorrelating flips will proliferate or not, because since every neuron is updated on average once per time constant, if $n_{\text{spawns}} < 1$ and the neuron carrying the original flip is updated again, it is most likely updated "correctly" again, and the average number of flips in the network has decreased. If $n_{\text{spawns}} > 1$ on the other hand, the average number of flips increases.

Being interested in the transition point, we can assume $n_{\text{spawns}} \approx 1$ so that we do not need to take the interaction of several flips into account. Then,

$$\bar{n}_{\text{spawns}} = N \langle p(\text{flip}(x)|\Delta h) \rangle_{\Delta h}, \qquad (A36)$$

and while we take the mean input R into account, we neglect the perturbation of the mean input $\langle \Delta h \rangle = \pm 2\bar{g}/N = \mathcal{O}(N^{-1}) \approx 0$ as it is small compared to the standard deviation $\sigma_{\Delta h} = 2g/\sqrt{N} = \mathcal{O}(N^{-\frac{1}{2}})$, allowing the simple calculation

$$\bar{n}_{\text{spawns}} = N \langle T'_p(h) \rangle_{h \sim \mathcal{N}(R,Q_0)} \langle |\Delta h| \rangle_{\Delta h \sim \mathcal{N}(0,\sigma^2_{\Delta h})}$$

$$= N \langle T'_p(h) \rangle_{h \sim \mathcal{N}(R,Q_0)} \sqrt{\frac{2}{\pi}} \sigma_{\Delta h}$$

$$= 2\sqrt{\frac{2N}{\pi}} g \langle T'_p(h) \rangle_{h \sim \mathcal{N}(R,Q_0)}.$$

Finally, accounting for $T'_p(h) = T'(h)/2$ given by Eq. (6), the chaos transition is expected at

$$1 \overset{!}{=} \bar{n}_{\text{spawns}} = \frac{\sqrt{2N}}{\sqrt{\pi}} g \langle T'(h) \rangle_{h \sim \mathcal{N}(R,Q_0)}, \qquad (A37)$$

which is exactly the result (15) derived via the completely different route of the replica calculation. While the derivation here is nicely and intuitively interpretable, the derivation via field theory and replica calculation allows for systematic generalizations. For example, it is not clear how to obtain the residual correlation (13) in the *ad hoc* approach.

a. Flux-tube size

The flux-tube diameter is not a very informative measure for a binary network, since the system trajectory in phase space is typically not in the middle of a "tube" but close to some of its boundaries (given by the thresholds). Therefore, the distance to a boundary strongly depends on the direction of perturbation. As a relatively informative measure, we consider smearing the trajectory in all directions with some variance $\text{Var}(\Delta h) = \sigma^2_{\text{fl}}$, which is chosen such that on average, one flux-tube boundary is crossed. This procedure makes sense insofar as it is similar to adding noise onto the input. It is important to be aware that σ_{fl} is

not strictly the average distance to the closest boundary, although the two quantities should covary.

The situation is analogous to the above calculation, because we again need to consider the flips occurring during an update in (on average) all N neurons, which is given by Eq. (A36) only with $\sigma_{\Delta h}$ replaced by σ_{fl}. Demanding $\bar{n}_{\text{spawns}} \overset{!}{=} 1$ then yields

$$1 \overset{!}{=} N \langle p(\text{flip}(x)|\Delta h) \rangle_{\Delta h \sim \mathcal{N}(0,\sigma^2_{\text{fl}})}$$

$$\Rightarrow \sigma_{\text{fl}} = \frac{\sqrt{2\pi}}{N \langle T'(h) \rangle_{h \sim \mathcal{N}(R,Q_0)}}.$$

Of course, the $1/N$ scaling needs to be taken with caution, since our perturbation goes into all N phase-space directions, resulting in a total length scaling as $1/\sqrt{N}$.

8. Equivalence of dynamical mean-field theories of binary and rate networks

The dynamics (21) can equivalently be written as

$$\tau \partial_t x = -x + T(h),$$
$$h = Jx + \sqrt{\tau}\eta, \qquad (A38)$$

where the noise η_i is an Ornstein-Uhlenbeck process [93], $(\tau\partial_t + 1)\eta_i = \xi_i$. This form allows the application of the model-independent field theory. The single-neuron, single-time-slice probability functional is $\rho[x(t)|h] = \delta[x(t) - \int_{-\infty}^{t} e^{-(t-t')/\tau} T(h_i(t'))(dt'/\tau)]$, and the noise term is taken into account in $\rho[h] = \delta[h - Jx - \sqrt{\tau}\eta]$. Plugging this expression into Eq. (2), we obtain the same Eq. (5) for the mean activity as for the binary neuron if we choose the strength of the noise σ_ξ such that $Q(t,t) = g^2$ as well. The reason for the equivalence is that the exponential function appearing in the convolution equation is the Green's function of $\tau\partial_t + 1$. In a stationary state, the saddle-point solution for $Q(\Delta t) := Q(t, t + \Delta t) = g^2 \langle x(t)x(t + \tau) \rangle$, moreover, follows the same Newtonian equation of motion (9) as for the binary model [see Ref. [21], Eq. (7)].

