Statistical physics of deep neural networks: Initialization toward optimal channels

Kangyu Weng,[1,*] Aohua Cheng,[1,†] Ziyang Zhang,[2,‡] Pei Sun,[3,§] and Yang Tian[3,‖]

[1]*Tsien Excellence in Engineering Program, Tsinghua University, Beijing 100084, China*

[2]*Laboratory of Advanced Computing and Storage, Central Research Institute, 2012 Laboratories, Huawei Technologies Co. Ltd., Beijing 100084, China*

[3]*Department of Psychology & Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing 100084, China*

(Received 6 December 2022; accepted 23 March 2023; published 13 April 2023)

In deep learning, neural networks serve as noisy channels between input data and its latent representation. This perspective naturally relates deep learning with the pursuit of constructing channels with optimal performance in information transmission and representation. While considerable efforts are concentrated on realizing optimal channel properties during network optimization, we study a frequently overlooked possibility that neural networks can be initialized toward optimal channels. Our theory, consistent with experimental validation, identifies primary mechanics underlying this unknown possibility and suggests intrinsic connections between statistical physics and deep learning. Unlike the conventional theories that characterize neural networks applying the classic mean-field approximation, we offer analytic proof that this extensively applied simplification scheme is not appropriate in studying neural networks as information channels. To fill this gap, we develop a restricted mean-field framework applicable for characterizing the limiting behaviors of information propagation in neural networks without strong assumptions on inputs. Based on it, we propose an analytic theory to prove that mutual information maximization is realized between inputs and propagated signals when neural networks are initialized at dynamic isometry, a case where information transmits via norm-preserving mappings. These theoretical predictions are validated by experiments on real neural networks, suggesting the robustness of our theory against finite-size effects. Finally, we analyze our findings with information bottleneck theory to confirm the precise relations among dynamic isometry, mutual information maximization, and optimal channel properties in deep learning. Our work may lay a cornerstone for promoting deep learning in terms of network initialization and suggest general statistical physics mechanisms underlying diverse deep learning techniques.

DOI: 10.1103/PhysRevResearch.5.023023

I. INTRODUCTION

A. Neural networks are information channels

In deep learning, neural networks attempt to identify an optimal latent representation (e.g., a low-dimensional feature space) of the data such that subsequent learning tasks can be solved more efficiently [1]. Below, we first review the latest advances in deep learning theories and then suggest a general perspective to unify them.

Let us consider a sample set $\mathbf{X}$ and an associated learning target set $\mathbf{Y} = \gamma(\mathbf{X})$ (e.g., labels), where γ is a mapping defined by the learning task. A neural network parameterized by ϕ is expected to optimize a representation $\phi(\mathbf{X})$ with $\dim(\phi(\mathbf{X})) < \dim(\mathbf{X})$ (here $\dim(\cdot)$ measures the dimensionality) such that an ideal mapping $\gamma_\phi : \phi(\mathbf{X}) \to \mathbf{Y}$ can be readily learned to solve the task. This objective requires an appropriate evaluation of the optimality of neural network representation $\phi(\mathbf{X})$.

Although the optimality of $\phi(\mathbf{X})$ can be evaluated by diverse metrics according to task demands (e.g., see instances in reinforcement [2], graph [3], and causal [4] representation learning frameworks), a mainstream idea is to consider the neural network as a noisy channel between $\mathbf{X}$ and its representation $\phi(\mathbf{X})$. This perspective naturally leads to the consideration of two cases:

(1) In unsupervised learning, the information of learning target $\mathbf{Y}$ is not known to the neural network [5]. To make the learning task resolvable, mapping γ is assumed as a bijective function from sample $\mathbf{X}$ to target $\mathbf{Y}$ such that the neural network can learn the distribution of $\mathbf{Y}$ by representing the distribution of $\mathbf{X}$ (e.g., the target distribution is exactly an optimal representation of sample distribution in clustering tasks [6,7] and unsupervised representation learning [8,9]). If

*wengky20@mails.tsinghua.edu.cn
†aohuacheng18@gmail.com
‡zhangziyang11@huawei.com
§Corresponding author: peisun@tsinghua.edu.cn
‖Corresponding author: tiany20@mails.tsinghua.edu.cn; Also at Laboratory of Advanced Computing and Storage, Central Research Institute, 2012 Laboratories, Huawei Technologies Co. Ltd., Beijing, 100084, China.

