


Optimal reinforcement learning near the edge of a synchronization transitionMahsa Khoshkhou and Afshin Montakhab *Department of Physics, College of Sciences, Shiraz University, Shiraz 71946-84795, Iran*

(Received 21 October 2021; accepted 30 March 2022; published 19 April 2022)

Recent experimental and theoretical studies have indicated that the putative criticality of cortical dynamics may correspond to a synchronization phase transition. The critical dynamics near such a critical point needs further investigation specifically when compared to the critical behavior near the standard absorbing state phase transition. Since the phenomena of learning and self-organized criticality (SOC) at the edge of synchronization transition can emerge jointly in spiking neural networks due to the presence of spike-timing dependent plasticity (STDP), it is tempting to ask the following: what is the relationship between synchronization and learning in neural networks? Further, does learning benefit from SOC at the edge of synchronization transition? In this paper, we intend to address these important issues. Accordingly, we construct a biologically inspired model of a cognitive system which learns to perform stimulus-response tasks. We train this system using a reinforcement learning rule implemented through dopamine-modulated STDP. We find that the system exhibits a continuous transition from synchronous to asynchronous neural oscillations upon increasing the average axonal time delay. We characterize the learning performance of the system and observe that it is optimized near the synchronization transition. We also study neuronal avalanches in the system and provide evidence that optimized learning is achieved in a slightly supercritical state.

DOI: [10.1103/PhysRevE.105.044312](https://doi.org/10.1103/PhysRevE.105.044312)**I. INTRODUCTION**

A cognitive system, either biological or artificial, is a continuously active complex system autonomously exploring and reacting to the environment with the capability to survive [1]. In contrast with the usual paradigm of artificial intelligence (AI) which mainly follows an all-in-one-step approach to intelligent systems [2–4], a cognitive system is not necessarily intelligent, but intelligence can be achieved once the system has been developed [5]. The universal principles necessary for the realization of a cognitive system resembling our own cognitive organ is not fully understood, yet. However, it is believed that learning within a biological cognitive system with self-induced dynamics is not supervised [6]. Accordingly, reinforcement learning (RL), i.e., learning how to map situations to actions so as to maximize a numerical reward signal, is the closest paradigm to the kind of learning that biological systems follow [7,8]. Trial-and-error search and delayed reward are the most important distinguishing characteristics of RL [7].

Spike-timing dependent plasticity (STDP) is an experimentally well-established rule that leads to synapse strengthening for correlated activity at the pre- and postsynaptic neurons and synapse weakening for decorrelated activity [9–13]. A modified version of this rule that is called dopamine-modulated STDP is the main candidate to explain the relationship between synaptic plasticity on the microscopic level, and the adaptive changes of behavior of biological organisms on the macroscopic level; i.e., linkage of dopamine signaling

with STDP triggered the development of phenomenological models for RL that could explain how behaviorally relevant adaptive changes in complex networks of spiking neurons could be achieved in a self-organizing manner [14–18].

STDP is also identified as a candidate mechanism underlying the emergence and maintenance of self-organized criticality (SOC) in neural circuits [19–21]. It is hypothesized that cortical neural circuits self-organize to a critical state which is associated with a transition point of a continuous phase transition [22–25]. Operating at the vicinity of this critical point has functional benefits for a neural system including optimal capacity of information processing, transfer, and storage as well as optimal dynamic range [25–28]. However, the relationship between SOC and learning is not clear. Criticality has been shown to be beneficial for learning in some models [29–31], but destructive to learning [32], or having task-dependent profit [33], in other models.

As opposed to the earlier models which have attributed the putative criticality of cortical dynamics to an activity transition point associated with an absorbing state phase transition [34–37], recent experimental and theoretical studies indicate that such critical dynamics corresponds to a synchronization phase transition, at which incipient oscillations and scale-free avalanches coexist [38–40]. In particular, we have previously shown that neurophysiological regulatory mechanisms such as STDP with suitable axonal conduction delays brings about and maintains SOC at the edge of synchronization transition in a biologically meaningful model of cortical networks [39]. The basic idea is that the positive feedback, where synchronization leads to synaptic potentiation which in turn leads to more synchronization, is broken with a time delay. Consequently, synaptic strengths depress and potentiate in turn as

*montakhab@shirazu.ac.ir

the system finds and maintains a critical state at the edge of synchronization with scale-free avalanches. The implications for the brain's performance at such a critical point has been investigated in recent studies [38–42].

We note that the phenomena of learning and SOC at the edge of synchronization are both rooted in the regulatory mechanism of STDP. Therefore, it is natural to ask what the relation between learning and the edge of synchronization is. In particular, can learning profit from the brain operating at or near a synchronization transition? This is exactly what we propose to address in this article.

In order to specify the relationship between synchronization, learning, and SOC in cognitive systems, we construct a biologically relevant model of cortical networks with parameters that are set according to empirical evidence. Constructing a biologically inspired model of a cognitive system is of crucial importance because it is an attempt toward building an artificial brain on one hand and toward understanding the relevance of various biological mechanisms in the brain on the other hand. This cognitive system learns to perform stimulus-response (SR) tasks with different levels of complication. We train this system using a RL rule which is implemented through dopamine-modulated STDP. We show that the system exhibits a continuous transition from synchronous to asynchronous neural oscillations upon increasing axonal time delays. The learning performance of the system depends on the amount of synchronization in neural activity. For all SR tasks of various complexity, optimal performance is achieved while the system is above but very close to the transition point of synchronization. More strictly, the optimal performance occurs at a slightly supercritical state of synchronization.

In the following section, we describe the model we use for our study. Results of our numerical study are presented in Sec. III and we close the paper with some concluding remarks in Sec. IV.

II. MODEL AND METHODS

We construct a biologically inspired cognitive system. A detailed description of this system is given below.

Neural dynamics. The system consists of N spiking Izhikevich neurons interacting by transition of chemical synaptic currents with axonal conduction delays. The dynamics of each neuron is described by a set of two differential equations [43]:

$$\dot{v}_i = 0.04v_i^2 + 5v_i + 140 - u_i + I_i^{DC} + I_i^{\text{syn}}, \quad (1)$$

$$\dot{u}_i = a(bv_i - u_i), \quad (2)$$

with the auxiliary after-spike reset:

$$\text{if } v_i \geq 30, \quad \text{then } v_i \rightarrow c \text{ and } u_i \rightarrow u_i + d \quad (3)$$

for $i = 1, 2, \dots, N$. Here, v_i is the membrane potential and u_i is the membrane recovery variable. When v_i reaches its apex ($v_{\max} = 30$ mV), v_i and u_i are reset according to Eq. (3). a , b , c , and d are adjustable parameters that determine the pattern of firing and are different for excitatory and inhibitory neurons (see Table I) [43]. The population density of inhibitory neurons is set to be $\rho = 0.2$.