a. Matching initial conditions

Knowing that the differential equations for the time-lagged autocorrelations are the same, we have to adjust their respective initial conditions to establish full equivalence. Here we use the subscripts b and r to refer to the quantities of the binary and rate model, respectively. Two initial conditions are needed for a unique solution. One is to require that $\lim_{t \to \infty} \dot{Q}(t) = 0$, which is the same in both cases. So, the autocorrelation for infinite time lags is described by a single value Q_∞, which vanishes for point-symmetric activation functions, but is in general nonzero and self-consistently determined by the static variability across neurons caused by the disorder (compare
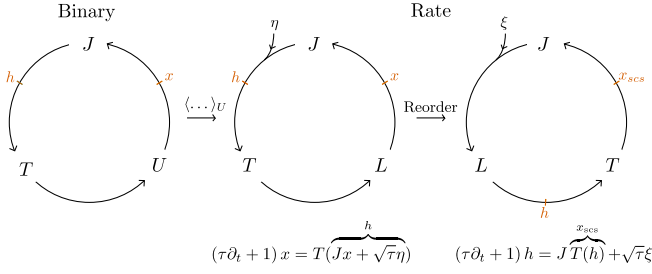
FIG. 14. Equivalence of binary (left) and rate models in the two different forms (A38) (middle) and (21) (right): mapping from output to input by identical matrix J; asynchronous update process U with rate τ^{-1} implies exponential convolution kernel leading to leaky integration [cf. Eq. (5)] identical to operator $L = (\tau\partial_t + 1)^{-1}$ present explicitly in Eqs. (A38) and (21). Transitions between discrete binary states effect red noise η in input h (middle), which corresponds to white noise ξ that is low-pass filtered by L (right). Rate models differ in the order of application of this kernel and the connectivity, which yields equivalent dynamics because the two operators commute.

Fig. 5). In the binary case, the second condition is $Q_b(0) = g^2$ because the zero-lag autocorrelation of a single spin is always 1. In a rate network, however, the input noise strength determines how quickly the autocorrelation decays, resulting in the condition on the derivative $\dot{Q}_r(0+) = -\sigma_\xi^2/2$ [23,94].

The idea is to choose the variance of the noise σ_ξ^2 in the rate network such that $Q_r(0) = Q_b(0)$, so that the time-lagged solutions for the variance $Q(\Delta t)$ match.

To do so, using $\dot{Q}(\infty) = 0$ and conservation of total "energy" $V_{Q_0} + \tau^2 \dot{Q}^2/2$ implied by the Newtonian form of Eq. (10), the condition $Q_0 := Q_r(0) \overset{!}{=} Q_b(0) = g^2$ can be expressed as a condition for the derivative and thus the noise amplitude

$$\frac{1}{2}\tau^2 \dot{Q}_0^2 + V_{Q_0}(Q_0) = V_{Q_0}(Q_\infty)\big|_{Q_0=g^2}. \quad \text{(A39)}$$

Plugging in $\dot{Q}_0 = -\sigma_\xi^2/2$ and solving for σ_ξ^2 yields the condition (22) in the main text. This proves that the binary and rate model with appropriate noise have equivalent mean activities and time-lagged autocorrelations in dynamical mean-field approximation.

b. Explanation of the result

Taking a step back, what is the intuition behind this result? When the binary neurons are averaged over realizations of the update time disorder, the Poisson update process with rate τ^{-1} becomes an exponential kernel corresponding to that of the rate network. The discrete jumps of the binary neurons around their mean become red noise [95,96] corresponding to the low-pass-filtered noise η of the rate network (A38). By nice conspiracy, this noise

corresponds to simple white noise ξ in Eq. (21), which is also the version treated in most works on rate networks with noise, such as Refs. [22,23,32,46,97,98]. This tight relation between the binary and rate models is summarized conceptually in Fig. 14.

9. Slope of correlation transmission in binary and rate neurons

Here we show that the difference between discrete signaling and continuous signaling leads to a qualitative difference in the slope of the correlation-transmission curve and thus the transition to chaos.

Assume, as an approximation, that two neurons receive inputs that are jointly Gaussian distributed as

$$(h_1, h_2) \sim \mathcal{N}(0, K),$$

where the covariance matrix is given by

$$K(c_{\text{in}}) := q\begin{pmatrix} 1 & c_{\text{in}} \\ c_{\text{in}} & 1 \end{pmatrix}.$$

Here, $c_{\text{in}} \in [-1, 1]$ controls the correlation between the inputs.

a. Continuous signaling

A neuron with continuous signaling has the output

$$y_i = T(h_i),$$

where $T \in [-1, 1]$ is an activation function. The mean output is thus,

$$\langle y_i \rangle = \langle T(h) \rangle_{h\sim\mathcal{N}(0,q)}. \quad \text{(A40)}$$

For a point-symmetric gain function that we assume in the following, the mean vanishes so that the variance of the outputs is

$$a = \langle T^2(h) \rangle_{h\sim\mathcal{N}(0,q)}.$$

The correlation coefficient between the outputs of a pair of neurons is

$$
\begin{aligned}
c_{\text{out}}^{\text{cont}}(c_{\text{in}}) &:= a^{-1}\langle y_1 y_2 \rangle \\
&= a^{-1}\langle T(h_1)T(h_2) \rangle_{(h_1,h_2)\sim\mathcal{N}(0,K(c_{\text{in}}))}, \quad \text{(A41)}
\end{aligned}
$$

which has the slope

$$\frac{dc_{\text{out}}^{\text{cont}}(c_{\text{in}})}{dc_{\text{in}}} = a^{-1}\langle T'(h_1)T'(h_2) \rangle_{(h_1,h_2)\sim\mathcal{N}(0,K(c_{\text{in}}))}$$

by Price's theorem [30]. Evaluated at $c_{\text{in}} = 1$, this is

$$\frac{dc_{\text{out}}^{\text{cont}}(1)}{dc_{\text{in}}} = a^{-1}\langle[T'(h)]^2\rangle_{h\sim\mathcal{N}(0,q)} \overset{T'<\infty}{<} \infty. \qquad (A42)$$

For activation functions T with finite slope $T' < \infty$, this slope is thus finite. For the signum function $T(x) = 2H(x) - 1$, we get $a = 1$ and

$$\frac{dc_{\text{out}}^{\text{cont}}(c_{\text{in}})}{dc_{\text{in}}} = 2\langle\delta(h_1)\delta(h_2)\rangle_{(h_1,h_2)\sim\mathcal{N}(0,K(c_{\text{in}}))}$$

$$= \frac{1}{\pi\sqrt{\det(K(c_{\text{in}}))}} = (q\pi)^{-1}(1 - c_{\text{in}}^2)^{-\frac{1}{2}}, \quad (A43)$$

where the latter line comes from the normalization condition of the two-dimensional Gaussian distribution. Thus, the slope diverges if and only if the output of the neuron becomes discrete.