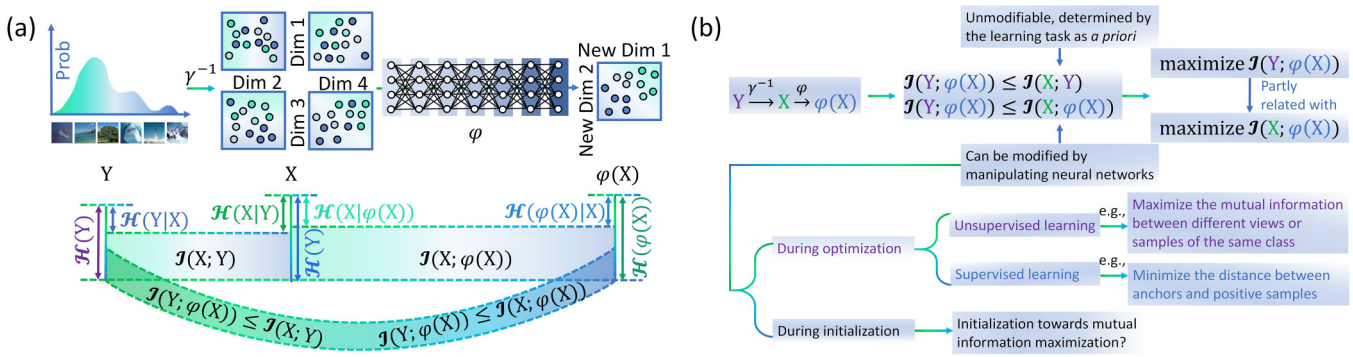

FIG. 1. Conceptual illustrations of research objectives. (a) The learning target $\mathbf{Y}$ (e.g., a distribution of different classes of objects) is represented by a sample set $\mathbf{X}$, where different classes of samples (denoted by nodes with different colors) are promiscuously distributed in the original sample space. A neural network is expected to learn an appropriate neural representation $\phi(\cdot)$ such that all classes are distributed following clear patterns in $\phi(\mathbf{X})$ to optimally capture the information of $\mathbf{Y}$. During this process, a joint channel consisting of two subchannels is defined between $\mathbf{Y}$ and $\phi(\mathbf{X})$. The fist subchannel, bridging between $\mathbf{Y}$ and $\mathbf{X}$, determines how two Shannon entropies, $\mathcal{H}(\mathbf{Y})$ and $\mathcal{H}(\mathbf{X})$, share a common part of information measured by $\mathcal{I}(\mathbf{X};\mathbf{Y})$ given the information loss measured by conditional entropies, $\mathcal{H}(\mathbf{Y}\mid\mathbf{X})$ and $\mathcal{H}(\mathbf{X}\mid\mathbf{Y})$. The second subchannel, bridging between $\mathbf{Y}$ and $\phi(\mathbf{X})$, is optimizable for neural networks. These subchannels jointly define the upper-bound of $\mathcal{I}(\phi(\mathbf{X});\mathbf{Y})$ shown in Eq. (2). (b) Because $\mathcal{I}(\phi(\mathbf{X});\mathbf{Y})$ is unmodifiable for neural networks, deep learning studies primarily explore maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$. This objective is partly related with, not directly equivalent to, maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{Y})$. While previous works mainly focus on maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ during optimization (e.g., training), we shall suggest a possibility to realize this objective during neural network initialization by proposing a statistical physics theory of neural networks.

samples and targets have distinct or irrelevant distributions, then the unsupervised learning of $\mathbf{Y}$ based on $\mathbf{X}$ is ill-posed.

(2) In supervised learning, the information of learning target $\mathbf{Y}$ is known to the neural network as supervision [10]. In an ideal situation where γ is a well-defined bijective mapping from $\mathbf{X}$ to $\mathbf{Y}$ (i.e., samples and targets are perfectly paired), learning the distribution of $\mathbf{Y}$ is principally equivalent to learning the distribution of $\mathbf{X}$ (e.g., consider the separable case where $\mathbf{X}$ can be subdivided into disjoint convex sets in Euclidean space according to the label information in $\mathbf{Y}$ [11,12]). In realistic situations where samples and targets are not perfectly paired, learning the distribution of $\mathbf{Y}$ is nontrivial and not necessarily consistent with representing the distribution of $\mathbf{X}$ (e.g., consider the case where noisy labels exist [13,14]).