The term I_i^{DC} is an external current which determines the intrinsic firing rate of uncoupled neurons [43,44]. The values

TABLE I. Values of constant parameters used in this study.

| | | | | |
|-------------------|---|--|--|--|
| Izhikevich neuron | $a^{\text{ex}} = 0.02$ $a^{\text{in}} = 0.1$ | $b^{\text{ex}} = 0.2$ $b^{\text{in}} = 0.2$ | $c^{\text{ex}} = -65$ $c^{\text{in}} = -65$ | $d^{\text{ex}} = 8$ $d^{\text{in}} = 2$ |
| Synaptic current | $\tau_f = 1$ | $\tau_s = 5$ | $V_0^{\text{ex}} = 0$ | $V_0^{\text{in}} = -75$ |
| Plasticity rule | $A_{\pm} = 0.05$ $\tau_x = 1000$ | $\tau_{\pm} = 30$ $\tau_y = 200$ | $w_{\min} = 0$ $y_0 = 2$ | $w_{\max} = 1$ |

of I_i^{DC} are chosen randomly from the range [3.8,4.5]. This choice leads to alpha-band intrinsic firing frequencies with a mean firing rate around 9 Hz. Alpha rhythms are typically observed during learning and task performance [45–48]. The term I_i^{syn} represents the chemical synaptic current that goes into each postsynaptic neuron i [49]:

$$I_i^{\text{syn}} = \frac{V_0 - v_i}{D_i} \sum_j w_{ji} \frac{\exp(-\frac{t-(t_j+\tau_{ji})}{\tau_s}) - \exp(-\frac{t-(t_j+\tau_{ji})}{\tau_f})}{\tau_s - \tau_f}. \quad (4)$$

Here, D_i is the in-degree of node i , t_j is the instance of last spike of presynaptic neuron j , and τ_{ji} is the axonal conduction delay from presynaptic neuron j to postsynaptic neuron i . The values of τ_{ji} are chosen randomly from a Poisson distribution with mean value $\tau = \langle \tau_{ji} \rangle$. $\tau = 0$ means that $\tau_{ji} = 0 \forall j \neq i$. τ_f and τ_s are the synaptic fast and slow time constants and V_0 is the reversal potential of the synapse. The most relevant variable in the context of learning is w_{ji} , which denotes the strength of synapse from neuron j to neuron i . w_{ji} 's are the elements of the adjacency matrix of the underlying network ($w_{ji} \neq 0$ if there is a directed edge from neuron j to neuron i and $w_{ji} = 0$ otherwise). The simulations are carried out using Erdős-Rényi (ER) random networks of size N and average connectivity $z = pN$, where we set the connection probability $p = 0.1$ [50]. The initial strength of excitatory synapses is $w_{ji}(t = 0) = w_0$. To enforce the excitation-inhibition balance in the system, the initial strength of inhibitory synapses is set to $w_{ji} = rw_0$, where r is the ratio of excitatory to inhibitory synapses.

Dopamine-modulated STDP rule. To train the network, we modify the strength of *excitatory* synapses using the simplest phenomenological model that captures the essence of dopamine modulation of STDP [14]. According to this model,

$$\dot{w}_{ji} = x_{ji}y, \quad (5)$$

$$\dot{x}_{ji} = -\frac{x_{ji}}{\tau_x} + \Gamma_{ji}(\Delta t)\delta(t - t_{\text{pre/post}}), \quad (6)$$

$$\dot{y} = -\frac{y}{\tau_y} + y_0\delta(t - t_{\text{rew}}), \quad (7)$$

$$\Gamma_{ji}(\Delta t) = \begin{cases} A_+(w_{\max} - w_{ji})e^{-\frac{\Delta t - \tau_{ji}}{\tau_+}} & \text{if } \Delta t > \tau_{ji}, \\ -A_-(w_{ji} - w_{\min})e^{-\frac{\Delta t - \tau_{ji}}{\tau_-}} & \text{if } \Delta t \leq \tau_{ji}. \end{cases} \quad (8)$$

Here, x_{ji} is the synaptic eligibility function or synaptic tag, y is the extracellular concentration of dopamine, which is assumed to be the same for all synapses, τ_x and τ_y are time constants, and $\delta(\cdot)$ is the Dirac delta function. t_{rew} is the time when the reward is delivered to the network and y_0 is the amount of dopamine released by the activity of dopaminergic neurons at $t = t_{\text{rew}}$. $\Gamma(\Delta t)$ is a nearest-neighbor and soft-band STDP function with upper and lower bands w_{\max} and w_{\min}

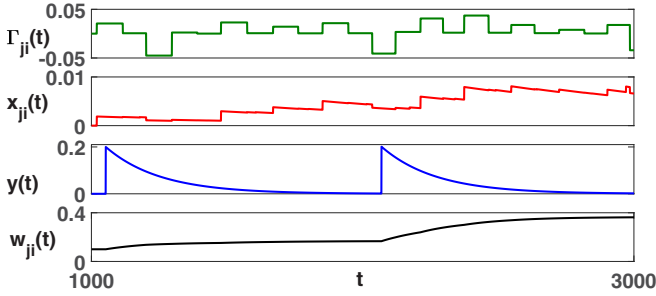


FIG. 1. Illustration of reward-modulated STDP rule to potentiate a synapse: temporal order of pre- and postsynaptic spikes determines the value of $\Gamma_{ji}(t)$ (green) and subsequently changes the synaptic tag $x_{ji}(t)$ (red). While the global reward $y(t)$ (blue) is released with a small delay after a correct network's response, if $x_{ji}(t)$ has a relatively large positive value, the synaptic strength $w_{ji}(t)$ (black) increases.

[9,11]. $\Delta t = t_{\text{post}} - t_{\text{pre}}$ is the time difference between the last post- and presynaptic spikes, A_{\pm} determine the maximum synaptic potentiation and depression, and τ_{\pm} determine the temporal extent of the STDP window for potentiation and depression. Numerical values of all of these parameters are listed in Table I. Figure 1 illustrates how dopamine-modulated STDP leads to potentiation of a synapse that plays a role in the correct network's response.