b. Discrete signaling

Now consider a neuron with discrete output but smooth activation function $T \in [-1, 1]$; a smooth function here corresponds to a probabilistic activation

$$y_i = \begin{cases} 1 & \text{with prob } [T(h) + 1]/2, \\ -1 & \text{with prob } 1 - [T(h) + 1]/2. \end{cases}$$

The mean output is thus,

$$\langle y_i \rangle = \langle 1 \times [T(h) + 1]/2 - 1 \times \{1 - [T(h) + 1]/2\}\rangle_{h\sim\mathcal{N}(0,q)}$$

$$= \langle T(h) \rangle_{h\sim\mathcal{N}(0,q)},$$

and the same as for the continuous signaling (A40). For a point-symmetric gain function that we assume in the following, the mean vanishes so that the variance of the outputs is $a = 1$. The correlation coefficient of the output is then identical to the second moment between the outputs of a pair of neurons

$$c_{\text{out}}^{\text{disc}}(c_{\text{in}}) := \langle y_1 y_2 \rangle$$

$$= 1 \times (1 - p_{\text{diff}}) - 1 \times p_{\text{diff}} = 1 - 2p_{\text{diff}}$$

$$= 1 - \langle|T(h_1) - T(h_2)|\rangle_{(h_1,h_2)\sim\mathcal{N}(0,K(c_{\text{in}}))}. \quad (A44)$$

The latter expression is related to the probability $p_{\text{diff}} = \frac{1}{2}\langle|T(h_1) - T(h_2)|\rangle$ that the two neurons are in different states. This expression is of course the same as found in Eq. (A27). In the limit of $c_{\text{in}} \to 1$, we thus have

$$c_{\text{out}}^{\text{disc}}(c_{\text{in}}) \overset{c_{\text{in}}=1-\hat{\epsilon}}{\simeq} 1 - 2\langle T'(h)\rangle_{h\sim\mathcal{N}(0,q)}\sqrt{\frac{\hat{\epsilon}}{\pi}} + \mathcal{O}(\hat{\epsilon}^{3/2}),$$

where $\hat{\epsilon} = (\epsilon/g^2)$. So the slope diverges for $c_{\text{in}} \to 1$ as

$$\frac{dc_{\text{out}}^{\text{disc}}(c_{\text{in}})}{dc_{\text{in}}}\Big|_{c_{\text{in}}=1-\hat{\epsilon}}$$

$$\overset{(A27)}{=} \frac{d}{d(-\hat{\epsilon})}\left[1 - 2\langle T'(h)\rangle_{h\sim\mathcal{N}(0,q)}\sqrt{\frac{\hat{\epsilon}}{\pi}} + \mathcal{O}(\hat{\epsilon}^{3/2})\right]$$

$$= \langle T'(h)\rangle_{h\sim\mathcal{N}(0,q)}(\pi\hat{\epsilon})^{-\frac{1}{2}} + \mathcal{O}(\hat{\epsilon}^{1/2}) \propto (1 - c_{\text{in}})^{-\frac{1}{2}}.$$

This divergence is present even if the gain function has a finite slope $T' < \infty$. This is in qualitative contrast to the finite slope found for the continuous signaling in Eq. (A42).

The infinite slope for continuous signaling in the limit of a sharp activation function (A43) can be shown to have the same form of divergence for $c_{\text{in}} \to 1$ when expanded for $c_{\text{in}} = 1 - (1/q)\epsilon$ in the limit of small $\epsilon \ll 1$.

10. Noisy binary pattern classification task

We implement a classification task by training one linear readout

$$S_{\alpha'}(t) = w_{\alpha'}(t)^T[x_\alpha(t) + \xi_{\text{pre}}] + \xi_{\text{post}} \qquad (A45)$$

of the network state $x_\alpha(t)$ at time t for each of the $\alpha' = 1, \ldots, P = 50$ patterns to be detected. Here, ξ_{pre} and ξ_{post} are additional Gaussian readout noises of standard deviation $\sigma_{\xi,\text{pre}}$ and $\sigma_{\xi,\text{post}}$, respectively. ξ_{pre} controls how precisely a single neuron's state can be read out. ξ_{post} represents a noise component of the classification mechanism. Training of the readout $w_{\alpha'}(t)$ is performed for each time point t by linear regression (see the Appendix 10 a) minimizing the quadratic error of detecting the stimulus identity, i.e., minimizing $(S_{\alpha'} - \delta_{\alpha\alpha'})^2$.

The patterns are presented to the network by initializing the first $L = 10$ of the $N = 500$ neurons to the stimulus. All other neurons are in an initial state corresponding to the stationary statistics. Each stimulus α is a random binary pattern of length L with $\{-1, 1\}$ appearing equally likely, superimposed with Gaussian noise of standard deviation σ. Note that because of the noise added to the binary values, these initial states are not strictly $\in \{-1, 1\}$. This freedom in the initial states is just a way to introduce the noise; after their first update, the neurons' states are strictly $\{-1, 1\}$ again. The resulting evolution of the network state given this initial condition is termed $x_\alpha(t)$.

a. Linear regression

Minimizing the quadratic error over all patterns amounts to linear regression; we consider a single scalar readout target value $y_\alpha \in \mathbb{R}$ for each pattern α. In the example above, $y_\alpha \in \{0, 1\}$. Then, $w, x \in \mathbb{R}^N$, and the quadratic error is

$$\epsilon := \min_w \sum_{\alpha=1}^P (w^T x_\alpha - y_\alpha)^2. \qquad (A46)$$