B. How can neural networks become optimal channels

How can neural networks become optimal channels favorable for deep learning tasks? This is the central question concerned in our research.

Mathematically, learning tasks are unified by a Markov chain of data processing in information theory [15] (see Fig. 1 for illustration)

$$\mathbf{Y} \xrightarrow{\gamma^{-1}} \mathbf{X} \xrightarrow{\phi} \phi(\mathbf{X}). \tag{1}$$

Note that the above Markov chain is different from $\mathbf{X} \xrightarrow{\phi} \phi(\mathbf{X}) \xrightarrow{\gamma_\phi} \mathbf{Y}$, the Markov chain of hidden variable models [16]. This is because the joint distribution between samples and targets is given as *a priori* rather than an adjustable setting in deep learning (e.g., neural networks cannot modify task designs). The Markov property in Eq. (1) implies zero conditional mutual information values $\mathcal{I}(\mathbf{Y};\phi(\mathbf{X})|\mathbf{X}) = \mathcal{I}(\phi(\mathbf{X});\mathbf{Y}|\mathbf{X}) = 0$ [15], leading to a generalized version of the data processing inequality [15,17,18]

$$\mathcal{I}(\phi(\mathbf{X});\mathbf{Y}) \leqslant \min\{\mathcal{I}(\mathbf{X};\mathbf{Y}), \mathcal{I}(\phi(\mathbf{X});\mathbf{X})\}, \tag{2}$$

which can be readily derived from $\mathcal{I}(\phi(\mathbf{X});\mathbf{Y}) + \mathcal{I}(\mathbf{Y};\mathbf{X}|\phi(\mathbf{X})) = \mathcal{I}(\mathbf{X};\mathbf{Y})$ and $\mathcal{I}(\phi(\mathbf{X});\mathbf{Y}) + \mathcal{I}(\phi(\mathbf{X});\mathbf{X}|\mathbf{Y}) = \mathcal{I}(\mathbf{X};\phi(\mathbf{X}))$ because conditional mutual information is nonnegative.

Equation (2) suggests a clear direction to evaluate the optimality of neural network representation $\phi(\mathbf{X})$. Because the mutual information between samples and targets, denoted by $\mathcal{I}(\mathbf{X};\mathbf{Y})$, is unoptimizable for the neural network, the optimality of $\phi(\mathbf{X})$ is, at least partly, determined by maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$, the mutual information between samples and their represented counterparts.

C. Previous studies on neural networks as optimal channels

Given the possible direction for neural networks to become optimal channels, we review previous efforts that devote to realize this condition during optimization (e.g., training).

In unsupervised learning, mutual information maximization has been demonstrated as a promising approach in recent works [19–22]. Although this approach arises from the idea of maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ in Eq. (2), it differs from the direct maximization because estimating the mutual information between the entire sample set and its neural network representation is statistically deficient [23] (i.e., distribution-free estimators (e.g., statistical bounds) of entropy-related quantities have high sample complexity [24]). In general, what mutual information maximization approach follows is a kind of multiview formulation [25]. Given each canonical sample X of $\mathbf{X}$ (e.g., an image) and its two different and potentially overlapping observations (e.g., two views

of the image), X^i and X^j, neural networks are optimized for $\max_\phi \mathcal{I}(\phi(X^i);\phi(X^j))$. Such an idea can date back to Refs. [26,27] and is valid for optimizing neural networks because $\mathcal{I}(\phi(X^i);\phi(X^j)) \leqslant \mathcal{I}(X;\phi(X^i),\phi(X^j))$ [25], which can be generally used as a lower bound of the mutual information maximization objective $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ [25,27].