To include the transmission time for a causal relation between pre- and postsynaptic firing, we employ a temporally shifted STDP window for which the boundary separating potentiation and depression does not occur for simultaneous pre- and postsynaptic spikes, but rather for spikes separated by a small time shift [51]. We set the value of this shift equal with the actual axonal delay for each synapse. This rule retrieves the conventional STDP rule when $\tau = 0$. This temporal shift causes synchronous or nearly synchronous pre- and postsynaptic spikes to induce long-term depression, which leads to intrinsic stability in the network [51–53]. Also to keep the excitation-inhibition ratio balanced during the simulations, the strength of each inhibitory synapse is set to be $w_{ji}(t) = r\bar{w}(t)$, while $\bar{w}(t)$ is the average strength of all excitatory synapses at time t .

Neural synchronization. We integrate the dynamical equations using the fourth-order Runge-Kutta method with a time step $h = 0.1$ ms and obtain $v_i(t)$ and the spike times of all neurons. We assign a phase $\phi_i(t) = 2\pi \frac{t - t_i^m}{t_i^{m+1} - t_i^m}$ to each neuron between each pair of successive spikes. t_i^m is the time that neuron i emits its m th spike. Next we evaluate a time-dependent order parameter [39,44,54]:

$$S(t) = \frac{2}{N(N-1)} \sum_{i \neq j} \cos^2\left(\frac{\phi_i(t) - \phi_j(t)}{2}\right). \quad (9)$$

This order parameter measures the collective phase synchronization at time t . $S(t)$ is bounded between 0.5 and 1. If neurons spike out of phase, then $S(t) \simeq 0.5$, when they spike completely in phase $S(t) \simeq 1$ and for states with partial synchrony $0.5 < S(t) < 1$. The global order parameter S^* is the long-time average of $S(t)$ at the stationary state after the influence of STDP [$S^* = \langle S(t) \rangle_t$].

Learning and performance. Reinforcement of specific firing patterns is possible by this cognitive system. In this paper, we reinforce the network to produce an appropriate response to a stimulus. We choose n random nonoverlapping group of five neurons, called S_1, \dots, S_n , which represent the input stimulus to the network, and n other random nonoverlapping group of five neurons, called R_1, \dots, R_n , which give rise to n distinct responses of the network. R_i is assumed to be the correct response to stimulus S_i ($i = 1, \dots, n$). Larger n corresponds to more complicated tasks learned by the network. A series of simulations is carried out with n pairs of stimulus response denoted as n-SR simulated experiment and we consider the cases $n = 2, 3, 4, 5$. Our simulations consist of trials separated by one second. At the onset of each trial the stimulus is delivered to the network by injection of a strong 2-ms pulse of current into the neurons in S_i . The sequence of S_i 's are delivered to the network randomly. The coincident firing of neurons in this group typically induces a few spikes in the other neurons in the network. During a 20-ms time window beginning τ ms after the stimulation, we count the number of spikes fired by neurons in all R groups. We say the network has exhibited response R_i if R_i has fired the largest number of spikes among R groups. If in a trial after delivering the stimulus S_i to the network, R_i and another group R_j simultaneously fire the largest number of spikes, we do not accept R_i as a correct response and label the case as *no response*. One might think of R groups as projecting, for example, to different motor areas in the brain. To produce a noticeable movement, one group has to fire more spikes than the other groups [14]. After each stimulation, we monitor the response of the circuit. If the circuit has exhibited the desired response, then we deliver a reward in the form of the increase of extracellular dopamine with a random delay in the range 10–50 ms [see Eq. (7)]. This delay is included because in biological neural networks reward typically comes with some delay after reward-triggering actions. The network performance is evaluated as [55]

$$p(T) = p(T-1)(1-\lambda) + \lambda q, \quad (10)$$

where $p(T)$ is the performance in the T th trial and λ is a parameter that controls the size of fluctuations of $p(T)$. Although the stationary state amount of $p(T)$ is independent of λ , it preferably should be a small value. We set $\lambda = 0.002$; q is a binary variable which is equal to 1 if the network response is correct and is 0 otherwise.

We note that the intricate details of the model along with the need to obtain long-time dynamics of the system limit our computational abilities. We have therefore performed simulations for $100 < N < 400$. We however note that our general results and conclusions are independent of the system size and will therefore present results for $N = 100$ and/or $N = 200$. A FORTRAN code is available upon request.

III. RESULTS

First, we establish the occurrence of a synchronization transition in this cognitive system. The amount of synchronization in neural networks with static synapses can be controlled upon changing the average synaptic strength [44,54]. In the present study synaptic strength is an

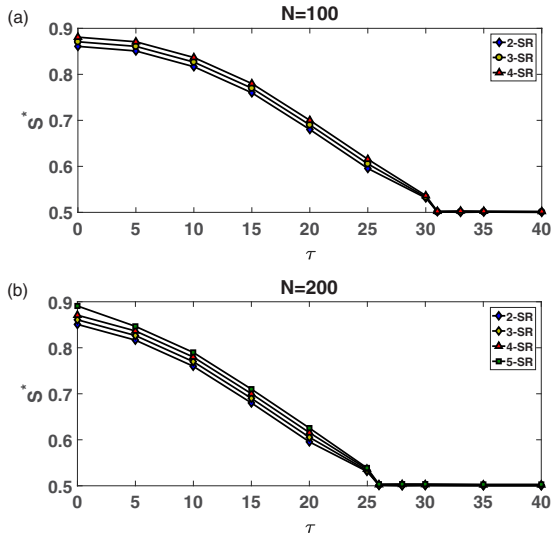


FIG. 2. Synchronization plots (S^* vs τ) for different n -SR simulated experiments for the system size (a) $N = 100$ and (b) $N = 200$. The dopamine-modulated STDP leads to the emergence of synchronization transition with a sharp, size-dependent transition point. $\tau_c(N = 100) = 31$ ms and $\tau_c(N = 200) = 26$ ms.

autonomous variable which is modified by the internal dynamics of the network through the RL process. Thus one cannot control the synaptic strength to tune the amount of synchronization. However, we have previously shown that, in a similar adaptive network of spiking Izhikevich neurons, synchronization is a self-organized emergent property which depends instead on the average time delay τ [39]. In particular, spike-timing dependent plasticity along with time delay provides an underlying mechanism where the causal effect (synapse) between strongly correlated neurons gets suppressed, while slightly correlated activity will be potentiated. Motivated by this insight, we monitor the amount of synchronization S^* for different values of τ in the cognitive system and find that increasing τ leads to a continuous phase transition from strong synchronization ($S^* \simeq 0.9$) to asynchronous neural oscillations ($S^* \simeq 0.5$). Synchronization diagrams of the circuits with $N = 100$ and $N = 200$ in different n -SR experiments are shown in Figs. 2(a) and 2(b), respectively. It is observed that the transition point τ_c depends on the system size and moves toward smaller τ values by increasing N . We note that $\tau_c = 31$ ms for $N = 100$ and $\tau_c = 26$ ms for $N = 200$ in all considered n -SR experiments. Increasing the system size further ($N = 400$) reduces the transition point by a lesser amount; however, computational limits ($N = 100, 200, 400$) do not allow us to obtain enough points to extrapolate to the thermodynamic limit.