Demanding stationarity with regard to w by differentiating by ∂_{w_i}, we get N equations

$$0 = \sum_{\alpha=1}^{P} 2(w^T x_\alpha - y_\alpha)x_{\alpha i} \quad \forall i$$

$$w^T \sum_{\alpha=1}^{P} x_\alpha x_\alpha^T = \sum_{\alpha=1}^{P} y_\alpha x_\alpha^T.$$

The value w^* to achieve stationarity is

$$w^* = C^{-1} \sum_{\alpha=1}^{P} y_\alpha x_\alpha^T$$

$$\text{with } C := \sum_{\alpha=1}^{P} x_\alpha x_\alpha^T, \tag{A47}$$

where we use the symmetry of C. Inserted into Eq. (A46)

$$\epsilon = w^{*T} C w^* - 2w^{*T} \sum_\alpha y_\alpha x_\alpha + \sum_\alpha y_\alpha^2$$

$$= \sum_{\alpha=1}^{P} y_\alpha^2 - S,$$

$$S = \left(\sum_{\alpha=1}^{P} y_\alpha x_\alpha^T\right) C^{-1} \left(\sum_{\alpha=1}^{P} y_\alpha x_\alpha\right). \tag{A48}$$

In the case of classification, the latter expression simplifies even further: The first term is a constant $\sum_{\alpha=1}^{P} y_\alpha^2 = 1$ for labels $y_\alpha \in \{0, 1\}$, where $y_\alpha = 1$ if the presented pattern α is the pattern α' to be detected and $y_\alpha = 0$ else, $y_\alpha = \delta_{\alpha\alpha'}$. The second term is then identical to the definition Eq. (A45) for $\xi = 0$ obtained by inserting w^* from Eq. (A47). The expression shows that the signal amplitude $S_{\alpha'}$ actually depends on the signal-to-noise ratio, the length of the selected vector $x_{\alpha'}$ measured with regard to the variability C across all patterns

$$S_{\alpha'} = x_{\alpha'}^T C^{-1} x_{\alpha'}. \tag{A49}$$

The generalization to stochastic realizations of x_α, for example, due to the presentation of noisy patterns, is straightforward. We need to replace $\sum_{\alpha=1}^{P} \ldots$ with $\sum_{\alpha=1}^{P} \langle \ldots \rangle$ in the measure for the error (A46) and thus throughout this calculation, where $\langle \ldots \rangle$ is the expectation over the noise realizations.

b. Approximation of orthogonal patterns and uniform noise

If the patterns x are sufficiently orthogonal in the signal subspace, we can think of the entries k of any state vector x_α to be drawn independently. So, the $x_{\alpha,k} \in \{-1, 1\}$ appear with equal probability for those entries that lie in the

subspace of dimension d_s. All remaining entries are assumed to be constant across patterns. We may thus restrict the space to the d_s informative components. The kl element of the covariance matrix for independently drawn entries is

$$C_{kl} = \sum_{\alpha=1}^{P} x_{\alpha k} x_{\alpha l}$$

$$\simeq P \delta_{kl}. \tag{A50}$$

The signal of the readout α' following from Eq. (A49) then takes the simple form

$$S_{\alpha'} \simeq P^{-1} \|x_{\alpha'}\|_{d_s}^2. \tag{A51}$$

If the signal is perfectly reliable, that is, if for all noise realizations i the response $x_{\alpha i}$ is equal to the stereotypical response $x_{\alpha i} = \bar{x}_\alpha$, and if the dimension of the informative subspace is d_s, so $\bar{x} \in \mathbb{R}^{d_s}$, we get with $\|\bar{x}\|_{d_s}^2 = d_s$,

$$S_{\alpha',\max} \lesssim \frac{d_s}{P}.$$

If noisy realizations of patterns cause flips in random entries of $x_{\alpha'}$, which is an approximation since spins are expected to differ in their susceptibility, the responses are not perfectly reliable, so we need to replace $x_{\alpha'}$ by $\langle x_{\alpha'} \rangle$ in Eq. (A49) and thus,

$$S_{\alpha'} \simeq P^{-1} \|\langle x_{\alpha'} \rangle\|_{d_s}^2.$$

The above expressions thereby link the readout signal to the dimensionality of the responses, as discussed in the main text in Sec. II G.

c. Nonzero plateau of the signal

In the simulations, the noise distance somewhat unintuitively saturates slightly below the signal distance. This is explainable by taking into account that not all the initial noise realizations actually cause a crossing of the flux-tube boundary. Instead, those realizations simply follow the unperturbed pattern trajectory, so that $d_{n,i} = 0$ in those cases. Then, it is clear that the average noise distance is smaller than the one predicted based on the assumption of diverging trajectories:

$$\langle d_{n,i} \rangle_i = (1 - p_{\text{no flip}}) d_n^{\text{residual}}.$$

We can estimate the probability that no flip occurred due to the noise by using the results from the Appendix 7, where we calculate the average number of flips in the network after one time constant given an additional (noise) variance in the input of the neurons $\sigma_{\Delta h}^2$. In our present case, the noise is given by adding $\xi_i \sim \mathcal{N}(0, \sigma^2)$ on the output

activities of the L original neurons of the pattern, so that the corresponding input variance felt by all neurons in the network is

$$\tilde{\sigma}_{\Delta h}^2 = \frac{g^2}{N}\sigma^2 L.$$

Now we need only to consider that actually the variance is not constant for a complete round of N updates, but linearly diminishes every time one of the L source neurons is updated until none is left. Since the flip probability depends on the square root of the variance, there is a corrective factor $c_{\text{dim}} = \sqrt{1 - (k/N)}$ in each term of the product:

$$
\begin{aligned}
p_{\text{no flip}} &= \prod_{k=0}^{N-1}[1 - p_{\text{single flip}}(k)]\\
&= \prod_{k=0}^{N-1}\left(1 - \frac{1}{N}\bar{n}_{\text{spawns}}(\tilde{\sigma}_{\Delta h})c_{\text{dim}}(k)\right)\\
&= \prod_{k=0}^{N-1}\left(1 - g\sigma\sqrt{\frac{2L}{\pi N}}\langle T'_p(h)\rangle_{h\sim\mathcal{N}(R,Q_0)}\sqrt{1 - \frac{k}{N}}\right).
\end{aligned}
$$
(A52)

This result fits well with the simulations, yielding the predicted offset of the asymptotic average signal and noise distances shown in Fig. 7(b) and the asymptotic plateau of the approximated average signal in Fig. 7(c).

