Universal scaling laws of absorbing phase transitions in artificial deep neural networks

Keiichi Tamai [1,*] Tsuyoshi Okubo [1] Truong Vinh Truong Duy [2] Naotake Natori,[2] and Synge Todo [3,1,4]

[1]*Institute for Physics of Intelligence, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

[2]*AI Laboratory, Aisin Corporation, 1-18-13 Sotokanda, Chiyoda-ku, Tokyo 101-0021, Japan*

[3]*Department of Physics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

[4]*Institute for Solid State Physics, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8581, Japan*

(Received 30 October 2024; accepted 10 June 2025; published 18 July 2025)

We demonstrate that conventional artificial deep neural networks operating near the phase boundary of the signal propagation dynamics—also known as the edge of chaos—exhibit universal scaling laws of absorbing phase transitions in nonequilibrium statistical mechanics. We exploit the fully deterministic nature of the propagation dynamics to elucidate an analogy between a signal collapse in the neural networks and an absorbing state (a state that the system can enter but cannot escape from). Our numerical results indicate that the multilayer perceptrons and the convolutional neural networks belong to the mean-field and the directed percolation universality classes, respectively. Also, the finite-size scaling is successfully applied, suggesting a potential connection to the depth-width trade-off in deep learning. Furthermore, our analysis of the training dynamics under the gradient descent reveals that hyperparameter tuning to the phase boundary is necessary but insufficient for achieving optimal generalization in deep networks. Remarkably, nonuniversal metric factors associated with the scaling laws are shown to play a significant role in concretizing the above observations. These findings highlight the usefulness of the notion of criticality for analyzing the behavior of artificial deep neural networks and offer new insights toward a unified understanding of the essential relationship between criticality and intelligence.

DOI: 10.1103/jp61-6sp2

I. INTRODUCTION

Critical phenomena at second-order phase transitions have long been hypothesized to be the key to the extraordinary computational power of living systems [1–3]. The idea behind the hypothesis is that information cannot propagate through ordered states of matter, and it rapidly decays to random noise in the disordered states due to the overly enhanced capability of the medium to convey disturbance [2]. Notably, the notion of *phases of matter* lies at the core of the hypothesis; this viewpoint focuses on collective aspects of the systems, complementary to more traditional reductionist ones [4,5]. Despite the experimental and theoretical challenges stemming from the many-body nature of the problem, pursuing the *computation at the criticality* hypothesis has proven to be a fruitful research direction [3,6]. In particular, brain dynamics has been intensively discussed [7,8] in the context of absorbing phase transitions [9,10] due to the theoretical relation [11] to self-organized criticality [12,13], not to mention the straightforward correspondence between death and an absorbing state.

The rapid progress in applying deep learning techniques [14–16] motivates us to ask whether artificial deep neural networks also utilize criticality for their performance. Recent theoretical studies on infinitely wide networks suggest this is the case. Under a specific setup (see also Sec. II), the signal propagation dynamics in untrained deep neural networks can be classified into two phases: the ordered phase and the chaotic phase, depending on the hyperparameters used for initialization [17,18]; see also Fig. 1. The network with sufficiently many hidden layers returns almost the same outputs for any inputs in the ordered phase, whereas decorrelated outputs in the chaotic phase. In either case, the network cannot remember the degree of similarity between different inputs, which limits the performance as a learning agent. As such, the network at the ordered phase suffers from vanishing gradient and untrainability, while at the chaotic phase from exploding gradient and ungeneralizability [18,19]. Consequently, the phase boundary, often called *the edge of chaos*, attracts considerable interest in deep learning research, even though some studies indicate that initialization at the edge alone does not necessarily lead to good generalization [20,21]. Remarkably, the characteristic depth of the network dynamics is suggested to diverge at the edge of chaos [18], highly reminiscent of critical phenomena at second-order phase transitions. Subsequent works extend similar results to various activation functions [22,23] and network architectures [24,25].

From a statistical mechanics viewpoint, universal scaling laws, if any, are relevant for characterizing the critical phenomena at the edge of chaos in deep neural networks. Even though the mean-field theory in a suitable limit [17,24] and its perturbative expansion [26,27] are available in simple cases, the scaling laws may provide complementary, flexible, and powerful insight into the network dynamics; thanks

*Contact author: tamai@phys.s.u-tokyo.ac.jp

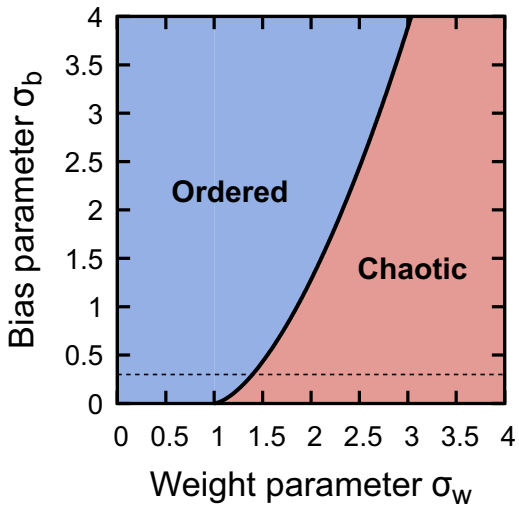

FIG. 1. The phase diagram of the signal propagation in the multilayer perceptions (1). The solid curve indicates the phase boundary in the case of tanh activation function. The dashed line indicates $\sigma_b = 0.3$, where the maximum Lyapunov exponent λ_1 [see Eq. (3)] and the order parameter $\rho^{(l)}$ [see Eq. (21)] are studied in Fig. 2 and Fig. 3(a), respectively.

to the universality of the critical phenomena, intuitive phenomenological considerations may result in at least partially quantitative predictions. Also, theoretical tools such as finite-size [28] or short-time [29] scaling shed further light on the dynamics beyond the limiting cases. Besides the benefits for our understanding of deep neural networks, embedding them better in statistical mechanics may provide clues for studying how living systems perform intellectual tasks.

Below, we demonstrate the connection between deep neural network dynamics at the edge of chaos and absorbing phase transitions [9,10]. After some preliminaries (Sec. II), we establish a correspondence between the ordered state of deep neural networks and an absorbing state by studying the linear stability of the former (Sec. III). Next, in the case of the multilayer perceptrons, we thoroughly investigate the scaling properties of the phase transition in the thermodynamic limit (Secs. IV A and IV B). Remarkably, the scaling properties of the neural tangent kernel (NTK) [30] provide a novel insight into the *curse of depth* reported in the literature [21]. The investigation is then extended to finite width and different architectures (Sec. IV C), although we content ourselves only with front propagation dynamics in these cases. In particular, we provide numerical evidence of the directed percolation universality in convolutional neural networks. We conclude the paper with a brief discussion on possible directions for future work (Sec. V).

II. PRELIMINARIES

To illustrate our view with a simple setup, we exclusively consider the multilayer perceptrons with the NTK parametrization [30] until Sec. IV B. Formally, the recurrence relation for the preactivation $z^{(l)}$ at the lth hidden layer ($l = 1, 2, \ldots, L$, where L is the depth of the network) and the output y, assumed to be of a single element for simplicity, are written as follows:

$$z_i^{(l+1)} = \begin{cases} \dfrac{\sigma_w}{\sqrt{n_{\text{in}}}} \displaystyle\sum_j W_{ij}^{(l+1)} x_j + \sigma_b b_i^{(l+1)} & l = 0; \\ \dfrac{\sigma_w}{\sqrt{n}} \displaystyle\sum_j W_{ij}^{(l+1)} h\left(z_j^{(l)}\right) + \sigma_b b_i^{(l+1)} & 1 \leqslant l < L, \end{cases} \tag{1}$$

$$y = \frac{\sigma_w}{\sqrt{n}} \sum_j W_{1j}^{(L+1)} h\left(z_j^{(L)}\right) + \sigma_b b^{(L+1)}. \tag{2}$$

Here $\boldsymbol{x}$ is the input with n_{in} elements; $W^{(l)}$ and $\boldsymbol{b}^{(l)}$ the weight matrix and the bias vector at the lth hidden layer, respectively; σ_w, σ_b the associated hyperparameters; and n the width of the hidden layers. The meanings of the symbols associated with the neural networks are also summarized in Table I for convenience. Every element of the weight and the bias is initialized according to the standard normal distribution $\mathcal{N}(0, 1)$. This parametrization differs slightly from the one commonly used in practice [31], but the difference is not essential for our study [32]. The activation function h is assumed to be in the $K^* = 0$ universality class in the sense of Roberts *et al.* [23], a few examples being tanh, erf, and sin, but not the rectifier linear unit (ReLU). Note, however, that similar observations can be made also for ReLU-like activation functions; see Appendix B.

Since the forward dynamics (1) is fully deterministic after initialization, a pair of signals may end up with a collapse: Once the two signals $z_1^{(l)}, z_2^{(l)}$ become identical at one hidden layer, they never deviate from each other again at deeper hidden layers. Interpreting the network depth as a temporal degree of freedom, the ordered state $z_1^{(l)} = z_2^{(l)}$ can be regarded as an absorbing state of the dynamics. Even though the exact order rarely occurs in practice because it requires an accidentally degenerate weight matrix, the difference between two signals can decay exponentially in some cases, especially when the network is narrow. Meanwhile, with sufficient width and a suitable choice of the hyperparameters (σ_w, σ_b), one can observe the opposite: magnification of the difference, even if initially tiny. In this case, the difference no longer converges

TABLE I. Glossary of the symbols associated with the neural networks. If accompanied by the superscript "(l)," then the symbol indicates the corresponding quantity at the lth hidden layer.

Symbol	Meaning
$\boldsymbol{x}$	Input vectors.
W	Weight matrices.
$\boldsymbol{b}$	Bias vectors.
σ_w, σ_b	Hyperparameters associated with the network, corresponding to standard deviation of the weight/bias.
n_{in}	The number of the elements of the input.
n	Width of the network.
L	Depth of the network.
z	Preactivation: the signal vector just before applying the activation function.
h	Activation function.
q	Preactivation variance in the mean-field theory.
C	Preactivation covariance in the mean-field theory.

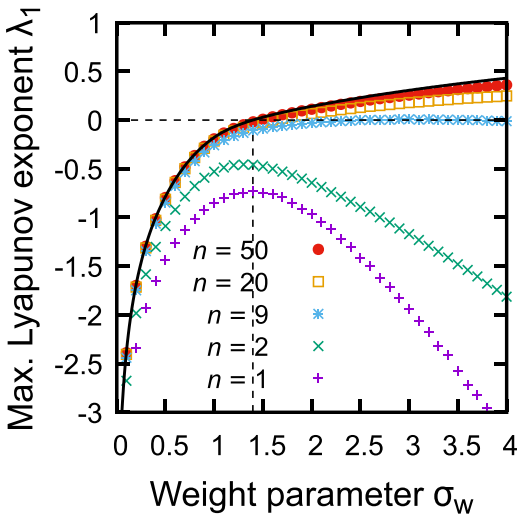

FIG. 2. The maximum Lyapunov exponent λ_1 [see Eq. (3)] as a function of the weight parameter σ_w for various widths n, estimated from the convergence of the finite counterpart of Eq. (3) at $l = 10^5$. The bias parameter σ_b was fixed to be 0.3, and tanh activation function was used. The vertical line indicates the position of the critical point ($\sigma_{w;c} \sim 1.39558$). The black curve is the result expected in the limit of $n \to \infty$, namely $\lambda_C/2$ [see Eq. (17)].

to a unique $l \to \infty$ limit but instead fluctuates randomly. Thus, the networks seem to exhibit a phase transition between a unique absorbing (ordered) phase and a fluctuating active (chaotic) phase. As we will see shortly, the transition can be further formalized by studying the linear stability of the exactly ordered state.