Next, we show that reinforcement of specific firing patterns is possible by this cognitive system. Specially, reward-based reinforcement of the network produces an appropriate response to a stimulus. Figures 3(a)–3(d) show the performance of the system $p(T)$ vs trail number T for different τ 's in each n -SR simulated experiment for $N = 200$. Over time, $p(T)$ increases quickly as the system learns to respond correctly to each stimulus as it reaches a stationary state after nearly 2000 trials. The stationary value of $p(T)$ depends on τ . It is seen

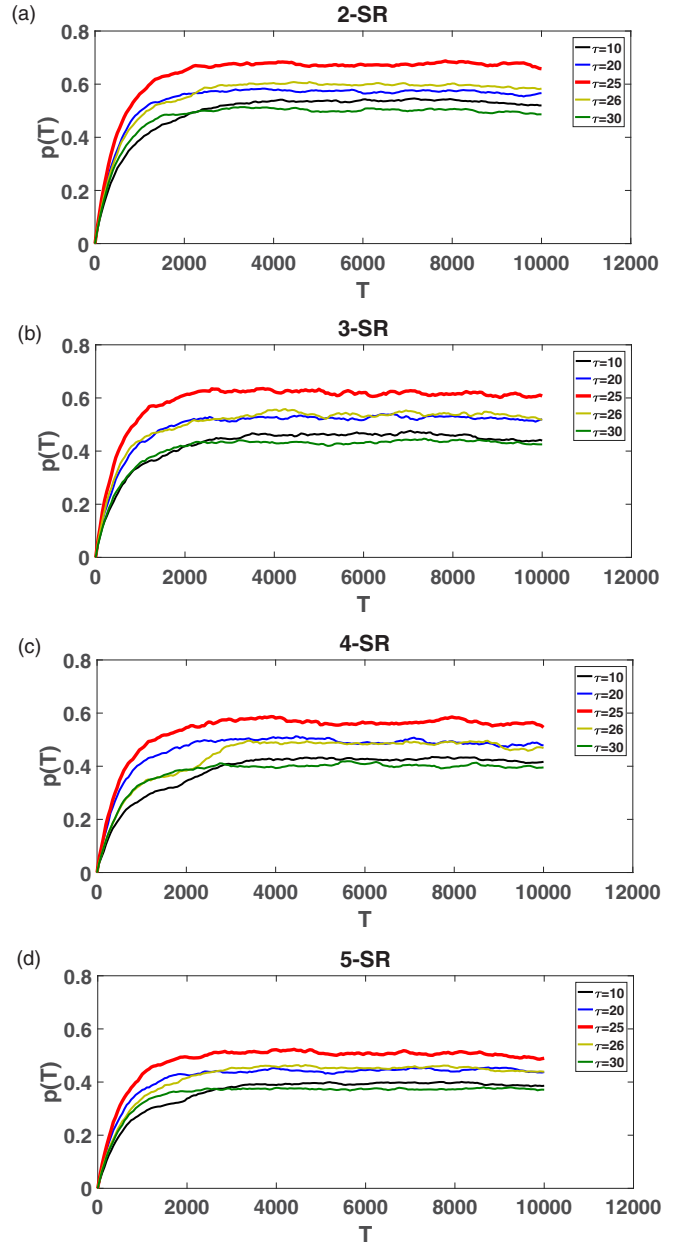


FIG. 3. Dynamics of network performance over time for the system size $N = 200$ and for different τ values in n -SR simulated experiments with $n = 2, \dots, 5$. Time is shown in units of T . To obtain each $p(T)$ vs T plot, we carried out an ensemble of 20 simulations with different stochasticity conditions including network realization. We coarse grained the raw data on T axis for each simulation and next averaged over the ensemble.

from Figs. 3 the $p(T)$ vs T plots for $\tau = \tau_m = 25$ ms (red curve) stand above the other curves for all n -SR simulated experiments. This is an important indication of the possible relation between synchronization and learning in the system. Clearly, the emergence of spontaneous synchronization is related to learning and its optimization, i.e., performance.

It is evident from Figs. 2 and 3 that variation of τ changes the state of synchronization in the system and subsequently influences the network's performance. Thus performance depends on synchronization. Now, one may ask “what amount of

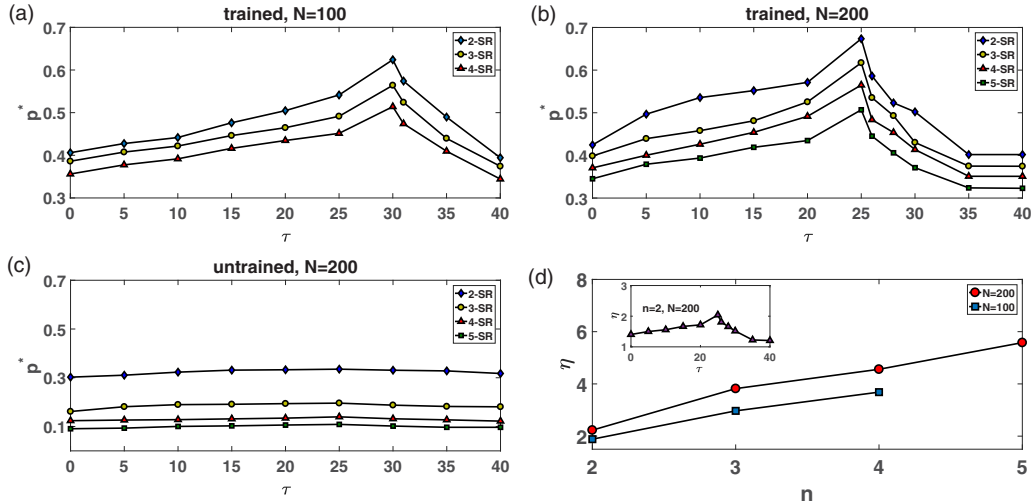


FIG. 4. p^* vs τ in n -SR simulated experiments for (a) trained system with $N = 100$ and $n = 2, 3, 4$, (b) trained system with $N = 200$ and $n = 2, 3, 4, 5$, (c) untrained system with $N = 200$ and $n = 2, 3, 4, 5$, and (d) η vs n plots that show the relative efficiency of the model improves with increasing n as well as system size N . (Inset: the relative efficiency of the model vs τ for the system with $n = 2$ and $N = 200$.)