11. Description of simulations

Simulations for Figs. 5, 4(b), and 4(d) are implemented using NEST [99]. NEST treats binary neurons in the bitlike $\{0, 1\}$ representation. To let every neuron "see" inputs from $\{-1, 1\}$ (Ising spins) we add to each neuron i a bias $\sum_j J_{ij}$ and then connect the neurons by the connections $2J_{ij}$ instead of J_{ij}, thereby effectively simulating an Ising system. To obtain the autocorrelations for the Ising case, Eq. (A13) is used, leading to the result shown in Fig. 5. Furthermore, we use a non-point-symmetric activation function $T(h) = \tanh(h - \Theta)$ by choosing a Θ to be nonzero for this plot. The reason is, first, that $\Theta = 0$ leads to the theoretical prediction of maximal output variance $\langle\!\langle x^2 \rangle\!\rangle = 1$ because the mean output activity $\langle x \rangle$ is 0. However, due to disorder, the time-averaged activity is actually a fluctuating quantity across the population, and therefore, the population-averaged variance is always below 1 for finite systems. This systematic underestimation of the peak of the autocorrelation at zero time lag can be avoided only by choosing an activation function that is not point symmetric. Second, this choice is also natural because mean activity 0 would imply that neurons are active half of the time on average (see Supplemental Material II B in Ref. [100]), which is considerably more than indicated by the low firing rates measured in cortex [101]. For Fig. 5, we

therefore shift the working point by numerically inverting $\langle x \rangle = \langle\tanh(h - \Theta)\rangle_{h\sim\mathcal{N}(R,Q)}$ to obtain the value for Θ, which in mean-field approximation corresponds to $\langle x \rangle = -0.5$. In the simulations, it turns out that $\langle x \rangle = -0.501$, which agrees well with mean-field theory.

For Figs. 4(b) and 4(d), for each point of the grid one simulation of two identical networks is performed. After 1000 ms, in one replica, the first two neurons are set to the active state, and the third and fourth neurons are set to the inactive state, after which, the simulation continues for 2500 ms. This method of perturbation entails the small probability that these four neurons are already in exactly this state, so that nothing is changed; this is the explanation for the scattered single green dots in Figs. 4(b) and 4(d). The advantage of the method is, however, that it guarantees the same state of the random number generators across both replicas.

The simulations for Figs. 7, 9, and 10 are performed using a custom FORTRAN kernel.

The simulations of the LIF network for Fig. 11 are implemented in NEST, using the "iaf_psc_delta" neuron with its default parameter settings $V_{\text{rest}} = V_{\text{reset}} = -70$ mV, $V_{\text{th}} = -55$ mV, $\tau_m = 10$ ms, $t_{\text{refr}} = 2$ ms, $C_m = 250$ pF. Asynchronous-irregular firing in the inhibitory network is evoked by supplying external excitatory Poisson input with rate of 3000 Hz and unit weight, leading to an average network activity of 19 Hz. After simulating for $10\tau_m$ to obtain a state with stationary statistics, the input pattern is applied by causing a spike in the corresponding neurons. Noise is added to the pattern by additional external input spikes perturbing the membrane potentials of the L pattern neurons. Further parameters as mentioned in the Fig. 11 caption are $N = 500, J_{ii} = -1$ mV, $K = 125, \tau_m = 10$ ms.

The LSTM network for Fig. 12 is simulated using the vanilla PYTORCH implementation with a hidden layer of size $N = 200$. To obtain chaotic fluctuations over a range of timescales, the hidden weights W_{HI}, W_{HF}, W_{HG} are initialized distributed as $\mathcal{N}(0, 5.8/N)$. This excludes the W_{HO} weights, which will cause very rapid, erratic dynamics. Instead, these and all remaining weights and biases used the default uniformly distributed $\mathcal{U}(-1/\sqrt{N}, 1/\sqrt{N})$ initialization. After simulating for 50 time steps to obtain stationary statistics, the input patterns are supplied using the standard input function.

Analysis of simulation data and numerical solutions are implemented in PYTHON. The code to generate all figures is available as a Zenodo archive [102].

---

[1] R. Kempter, W. Gerstner, and J. L. Van Hemmen, in *Proceedings of the 11th International Conference on Neural Information Processing Systems, 1998* (MIT Press, Cambridge, MA, 1998), https://papers.nips.cc/paper/1998/file/42a3964579017f3cb42b26605b9ae8ef-Paper.pdf.

[2] M. Pfeiffer and T. Pfeil, *Deep Learning with Spiking Neurons: Opportunities and Challenges*, Front. Neurosci. **12,** 00774 (2018).

[3] S. Laughlin, *Energy as a Constraint on the Coding and Processing of Sensory Information*, Curr. Opin. Neurobiol. **11,** 475 (2001).

[4] M. Abeles, *Corticonics: Neural Circuits of the Cerebral Cortex* (Cambridge University Press, Cambridge, England, 1991).

[5] W. R. Softky and C. Koch, *The Highly Irregular Firing of Cortical Cells Is Inconsistent with Temporal Integration of Random EPSPs*, J. Neurosci. **13,** 334 (1993).

[6] P. König, A. K. Engel, and W. Singer, *Integrator or Coincidence Detector? The Role of the Cortical Neuron Revisited*, Trends Neurosci. **19,** 130 (1996).