III. ABSORBING PROPERTY OF THE ORDERED STATE

To address the issue of the linear stability of the ordered state, we study (with a slight abuse of language) the maximum Lyapunov exponent for the front propagation dynamics (1):

$$\lambda_1 := \lim_{l\to\infty} \frac{1}{l} \log \frac{\|J^{(l+1)}(z^{(l)}) \cdots J^{(2)}(z^{(1)})\boldsymbol{u}_0\|_2}{\|\boldsymbol{u}_0\|_2}, \tag{3}$$

where $\boldsymbol{u}_0 \in \mathbb{R}^n$ is an arbitrary nonzero vector and $J^{(l)}$ is the layerwise input-output Jacobian

$$J^{(l)}(z) = \frac{\sigma_w}{\sqrt{n}} \begin{pmatrix} J_{11}^{(l)}(z) & \cdots & J_{1n}^{(l)}(z) \\ \vdots & \ddots & \vdots \\ J_{n1}^{(l)}(z) & \cdots & J_{nn}^{(l)}(z) \end{pmatrix} \tag{4}$$

with [33]

$$J_{ij}^{(l)}(z) := W_{ij}^{(l)} h'(z_j). \tag{5}$$

By doing so, we can directly see how the notion of the order-to-chaos transition emerges as a many-body effect in the neural networks. For instance, λ_1 as a function of σ_w for tanh activation function is shown in Fig. 2. In this case, λ_1 is negative in the entire domain for small width n, which suggests that the ordered state is always stable against infinitesimal discrepancy. However, λ_1 increases as n becomes larger, and eventually, λ_1 changes its sign at some σ_w for large n, indicating loss of linear stability. Naturally, the position of the onset of the linear instability is very close to that of the critical point predicted from the mean-field theory [17] when n is large and is expected to coincide with the limit of $n \to \infty$. One can also empirically see that the loss of the stability of the ordered state at some σ_w for sufficiently large n is a robust feature of the neural networks under the current consideration, although how λ_1 as a function of σ_w behaves for small n is somewhat more sensitive to the choice of h.

Thus, the maximum Lyapunov exponent λ_1 successfully captures the well-defined transition from the ordered phase to the chaotic phase, even for finite networks. In the ordered phase, once a pair of preactivations (z_1, z_2) reach reasonably close to the ordered state, they are hard to escape from it. Meanwhile, in the chaotic phase, a pair of preactivations are allowed to get away from the vicinity of the ordered state, although the ordered state itself is still absorbing. This scenario, a transition from a nonfluctuating absorbing phase to a fluctuating active phase, is highly reminiscent of an absorbing phase transition in statistical mechanics.

IV. UNIVERSAL SCALING AROUND THE ORDER-TO-CHAOS TRANSITION

Having seen that the order-to-chaos transition is at least conceptually analogous to absorbing phase transitions, the next step is to seek a deeper connection between these two by further quantitative characterization.

A. Mean-field theory of signal propagation

The phase transition between the ordered and the chaotic phases can be quantitatively studied in the limit of wide networks. In this limit, the neural network becomes equivalent to a Gaussian process [34–37]. That is, the preactivations $(z_i^{(l)}(\boldsymbol{x}_1), z_i^{(l)}(\boldsymbol{x}_2))$ on a same neuron for different inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$ follows a multivariate normal distribution

$$\left(z_i^{(l)}(\boldsymbol{x}_1), z_i^{(l)}(\boldsymbol{x}_2)\right) \sim \mathcal{N}(\boldsymbol{0}, \Sigma) \text{ with } \Sigma := \begin{pmatrix} q_1^{(l)} & C^{(l)} \\ C^{(l)} & q_2^{(l)} \end{pmatrix}, \tag{6}$$

where $q_i^{(l)}$ $(i = 1, 2)$ and $C^{(l)}$ for each hidden layer can be recursively described by the mean-field theory [17]:

$$q_i^{(l+1)} = \sigma_w^2 \int \mathcal{D}z\, h^2\left(\sqrt{q_i^{(l)}} z\right) + \sigma_b^2; \tag{7}$$

$$C^{(l+1)} = \sigma_w^2 \int \mathcal{D}z_1 \int \mathcal{D}z_2\, h\left(u_1^{(l)}\right) h\left(u_2^{(l)}\right) + \sigma_b^2; \tag{8}$$

$$\begin{aligned} u_1^{(l)} &:= \sqrt{q_1^{(l)}} z_1; \\ u_2^{(l)} &:= \sqrt{q_2^{(l)}} \left(c^{(l)} z_1 + \sqrt{1 - (c^{(l)})^2} z_2\right), \end{aligned} \tag{9}$$

with the initial conditions

$$q_i^{(1)} = \sigma_w^2 \frac{\|\boldsymbol{x}_i\|_2^2}{n_{\text{in}}} + \sigma_b^2, \quad C^{(1)} = \sigma_w^2 \frac{\boldsymbol{x}_1 \cdot \boldsymbol{x}_2}{n_{\text{in}}} + \sigma_b^2. \tag{10}$$

Here,

$$c^{(l)} := \frac{C^{(l)}}{\sqrt{q_1^{(l)} q_2^{(l)}}} \tag{11}$$

is the Pearson correlation coefficient and

$$\int \mathcal{D}z := \int_{-\infty}^{\infty} \mathrm{d}z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \tag{12}$$

Let us recall some basic results of this theory. One can see that $q^{(l)}$ rapidly converges to a fixed point $q^* := \lim_{l\to\infty} q^{(l)}$ as the depth l tends to infinity [17,18], generally without a sign of a phase transition (unless $\sigma_b = 0$, where q^* vanishes at the ordered phase). By substituting $q_1^{(l)} = q_2^{(l)} = q^*$ to Eqs. (8) and (9) and observing Eq. (11), we obtain an approximate closed-form description for $c^{(l)}$ also known as the iterative $\mathcal{C}$-map, valid for large l:

$$c^{(l+1)} = \frac{1}{q^*}\left[\sigma_w^2 \int \mathcal{D}z_1 \int \mathcal{D}z_2\, h\big(u_1^{*(l)}\big)h\big(u_2^{*(l)}\big) + \sigma_b^2\right], \tag{13}$$

where $u_1^{*(l)} := \sqrt{q^*}z_1$, $u_2^{*(l)} := \sqrt{q^*}(c^{(l)}z_1 + \sqrt{1-(c^{(l)})^2}z_2)$. Linear stability of the trivial fixed point $c^{(l)} = c^{(l+1)} = 1$ of Eq. (13) determines the phase for a given pair of hyperparameters (σ_w, σ_b), as depicted in Fig. 1: stable at the ordered phase, whereas unstable at the chaotic phase. It can be shown that, for a fixed σ_b, the discrepancy from the fixed point c^* asymptotically decays exponentially with a suitable correlation depth $\xi_\parallel$ [18]:

$$\lim_{l\to\infty} \frac{\log|c^{(l)} - c^*|}{l} = -\frac{1}{\xi_\parallel} \tag{14}$$

with $\xi_\parallel$ given by the following:

$$e^{-\frac{1}{\xi_\parallel}} = \begin{cases} \sigma_w^2 \int \mathcal{D}z\, h'^2(\sqrt{q^*}z) & \sigma_w < \sigma_{w;c}; \\ \sigma_w^2 \int \mathcal{D}z_1 \int \mathcal{D}z_2\, h'(u_1^*)h'(u_2^*) & \sigma_w > \sigma_{w;c}, \end{cases} \tag{15}$$

where $\sigma_{w;c}$ is the critical point for the specified σ_b, satisfying

$$\sigma_{w;c}^2 \int \mathcal{D}z\, h'^2(\sqrt{q^*}z) = 1. \tag{16}$$

Note also that half of the maximum Lyapunov exponent for the trivial fixed point of the iterative $\mathcal{C}$-map

$$\lambda_{\mathcal{C}} := \log\left(\sigma_w^2 \int \mathcal{D}z\, h'^2(\sqrt{q^*}z)\right) \tag{17}$$

equals the maximum Lyapunov exponent λ_1 of the ordered state [see Eq. (3)] in the limit of infinite width n. Encouragingly, the behavior of λ_1 for $n = 50$ as a function of σ_w is already close to the $n \to \infty$ limit [Fig. 1(b)]; this suggests that the infinitely wide neural network may serve as a good starting point for understanding the behavior of the neural networks of practical width.

B. Scaling results for the infinitely wide networks

One of the most common strategies for studying systems with absorbing phase transition is to examine universal scaling laws [9,10]. For instance, the time evolution of the order parameter $\rho(t)$ in the thermodynamic limit admits the following scaling ansatz:

$$\rho(t;\tau) \sim (\kappa t)^{-\beta/\nu_\parallel} f((\kappa t)^{1/\nu_\parallel}\zeta\tau), \tag{18}$$

where τ denotes the discrepancy from the critical point and $\beta, \nu_\parallel$ are the critical exponents associated with the onset of order parameter ρ and the correlation time $\xi_\parallel$ of the steady state, respectively, that is,

$$\rho(t\to\infty) \sim (\zeta\tau)^\beta, \quad \xi_\parallel \sim |\gamma\tau|^{-\nu_\parallel} \quad \text{as} \quad \tau \to 0; \tag{19}$$

$$\gamma := \zeta\kappa^{1/\nu_\parallel}. \tag{20}$$

Remarkably, the critical exponents and the scaling function f are the same for all systems in a given universality class, while the specific details are summarized in the nonuniversal metric factors κ, ζ, γ [38], two of which are independent.

Let us investigate the universal scaling laws in the signal propagation dynamics. In the present context, we define the order parameter ρ to be the Pearson correlation coefficient between preactivations for different inputs, which is then subtracted from unity so that ρ vanishes in the ordered phase. In particular, when the network is infinitely wide, the order parameter $\rho^{(l)}$ at each hidden layer is directly related to $c^{(l)}$ [see Eq. (11)] in the mean-field theory:

$$\rho^{(l)} := 1 - c^{(l)}. \tag{21}$$

We can show that the multilayer perceptrons exhibit the universal scaling laws identical to those of the mean-field theory for absorbing phase transitions [9,10]. Specifically, $\rho^{(l)}$ and the correlation depth $\xi_\parallel$ [see Eq. (14)] exhibits the power-law scaling (19) with

$$\beta = 1, \quad \nu_\parallel = 1. \tag{22}$$

This can be shown by considering an infinitesimally small deviation from the critical point and expanding the mean-field theory (7) and (8) to track the change of the position of the fixed point and of the correlation depth (15) up to the lowest relevant order of the deviation; see Appendix A for details. An interesting corollary is that the nonuniversal metric factors κ, γ in the sense of Eq. (18) can be evaluated theoretically in the present case. For instance, if we choose to fix the bias parameter σ_b and vary the weight parameter σ_w (the discrepancy τ from the critical point is defined to be $\sigma_w - \sigma_{w;c}$), then we find

$$\gamma_{\leftrightarrow} = \frac{2}{\sigma_{w;c}}\left(1 - \frac{(q_c^* - \sigma_b^2)\int \mathcal{D}z\, zh'(\sqrt{q_c^*}z)h''(\sqrt{q_c^*}z)}{\sqrt{q_c^*}\int \mathcal{D}z\, h(\sqrt{q_c^*}z)h''(\sqrt{q_c^*}z)}\right), \tag{23}$$

$$\kappa = \frac{q_c^* \int \mathcal{D}z\, h''^2(\sqrt{q_c^*}z)}{2\int \mathcal{D}z\, h'^2(\sqrt{q_c^*}z)}, \tag{24}$$

where q_c^* is the fixed point of Eq. (7) at the critical point and the arrow symbol $\leftrightarrow$ indicates the direction in which we cross the boundary in the phase diagram (Fig. 1). We empirically validate the results in Fig. 3. In particular, we see that the order parameter dynamics $\rho^{(l)}$ for various activation functions collapse into a single universal curve, as predicted by the scaling ansatz (18), except when l is small just as expected. These observations further support the view that the network exhibits an absorbing phase transition into the ordered state. Similar results (albeit with different metric factor $\gamma_\updownarrow$) can be obtained