In supervised learning, although mutual information is not explicitly used as an optimization objective, it has been discovered as intrinsically related to deep metric learning [25,28]. Let us consider (X, Y, Z), a triplet where X is an anchor, Y is a positive sample (e.g., belongs to the same class as X), and Z is a negative sample (e.g., belongs to a different class compared with X). In deep metric learning, neural networks are optimized to learn a representation ϕ such that $d(\phi(X),\phi(Y)) < d(\phi(X),\phi(Z))$ for any (X, Y, Z), where $d(\cdot,\cdot)$ denotes a distance measure [28,29]. Meanwhile, InfoNCE, an extensively applied lower bound of mutual information [19], can be derived to support maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ in Eq. (2) if all negative samples are drawn from the true marginal distribution [23]. As proven by Ref. [25], InfoNCE can be equivalently reformulated as an expectation of the multiclass n-pair loss [30], which is a standard triplet loss in deep metric learning. In the case where negative samples are not independently drawn, InfoNCE is not valid in estimating mutual information [25].

In sum, existing deep learning approaches, irrespective of being unsupervised or supervised, primarily focus on driving neural networks toward optimal channels during optimization. Maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ is used as, or at least coincides with, a part of the optimization objectives of mainstream frameworks.

D. Open questions on neural networks as optimal channels

Given the review and unified formalization presented above, one may expect that the properties of neural networks as optimal channels have been completely confirmed. However, this is generally not true because the studies on neural networks as optimal channels still remain at their early stages. Below, we summarize three critical open questions in this direction:

(I) Is it possible to maximize $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ during neural network initialization such that the subsequent training process can be improved or at least be accelerated?

(II) If the properties of neural networks as optimal channels are determined by the dynamics of information propagation within neural networks, then is it possible to bridge between the static framework of Shannon theory [15] and the dynamic characterization of information propagation via existing statistical physics theories of neural networks (e.g., mean-field approximation [31,32] and neural tangent kernel [32–34] theories)? If it is not possible, then what are the main limitations of existing theories and how to resolve them?

(III) What is the precise relation between maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ during initialization and driving neural networks toward optimal channels? Can we further relate this relation with other deep learning theories (e.g., information bottleneck [35–37]) to explore a unified view?

E. Our framework and contribution

Motivated by these open questions, we attempt to develop general theories to verify the possibility of initializing neural networks at optimal states for maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$, based on which, we aim at exploring the underlying connections between information theory and statistical physics in deep learning. Specifically, our framework and contributions are summarized as the following.

In Sec. II, we review and unify existing works about characterizing neural networks as information channels and analyzing information propagation within them [see Fig. 2(a)]. To make analytic derivations possible, our work primarily focuses on infinite-width neural networks, whose formal definitions are presented in Sec. II A. Infinite-width neural networks, irrespective of being optimized in a Bayesian manner [38–40] or by gradient descent approaches [33,41,42], are favorable in analytic formulations because they become Gaussian processes defined with specific kernels (e.g., neural tangent kernel [32–34]) at the limits. This property holds across different network architectures (e.g., fully connected layers [38], convolutional layers [40,43], residual connections [40], and recurrent networks [44]), enabling our theory to analyze mainstream deep learning models. Given an infinite-width neural network, we unify previous studies to formalize how information propagates within the network and analytically measure diverse properties of propagated signals (e.g., the second moment) on each layer in Secs. II A and II B. This formalization relates our analysis with the studies on edge of chaos and dynamic isometry in deep neural networks [45–51], whose formal definitions are presented in Sec. II C. The unified framework lays the foundation of our subsequent analysis.

Given the fundamental definitions in Sec. II, we do not limit ourselves to adapting classic theories completely. On the contrary, we show that the independent and identical assumption held by the classic mean-field approximation of infinite-width neural networks is invalid in analyzing neural networks as information channels in Sec. III [see Fig. 2(a)]. Specifically, the independent and identical assumption is proven to imply an Gaussian distribution of the correlation between propagated signal and its original form in inputs in Sec. III A. In other words, the classic mean-field approximation creates a nonzero probability for the correlation to be larger than 1 or smaller than -1, which contradicts the definition of correlation (i.e., a correlation must belongs to the interval of $[-1, 1]$). Although a correlation quantity can be empirically treated as a constant and may still be valid (e.g., when the constant is located within $[-1, 1]$) if it follows a Gaussian distribution with an infinitesimal variance approaching to 0 or a strictly zero variance, we prove that the correlation measured under the classic mean-field framework dissatisfies these two conditions in Sec. III B. Therefore, the classic mean-field approximation is invalid in correlation measurement even from an empirical perspective.