[7] M. N. Shadlen and W. T. Newsome, *The Variable Discharge of Cortical Neurons: Implications for Connectivity, Computation, and Information Coding*, J. Neurosci. **18,** 3870 (1998).

[8] M. N. Shadlen and A. J. Movshon, *Synchrony Unbound: A Critical Evaluation of the Temporal Binding Hypothesis*, Neuron **24,** 67 (1999).

[9] E. T. Rolls and G. Deco, *The Noisy Brain: Stochastic Dynamics as a Principle* (Oxford University Press, New York, 2010).

[10] R. Brette, *Philosophy of the Spike: Rate-Based vs. Spike-Based Theories of the Brain*, Front. Syst. Neurosci. **9,** 151 (2015).

[11] S. Denève and C. K. Machens, *Efficient Codes and Balanced Networks*, Nat. Neurosci. **19,** 375 (2016).

[12] D. V. Buonomano and M. M. Merzenich, *Temporal Information Transformed into a Spatial Code by a Neural Network with Realistic Properties*, Science **267,** 1028 (1995).

[13] H. Jaeger, *German National Research Center for Information Technology Technical* Report No. 148, 2001.

[14] W. Maass, T. Natschläger, and H. Markram, *Real-Time Computing without Stable States: A New Framework for Neural Computation Based on Perturbations*, Neural Comput. **14,** 2531 (2002).

[15] V. N. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

[16] T. M. Cover, *Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition*, IEEE Trans. Electron. Comput. **EC-14,** 326 (1965).

[17] 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).

[18] N. Bertschinger and T. Natschläger, *Real-Time Computation at the Edge of Chaos in Recurrent Neural Networks*, Neural Comput. **16,** 1413 (2004).

[19] R. Legenstein and W. Maass, in *New Directions in Statistical Signal Processing: From System to Brains*, edited by S. Haykin, J. C. Principe, T. J. Sejnowski, and J. G. McWhirter (MIT Press, Cambridge, MA, 2007), pp. 127–154.

[20] R. Legenstein and W. Maass, *Edge of Chaos and Prediction of Computational Performance for Neural Circuit Models*, Neural Netw. **20,** 323 (2007).

[21] H. Sompolinsky, A. Crisanti, and H. J. Sommers, *Chaos in Random Neural Networks*, Phys. Rev. Lett. **61,** 259 (1988).

[22] A. Crisanti and H. Sompolinsky, *Path Integral Approach to Random Neural Networks*, Phys. Rev. E **98,** 062120 (2018).

[23] J. Schuecker, S. Goedeke, and M. Helias, *Optimal Sequence Memory in Driven Random Networks*, Phys. Rev. X **8,** 041029 (2018).

[24] C. van Vreeswijk and H. Sompolinsky, *Chaos in Neuronal Networks with Balanced Excitatory and Inhibitory Activity*, Science **274,** 1724 (1996).

[25] C. van Vreeswijk and H. Sompolinsky, *Chaotic Balanced State in a Model of Cortical Circuits*, Neural Comput. **10,** 1321 (1998).

[26] P. H. Baxendale, in *Diffusion Processes and Related Problems in Analysis* (Birkhäuser, Boston, 1992), Vol. 2, pp. 3–35.

[27] R. Glauber, *Time-Dependent Statistics of the Ising Model*, J. Math. Phys. (N.Y.) **4,** 294 (1963).

[28] I. Ginzburg and H. Sompolinsky, *Theory of Correlations in Stochastic Neural Networks*, Phys. Rev. E **50,** 3171 (1994).

[29] D. R. Cox, *Renewal Theory* (Chapman and Hall, London, 1962).

[30] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. (McGraw-Hill, Boston, 2002).

[31] D. Sussillo and L. F. Abbott, *Generating Coherent Patterns of Activity from Chaotic Neural Networks*, Neuron **63,** 544 (2009).

[32] J. Kadmon and H. Sompolinsky, *Transition to Chaos in Random Neuronal Networks*, Phys. Rev. X **5,** 041030 (2015).

[33] B. Derrida and Y. Pomeau, *Random Networks of Automata: A Simple Annealed Approximation*, Europhys. Lett. **1,** 45 (1986).

[34] M. P. Touzel and F. Wolf, *Statistical Mechanics of Spike Events Underlying Phase Space Partitioning and Sequence Codes in Large-Scale Models of Neural Circuits*, Phys. Rev. E **99,** 052402 (2019).

[35] S. Hwang, V. Folli, E. Lanza, G. Parisi, G. Ruocco, and F. Zamponi, *On the Number of Limit Cycles in Asymmetric Neural Networks*, J. Stat. Mech. (2019) 053402.

[36] M. E. Turner, E. L. Bradley, K. A. Kirk, and K. M. Pruitt, *A Theory of Growth*, Math. Biosci. **29,** 367 (1976).

[37] G. Wainrib and J. Touboul, *Topological and Dynamical Complexity of Random Neural Networks*, Phys. Rev. Lett. **110,** 118101 (2013).

[38] R. Engelken, F. Wolf, and L. F. Abbott, *Lyapunov Spectra of Chaotic Recurrent Neural Networks*, arXiv:2006.02427.

[39] S. Hochreiter and J. Schmidhuber, *Long Short-Term Memory*, Neural Comput. **9,** 1735 (1997).

[40] M. Monteforte and F. Wolf, *Dynamic Flux Tubes Form Reservoirs of Stability in Neuronal Circuits*, Phys. Rev. X **2,** 041007 (2012).

[41] S. A. Kauffman, *The Origins of Order Self-Organization and Selection in Evolution* (Oxford University Press, New York, 1993).

[42] D. Snyder, A. Goudarzi, and C. Teuscher, in *Proceedings of the 13th International Conference on the Synthesis and*

Simulation of Living Systems, ALIFE 2012 (MIT Press, Cambridge, MA, 2012).