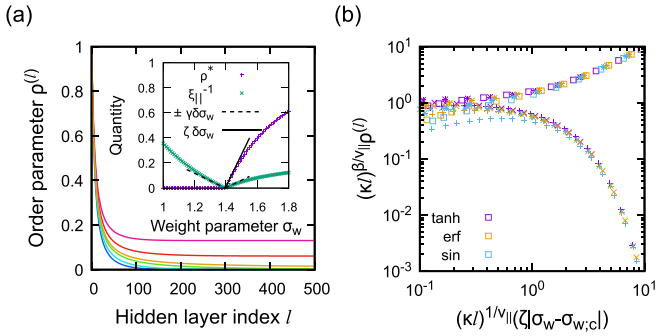

FIG. 3. The universal scaling laws in the order parameter $\rho^{(l)}$ [see Eqs. (11) and (21)] for the signal propagation dynamics in the infinitely wide multilayer perceptrons (1). (a) The main panel shows $\rho^{(l)}$ as a function of l for various σ_w [from 1.35 (blue; the ordered phase) to 1.45 (magenta; the chaotic phase)], calculated from the mean-field theory (7) and (8). The inset shows the stationary value ρ^* of the order parameter and the reciprocal correlation depth $\xi_\parallel^{-1}$ [see Eqs. (14) and (15)] as a function of σ_w. The dashed line and the solid line are guides to the eye for linear onset with a slope of $\gamma_\leftrightarrow$ and $\zeta_\leftrightarrow := \gamma_\leftrightarrow/\kappa^{1/\nu_\parallel}$, respectively ($\delta\sigma_w := \sigma_w - \sigma_{w;c}$). The bias parameter σ_b was fixed to be 0.3, and tanh activation function was used in the both panels. (b) The order parameter $\rho^{(l)}$ as a function of l for various activation functions, rescaled according to the universal scaling ansatz (18) with the mean-field critical exponents (22). Different symbols with the same color correspond to different σ_w: $+$, $\times$, $*$, $\square$ in ascending order. The bias parameter σ_b was fixed to be 0.3. The nonuniversal metric factors $\gamma_\leftrightarrow$ and κ are calculated from Eq. (23) and Eq. (24), respectively, for each case.

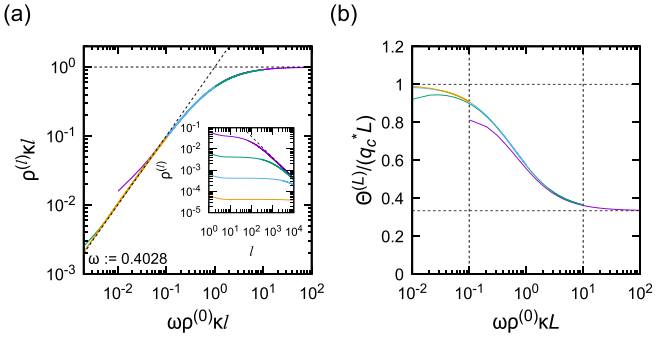

FIG. 4. Critical dynamic scaling of the order parameter $\rho^{(l)}$ and its consequence on NTK. (a) The order parameter $\rho^{(l)}$ at a critical point (specifically $(\sigma_w, \sigma_b) \sim (1.23367, 0.3)$ with erf activation; $\kappa \sim 0.252674$) for various cosine distances $\rho^{(0)}$ [ranging from 10^{-4} (orange) to 10^{-1} (purple); see Eq. (27)], calculated from the mean-field theory (7) and (8) and rescaled according to the scaling ansatz (28). The inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$ with $n_{\text{in}} = 10$ elements were normalized so that $\|\boldsymbol{x}_1\|_2 = \|\boldsymbol{x}_2\|_2 = 1$. The dashed lines are guides to the eye for the asymptotic behavior (29) of the scaling function g. The raw $\rho^{(l)}$ is shown in the inset with a guide to eye for $\rho^{(l)} = (\kappa l)^{-1}$. (b) The NTK $\Theta^{(L)}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ [see Eq. (34)] as a function of the network depth L, rescaled according to the scaling ansatz (37). The same $(\sigma_w, \sigma_b, \rho^{(0)})$ and activation function as (a) were used. The two horizontal lines are guides to the eye for the asymptotic behavior (40) as $L \to \infty$: the upper one for $\boldsymbol{x}_1 = \boldsymbol{x}_2$ [$\rho^{(0)} = 0$] and the lower one for $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ [$\rho^{(0)} > 0$].

if σ_w is fixed and σ_b is varied, provided that $\sigma_w > (h'(0))^{-1}$:

$$\gamma_\updownarrow = \frac{2\sigma_{b;c}\int \mathcal{D}z\, z h'(\sqrt{q_c^*}z)h''(\sqrt{q_c^*}z)}{\sqrt{q_c^*}\int \mathcal{D}z\, h(\sqrt{q_c^*}z)h''(\sqrt{q_c^*}z)}, \tag{25}$$

where $\sigma_{b;c}$ is the critical point for the specified σ_w.

The nonuniversal metric factor κ deserves special attention because it serves as an intrinsic characterizer of a critical point. That is, κ is uniquely determined once a critical point is specified, in contrast with γ and ζ, which also depend on how we approach the critical point. Formally κ is the reciprocal amplitude of the power-law decay at a critical point

$$\rho^{(l)} \sim (\kappa l)^{-\beta/\nu_\parallel} \quad \text{for} \quad l \gg 1, \tag{26}$$

but the readers might ask for a more intuitive meaning. The scaling laws for a critical initial slip [29] can be utilized to address this question. At a critical point, we find that the order parameter $\rho^{(l)}$ for various cosine distances

$$\rho^{(0)} := 1 - \frac{\boldsymbol{x}_1 \cdot \boldsymbol{x}_2}{\|\boldsymbol{x}_1\|_2 \|\boldsymbol{x}_2\|_2} \tag{27}$$

of the normalized inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$ exhibits the universal scaling law described by the following scaling ansatz [Fig. 4(a)]:

$$\rho^{(l)} \simeq (\kappa l)^{-1} g(\omega \rho^{(0)} \kappa l), \tag{28}$$

where ω is a metric factor associated with an initial condition depending on a critical point of interest and the inputs [39]. Notably, the scaling function g shows a crossover between two asymptotic behaviors:

$$g(x) \sim \begin{cases} x & x \ll 1; \\ 1 & x \gg 1. \end{cases} \tag{29}$$

Which asymptotic regime the signal propagation dynamics belongs to is a matter of comparison between $\rho^{(0)}$ and $(\omega\kappa l)^{-1}$. This suggests a striking resemblance to the cosine distance scoring [40], where a simple thresholding on the cosine distance of the feature vectors yields fast and robust speaker verification. In the case of the multilayer perceptrons, the threshold for the crossover is determined implicitly by specifying a critical point (κ, ω), the depth of the network (l), and how we design the inputs (ω). In other words, the metric factor κ, *combined with the depth*, characterizes the network's sensitivity against input differences.

To gain a deeper insight into the implications of the scaling laws for the training dynamics of the neural networks, let us study the NTK [30]

$$\Theta^{(L)}(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}_0) := \sum_j \frac{\partial y}{\partial \theta_j}(\boldsymbol{x}_1; \boldsymbol{\theta}_0)\frac{\partial y}{\partial \theta_j}(\boldsymbol{x}_2; \boldsymbol{\theta}_0) \tag{30}$$

of the initialized networks [here $y(\boldsymbol{x}; \boldsymbol{\theta})$ is the output and $\boldsymbol{\theta}_0$ is an initial value of the trainable parameters, namely $\{W^{(l)}, \boldsymbol{b}^{(l)}\}_{l=1}^{L+1}$]. NTK is directly related to the time evolution of the output $y(\boldsymbol{x}; \boldsymbol{\theta})$ under the gradient descent using mean-squared error,

$$\frac{\mathrm{d}y(\boldsymbol{x}; \boldsymbol{\theta})}{\mathrm{d}t} = -\frac{2\eta}{N}\sum_{i=1}^{N} \Theta^{(L)}(\boldsymbol{x}, \boldsymbol{x}_i; \boldsymbol{\theta})(y(\boldsymbol{x}_i; \boldsymbol{\theta}) - y_i), \tag{31}$$

where η is a learning rate and $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ a training dataset. Remarkably, in the limit of infinite width, the initial value of NTK is deterministic despite the randomness of $\boldsymbol{\theta}_0$ themselves, and also NTK stays constant during the training if the learning rate is small enough [32]. Consequently, the dynamics of the output $y(\boldsymbol{x};\boldsymbol{\theta})$ under such circumstances can be reduced to a linear ordinary differential equation. In particular, a collection of the residual errors $\Delta\boldsymbol{y}(\boldsymbol{\theta}) := (\Delta y(\boldsymbol{x}_1;\boldsymbol{\theta}), \ldots, \Delta y(\boldsymbol{x}_N;\boldsymbol{\theta}))^T$ for training inputs $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ is governed by

$$\frac{\mathrm{d}\Delta\boldsymbol{y}(\boldsymbol{\theta}(t))}{\mathrm{d}t} = -\frac{2\eta}{N}\Theta^{(L)}_{\mathrm{train}}\Delta\boldsymbol{y}(\boldsymbol{\theta}(t)), \tag{32}$$

where $\Theta^{(L)}_{\mathrm{train}}$ is the following Gram matrix (the explicit $\boldsymbol{\theta}_0$ dependence is dropped due to the deterministic property):

$$\Theta^{(L)}_{\mathrm{train}} := \begin{pmatrix} \Theta^{(L)}(\boldsymbol{x}_1,\boldsymbol{x}_1) & \cdots & \Theta^{(L)}(\boldsymbol{x}_1,\boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ \Theta^{(L)}(\boldsymbol{x}_N,\boldsymbol{x}_1) & \cdots & \Theta^{(L)}(\boldsymbol{x}_N,\boldsymbol{x}_N) \end{pmatrix}. \tag{33}$$

NTK also plays a significant role in characterizing generalizability of the infinitely wide neural networks trained under the stochastic gradient descent [41]. Thus, the initial value of NTK is relevant for understanding the training dynamics and performance of the neural networks.

The connection between the initialized NTK and the universal scaling laws can be seen by observing that the closed-form expression [42] of the NTK for the present case is described in terms of $u_1^{(l)}$, $u_2^{(l)}$, and $C^{(l)}$ in the mean-field theory:

$$\Theta^{(L)}(\boldsymbol{x}_1,\boldsymbol{x}_2) = \sum_{l=1}^{L+1} C^{(l)} \prod_{l'=l}^{L} \left(\sigma_w^2 \int \mathcal{D}z_1 \int \mathcal{D}z_2\, h'\left(u_1^{(l')}\right) h'\left(u_2^{(l')}\right) \right). \tag{34}$$

The both terms in the right-hand side of Eq. (34) at a critical point is directly related to the order parameter $\rho^{(l)}$ [to find Eq. (36), perform a similar infinitesimal expansion to the one demonstrated in Appendix A]

$$C^{(l)} \simeq q_c^*(1-\rho^{(l)}), \tag{35}$$

$$\sigma_{w;c}^2 \int \mathcal{D}z_1 \int \mathcal{D}z_2\, h'\left(u_1^{(l)}\right) h'\left(u_2^{(l)}\right) \simeq 1 - 2\kappa\rho^{(l)}. \tag{36}$$

As such, we naturally expect a universal scaling ansatz akin to Eq. (28) for the NTK at a critical point:

$$\Theta^{(L)}(\boldsymbol{x}_1,\boldsymbol{x}_2) \simeq q_c^* L \tilde{g}(\omega\rho^{(0)}\kappa L) \tag{37}$$

with a suitable scaling function $\tilde{g}$. Here the prefactor $q_c^* L$ reflects the asymptotic proportionality to L established in the literature [19,21]; the limits of the scaling function $\tilde{g}$ can also be deduced from the prior works

$$\tilde{g}(x) \to \begin{cases} 1 & \text{as} \quad x \to 0; \\ 1/3 & \text{as} \quad x \to \infty. \end{cases} \tag{38}$$

Alternatively, one may heuristically understand these limits by substituting $\rho^{(l)} = 0$ to Eqs. (35) and (36) (for $x \to 0$) or by plugging the asymptote $\rho^{(l)} \simeq (\kappa l)^{-1}$ into Eq. (36) and then resorting to an arithmetic formula (for $x \to \infty$)

$$\lim_{L\to\infty}\left[\frac{1}{L}\sum_{l=1}^{L}\prod_{l'=l}^{L}\left(1-\frac{2}{l'}\right)\right] = \frac{1}{3}. \tag{39}$$

Either way, reasonable scaling collapse (except when L is small, just as expected) shown in Fig. 4(b) empirically validates the scaling ansatz (37).