To resolve the limitation of classic mean-field approximation suggested in Sec. III, we develop a new mean-field-like theory with restricted independent and identical assumption in Sec. IV A [see Fig. 2(b)]. Different from the classic one, our restricted mean-field approximation does not imply a

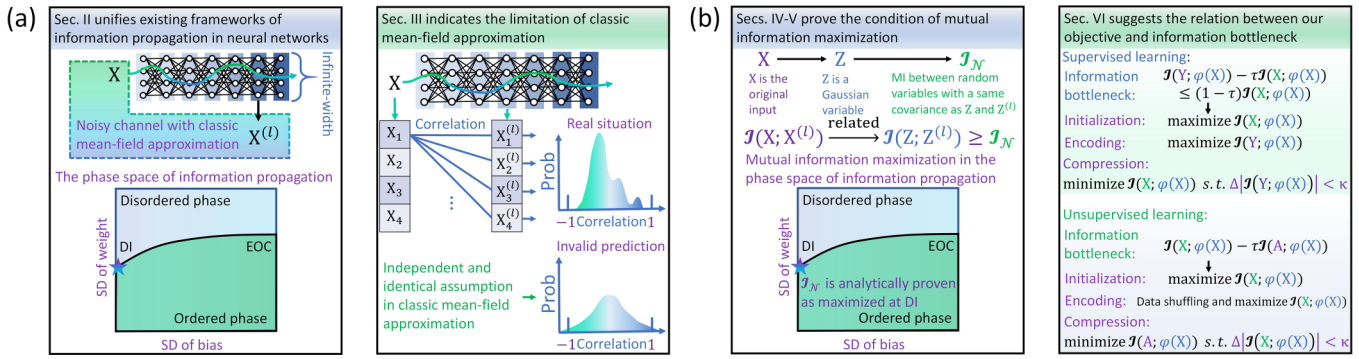

FIG. 2. Summary of our main contributions. (a) In Sec. II, we review and unify classic mean-field theories of neural networks, measure signal moments and correlation, construct the phase space of information propagation, and define edge of chaos (EOC) as well as dynamic isometry (DI). Based on this framework, we indicate the limitation of class mean-field approximation in characterizing neural networks as information channels in Sec. III. This limitation arises from the independent and identical assumption held by class mean-field approximation, which enables the correlation between propagated signal and its original form to be larger than 1 or smaller than -1. (b) In Secs. IV and V, we overcome the limitation of existing framework by proposing a restricted mean-field approximation theory of neural networks. We relate our analysis with Gaussian information bottleneck, based on which, we can derive an important lower bound of mutual information that is analytically proven as maximized at dynamic isometry. Apart from analytic proofs, our theory is also computationally validated on real neural networks. In Sec. VI, we suggest the relation between our objective of maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ and driving neural networks toward optimal channels from the perspective of information bottleneck theory, revealing the insights of our theory on deep learning.

Gaussian distribution of correlation with nonzero constant variance because it excludes the independent and identical assumption on input $\mathbf{X}$. Certainly, the loose constraints held by the restricted mean-field approximation propose challenges to analytic characterization of information channel properties. To overcome these challenges, we introduce Gaussian information bottleneck into our analysis to relate optimizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ with maximizing a certain lower bound of mutual information in Sec. IV B. In the phase space of information propagation, the lower bound is analytically proven as maximized at dynamic isometry point, a case where each layer serves as a random mapping with norm-preserving property during information transmission, in Secs. IV B and IV C. In other situations where dynamic isometry is absent, neural networks become highly noisy channels with high information dissipation rates. In Sec. V, our theory is computationally validated on real neural networks. Although our theory is initially developed on infinite-width neural networks, we demonstrate its general applicability on real deep learning models by showing consistency between our theoretical predictions and empirical observations on finite-width neural networks with diverse settings (e.g., different layer widths, layer quantities, and activation functions).