[43] D. Grytskyy, T. Tetzlaff, M. Diesmann, and M. Helias, *A Unified View on Weakly Correlated Recurrent Networks*, Front. Comput. Neurosci. **7,** 131 (2013).

[44] T. Kühn, Ph.D. thesis, *RWTH Aachen University*, 2020, https://doi.org/10.18154/RWTH-2020-01833.

[45] T. Tetzlaff, M. Helias, G. T. Einevoll, and M. Diesmann, *Decorrelation of Neural-Network Activity by Inhibitory Feedback*, PLoS Comput. Biol. **8,** e1002596 (2012).

[46] J. Aljadeff, M. Stern, and T. Sharpee, *Transition to Chaos in Random Networks with Cell-Type-Specific Connectivity*, Phys. Rev. Lett. **114,** 088101 (2015).

[47] Ł. Kuśmierz, S. Ogawa, and T. Toyoizumi, *Edge of Chaos and Scale-Free Avalanches in Neural Networks with Heavy-Tailed Synaptic Disorder*, Phys. Rev. Lett. **125,** 028101 (2020).

[48] P. Martin, E. Siggia, and H. Rose, *Statistical Dynamics of Classical Systems*, Phys. Rev. A **8,** 423 (1973).

[49] C. De Dominicis, *Techniques de Renormalisation de la Théorie des Champs et Dynamique des Phénomènes Critiques*, J. Phys. (Paris), Colloq. **37,** C1 (1976).

[50] C. Chow and M. Buice, *Path Integral Methods for Stochastic Differential Equations*, J. Math. Neurosci. **5,** 8 (2015).

[51] J. A. Hertz, Y. Roudi, and P. Sollich, *Path Integral Methods for the Dynamics of Stochastic and Disordered Systems*, J. Phys. A **50,** 033001 (2017).

[52] M. Helias and D. Dahmen, *Statistical Field Theory for Neural Networks* (Springer International Publishing, 2020), Vol. 970, p. 203, https://doi.org/10.1007/978-3-030-46444-8.

[53] M. Doi, *Second Quantization Representation for Classical Many-Particle System*, J. Phys. A **9,** 1465 (1976).

[54] L. Peliti, *Path Integral Approach to Birth-Death Processes on a Lattice*, J. Phys. (Paris) **46,** 1469 (1985).

[55] M. A. Buice and J. D. Cowan, *Field-Theoretic Approach to Fluctuation Effects in Neural Networks*, Phys. Rev. E **75,** 051919 (2007).

[56] H. J. Sommers, *Path-Integral Approach to Ising Spin-Glass Dynamics*, Phys. Rev. Lett. **58,** 1268 (1987).

[57] A. Andreanov, G. Biroli, J.-P. Bouchaud, and A. Lefevre, *Field Theories and Exact Stochastic Equations for Interacting Particle Systems*, Phys. Rev. E **74,** 030101(R) (2006).

[58] A. Lefevre and G. Biroli, *Dynamics of Interacting Particle Systems: Stochastic Process and Field Theory*, J. Stat. Mech. (2007) P07024.

[59] L. Büsing, B. Schrauwen, and R. Legenstein, *Connectivity, Dynamics, and Memory in Reservoir Computing with Binary and Analog Neurons*, Neural Comput. **22,** 1272 (2010).

[60] T. Tetzlaff, A. Morrison, T. Geisel, and M. Diesmann, *Consequences of Realistic Network Size on the Stability of Embedded Synfire Chains*, Neurocomputing;Variable Star Bulletin **58–60,** 117 (2004).

[61] E. Shea-Brown, K. Josic, J. de la Rocha, and B. Doiron, *Correlation and Synchrony Transfer in Integrate-and-Fire Neurons: Basic Properties and Consequences for Coding*, Phys. Rev. Lett. **100,** 108102 (2008).

[62] T. Tchumatchenko, A. Malyshev, T. Geisel, M. Volgushev, and F. Wolf, *Correlations and Synchrony in Threshold Neuron Models*, Phys. Rev. Lett. **104,** 058102 (2010).

[63] T. Tchumatchenko, T. Geisel, M. Volgushev, and F. Wolf, *Spike Correlations—What Can They Tell about Synchrony?* Front. Neurosci. **5,** 00068 (2011).

[64] T. Tetzlaff, M. Buschermöhle, T. Geisel, and M. Diesmann, *The Spread of Rate and Correlation in Stationary Cortical Networks*, Neurocomputing;Variable Star Bulletin **52–54,** 949 (2003).

[65] M. Schultze-Kraft, M. Diesmann, S. Gruen, and M. Helias, *Noise Suppression and Surplus Synchrony by Coincidence Detection*, PLoS Comput. Biol. **9,** e1002904 (2013).

[66] T. Deniz and S. Rotter, *Solving the Two-Dimensional Fokker-Planck Equation for Strongly Correlated Neurons*, Phys. Rev. E **95** (2017).

[67] G. Lajoie, K. K. Lin, and E. Shea-Brown, *Chaos and Reliability in Balanced Spiking Networks with Temporal Drive*, Phys. Rev. E **87,** 052901 (2013).

[68] G. Lajoie, J.-P. Thivierge, and E. Shea-Brown, *Structured Chaos Shapes Spike-Response Noise Entropy in Balanced Neural Networks*, Front. Comput. Neurosci. **8,** 00123 (2014).

[69] E. Gardner and B. Derrida, *Optimal Storage Properties of Neural Network Models*, J. Phys. A **21,** 271 (1988).

[70] A. Litwin-Kumar, M. J. Chacron, and B. Doiron, *The Spatial Structure of Stimuli Shapes the Timescale of Correlations in Population Spiking Activity*, PLoS Comput. Biol. **8,** e1002667 (2012).

[71] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, in *Proceedings of Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2016), pp. 3360–3368, https://papers.nips.cc/paper/2016/hash/148510031349642de5ca0c544f31b2ef-Abstract.html.