The ansatz (37), together with the asymptotics (38), suggests that κL is a crucial factor for the properties of NTK at a critical point. If κL is too small, then the resulting $\Theta^{(L)}_{\mathrm{train}}$ becomes nearly rank-1, which implies slow training of the network in general. Conversely, if κL is too large, then NTK asymptotically behaves like an indicator function

$$\Theta^{(L)}(\boldsymbol{x}_1,\boldsymbol{x}_2) \simeq \begin{cases} q_c^* L & \boldsymbol{x}_1 = \boldsymbol{x}_2; \\ q_c^* L/3 & \text{otherwise}, \end{cases} \tag{40}$$

which seriously deteoriates the network. For instance, the dynamics of an output $y(\boldsymbol{x})$ for an input $\boldsymbol{x}$ outside the training dataset under the gradient descent, namely,

$$\frac{\mathrm{d}y(\boldsymbol{x};\boldsymbol{\theta}(t))}{\mathrm{d}t} = -\frac{2\eta}{N}\Theta^{(L)}_{\mathrm{test}}(\boldsymbol{x})\Delta\boldsymbol{y}(\boldsymbol{\theta}(t)) \tag{41}$$

with

$$\Theta^{(L)}_{\mathrm{test}}(\boldsymbol{x}) := (\Theta^{(L)}(\boldsymbol{x},\boldsymbol{x}_1), \cdots, \Theta^{(L)}(\boldsymbol{x},\boldsymbol{x}_N)), \tag{42}$$

becomes almost independent of $\boldsymbol{x}$, which indicates a poor generalization performance. To put it differently, even if initialized at a critical point, the networks with too small κL behave as if they were in the ordered phase, whereas those with too large κL in the chaotic phase (see also Xiao *et al.* [19]). Thus, κL should be properly chosen to fully exploit the benefit of initialization at a critical point. Along this line of thinking, the *curse of depth* reported by Hayou *et al.* [21] can be understood as a devastating consequence of infinite κL, rather than as an intrinsic limitation of the infinitely wide networks. One caveat is that the range within which κL should be tuned as suggested from Fig. 4(b) alone, namely [below, $\rho^{(0)}_{\max}$ and $\rho^{(0)}_{\min}$ denote the maximum and minimum nonzero cosine distance $\rho^{(0)}$ achieved in a training dataset, respectively]

$$0.1/\omega\rho^{(0)}_{\max} \lesssim \kappa L \lesssim 10/\omega\rho^{(0)}_{\min} \tag{43}$$

so that $\Theta^{(L)}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ for all the pairs of training inputs $\boldsymbol{x}_i, \boldsymbol{x}_j$ $(i \neq j)$ do not fall into the same asymptotic regime, is rather loose. We might be able to tighten the range by a more thorough analysis of NTK for a dataset at hand, although we cannot expect such a tighter range to be carried over different datasets. We plan to investigate this point more in the near future.

C. Scaling results for the finite networks and different architectures

Another virtue of the universal scaling laws is that they give us useful intuition even into the networks of finite width, where quantitatively tracking the deviation from the Gaussian process can be cumbersome (if not impossible [26,27]). To illustrate this point, let us consider the finite-size scaling of the neural network at a critical point. Since the order parameter $\rho^{(l)}$ in the mean-field theory is defined through the Pearson

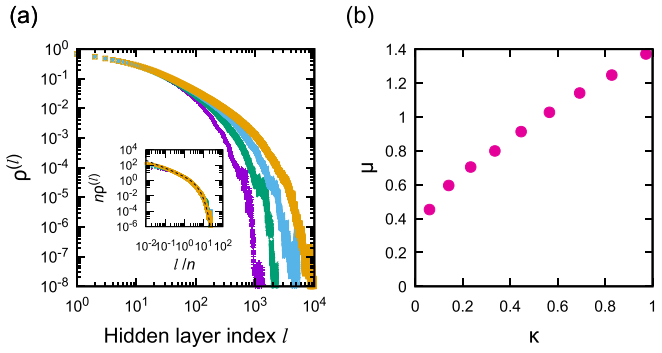

FIG. 5. Finite-size scaling of the order-to-chaos transition in the multilayer perceptrons (1). (a) The order parameter $\rho^{(l)}$ [see Eq. (44)] at a critical point $[(\sigma_{w;c}, \sigma_b) \sim (1.39558, 0.3)$ with tanh activation; $\kappa \sim 0.233498]$ for various widths n [ranging from 50 (purple) to 400 (orange)], empirically averaged over 10^4 independent runs. Two orthogonal inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$ (with $\|\boldsymbol{x}_1\|_2 = \|\boldsymbol{x}_2\|_2 = 1$) of size $n_{\text{in}} = 10$ were given. The inset shows the same data rescaled according to the univesal scaling ansatz (45). The black dashed curve indicates the solution (47) of the phenomenological description (46) with $(\kappa, \mu) \sim (0.233498, 0.6601)$, where μ was chosen by fitting the solution to the empirical result for $n = 400$. (b) The nonuniversal metric factor μ as a function of κ in the case of tanh activation, where μ for each κ was estimated from the same fitting as the inset of (a) using the empirical $\rho^{(l)}$ for $n = 200$.

correlation coefficient $c^{(l)}$ [see Eq. (11)], definition of the finite-width counterpart is straightforward:

$$\rho^{(l)} := 1 - \frac{\sum_i \left(z_{1;i}^{(l)} - Z_1^{(l)}\right)\left(z_{2;i}^{(l)} - Z_2^{(l)}\right)}{\sqrt{\sum_i \left(z_{1;i}^{(l)} - Z_1^{(l)}\right)^2 \sum_i \left(z_{2;i}^{(l)} - Z_2^{(l)}\right)^2}}, \tag{44}$$

where $\boldsymbol{z}_1^{(l)}, \boldsymbol{z}_2^{(l)} \in \mathbb{R}^n$ are the preactivations at the lth hidden layer for different inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$, respectively, $z_{j;i}^{(l)}$ the ith element of $\boldsymbol{z}_j^{(l)}$, and $Z_j^{(l)} := \frac{1}{n}\sum_i z_{j;i}^{(l)}$. We empirically find that the order parameter $\rho^{(l)}$ at a critical point for various widths n admits the following universal scaling ansatz [Fig. 5(a)]:

$$\rho^{(l)}[\sigma_w = \sigma_{w;c}; n] \simeq n^{-1} f(n^{-1} l). \tag{45}$$

The empirical finding above can be heuristically understood by considering a finite-width correction to the mean-field theory. In the case of the finite-width networks, the fourth-order (and other even-order) cumulants come into play, while the third-order (and other odd-order) ones vanish just as the odd-order moments do [43]. In the spirit of the asymptotic expansion of the probability distribution [44], this observation indicates that the leading correction to the Gaussian process is an order of n^{-1}, the reciprocal of the width. Thus, together with an obvious fact that $\rho = 0$ is an absorbing state also for the finite networks, we are led to the following modified phenomenological description:

$$\frac{\mathrm{d}\rho}{\mathrm{d}l} = -\frac{\mu}{n}\rho - \kappa\rho^2, \tag{46}$$

where μ is a new nonuniversal metric factor. Unfortunately, theoretical calculation of μ would be challenging, since this requires us to analyze the approximate recursion relation up to $O(n^{-1})$ for the covariance, which is no longer closed within the variance and the covariance [as opposed to the mean-field theory (7) and (8)]. Still, one can *measure* it by fitting the empirical $\rho^{(l)}$ to the analytical solution of the phenomenological description (46)

$$n\rho^{(l)} = \frac{\rho_0 \mu}{\rho_0 \kappa (e^{\frac{\mu l}{n}} - 1) + (\mu/n) e^{\frac{\mu l}{n}}}, \tag{47}$$

as demonstrated in the inset of Fig. 5(a).

While it is nowadays well established that the depth-to-width ratio L/n is a key quantity for describing the multilayer perceptrons with finite n [23,45,46], the metric factor μ enriches this insight by providing the means to quantitatively characterize the sensitivity of the network to the width. Specifically, the width n of the network should satisfy $n \gtrsim \mu L$ so that the signal propagation dynamics therein is reasonably well approximated by the infinitely wide limit. This introduces another design consideration for the neural networks. Recalling the relevance of κL for the training dynamics of the networks, one would be tempted to use a critical network with larger κ to obtain good generalization with smaller L. However, Fig. 5(b) empirically suggests that larger κ comes with a cost of larger μ, which imposes an extra computational burden for larger n. Hence, if one chooses to operate the network near the infinitely wide limit, then one needs to make a trade-off between these two factors for cost-effective training. Theoretically, a more precise formulation of this idea might be achieved by studying the finite-width corrections [47] to the training dynamics, which is beyond the scope of the present work.

Finally, we briefly discuss the convolutional neural networks to see how the analogy to absorbing phase transitions carries over different architectures. Formally, the recurrence relations for the preactivation $\boldsymbol{z}^{(l;\alpha)}$ of a d-dimensional convolutional neural network (a periodic boundary condition, also known as circular padding, is assumed for simplicity) is described as follows (below c and k denote the number of channels and the width of the convolution filter, respectively):

$$\boldsymbol{z}^{(l+1;\alpha)} = \frac{\sigma_w}{\sqrt{ck^d}} \sum_{m=1}^{c} w^{(l+1;\alpha,m)} \star h(\boldsymbol{z}^{(l;m)}) + \sigma_b \boldsymbol{b}^{(l+1;\alpha)}, \tag{48}$$

where $\star$ denotes the cross-correlation operator [the summation below is taken over the range $\{(k-1)/2, \cdots, -1, 0, 1, \cdots, (k-1)/2\}$ for each $j_1, \cdots, j_d$]

$$(A \star B)_{i_1,\cdots,i_d} := \sum_{j_1,\cdots,j_d} A_{j_1+\frac{k+1}{2},\cdots,j_d+\frac{k+1}{2}} B_{i_1+j_1,\cdots,i_d+j_d}, \tag{49}$$

and $h(z)$ is a shorthand for element-wise application of h to z. Each element of the convolutional filter $w^{(l;\alpha,m)} \in \mathbb{R}^{k^d}$ and the bias $\boldsymbol{b}^{(l;\alpha)}$ is initialized according to the standard normal distribution $\mathcal{N}(0, 1)$. In the present case, $\rho^{(l)}$ is obtained by first calculating the correlation coefficient (44) for each channel and then taking the average over all the channels.