In Sec. V, we relate our theories with information bottleneck theory [35–37], a special case of rate distortion theory [52] and sufficient statistics theory [53], to present a unified discussion on the role of maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ in driving neural networks toward optimal channels in deep learning [see Fig. 2(b)]. We show that supervised learning and unsupervised learning share similar optimization objectives in terms of information bottleneck. When neural networks are trained with random data shuffling tricks [54,55], we theoretically suggests the possibility that maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ in Eq. (2) serves as a conditional mechanism for neural networks to become optimal in both supervised learning and unsupervised learning. Finally, we discuss the potential insights of our work on deep learning and statistical physics.

II. INFINITE-WIDTH NEURAL NETWORKS AS INFORMATION CHANNELS

In this section, we introduce a framework to characterize infinite-width neural networks as information channels by summarizing and reformulating existing studies on mean-field behaviors and dynamic isometry of neural networks [46,47,49–51,56–58].

A. Mean-field behaviors of infinite-width neural networks

Let us consider an arbitrary deep neural network with multiple layers. The dynamics of cross-layer information propagation (i.e., a signal propagates from the lth layer to the $(l+1)$th layer) is characterized as

$$\mathbf{X}^{(l+1)} = \mathbf{W}^{(l+1)}\psi(\mathbf{X}^{(l)}) + \varepsilon^{(l+1)}, \tag{3}$$

where $\mathbf{X}^{(l)} = (\mathbf{X}_1^{(l)}, \ldots, \mathbf{X}_{N_l}^{(l)}) \in \mathbb{R}^{N_l}$ denotes the vector of pre-activation signals in the lth layer, parameter $N_l \in \mathbb{N}^+$ denotes the width of the lth layer, mapping $\psi(\cdot)$ denotes a nonlinear activation function, matrix $\mathbf{W}^{(l+1)} \in \mathbb{R}^{N_{l+1}} \times \mathbb{R}^{N_l}$ defines the weights of all connections between the lth layer and the $(l+1)$th layer, and $\varepsilon^{(l+1)} = (\varepsilon_1^{(l+1)}, \ldots, \varepsilon_{N_{l+1}}^{(l+1)}) \in \mathbb{R}^{N_{l+1}}$ denotes the associated residuals. In common cases, each residual $\varepsilon_i^{(l+1)}$ is frequently assumed as a Gaussian variable [56,57]. Please see Fig. 3(a) for illustrations.

To offer a clear vision, we begin with formalizing the classic mean-field approximation of the above neural network before we analyze its limitations in Sec. III A. Under the independent and identical assumption of $\mathbf{W}^{(l+1)}$ and $\mathbf{X}^{(l)}$ (e.g., each element $\mathbf{W}_{i,j}^{(l+1)}$ in $\mathbf{W}^{(l+1)}$ is independently and identically distributed), we can derive

$$\mathbf{X}_i^{(l+1)} = \left\langle \mathbf{W}_i^{(l+1)}, \psi\left(\mathbf{X}^{(l)}\right)\right\rangle + \varepsilon_i^{(l+1)} \xrightarrow{d} \mathcal{N}\left(\mu_i, \sigma_i^2\right) \tag{4}$$

as $N_{l+1} \to \infty$, where $\mathbf{X}_i^{(l+1)}$ and $\mathbf{W}_i^{(l+1)}$, respectively, denote the ith rows of $\mathbf{X}^{(l+1)}$ and $\mathbf{W}^{(l+1)}$ for any $i \in \{1, \ldots, N_{l+1}\}$,