[72] S. Recanatesi, M. Farrell, M. Advani, T. Moore, G. Lajoie, and E. Shea-Brown, *Dimensionality Compression and Expansion in Deep Neural Networks*, arXiv:1906.00443.

[73] M. Farrell, S. Recanatesi, T. Moore, G. Lajoie, and E. Shea-Brown, *Recurrent Neural Networks Learn Robust Representations by Dynamically Balancing Compression and Expansion*, https://www.biorxiv.org/content/10.1101/564476v2.

[74] G. Hennequin, T. P. Vogels, and W. Gerstner, *Non-Normal Amplification in Random Balanced Neuronal Networks*, Phys. Rev. E **86,** 011909 (2012).

[75] G. Kerg, K. Goyette, M. P. Touzel, G. Gidel, E. Vorontsov, Y. Bengio, and G. Lajoie, in *Proceedings of Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2019), pp. 13613–13623, https://papers.nips.cc/paper/2019/hash/9d7099d87947faa8d07a272dd6954b80-Abstract.html.

[76] G. Bondanelli and S. Ostojic, *Coding with Transient Trajectories in Recurrent Neural Networks*, PLoS Comput. Biol. **16,** e1007655 (2020).

[77] W. Tarnowski, *Transient Amplification in Balanced Neural Networks*, arXiv:2011.08215.

[78] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, *Statistical Mechanics of Deep Learning*, Annu. Rev. Condens. Matter Phys. **11**, 501 (2020).

[79] S. Thorpe, D. Fize, and C. Marlot, *Speed of Processing in the Human Visual System*, Nature (London) **381**, 520 (1996).

[80] C. P. Hung, G. Kreiman, T. Poggio, and J. J. DiCarlo, *Fast Readout of Object Identity from Macaque Inferior Temporal Cortex*, Science **310**, 863 (2005).

[81] G. Lajoie, K. K. Lin, J.-P. Thivierge, and E. Shea-Brown, *Encoding in Balanced Networks: Revisiting Spike Patterns and Chaos in Stimulus-Driven Systems*, PLoS Comput. Biol. **12**, e1005258 (2016).

[82] R. Friedrich and G. Laurent, *Dynamic Optimization of Odor Representations by Slow Temporal Patterning of Mitral Cell Activity*, Science **291**, 889 (2001).

[83] O. Mazor and G. Laurent, *Transient Dynamics versus Fixed Points in Odor Representations by Locust Antennal Lobe Projection Neurons*, Neuron **48**, 661 (2005).

[84] K. M. Cury and N. Uchida, *Robust Odor Coding via Inhalation-Coupled Transient Activity in the Mammalian Olfactory Bulb*, Neuron **68**, 570 (2010).

[85] P. Grassberger and I. Procaccia, *Estimation of the Kolmogorov Entropy from a Chaotic Signal*, Phys. Rev. A **28**, 2591 (1983).

[86] D. Toker, F. T. Sommer, and M. D'Esposito, *A Simple Method for Detecting Chaos in Nature*, Commun. Biol. **3**, 11 (2020).

[87] A. Celletti and A. E. P. Villa, *Low-Dimensional Chaotic Attractors in the Rat Brain*, Biol. Cybern. **74**, 387 (1996).

[88] M. M. Churchland *et al.*, *Stimulus Onset Quenches Neural Variability: A Widespread Cortical Phenomenon*, Nat. Neurosci. **13**, 369 (2010).

[89] L. Mazzucato, A. Fontanini, and G. La Camera, *Stimuli Reduce the Dimensionality of Cortical Activity*, Front. Syst. Neurosci. **10**, 11 (2016).

[90] P. Gao, E. Trautmann, B. M. Yu, G. Santhanam, S. Ryu, K. Shenoy, and S. Ganguli, *A Theory of Multineuronal Dimensionality, Dynamics and Measurement*, https://www.biorxiv.org/content/10.1101/214262v2.

[91] A. C. C. Coolen, *Statistical Mechanics of Recurrent Neural Networks II. Dynamics*, in *Handbook of Biological Physics* (North-Holland, Amsterdam, 2001), Vol. 4, pp. 619–684, https://arxiv.org/abs/cond-mat/0006011.

[92] M. Helias, T. Tetzlaff, and M. Diesmann, *The Correlation Structure of Local Cortical Networks Intrinsically Results from Recurrent Dynamics*, PLoS Comput. Biol. **10**, e1003428 (2014).

[93] G. E. Uhlenbeck and L. S. Ornstein, *On the Theory of the Brownian Motion*, Phys. Rev. **36**, 823 (1930).

[94] Note that our notational convention differs by the factor 1/2.

[95] B. Lindner, in *Stochastic Methods in Neuroscience*, edited by C. Laing and G. J. Lord (Oxford University Press, Oxford, 2009).

[96] E. Frey and M. F. Weber, *Master Equations and the Theory of Stochastic Path Integrals*, Rep. Prog. Phys. **80**, 046601 (2017).

[97] K. Rajan, L. F. Abbott, and H. Sompolinsky, *Stimulus-Dependent Suppression of Chaos in Recurrent Neural Networks*, Phys. Rev. E **82**, 011903 (2010).

[98] D. Martí, N. Brunel, and S. Ostojic, *Correlations between Synapses in Pairs of Neurons Slow Down Dynamics in Randomly Connected Neural Networks*, Phys. Rev. E **97**, 062314 (2018).

[99] C. Linssen *et al.*, NEST 2.16.0, 2018.

[100] T. Kühn and M. Helias, *Locking of Correlated Neural Activity to Ongoing Oscillations*, PLoS Comput. Biol. **13**, e1005534 (2017).

[101] A. Roxin, N. Brunel, D. Hansel, G. Mongillo, and C. van Vreeswijk, *On the Distribution of Firing Rates in Networks of Cortical Neurons*, J. Neurosci. **31**, 16217 (2011).

[102] https://doi.org/10.5281/zenodo.4705262.