The key difference compared to the multilayer perceptrons is the locality of the interaction between the neurons. The neurons within a convolutional layer interact only locally through the convolutional filters, in contrast with the multilayer perceptions, where the network admits the fully connected structure. The richer dynamics due to the spatial degrees of freedom can be partly grasped by studying the limit

of infinitely many channels $c \to \infty$. In this limit, the neural network is again equivalent to a Gaussian process [48] and the phase diagram of the mean-field theory remains exactly the same [24] as the multilayer perceptrons (Fig. 1), making it easy to compare between the two architectures. Different phases in the convolutional networks are characterized by how a noise in a single pixel spatially spread in the course of the signal propagation, in addition to the asymptotic behavior of the order parameter ρ. One can empirically check that the noise eventually decays in the ordered phase, whereas it spreads ballistically in the chaotic phase. At a critical point, the spreading process is diffusive, whose characteristic width n_* scales with the network depth l as $n_* \sim l^{\nu_\perp/\nu_\parallel} = \sqrt{l}$, which induces a new critical exponent

$$\nu_\perp = 1/2. \tag{50}$$

This is exactly what happens in the mean-field theory of absorbing phase transitions with the spatial degrees of freedom [10] at a critical point,

$$\frac{\partial \rho}{\partial l} = -\kappa \rho^2 + D\nabla^2 \rho. \tag{51}$$

Thus, the signal propagation dynamics of the convolutional neural networks with infinitely many channels and input pixels at the critical point is characterized by two independent metric factors (κ, D). Theoretical calculation of the new metric factor D could perhaps be done using the similar techniques [49,50] for studying the dynamics of wavefronts in coupled map lattices [51], although this is substantially more challenging than κ. Since an exact and efficient algorithm to compute NTK is also available for the convolutional networks [42], it would be interesting to investigate the role of D in the training dynamics. At any rate, we believe it is safe to say that the analogy to absorbing phase transitions is a promising insight for studying deep neural networks outside the multilayer perceptrons.

The analogy to absorbing phase transitions also gives us a nontrivial and yet intuitive insight into the signal propagation dynamics of the convolutional neural networks with finite channels c, although the implications to the training dynamics may be less direct, just as in the case of the finite-width multilayer perceptrons. Empirical evidence we show in Fig. 6 suggests the following phenomenology. If c is finite, then the dynamics of the covariance (and hence of the order parameter $\rho^{(l)}$) is no longer deterministic but is accompanied by a multiplicative noise, whose amplitude is asymptotically proportional to $\sqrt{\rho^{(l)}}$ as normally expected for models with microscopic stochastic elements [52]. The noise works as a relevant perturbation to the mean-field theory (51) in the sense of the renormalization group, and it changes the asymptotic scaling behavior of the network at a large scale to that of the directed percolation (DP) universality class [53]. As such, the universal scaling ansatz (18) remains valid with different critical exponents. For instance, reasonable scaling collapse can be found for the order parameter $\rho^{(l)}$ in the spatially one-dimensional convolutional networks [Fig. 6(a)] using the exponents for the $(1+1)$-dimensional DP universality class [54]

$$\beta_{1\mathrm{DDP}} \sim 0.27649, \quad \nu_{\parallel 1\mathrm{DDP}} \sim 1.73385, \tag{52}$$

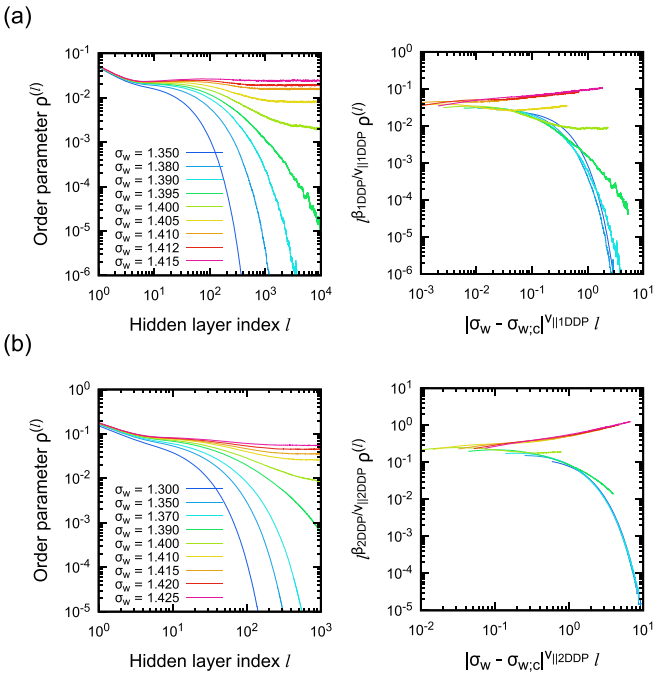

FIG. 6. Directed percolation scaling in the order-to-chaos transition in the convolutional neural networks (48). (a) The order parameter $\rho^{(l)}$ with $d = 1$, $n = 400$, $k = 5$, and $c = 10$, empirically averaged over 10^6 independent runs. The raw data are shown in the left panel, which is then rescaled according to the scaling ansatz (18) with the critical exponents of $(1+1)$-dimensional DP (52) in the right. Two orthogonal inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$ (with $\|\boldsymbol{x}_1\|_2 = \|\boldsymbol{x}_2\|_2 = 1$) of size n were given. $\sigma_{w;c} := 1.408$ is chosen to find the scaling collapse. (b) Similarly with (a), but with $d = 2$, $n = 100$, $k = 3$, and $c = 5$, averaged over 4000 independent runs. The critical exponents of $(2+1)$-dimensional DP (53) and $\sigma_{w;c} := 1.404$ were used to find the scaling collapse in this case. In both (a) and (b), tanh activation was used and the bias parameter σ_b was fixed to be 0.3.

although the order parameter can saturate to a nonzero value slightly below the critical point (causing the tilts in the rescaled plot) mainly because the perfect order $[\rho^{(l)} = 0]$ is virtually unachievable; the imperfection acts as a small fluctuation, to which the setups near the critical point is particularly sensitive [55]. We also checked that the essentially same scenario holds true for the two-dimensional convolutional networks [Fig. 6(b)], where the critical exponents $\beta, \nu_\parallel$ are replaced [56,57] with

$$\beta_{2\mathrm{DDP}} \sim 0.58, \quad \nu_{\parallel 2\mathrm{DDP}} \sim 1.29. \tag{53}$$

Remarkably, the deviation from the scaling collapse is less prominent in the two-dimensional networks than in their one-dimensional counterparts, which is consistent with our interpretation of the tilts since fluctuations generally become less relevant as the spatial dimensionality goes up [10].

The correspondence to the DP universality class suggest that the signal propagation dynamics in the convolutional networks is highly nontrivial, especially given the notorious difficulty of exactly solving DP [58]. Yet, thanks to the universality of the scaling laws of absorbing phase transitions, semiquantitative predictions can be gained via simple phenomenological considerations. In particular, the most

informative combination of the width n and the depth L for describing the behavior of the critically initialized deep convolutional networks may be changed from L/n to $L/n^{\nu_\parallel/\nu_\perp}$ using the corresponding critical exponents

$$\nu_{\perp 1\mathrm{DDP}} \sim 1.096854, \quad \nu_{\perp 2\mathrm{DDP}} \sim 0.73. \tag{54}$$

V. DISCUSSION

To summarize, we pursued the analogy between the behavior of the conventional deep neural networks and absorbing phase transitions in the present work. During the pursuit, we demonstrated that the signal propagation dynamics in the untrained neural networks follows the universal scaling laws, while the specific details are summarized using the associated nonuniversal metric factors. In particular, the nonuniversal metric factor κ was shown to play a significant role in the training dynamics of the multilayer perceptrons: Its product with the network depth L should be tuned for optimal generalization. Thus, the present work provides useful insights into the neural networks with many but finite hidden layers, which complements our understanding of two-layer [59,60] or infinitely deep networks [19,21]. The framework can be readily extended to ReLU-like activation functions (albeit with different exponents), which consequently underlines the significance of properly choosing the amount of leak; see Appendix B. Furthermore, we provided numerical evidence suggesting that the analogy to absorbing phase transitions well captures the signal propagation dynamics in the neural networks with finite width or different architecture, holding great promise for future developments.

Let us emphasize that successful deep learning can only be achieved via a complicated interplay among various setups, even in one of the simplest cases where NTK describes the training dynamics reasonably well. In addition to the relevance of initialization at criticality [18,21] and of proper scaling of a learning rate with respect to depth [19], the present work demonstrates the necessity of a more specific, depth-dependent choice of hyperparameters. Furthermore, in the case of a large but finite width, one should also strike a balance between width and depth for efficiency. The fact that all these setups need to be considered simultaneously highlights the major challenge in deep learning, which necessitates extensive study on hyperparameter optimization [61]. To put it the other way around, considerable theoretical insight into the mechanism behind the recent success of deep learning may be obtained by studying the neural networks near the realm of NTK, contrary to a common belief [62,63].

In a broader context, the present work hopefully exemplifies a subtle relationship between criticality and intelligence. In the case of the artificial neural networks, being at criticality alone is not sufficient for successful learning, although it is likely to be necessary. Intriguingly, this theoretical insight is consistent with the experimental findings on real neural systems: While deviation from criticality often results in an altered or abnormal state of consciousness [64,65], a sign of criticality does not necessarily imply presumable capability of performing intellectual tasks [66–69]. The effectiveness of dynamic scaling [29,70] for capturing the training dynamics of the deep neural networks suggests that intrinsic memory characteristics of the network arising from the criticality (such as the metric factor κ in the present work) plays a pivotal role in information processing within the artificial networks, which again draws an interesting parallel with brains [71,72]. As such, understanding of the relationship between criticality and intelligence may benefit from further study on how the memory characteristics should be adjusted for specific datasets.

We foresee some interesting directions for future work. Apart from the ones mentioned in the previous section [73], one of the most natural directions is to extend the present framework to more modern architectures [25,74,75]. In particular, the skip connections employed in the residual neural networks (ResNet [76]) may be seen as a stimulus kicking the system out of an absorbing state: Even if the signals are collapsed during the propagation within a residual block, the skip connection breaks the order before entering the new block. Given that a combination of external drive and dissipation into absorbing states has been conjectured to be a key ingredient of self-organized criticality in physical systems [3,7], the skip connections may have their unique benefits in deep learning as well. Another, albeit less straightforward, direction is to provide thermodynamic foundations of the speed-accuracy trade-off in deep learning. The deep neural networks in the ordered phase sacrifice speed for accuracy and vice versa in the chaotic phase. Since the speed-accuracy trade-off has been extensively studied in the context of living systems [77–80], thermodynamical insights developed therein are likely to be helpful (and indeed, very recently, the trade-off in the diffusion models has been studied from a thermodynamic viewpoint [81]), although pursuing this direction would call for an improved understanding of the thermodynamics of absorbing phase transitions [82,83]. We believe that further investigation into a parallel between intelligence in living systems and that in artificial neural networks will lead us to a lot of exciting developments, beneficial for both physics and machine learning communities.

ACKNOWLEDGMENTS

The numerical experiments for supporting our arguments in this work were performed on the cluster machine provided by Institute for Physics of Intelligence, The University of Tokyo. This work was supported by the Center of Innovations for Sustainable Quantum AI (JST Grant No. JPMJPF2221) and JSPS KAKENHI Grant No. JP23H03818. T.O. and S.T. acknowledge support by the Endowed Project for Quantum Software Research and Education, The University of Tokyo [84].

APPENDIX A: DERIVATION OF THE CRITICAL EXPONENTS FOR THE MULTILAYER PERCEPTRONS

Here we derive the critical exponents (22) of the multilayer perceptrons in the main text. That is, we show that the order parameter ρ^* [see Eq. (21)] at the fixed point of the iterative $\mathcal{C}$-map (13) and the correlation depth $\xi_\parallel$ as defined by Eq. (14) respectively exhibits linear onset in the vicinity of the critical point [although we have already seen it empirically in Fig. 1(c)]. To achieve this goal, we expand the mean-field

theory (7) and (8) with respect to infinitesimally small deviation $\delta\sigma_w$ from the critical point $\sigma_{w;c}$.