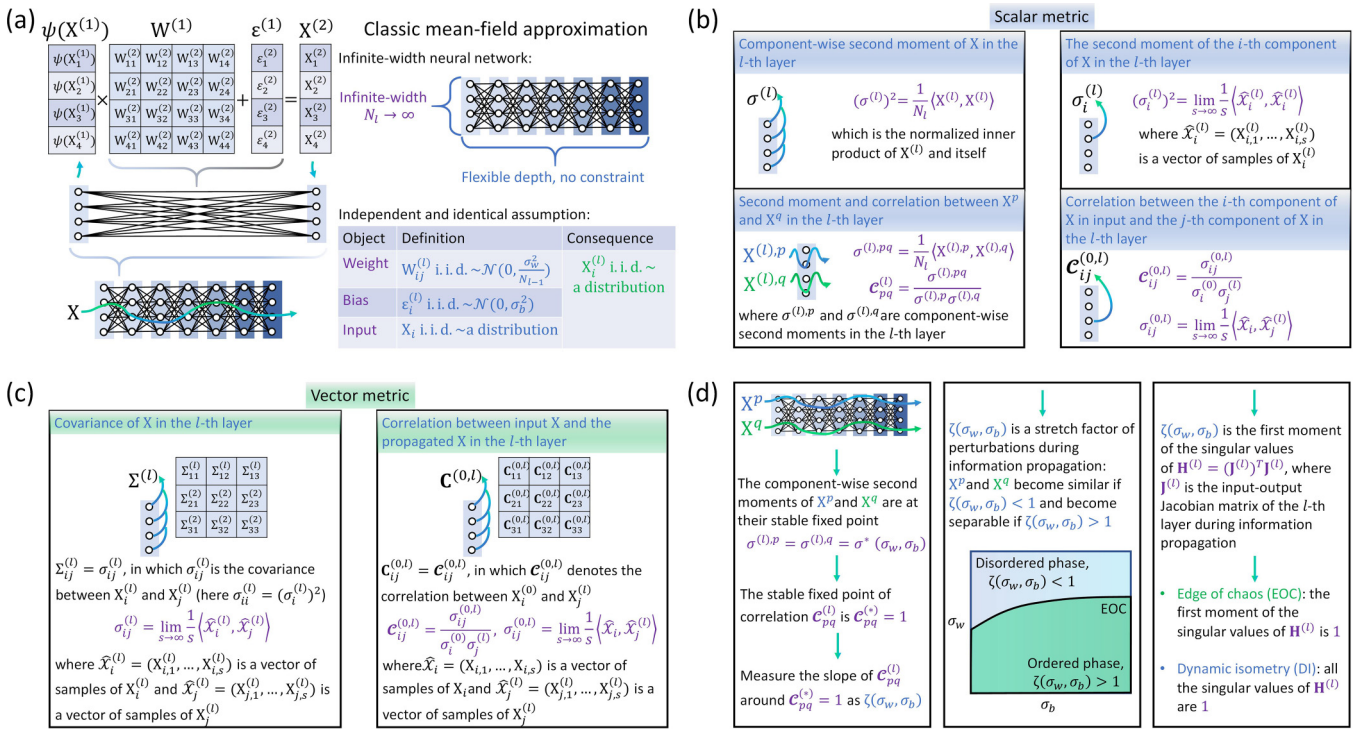

FIG. 3. Conceptual illustrations of the classic mean-field approximation of neural networks. (a) Key settings of classic mean-field approximation, such as infinite-width condition as well as independent and identical assumption, are summarized. (b, c) Illustrations of the scalar and vector metrics used in characterizing information propagation processes are presented. (d) Main steps of the derivation processes of edge of chaos (EOC) and dynamic isometry (ID) in the phase space of information propagation are shown.

notion $\langle \cdot, \cdot \rangle$ defines the inner product, and $\mathcal{N}(\mu_i, \sigma_i)$ defines a specific Gaussian variable.

In general, Eq. (4) approximates each signal propagating in the infinite-width neural network as a certain Gaussian variable under the central limit theorem, enabling us to study an ensemble of random neural networks associated with the original neural network [58] [see Fig. 3(a)]. This idea has been demonstrated as effective in characterizing the mean-field behaviors of two-layer [56] and multilayer [57] neural networks.

B. Information propagation in infinite-width neural networks

As we have explained above, our motivation to consider infinite-width neural networks is to analytically study information propagation dynamics within them. Because the signal propagating across layers in an infinite-width neural network has become a Gaussian variable, we can capture most of its dynamics by studying the first two moments of it. In our research, we primarily focus on the second moment since it has been demonstrated as relevant with the expressivity of neural networks (i.e., the second moment of a signal coincides with the length of its internal Riemannian manifold in downstream layers) [58], enabling us to analyze the order-to-chaos expressivity phase transition [58].