synchronization optimizes the performance of the system for a given SR task?" To address this question, we define p^* as the average of $p(T)$ in the stationary state [$p^* = \langle p(T) \rangle_{T > 2000}$]. Figures 4(a) and 4(b) illustrate p^* as a function of τ for all n -SR simulated experiments in the network with $N = 100$ and $N = 200$, respectively. One observes a significant peak at the value of τ_m regardless of the value of n or the size of the system. Interestingly, the value at which performance is optimized is not exactly at the critical point, τ_c , where synchronization begins to emerge, but a slightly supercritical state where a small amount of order is present in the system $S^*(\tau_m) \approx 0.52$. For the record, for $N = 100$ $\tau_m = 30$, where $\tau_c = 31$, and for $N = 200$ network, $\tau_m = 25$, where $\tau_c = 26$, within our numerical resolution. Also, note that, for $N = 100$, $n \leq 4$ due to limited network size.

Clearly, if one turns off the mechanism for plasticity, no learning occurs despite the fact that τ can change and thus influence the amount of synchronization in the system. The performance of the system in this scenario is a random performance which tends to decrease as the learning task gets more complicated, i.e., increasing n . However, for a given n one can evaluate the performance of the system in order to compare how plasticity plays a role in the learning performance. This is shown in Fig. 4(c) for $N = 200$. In order to better characterize the relative performance of the system, we define $\eta(n) = p_m^*(n)/p_{un}^*(n)$, where $p_m^*(n) = p^*(\tau_m)$ and p_{un}^* is the maximum of the untrained, random performance depicted in Fig. 4(c) for each n task. In Fig. 4(d) we show the results for the relative performance as a function of n for the two system sizes we have studied. The results indicate that the relative performance increases with increasing n and for a given n with increasing system size. Therefore, optimization may lead to significant improvement in learning of complicated tasks in large networks. We also note that one can simply define $\eta(\tau) = p^*(\tau)/p_{un}^*(\tau)$, for various n -task performance [see inset of Fig. 4(d)]. Incidentally, we note that the relative performance due to plasticity can increase from $\eta(\tau = 40) \approx 1.3$ to $\eta(\tau_m) \approx 2.1$, i.e., a difference between a

30% improvement and a more than 100% improvement for a given task by taking advantage of criticality.

Many recent studies have indicated the functional advantages of operating near the critical point associated with second order, critical phase transition. Such studies both include computational models [56] as well as clinical studies [57]. Our results seem to indicate such functional advantage in a biologically relevant spiking neural network. However, our results also indicate a slightly supercritical state in a synchronization transition as the optimal point of reinforcement learning. A standard method to characterize the critical state is to look at neuronal avalanches [22–24,38,58,59]. We next propose to study neuronal avalanches in order to better correlate critical dynamics with reinforcement learning.

Neural avalanches are heterogeneous outbursts of activity interspersed by brief periods of quiescence [22–24]. To study neural avalanches, we record the raster plot in the stationary state of the system in each simulated experiment. Next, we divide it into temporal bins of length 5 ms and count the number of spikes in each bin to extract the time series of network activity M ; see Figs. 5(a) and 5(b). By monitoring the network activity M we can identify outbursts of spikes the number of which is associated with the size s and the lifetime with the duration d of avalanches. An avalanche begins when M exceeds a threshold M_{th} and ends when it turns back below that threshold, Fig. 5(b). Here, we set the threshold to be $M_{th} = \bar{M} - \sigma$, while \bar{M} and σ are the mean value and standard deviation of M , respectively. Our consideration shows that displacement of the threshold in the interval $[\bar{M} - \sigma, \bar{M}]$ does not influence the avalanche statistics.

We consider neural avalanches for different values of τ and different n -SR simulated experiments. For any given set of parameters the network is simulated for a considerably long time, producing nearly five million avalanches. Probability distribution of avalanche sizes and avalanche durations for 2-SR simulations of the network with $N = 200$ are illustrated in Figs. 5(c) and 5(d), respectively. $P(s)$ and $P(d)$ for other SR tasks are qualitatively the same (not shown). The dotted lines

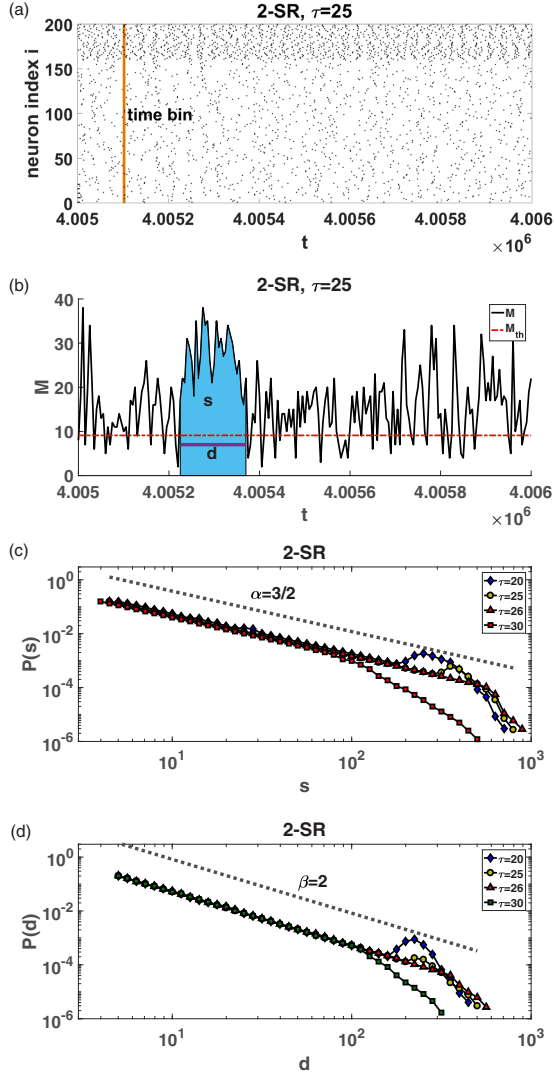


FIG. 5. (a) 1 s segment of the raster plot for the system that learns 2-SR task. The orange area illustrates a typical time bin used to extract M time series. (b) A typical snapshot of the network activity M and the threshold M_{th} . The blue area shows the size s of an avalanche and the length of purple line shows its duration d . (c),(d) Probability distribution function of avalanche size $P(s)$ and avalanche duration $P(d)$ for the trained networks performing 2-SR task. The network size is $N = 200$. Avalanches are measured only when the performance has reached a stationary state. The dashed lines are meant to help the eye.