First, let us show the continuity of q^* as a function of σ_w at the edge of chaos for later convenience. Consider the fixed point q^* of the mean-field theory (7) for infinitesimally different σ_w, and let $\delta\sigma_w$ and δq^* respectively denote the increment in σ_w and q^*. Then we compare the equality for the fixed point of q as follows:

$$q^* = \sigma_w^2 \int \mathcal{D}z\, h^2(\sqrt{q^*}z) + \sigma_b^2; \tag{A1}$$

$$\begin{aligned} q^* + \delta q^* &= (\sigma_w + \delta\sigma_w)^2 \int \mathcal{D}z\, h^2(\sqrt{q^* + \delta q^*}z) + \sigma_b^2 \\ &\simeq \sigma_w^2 \int \mathcal{D}z\, h^2(\sqrt{q^*}z) + \sigma_b^2 \\ &\quad + \delta q^*\, \sigma_w^2 \int \mathcal{D}z\, \frac{z}{\sqrt{q^*}} h(\sqrt{q^*}z) h'(\sqrt{q^*}z) \\ &\quad + 2\delta\sigma_w\, \sigma_w \int \mathcal{D}z\, h^2(\sqrt{q^*}z), \end{aligned} \tag{A2}$$

where we have neglected difference of $O((\delta q^*)^2)$, $O((\delta\sigma_w)^2)$, or $O(\delta q^* \delta\sigma_w)$. By subtracting Eq. (A1) from Eq. (A2), we find

$$\begin{aligned} \delta q^* &\simeq \frac{2\sigma_w \int \mathcal{D}z\, h^2(\sqrt{q^*}z)}{1 - \sigma_w^2 \int \mathcal{D}z\, \frac{z}{\sqrt{q^*}} h(\sqrt{q^*}z) h'(\sqrt{q^*}z)} \delta\sigma_w \\ &=: \alpha \delta\sigma_w. \end{aligned} \tag{A3}$$

In particular at the critical point $\sigma_{w;c}$, the coefficient α can be further simplified to

$$\alpha = \frac{2 \int \mathcal{D}z\, h^2(\sqrt{q_c^*}z)}{-\sigma_{w;c} \int \mathcal{D}z\, h(\sqrt{q_c^*}z) h''(\sqrt{q_c^*}z)}, \tag{A4}$$

where q_c^* is the fixed point of Eq. (7) at the edge of chaos. It turns out that the numerator and the denominator of the right-hand side of Eq. (A3) converge to a finite value, so does α itself.

Next, we study the behavior of $\xi_\parallel^{-1}$ slightly below the critical point. To do this, we expand Eq. (15) with respect to an infinitesimal deviation $\delta\sigma_w$ from the critical point $\sigma_{w;c}$:

$$\begin{aligned} & e^{-\frac{1}{\xi_\parallel(\sigma_{w;c} - \delta\sigma_w)}} \\ &= (\sigma_{w;c} - \delta\sigma_w)^2 \int \mathcal{D}z\, h'^2(\sqrt{q_c^* - \alpha\delta\sigma_w}z) \\ &\simeq 1 - \left[\frac{2}{\sigma_{w;c}} + \frac{\alpha\sigma_{w;c}^2}{\sqrt{q_c^*}} \int \mathcal{D}z\, z h'(\sqrt{q_c^*}z) h''(\sqrt{q_c^*}z) \right] \delta\sigma_w \\ &= 1 - \gamma_1 \delta\sigma_w, \end{aligned} \tag{A5}$$

where

$$\gamma_1 := \frac{2}{\sigma_{w;c}} \left(1 - \frac{(q_c^* - \sigma_b^2) \int \mathcal{D}z\, z h'(\sqrt{q_c^*}z) h''(\sqrt{q_c^*}z)}{\sqrt{q_c^*} \int \mathcal{D}z\, h(\sqrt{q_c^*}z) h''(\sqrt{q_c^*}z)} \right). \tag{A6}$$

The coefficient γ_1 remains finite for activation functions in the $K^* = 0$ universality class, and hence $\xi_\parallel^{-1}$ decreases to 0 as $\sigma_w \uparrow \sigma_{w;c}$ in an asymptotically linear manner.

The correlation depth $\xi_\parallel^{-1}$ slightly above the critical point [see Eq. (9) for the definitions of u_1^*, u_2^*]

$$\begin{aligned} e^{-\frac{1}{\xi_\parallel(\sigma_{w;c} + \delta\sigma_w)}} &= (\sigma_{w;c} + \delta\sigma_w)^2 \int \mathcal{D}z_1 \int \mathcal{D}z_2\, h' \\ &\quad \times (u_1^* + \delta u_1^*) h'(u_2^* + \delta u_2^*) \end{aligned} \tag{A7}$$

can be studied similarly, but we need first to analyze the behavior of the fixed point c^* of the iterative $\mathcal{C}$-map (13) as a function of σ_w, due to the c dependence of u_2. Hence we expand the $\mathcal{C}$-map slightly above the critical point (that is, $\sigma_w = \sigma_{w;c} + \delta\sigma_w$) around the trivial fixed point $c^{(l)} = 1$

$$\begin{aligned} c^{(l+1)} - c^{(l)} &= \left(\left. \frac{\mathrm{d}c^{(l+1)}}{\mathrm{d}c^{(l)}} \right|_{c^{(l)}=1} - 1 \right) (c^{(l)} - 1) \\ &\quad + \frac{1}{2} \left. \frac{\mathrm{d}^2 c^{(l+1)}}{\mathrm{d}c^{(l)2}} \right|_{c^{(l)}=1} (c^{(l)} - 1)^2 + \cdots. \end{aligned} \tag{A8}$$

Notice that essentially the same calculation as the one for analyzing linear stability of the trivial fixed point [18] can be repeated to inductively see

$$\left. \frac{\mathrm{d}^n c^{(l+1)}}{\mathrm{d}c^{(l)n}} \right|_{c^{(l)}=1} = \sigma_w^2 q^{*n-1} \int \mathcal{D}z \left(\frac{\mathrm{d}^n h}{\mathrm{d}z^n}(\sqrt{q^*}z) \right)^2, \tag{A9}$$

which implies these derivatives are positive and finite at any order. Particularly in the vicinity of the critical point, we have, from Eq. (A5),

$$\left. \frac{\mathrm{d}c^{(l+1)}}{\mathrm{d}c^{(l)}} \right|_{c^{(l)}=1} - 1 = \gamma_1 \delta\sigma_w + o(\delta\sigma_w). \tag{A10}$$

By taking the first two terms of the expansion (A8) into account and solving it with respect to $\delta\rho := 1 - c^*$ at the fixed point, one can see that the leading contribution for $\delta\rho$ is of order $\delta\sigma_w$, more specifically

$$\delta\rho = \gamma_1 \frac{2 \int \mathcal{D}z\, h'^2(\sqrt{q_c^*}z)}{q_c^* \int \mathcal{D}z\, h''^2(\sqrt{q_c^*}z)} \delta\sigma_w =: \zeta\, \delta\sigma_w. \tag{A11}$$

This result implies that the critical exponent β associated with the onset of the order parameter is 1.

Now we are in the position of studying $\xi_\parallel^{-1}$ slightly above the critical point (A7):

$$\begin{aligned} & 1 - e^{-\frac{1}{\xi_\parallel(\sigma_{w;c} + \delta\sigma_w)}} \\ &\simeq \left[\zeta \sigma_{w;c}^2 \sqrt{q_c^*} \left(\int \mathcal{D}z\, z h'(\sqrt{q_c^*}z) h''(\sqrt{q_c^*}z) \right.\right. \\ &\quad \left. - \int \mathcal{D}z_1 \int \mathcal{D}z_2 \sqrt{q_c^*} z_2^2 h'(\sqrt{q_c^*}z_1) h'''(\sqrt{q_c^*}z_1) \right) \\ &\quad \left. - \frac{\alpha\sigma_{w;c}^2}{\sqrt{q_c^*}} \int \mathcal{D}z\, z h'(\sqrt{q_c^*}z) h''(\sqrt{q_c^*}z) - \frac{2}{\sigma_{w;c}} \right] \delta\sigma_w \\ &= \gamma_2\, \delta\sigma_w, \end{aligned} \tag{A12}$$

where

$$\gamma_2 := \zeta \frac{q_c^* \int \mathcal{D}z\, h''^2(\sqrt{q_c^*}z)}{\int \mathcal{D}z\, h'^2(\sqrt{q_c^*}z)} - \gamma_1 = \gamma_1 =: \gamma. \tag{A13}$$

This indicates that $\xi_\parallel^{-1}$ decreases to 0 as $\sigma_w \downarrow \sigma_{w;c}$ in an asymptotically linear manner. Thus, it is confirmed that $\nu_\parallel = 1$. Note that the contribution of order $\delta\sigma_w^{\frac{1}{2}}$ vanishes because

$$\int_{-\infty}^{\infty} \mathrm{d}z \frac{z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} = 0. \tag{A14}$$

Although the main purpose of this Appendix, namely the derivation of the critical exponents $\beta, \nu_\parallel$, has already been completed, let us discuss the nonuniversal metric factors γ, κ introduced in Eq. (18). By comparing the results (A5), (A11), and (A12) with the solution of the mean-field theory of absorbing phase transition [10]

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \gamma_\leftrightarrow(\sigma_w - \sigma_{w;c})\rho - \kappa\rho^2, \tag{A15}$$

we find the following results:

$$\gamma_\leftrightarrow = \gamma = \frac{2}{\sigma_{w;c}}\left(1 - \frac{\left(q_c^* - \sigma_b^2\right)\int \mathcal{D}z\, zh'(\sqrt{q_c^*}z)h''(\sqrt{q_c^*}z)}{\sqrt{q_c^*}\int \mathcal{D}z\, h(\sqrt{q_c^*}z)h''(\sqrt{q_c^*}z)}\right); \tag{A16}$$

$$\kappa = \gamma/\zeta = \frac{q_c^* \int \mathcal{D}z\, h''^2(\sqrt{q_c^*}z)}{2\int \mathcal{D}z\, h'^2(\sqrt{q_c^*}z)}. \tag{A17}$$

By repeating the same argument for a fixed $\sigma_w > (h'(0))^{-1}$, one can arrive at the same critical exponents with the different metric factor $\gamma_\updownarrow$; see Eq. (25).

APPENDIX B: SCALE-INVARIANT ACTIVATION FUNCTIONS

The purpose of this Appendix is to study the signal propagation dynamics of the infinitely wide multilayer perceptrons with scale-invariant activation functions (in the following, a is a non-negative parameter often referred to as *leak*)

$$h(x) = \begin{cases} ax & x < 0; \\ x & x \geqslant 0. \end{cases} \tag{B1}$$

The order-to-chaos transition in the neural networks of this kind is slightly different from the one discussed in the main text: qualitative change of the behavior can be found in the variance $q^{(l)}$ rather than in the covariance $C^{(l)}$. One can see, by carrying out the integration in Eq. (7), that $q^{(l)}$ exhibits the transition between convergence to some constant q^* and exponential divergence at $\sigma_w = \sigma_{w;c} := \sqrt{2/(1+a^2)}$, regardless of σ_b [22]. In particular, $q^{(l)}$ stays constant throughout the network if

$$(\sigma_w, \sigma_b) = \left(\sqrt{\frac{2}{1+a^2}}, 0\right), \tag{B2}$$

while it diverges linearly if $\sigma_b \neq 0$ [Fig. 7(a)]. Note that a special case of Eq. (B2) for $a = 0$ is nothing but the well-known He initialization [85] for ReLU activation. With this initialization scheme, we have a well-defined iterative $\mathcal{C}$-map (13) and hence we can study the order parameter $\rho^{(l)}$ defined in Eq. (21); Cho and Saul [86] provide technical details on how to analytically deal with the integration appearing in Eq. (13) in the case of ReLU activation. Notice also that q^* vanishes for $\sigma_w < \sigma_{w;c}$ if $\sigma_b = 0$. In other words, the neurons within sufficiently deep hidden layers *die out* [87], which is reminiscent of an absorbing phase transition discussed in the main text.

A natural question is whether one can apply the universal scaling of absorbing phase transitions in the present case, which we will address below. As visualized in Fig. 7(b), we find that the order parameter $\rho^{(l)}$ at the edge of chaos (B2) follows the universal scaling ansatz,

$$\rho^{(l)} \simeq (\kappa l)^{-2} g(\rho^{(0)}(\kappa l)^2), \quad \kappa = \frac{\sqrt{2}(1-a)^2}{3(1+a^2)\pi}, \tag{B3}$$

with a suitable scaling function g having the same asymptotic behavior as Eq. (29), where we dropped the metric factor ω associated with an initial condition because $\omega = 1$ in this case. The difference in the scaling exponent compared to the result (28) in the main text stems from the second-dominant term in the iterative $\mathcal{C}$-map (13). Specifically, one can see that the difference of $\rho^{(l)}$ in the adjacent layers is asymptotically of $\rho^{(l)\frac{3}{2}}$, rather than $\rho^{(l)2}$:

$$\begin{aligned}\rho^{(l+1)} - \rho^{(l)} &= -\frac{\sigma_{w;c}^2(1-a)^2}{2\pi}\left(\sqrt{1-c^{(l)2}} - c^{(l)}\cos^{-1}c^{(l)}\right) \\ &= -\frac{2\sqrt{2}(1-a)^2}{3(1+a^2)\pi}\rho^{(l)\frac{3}{2}} + O(\rho^{(l)\frac{5}{2}}).\end{aligned} \tag{B4}$$

Since the argument for a finite-width correction [see just above Eq. (46)] is not sensitive to a selection of the activation function, the correction can be made in the same manner as the $K^* = 0$ activation functions

$$\frac{\mathrm{d}\rho}{\mathrm{d}l} = -\frac{\mu}{n}\rho - 2\kappa\rho^{\frac{3}{2}}, \tag{B5}$$

although the resulting l dependence of the order parameter $\rho^{(l)}$ and the exponents for the finite-size scaling are different:

$$n^2\rho^{(l)} \simeq \frac{\mu^2\rho_0}{\left(2\kappa\sqrt{\rho_0}\left(e^{\frac{\mu l}{2n}} - 1\right) + (\mu/n)e^{\frac{\mu l}{2n}}\right)^2}, \tag{B6}$$

$$\rho^{(l)} \simeq n^{-2} f(n^{-1}l). \tag{B7}$$

Numerical results shown in Fig. 7(c) validate the finite-size scaling ansatz (B7) and suggest that the positive correlation between the two metric factors κ, μ holds true for scale-invariant activation functions (see the inset). Similarly, the universal scaling for NTK holds for small $\rho^{(0)}$ or large L

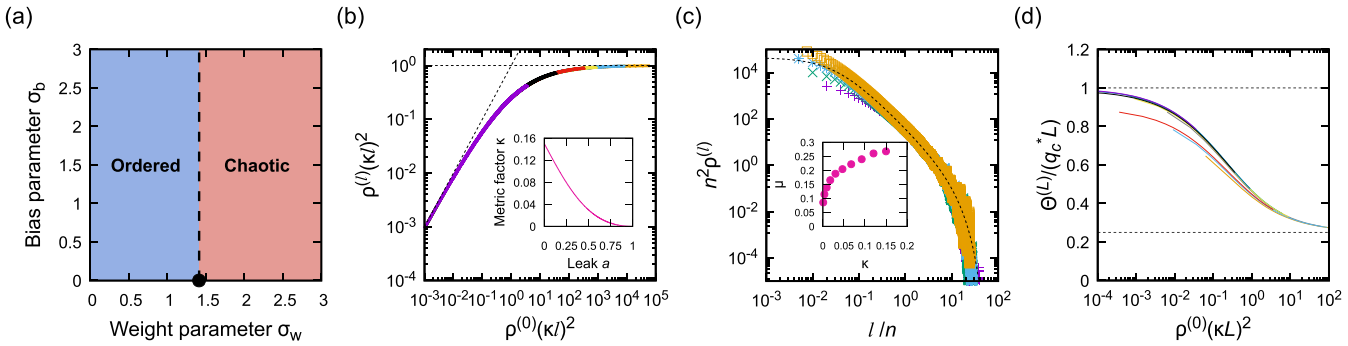

FIG. 7. Universal scaling in the infinitely wide multilayer perceptrons (1) with scale-invariant activation functions (B1). (a) The phase diagram of the signal propagation for ReLU activation ($a = 0$). A dashed line is used to indicate the phase boundary because the preactivation variance diverges at the boundary unless $\sigma_b = 0$ (a black point). (b) The order parameter $\rho^{(l)}$ at the edge of chaos (B2) for various combinations of cosine distances $\rho^{(0)}$ and leak parameters a [the resulting $\rho^{(0)}\kappa^2$ ranges from 3.6×10^{-8} (purple) to 6.3×10^{-4} (orange)], calculated from the mean-field theory (7), (8) and then rescaled according to the scaling ansatz (B3). The dashed lines are guides to the eye for the asymptotic behavior (29) of the scaling function g. The inset shows the nonuniversal metric factor κ as a function of a. (c) The main panel shows the order parameter $\rho^{(l)}$ [see Eq. (44)] at a critical point [$(\sigma_{w;c}, \sigma_b) \sim (\sqrt{2}, 0)$ with ReLU activation; $\kappa = \sqrt{2}/(3\pi) \sim 0.150053$] for various widths n [ranging from 50 (purple) to 400 (orange)], empirically averaged over 4×10^4 independent runs and rescaled according to the universal scaling ansatz (B7). Two orthogonal inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$ (with $\|\boldsymbol{x}_1\|_2 = \|\boldsymbol{x}_2\|_2 = 1$) of size $n_{\text{in}} = 10$ were given. The black dashed curve indicates the solution (B6) of the phenomenological description (B5) with $(\kappa, \mu) \sim (0.150053, 0.2676)$, where μ was chosen by fitting the solution to the empirical result for $n = 200$. The inset shows the nonuniversal metric factor μ as a function of κ in the case of scale-invariant activation functions, where μ for each κ was estimated from the same fitting as the main panel using the empirical $\rho^{(l)}$ for $n = 200$. (d) The NTK $\Theta^{(L)}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ [see Eq. (34)] for various network depths, rescaled according to the universal scaling ansatz (B8). The same combinations of $(\rho^{(0)}, a)$ as (b) were used. The two horizontal lines are guides to the eye for the asymptotic behavior as $L \to \infty$ [21]: the upper one for $\boldsymbol{x}_1 = \boldsymbol{x}_2$ [$\rho^{(0)} = 0$] and the lower one for $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ [$\rho^{(0)} > 0$].

[Fig. 7(d)]:

$$\Theta^{(L)}(\boldsymbol{x}_1, \boldsymbol{x}_2) \simeq q_c^* L \tilde{g}(\rho^{(0)}(\kappa L)^2). \tag{B8}$$

Thus, essentially the same scenario as the $K^* = 0$ activation functions is applicable to the scale-invariant functions, with a suitable change of the scaling exponents.

The analysis above potentially provides theoretical foundations of some empirical insights in the literature. First, we can correctly anticipate that a small leak of $a = 0.01$, commonly referred to as leaky ReLU in the literature, is unlikely to have a significant impact on the network performance [88,89], since the metric factor κ changes only by 2%. With larger a, however, κ noticeably decreases (for instance, it becomes half of the original ReLU at $a = 2 - \sqrt{3} \sim 0.27$) and the optimal depth for training increases in a reciprocal manner (see the final paragraph of Sec. IV B). Consequently, it is possible that the networks with suitably chosen leak works better for a fixed task and other network structures, in particular when the original ReLU network tends to overfit the training data; the superior performance of very leaky ($a \sim 0.18$) ReLU reported by Xu *et al.* [90] may be seen as a remarkable manifestation of such phenomenology, although the differences in the network architecture must be taken into account for a direct comparison.

[1] A. M. Turing, I.—Computing machinery and intelligence, Mind **LIX**, 433 (1950).
[2] C. G. Langton, Computation at the edge of chaos: Phase transitions and emergent computation, Physica D **42**, 12 (1990).
[3] M. A. Muñoz, Colloquium: Criticality and dynamical scaling in living systems, Rev. Mod. Phys. **90**, 031001 (2018).
[4] J. Bickle, Reducing mind to molecular pathways: Explicating the reductionism implicit in current cellular and molecular neuroscience, Synthese **151**, 411 (2006).
[5] E. R. Kandel, J. D. Koester, S. H. Mack, and S. A. Siegelbaum, *Principles of Neural Science, Sixth Edition* (McGraw–Hill, New York, 2021).
[6] J. M. Beggs, Addressing skepticism of the critical brain hypothesis, Front. Comput. Neurosci. **16**, 703865 (2022).
[7] M. Girardi-Schappo, Brain criticality beyond avalanches: open problems and how to approach them, J. Phys. Complex. **2**, 031003 (2021).
[8] C. Gros, A devil's advocate view on 'self-organized' brain criticality, J. Phys. Complex. **2**, 031001 (2021).
[9] H. Hinrichsen, Non-equilibrium critical phenomena and phase transitions into absorbing states, Adv. Phys. **49**, 815 (2000).
[10] M. Henkel, H. Hinrichsen, and S. Lübeck, *Non-Equilibrium Phase Transitions*, 1st ed. (Springer, Berlin, 2008).
[11] R. Dickman, A. Vespignani, and S. Zapperi, Self-organized criticality as an absorbing-state phase transition, Phys. Rev. E **57**, 5095 (1998).
[12] P. Bak, C. Tang, and K. Wiesenfeld, Self-organized criticality: An explanation of the 1/f noise, Phys. Rev. Lett. **59**, 381 (1987).
[13] J. Hesse and T. Gross, Self-organized criticality as a fundamental property of neural systems, Front. Syst. Neurosci. **8**, 166 (2014).
[14] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham,

N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, Mastering the game of go with deep neural networks and tree search, Nature (London) **529**, 484 (2016).

[15] F. Stahlberg, Neural machine translation: A review, J. Artif. Intell. Res. **69**, 343 (2020).

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)* (IEEE, Los Alamitos, CA, 2022), pp. 10684–10695.

[17] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, in *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2016), Vol. 29.

[18] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, Deep information propagation, in *The Fifth International Conference on Learning Representations* (2017).

[19] L. Xiao, J. Pennington, and S. Schoenholz, Disentangling trainability and generalization in deep neural networks, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, edited by H. D. III and A. Singh (Journal of Machine Learning Research Inc., Cambridge, MA, 2020), Vol. 119, pp. 10462–10472.

[20] S. Peluchetti and S. Favaro, Infinitely deep neural networks as diffusion processes, in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research Vol. 108, edited by S. Chiappa and R. Calandra (Journal of Machine Learning Research Inc., Cambridge, MA, 2020), pp. 1126–1136.

[21] S. Hayou, A. Doucet, and J. Rousseau, The curse of depth in kernel regime, in *Proceedings on "I (Still) Can't Believe It's Not Better!" at NeurIPS 2021 Workshops*, Proceedings of Machine Learning Research Vol. 163, edited by M. F. Pradier, A. Schein, S. Hyland, F. J. R. Ruiz, and J. Z. Forde (Journal of Machine Learning Research Inc., Cambridge, MA, 2022), pp. 41–47.

[22] S. Hayou, A. Doucet, and J. Rousseau, On the impact of the activation function on deep neural networks training, in *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 97, edited by K. Chaudhuri and R. Salakhutdinov (Journal of Machine Learning Research Inc., Cambridge, MA, 2019), pp. 2672–2680.

[23] D. A. Roberts, S. Yaida, and B. Hanin, *The Principles of Deep Learning Theory* (Cambridge University Press, Cambridge, UK, 2022).

[24] L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. Schoenholz, and J. Pennington, Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks, in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 80, edited by J. Dy and A. Krause (Journal of Machine Learning Research Inc., Cambridge, MA, 2018), pp. 5393–5402.

[25] G. Yang and S. Schoenholz, Mean field residual networks: On the edge of chaos, in *Advances in Neural Information Processing Systems* edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2017), Vol. 30.

[26] K. T. Grosvenor and R. Jefferson, The edge of chaos: Quantum field theory and deep neural networks, SciPost Phys. **12**, 081 (2022).