in Figs. 5(c) and 5(d) depict power-law functions $P(s) \sim s^{-\alpha}$ and $P(d) \sim d^{-\beta}$ with $\alpha = 3/2$ and $\beta = 2$. These exponents are associated with the mean-field branching process which is thought to underlie the critical dynamics of such networks [60,61]. For both size and duration statistics [Figs. 5(c) and 5(d)], we show avalanche statistics for critical ($\tau = \tau_c = 26$) and subcritical ($\tau = 30$), as well as supercritical ($\tau = 20$) states. The critical state is associated with a power law limited by the finite size of the network at large events. The subcritical state is associated with a cutoff and thus decaying large events, while the supercritical state is associated with a significant “bump” at large events. We have also plotted the

avalanche statistics for the corresponding value of time delay which optimizes RL performance, i.e., $\tau = \tau_m$. As expected, it displays a very similar statistics to the critical state with a slight bump towards the large events. We therefore conclude that being near the critical point offers the cognitive system functional advantages which are not present when operation occurs away from criticality. The fact that such optimization occurs in a slightly supercritical state could be related to the occurrence of more frequent large avalanches (i.e., small bump) which may play a role in correlated activity at large distances which facilitate pathways between S and R neurons, while still keeping the usual advantages associated with the actual critical point. We finally note that similar changes in the profile of avalanche statistics due to task performance have been observed in various experimental studies [62–64]. However, a clear method to measure “the distance” to criticality is still lacking.

IV. CONCLUDING REMARKS

In this work we have proposed to study a biologically relevant dynamical system which learns to exhibit specific responses to stimuli via reinforcement learning. The Izhikevich spiking neuronal network is supplemented by time-delayed chemical synapses as well as dopamine-modulated, time-delayed, spike-timing dependent plasticity. We observe that the system can show various amounts of synchronization depending on the average time delay, with a continuous phase transition to an asynchronous state. Learning performance is then evaluated in the steady state and is shown to depend on the amount of synchronization in the system. Interestingly, it is found that, regardless of the task complication (n) or network size studied (N), optimized performance for learning occurs near the critical transition point. In order to better understand such behavior we studied avalanche statistics as a way to probe the critical behavior of the system. We observe subcritical and supercritical avalanche statistics away from the critical point which itself exhibits clear power-law behavior with experimentally relevant exponents. We observed that the optimal learning state corresponds to a near critical state with a slight bump in the tail of the statistics, indicating the significance of more probable large events which may help correlate the distinct areas of the network associated with stimulus and response. This type of correlated activity provides more frequent long-range pathways which facilitate efficient learning.

A few points are of note. First, the reinforcement learning model which we have used [14] is based on reward modulation of a purely Hebbian STDP rule with a reward signal that is always positive. Another possible approach to implement reinforcement learning through reward-modulated STDP is given by Florian [16], with Hebbian STDP when the reward is positive and anti-Hebbian STDP when the reward is negative. A neural network trained using a simplified version of the Florian algorithm has shown to be a competitive alternative to the optimized state-of-the-art models in the field of machine learning [65].

Second, in comparison to the typical models employed in physics, the model considered in the current study is a detailed model which is quite complicated, in the sense that it is high dimensional and possesses many parameters. However,

several parameters, including the parameters of the Izhikevich neurons (Table I) which are set to produce a regular spiking pattern of firing as well as the parameters of synaptic currents (Table I) and dopamine signaling [Eqs. (6) and (7)] that are set to be compatible with empirical findings, are fixed in all simulations. We carried out intensive computer simulations to examine the effect of other parameters on the system's behavior. We found that the above results are robust upon changing STDP parameters (Table I) and upon changing the average connectivity of the initial ER network. We assessed the statistical properties of the networks before and after learning and observed that the trained networks still remain as random (ER) networks with Poisson degree distribution functions. The average connectivity, p , however typically decreases as a result of pruning caused by learning. Additionally, we observed that the above results are preserved while we start with a complete network rather than an ER network.

Third, critical behavior and its associated benefits in neural networks do not necessarily come by operating exactly at the critical point. Various mechanisms exist which can extend the "critical regime." Griffith's phase is the most well-known such

mechanism which has structural origins [66,67]. There are also dynamical mechanisms such as refractoriness in neuronal oscillations which can extend the critical regime [68]. However, what we have observed is that a slight deviation from the standard critical scaling regime leads to a more efficient learning.

Lastly, further studies are needed to better understand the relevance of synchronization and criticality, with reinforcement learning. Specifically, with regards to the functional advantages associated with the critical dynamics of neuronal networks, and recent intensive studies in machine learning, one would expect that insights from criticality literature may find important ramifications in the field of machine learning. However, neuromorphic networks may provide an ideal system to experimentally study oscillations, learning, and avalanches in a controlled manner [69].

ACKNOWLEDGMENT

Support from Shiraz University research council is acknowledged.

-
- [1] C. Gros, *Complex and Adaptive Dynamical Systems*, 4th ed. (Springer-Verlag, Berlin, 2015).
 - [2] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* (Prentice-Hall, Englewood Cliffs, NJ, 2002).
 - [3] M. I. Jordan and T. M. Mitchell, Machine learning: Trends, perspectives and prospects, *Science* **349**, 255 (2015).
 - [4] M. Mohri, A. Rostamzadeh, and A. Talwalkar, *Foundations of Machine Learning* (MIT Press, Cambridge, MA, 2018).
 - [5] C. Gros, Cognitive computation with autonomously active neural networks: an emerging field, *Cog. Comput.* **1**, 77 (2009).
 - [6] P. Slijepcevic, Principles of information processing and natural learning in biological system, *J. Gen. Philos. Sci.* **52**, 227 (2021).
 - [7] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, UK, 2019).
 - [8] E. O. Neffci and B. B. Averbeck, Reinforcement learning in artificial and biological systems, *Nat. Mach. Int.* **1**, 133 (2019).
 - [9] G. G. Bi and M. Poo, Synaptic modification by correlated activity: Hebb's postulate revisited, *Annu. Rev. Neurosci.* **24**, 139 (2001).
 - [10] N. Caporale and Y. Dan, Spike-timing-dependent plasticity: a Hebbian learning rule, *Annu. Rev. Neurosci.* **31**, 25 (2008).
 - [11] H. Markram, W. Gerstner, and P. J. Sjöström, Spike-timing-dependent plasticity: a comprehensive overview, *Front. Syn. Neurosci.* **4**, 1 (2012).
 - [12] Z. Brzosko, S. B. Mierau, and O. Paulsen, Neuromodulation of spike-timing-dependent plasticity: past, present, and future, *Neuron* **103**, 563 (2019).
 - [13] C. W. Dickey, A. Sargsyan, J. R. Madsen, E. N. Eskandar, S. S. Cash, and E. Halgren, Travelling spindles create necessary conditions for spike-timing-dependent plasticity in humans, *Nat. Commun.* **12**, 1027 (2021).
 - [14] E. M. Izhikevich, Solving the distal reward problem through linkage of STDP and dopamine signaling, *Cereb. Cortex* **17**, 2443 (2007).
 - [15] R. Legenstein, D. Pecevski, and W. Maass, A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback, *PLoS Comput. Biol.* **4**, e1000180 (2008).
 - [16] R. V. Florian, Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity, *Neural Comput.* **19**, 1468 (2007).
 - [17] M. A. Fairies and M. I. Fairhall, Reinforcement learning with modulated spike timing dependent synaptic plasticity, *J. Neurophysiol.* **98**, 3648 (2007).
 - [18] E. Gireesh and D. Plenz, Neuronal avalanches organize as nested theta- and beta/gamma-oscillations during development of cortical layer 2/3, *Proc. Natl. Acad. Sci. USA* **105**, 7576 (2008).
 - [19] M. Rubinov, O. Sporns, J. P. Thivierge, and M. Breakspear, Neurobiologically realistic determinants of self-organized criticality in networks of spiking neurons, *PLoS Comput. Biol.* **7**, e1002038 (2013).
 - [20] C. Meisel and T. Gross, Adaptive self-organization in a realistic neural network model, *Phys. Rev. E* **80**, 061917 (2009).
 - [21] F. Y. K. Kossio, S. Goedeke, B. van Den Akker, B. Ibarz, and R. M. Memmesheimer, Growing Critical: Self-Organized Criticality in a Developing Neural System, *Phys. Rev. Lett.* **121**, 058301 (2018).
 - [22] J. M. Beggs and D. Plenz, Neuronal avalanches in neocortical circuits, *J. Neurosci.* **23**, 11167 (2003).
 - [23] J. M. Beggs and D. Plenz, Neuronal avalanches are diverse and precise activity patterns that are stable for many hours in cortical slice cultures, *J. Neurosci.* **24**, 5216 (2004).
 - [24] D. Plenz, *Criticality in Neural Systems* (Wiley VCH, New York, 2014).
 - [25] R. Legenstein and W. Maass, Edge of chaos and prediction of computational performance for neural circuit models, *Neural Networks* **20**, 323 (2007).