[27] S. Yaida, Non-Gaussian processes and neural networks at finite widths, in *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, Proceedings of Machine Learning Research Vol. 107, edited by J. Lu and R. Ward (Journal of Machine Learning Research Inc., Cambridge, MA, 2020), pp. 165–192.

[28] A. E. Ferdinand and M. E. Fisher, Bounded and inhomogeneous ising models. I. Specific-heat anomaly of a finite lattice, Phys. Rev. **185**, 832 (1969).

[29] H. K. Janssen, B. Schaub, and B. Schmittmann, New universal short-time scaling behaviour of critical relaxation processes, Z. Phys. B **73**, 539 (1989).

[30] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2018), Vol. 31.

[31] J. Sohl-Dickstein, R. Novak, S. S. Schoenholz, and J. Lee, On the infinite width limit of neural networks with a standard parameterization, arXiv:2001.07301.

[32] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, Wide neural networks of any depth evolve as linear models under gradient descent, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2019), Vol. 32.

[33] The standard Lagrange's (prime) notation for differentiation is used throughout the paper.

[34] R. M. Neal, Priors for infinite networks, in *Bayesian Learning for Neural Networks* (Springer New York, New York, 1996), pp. 29–53.

[35] C. Williams, Computing with infinite networks, in *Advances in Neural Information Processing Systems*, edited by M. Mozer, M. Jordan, and T. Petsche (MIT Press, Cambridge, MA, 1996), Vol. 9.

[36] J. Lee, J. Sohl-dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri, Deep neural networks as Gaussian processes, in *The Sixth International Conference on Learning Representations* (2018).

[37] A. G. de G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani, Gaussian process behaviour in wide deep neural networks, in *The Sixth International Conference on Learning Representations* (2018).

[38] V. Privman and M. E. Fisher, Universal critical amplitudes in finite-size scaling, Phys. Rev. B **30**, 322 (1984).

[39] In particular, if we normalize the input so that $q^{(1)}$ matches the fixed point q^* of the mean-field theory (7), then ω equals the proportionality constant between the input cosine distance $\rho^{(0)}$ and the initial condition $\rho^{(1)}$ of the mean-field theory: $\rho^{(1)} = \omega\rho^{(0)}$ with $\omega = 1 - \sigma_b^2/q^*$.

[40] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. Aud. Speech Lang. Process. **19**, 788 (2011).

[41] Y. Cao and Q. Gu, Generalization bounds of stochastic gradient descent for wide and deep neural networks, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2019), Vol. 32.

[42] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang, On exact computation with an infinitely wide neural net, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2019), Vol. 32.

[43] This peculiarity explains why the finite size scaling in the multilayer perceptrons is different from that in the contact process [91] on a complete graph, where one finds the same β and $\nu_{\parallel}$ [Eq. (22)] but the exponent for finite-size scaling (45) is replaced with $-1/2$. If the third-order cumulant remained nonzero, then the leading order for the correction would be an order of $n^{-\frac{1}{2}}$.

[44] J. E. Kolassa, Edgeworth series, in *Series Approximation Methods in Statistics* (Springer New York, New York, 2006), pp. 31–62.

[45] Y. Bahri, B. Hanin, A. Brossollet, V. Erba, C. Keup, R. Pacelli, and J. B. Simon, Les Houches lectures on deep learning at large and infinite width, J. Stat. Mech. (2024) 104012.

[46] B. Hanin, Random fully connected neural networks as perturbatively solvable hierarchies, J. Mach. Learn. Res. **25**, 1 (2024).

[47] J. Huang and H.-T. Yau, Dynamics of deep neural networks and neural tangent hierarchy, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 119, edited by H. Daumé III and A. Singh (Journal of Machine Learning Research Inc., Cambridge, MA, 2020), pp. 4542–4551.

[48] A. Garriga-Alonso, C. E. Rasmussen, and L. Aitchison, Deep convolutional networks as shallow gaussian processes, in *The Seventh International Conference on Learning Representations* (2019).

[49] R. Coutinho and B. Fernandez, Extended symbolic dynamics in bistable CML: Existence and stability of fronts, Physica D **108**, 60 (1997).

[50] B. Fernandez and L. Raymond, Propagating fronts in a bistable coupled map lattice, J. Stat. Phys. **86**, 337 (1997).

[51] K. Kaneko, Period-doubling of kink-antikink patterns, quasiperiodicity in antiferro-like structures and spatial intermittency in coupled logistic lattice: Towards a prelude of a "field theory of chaos," Prog. Theor. Phys. **72**, 480 (1984).

[52] G. Grinstein and M. A. Muñoz, The statistical mechanics of absorbing states, in *Fourth Granada Lectures in Computational Physics*, edited by P. L. Garrido and J. Marro (Springer, Berlin, 1997), pp. 223–270.

[53] S. R. Broadbent and J. M. Hammersley, Percolation processes: I. Crystals and mazes, Math. Proc. Cambr. Philos. Soc. **53**, 629 (1957).

[54] I. Jensen, Low-density series expansions for directed percolation: I. A new efficient algorithm with applications to the square lattice, J. Phys. A: Math. Gen. **32**, 5233 (1999).

[55] Other possible (albeit relatively minor) reasons for the deviation from the scaling collapse include the finite-size effect and deviation of the noise from the asymptotic behavior (which is particularly true if the number c of channels is small).

[56] C. A. Voigt and R. M. Ziff, Epidemic analysis of the second-order transition in the Ziff-Gulari-Barshad surface-reaction model, Phys. Rev. E **56**, R6241(R) (1997).

[57] J. Wang, Z. Zhou, Q. Liu, T. M. Garoni, and Y. Deng, High-precision Monte Carlo study of directed percolation in $(d+1)$ dimensions, Phys. Rev. E **88**, 042102 (2013).

[58] A. Guttmann, Indicators of solvability for lattice models, Discr. Math. **217**, 167 (2000).

[59] Z. Chen, Y. Cao, Q. Gu, and T. Zhang, A generalized neural tangent kernel analysis for two-layer neural networks, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., Red Hook, NY, 2020), Vol. 33, pp. 13363–13373.

[60] P. Ju, X. Lin, and N. Shroff, On the generalization power of overfitted two-layer neural tangent kernel models, in *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, edited by M. Meila and T. Zhang (Journal of Machine Learning Research Inc., Cambridge, MA, 2021), Vol. 139, pp. 5137–5147.

[61] T. Yu and H. Zhu, Hyper-parameter optimization: A review of algorithms and applications, arXiv:2003.05689.

[62] L. Chizat, E. Oyallon, and F. Bach, On lazy training in differentiable programming, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2019), Vol. 32.

[63] E. Nichani, Y. Bai, and J. D. Lee, Identifying good directions to escape the NTK regime and efficiently learn low-degree plus sparse polynomials, in *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., Red Hook, NY, 2022), Vol. 35, pp. 14568–14581.

[64] C. Gervais, L.-P. Boucher, G. M. Villar, U. Lee, and C. Duclos, A scoping review for building a criticality-based conceptual framework of altered states of consciousness, Front. Syst. Neurosci. **17**, 1085902 (2023).

[65] Y. Xu, A. Schneider, R. Wessel, and K. B. Hengen, Sleep restores an optimal computational regime in cortical networks, Nat. Neurosci. **27**, 328 (2024).

[66] N. Friedman, S. Ito, B. A. W. Brinkman, M. Shimono, R. E. L. DeVille, 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).

[67] W. L. Shew, W. P. Clawson, J. Pobst, Y. Karimipanah, N. C. Wright, and R. Wessel, Adaptation to sensory input tunes visual cortex to criticality, Nat. Phys. **11**, 659 (2015).

[68] E. K. Kosmidis, Y. F. Contoyiannis, C. Papatheodoropoulos, and F. K. Diakonos, Traits of criticality in membrane potential fluctuations of pyramidal neurons in the CA1 region of rat hippocampus, Eur. J. Neurosci. **48**, 2343 (2018).

[69] 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. a. Rodrigues, N. Sousa, P. V. Carelli, and M.

Copelli, Criticality between Cortical States, Phys. Rev. Lett. **122**, 208101 (2019).

[70] B. Zheng, Monte Carlo simulations and numerical solutions of short-time critical dynamics, Physica A **283**, 80 (2000).

[71] U. Hasson, J. Chen, and C. J. Honey, Hierarchical process memory: Memory as an integral component of information processing, Trends Cogn. Sci. **19**, 304 (2015).

[72] M. Golesorkhi, J. Gomez-Pilar, F. Zilio, N. Berberian, A. Wolff, M. C. E. Yagoub, and G. Northoff, The brain and its time: Intrinsic neural timescales are key for input processing, Commun. Biol. **4**, 970 (2021).

[73] Namely, (a) derivation of more precise design principles of the neural networks from the universal scaling laws, possibly with prior knowledge about a dataset at hand (Sec. IV B) and (b) uncovering the details of the trade-off between the width and the depth (Sec. IV C).

[74] N. Bertschinger, T. Natschläger, and R. Legenstein, At the edge of chaos: Real-time computations and self-organized criticality in recurrent neural networks, in *Advances in Neural Information Processing Systems*, edited by L. Saul, Y. Weiss, and L. Bottou (MIT Press, Cambridge, MA, 2004), Vol. 17.

[75] S. Takasu and T. Aoyagi, Suppression of chaos in a partially driven recurrent neural network, Phys. Rev. Res. **6**, 013172 (2024).

[76] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)* (IEEE, Los Alamitos, CA, 2016), pp. 770–778.

[77] G. Lan, P. Sartori, S. Neumann, V. Sourjik, and Y. Tu, The energy–speed–accuracy trade-off in sensory adaptation, Nat. Phys. **8**, 422 (2012).

[78] R. Rao and L. Peliti, Thermodynamics of accuracy in kinetic proofreading: Dissipation and efficiency trade-offs, J. Stat. Mech. (2015) P06001.

[79] K. Banerjee, A. B. Kolomeisky, and O. A. Igoshin, Elucidating interplay of speed and accuracy in biological error correction, Proc. Natl. Acad. Sci. USA **114**, 5183 (2017).

[80] T. E. Ouldridge, C. C. Govern, and P. R. ten Wolde, Thermodynamics of computational copying in biochemical systems, Phys. Rev. X **7**, 021004 (2017).

[81] K. Ikeda, T. Uda, D. Okanohara, and S. Ito, Speed-accuracy trade-off for the diffusion models: Wisdom from nonequilibrium thermodynamics and optimal transport, arXiv:2407.04495 [Phys. Rev. X (to be published)].

[82] S. Zeraati, F. H. Jafarpour, and H. Hinrichsen, Entropy production of nonequilibrium steady states with irreversible transitions, J. Stat. Mech. (2012) L12001.

[83] K. Harada and N. Kawashima, Entropy governed by the absorbing state of directed percolation, Phys. Rev. Lett. **123**, 090601 (2019).

[84] https://qsw.phys.s.u-tokyo.ac.jp/.

[85] K. He, X. Zhang, S. Ren, and J. Sun, Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)* (IEEE, Los Alamitos, CA, 2015).

[86] Y. Cho and L. Saul, Kernel methods for deep learning, in *Advances in Neural Information Processing Systems*, edited by Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Curran Associates, Inc., Red Hook, NY, 2009), Vol. 22.

[87] L. Lu, Y. Shin, Y. Su, and G. Em Karniadakis, Dying ReLU and initialization: Theory and numerical examples, Commun. Comput. Phys. **28**, 1671 (2020).

[88] A. L. Maas, A. Y. Hannun, and A. Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing* (WDLASL'13) (2013).

[89] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, Activation functions in deep learning: A comprehensive survey and benchmark, Neurocomputing **503**, 92 (2022).

[90] B. Xu, N. Wang, T. Chen, and M. Li, Empirical evaluation of rectified activations in convolutional network, arXiv:1505.00853.

[91] T. E. Harris, Contact interactions on a lattice, Ann. Probab. **2**, 969 (1974).