- [26] W. L. Shew and D. Plenz, The functional benefits of criticality in the cortex, *Neuroscientist* **19**, 88 (2013).
- [27] J. Boedecker, O. Obst, J. T. Lizier, N. M. Mayer, and M. Asada, Information processing in echo state networks at the edge of chaos, *Theory Biosci.* **131**, 205 (2012).
- [28] O. Kinouchi and M. Copelli, Optimal dynamical range of excitable networks at criticality, *Nat. Phys.* **2**, 348 (2006).
- [29] L. de Arcangelis and H. J. Herrmann, Learning as a phenomenon occurring in a critical state, *Proc. Natl. Acad. Sci. USA* **107**, 9 (2010).
- [30] D. L. Berger, L. de Arcangelis, and H. J. Herrmann, Spatial features of synaptic adaptation affecting learning performance, *Sci. Rep.* **7**, 11016 (2017).
- [31] L. Michiels van Kessenich, D. Berger, L. de Arcangelis, and H. J. Herrmann, Pattern recognition with neural avalanche dynamics, *Phys. Rev. E* **99**, 010302(R) (2019).
- [32] B. Del Papa, V. Priesemann, and J. Triesch, Criticality meets learning: criticality signatures in a self-organizing recurrent neural network, *PLoS One* **12**, e0178683 (2017).
- [33] B. Cramer, D. Stöckel, M. Krefl, M. Wibral, J. Schemmel, K. Meier, and V. Priesemann, Control of criticality and computation in spiking neuromorphic networks with plasticity, *Nat. Commun.* **11**, 2853 (2020).
- [34] A. Levina, J. M. Herrmann, and T. Geisel, Dynamical synapses causing self-organized criticality in neural networks, *Nat. Phys.* **3**, 857 (2007).
- [35] L. de Arcangelis, C. Perrone-Capano, and H. J. Herrmann, Self-Organized Criticality Model for Brain Plasticity, *Phys. Rev. Lett.* **96**, 028107 (2006).
- [36] D. Millman, S. Mihalas, A. Kirkwood, and E. Niebur, Self-organized criticality occurs in non-conservative neuronal networks during up states, *Nat. Phys.* **6**, 801 (2010).
- [37] N. Stepp, D. Plenz, and N. Srinivasa, Synaptic plasticity enables adaptive self-tuning critical networks, *PLoS Comput. Biol.* **11**, e1004043 (2015).
- [38] A. J. Fontenele, N. A. P. de Vasconcelos, T. Feliciano, L. A. A. Aguiar, C. Soares-Cunha, B. Coimbra, L. Dalla Porta, S. Ribeiro, A. J. Rodrigues, N. Sousa, P. V. Carelli, and M. Copelli, Criticality between Cortical States, *Phys. Rev. Lett.* **122**, 208101 (2019).
- [39] M. Khoshkhou and A. Montakhab, Spike-timing-dependent plasticity with axonal delay tunes networks of Izhikevich neurons to the edge of synchronization transition with scale-free avalanches, *Front. Syst. Neurosci.* **13**, 73 (2019).
- [40] S. di Santo, P. Villegas, R. Burioni, and M. A. Munoz, Landau-Ginzburg theory of cortex dynamics: Scale-free avalanches emerge at the edge of synchronization, *Proc. Natl. Acad. Sci. USA* **115**, 1356 (2018).
- [41] S. S. Poil, R. Hardstone, H. D. Mansvelder, and K. Linkenkaer-Hansen, Critical-state dynamics of avalanches and oscillations jointly emerge from balanced excitation/inhibition in neuronal networks, *J. Neurosci.* **32**, 9817 (2012).
- [42] L. Dalla Porta and M. Copelli, Modeling neuronal avalanches and long-range temporal correlations at the emergence of collective oscillations: Continuously varying exponents mimic M/EEG results, *PLoS Comput. Biol.* **15**, e1006924 (2019).
- [43] E. M. Izhikevich, Simple model of spiking neurons, *IEEE Trans. Neural Netw.* **14**, 1569 (2003).
- [44] M. Khoshkhou and A. Montakhab, Beta-rhythm oscillations and synchronization transition in network models of Izhikevich neurons: effect of topology and synaptic type, *Front. Comput. Neurosci.* **12**, 59 (2018).
- [45] R. Chamberlain, K. M. Visscher, C. C. Le Danter, and A. R. Seitz, Alpha-band EEG activity in perceptual learning, *Journal of Vision* **15**, 112 (2015).
- [46] T. A. de Graaf, J. Gross, G. Paterson, T. Rusch, A. T. Sack, and G. Thut, Alpha-band rhythms in visual task performance: Phase-locking by rhythmic sensory stimulation, *PLoS One* **8**, e60035 (2013).
- [47] R. H. Grabner, A. Fink, A. Stipacck, C. Neuper, and A. C. Neubauer, Intelligence and working memory systems: evidence of neural efficiency in alpha band ERD, *Cog. Bra. Res.* **20**, 213 (2004).
- [48] J. J. Foxe and A. C. Snyder, The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention, *Front. Physiol.* **2**, 154 (2011).
- [49] A. Roth and M. van Rossum, in *Computational Modeling Methods for Neuroscientists*, edited by E. De Schutter (MIT Press, Cambridge, MA, 2009).
- [50] P. Erdos and A. Renyi, On random graphs, *Publ. Math.* **6**, 290 (1959).
- [51] B. Babadi and L. F. Abbott, Intrinsic stability of temporally shifted spike-timing dependent plasticity, *PLoS Comput. Biol.* **6**, e1000961 (2010).
- [52] M. M. Asl, A. Valizadeh, and P. A. Tass, Propagation delays determine neuronal activity and synaptic connectivity patterns emerging in plastic neuronal networks, *Chaos* **28**, 106308 (2018).
- [53] A. Pikovsky, M. Rosenblum, and G. Osipov, Phase synchronization of chaotic oscillators by external driving, *Physica D* **104**, 219 (1997).
- [54] M. Khoshkhou and A. Montakhab, Explosive, continuous and frustrated synchronization transition in spiking Hodgkin-Huxley neuronal networks: the role of topology and synaptic interaction, *Physica D* **405**, 132399 (2020).
- [55] S. Skorheim, P. Lonjers, and M. Bazhenov, A spiking network model of decision making employing rewarded STDP, *PLoS One* **9**, e90821 (2014).
- [56] M. A. Muñoz, Colloquium: criticality and dynamical scaling in living systems, *Rev. Mod. Phys.* **90**, 031001 (2018).
- [57] V. Zimmern, Why brain criticality is clinically relevant: A scoping review, *Front. Neural Circuits* **14**, 54 (2020).
- [58] N. Friedman, S. Ito, B. A. W. Brinkman, M. Shimono, R. E. LeeDeVile, K. A. Dahmen, J. M. Beggs, and T. C. Butler, Universal Critical Dynamics in High Resolution Neuronal Avalanche Data, *Phys. Rev. Lett.* **108**, 208102 (2012).
- [59] A. Haimovici, E. Taglizucchi, P. Balenzuela, and D. R. Chialvo, Brain Organization into Resting State Networks Emerges at Criticality on a Model of the Human Connectome, *Phys. Rev. Lett.* **110**, 178101 (2013).
- [60] S. A. Moosavi and A. Montakhab, Mean-field behavior as a result of noisy local dynamics in self-organized criticality: neuroscience implications, *Phys. Rev. E* **89**, 052139 (2014).
- [61] S. A. Moosavi and A. Montakhab, Structural versus dynamical origins of mean-field behavior in a self-organized critical model of neuronal avalanches, *Phys. Rev. E* **92**, 052804 (2015).

- [62] S. Yu, T. L. Ribeiro, C. Meisel, S. Chou, A. Mitz, R. Saunders, and D. Plenz, Maintained avalanche dynamics during task-induced changes of neuronal activity in nonhuman primates, *eLife* **6**, e27119 (2017).
- [63] W. L. Shew, W. P. Clawson, J. Pobst, Y. Karimipناه, N. C. Wright, and B. Wessel, Adaptation to sensory input tunes visual cortex to criticality, *Nat. Phys.* **11**, 659 (2015).
- [64] O. Arviv, A. Goldstein, and O. Shriki, Near-critical dynamics in stimulus-evoked activity of the human brain and its relation to spontaneous resting-state activity, *J. Neurosci.* **35**, 13927 (2015).
- [65] S. F. Chevtchenko and T. B. Ludermir, Combining STDP and binary networks for reinforcement learning from images and sparse rewards, *Neural Networks* **144**, 496 (2021).
- [66] M. A. Muñoz, R. Juhasz, C. Castellano, and G. Odor, Griffiths Phases on Complex Networks, *Phys. Rev. Lett.* **105**, 128701 (2010).
- [67] V. Buenda, P. Villegas, R. Burioni, and M. A. Munoz, The broad edge of synchronisation: Griffiths effects and collective phenomena in brain networks, [arXiv:2109.11783](https://arxiv.org/abs/2109.11783).
- [68] S. A. Moosavi, A. Montakhab, and A. Valizadeh, Refractory period in network models of excitable nodes: self-sustaining stable dynamics, extended scaling region and oscillatory behavior, *Sci. Rep.* **7**, 7107 (2017).
- [69] J. Hochstetter, R. Zhu, A. Loeffler, A. Diaz-Alvarez, T. Nakayama, and Z. Kuncic, Avalanches and edge-of-chaos learning in neuromorphic nanowire networks, *Nat. Commun.* **12**, 4008 (2021).