For convenience, we consider the case studied by Ref. [58], where each $\mathbf{W}_{ij}^{(l)}$ i.i.d. $\sim \mathcal{N}(0, \frac{\sigma_w^2}{N_{l-1}})$ and each $\varepsilon_i^{(l)}$ i.i.d. $\sim \mathcal{N}(0, \sigma_b^2)$ (note that "i.i.d." stands for being independently and identically distributed). In the infinite-width limit (i.e., $N_l \to \infty$), the second moment (i.e., variance) of signals in the l-layer, denoted by $(\sigma^{(l)})^2$, can be calculated as

$$(\sigma^{(l)})^2 = \frac{1}{N_l} \sum_{i=1}^{N_l} \left(\mathbf{X}_i^{(l)}\right)^2, \tag{5}$$

$$= \mathbb{E}\left(\left\langle \mathbf{W}_i^{(l)}, \psi\left(\mathbf{X}^{(l-1)}\right)\right\rangle + \varepsilon_i^{(l)}\right)^2, \tag{6}$$

where $\mathbb{E}(\cdot)$ denotes the expectation [see Fig. 3(b)]. Note that the second moment in Eq. (5) reduces to the second origin moment because $\mathbb{E}(\mathbf{X}_i^{(l)}) = 0$ holds for each $\mathbf{X}_i^{(l)}$. Equation (6) can be further reformulated as

$$(\sigma^{(l)})^2 = \mathbb{E}\left\{\left[\sum_{j=1}^{N_{l-1}} \mathbf{W}_{ij}^{(l)} \psi\left(\mathbf{X}_j^{(l-1)}\right) + \varepsilon_i^{(l)}\right]^2\right\}, \tag{7}$$

$$= \sum_{j=1}^{N_{l-1}} \mathbb{E}\left[\left(\mathbf{W}_{ij}^{(l)}\right)^2\right] \mathbb{E}\left[\psi\left(\mathbf{X}_j^{(l-1)}\right)^2\right] + \mathbb{E}\left[\left(\varepsilon_i^{(l)}\right)^2\right], \tag{8}$$

$$= \sigma_w^2 \int_{\mathbb{R}} \psi(\sigma^{(l-1)} x)^2 \mathsf{D}x + \sigma_b^2, \tag{9}$$

where Eq. (8) is derived from the fact that $\mathbb{E}(\mathbf{W}_{ij}^{(l)} \mathbf{W}_{ik}^{(l)}) = \mathbb{E}(\mathbf{W}_{ij}^{(l)}) \mathbb{E}(\mathbf{W}_{ik}^{(l)})$ for any $j \neq k$ under the independent and identical assumption. Equation (9) is derived using $\mathbb{E}[\sum_{j=1}^{N_{l-1}} (\mathbf{W}_{ij}^{(l)})^2] = \sum_{j=1}^{N_{l-1}} \mathbb{E}[(\mathbf{W}_{ij}^{(l)})^2] = N_{l-1} \frac{\sigma_w^2}{N_{l-1}} = \sigma_w^2$ (i.e., weights are independently and identically distributed). In Eq. (9), notion $\mathsf{D}x = \frac{\mathrm{d}x}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$ is a standard Gaussian measure, where $x \in \mathbb{R}$. In general, Eq. (9) defines an iterative

dynamic process of the second moment of signals in each layer [58], whose initial condition is

$$(\sigma^{(1)})^2 = \frac{\sigma_w^2}{N_1}\langle \mathbf{X}, \mathbf{X}\rangle + \sigma_b^2, \tag{10}$$

where $\mathbf{X}$ is the input vector to the neural network, which is assumed to obey the independent and identical assumption, i.e., each $\mathbf{X}_i$ i.i.d. $\sim$ a certain distribution, to ensure that $\mathbf{X}_i^{(l)}$ i.i.d. $\sim$ a certain distribution in Eq. (4). One can see Ref. [58] for more analyses of Eq. (9).

Apart from the second moment of a single signal, we can also consider the relation between two propagating signals. For two input signals, $\mathbf{X}^p$ and $\mathbf{X}^q$, correlated with each other because they propagate within the same neural network, we can mark their representations in the l-layer as $\mathbf{X}^{(l),p}$ and $\mathbf{X}^{(l),q}$ to measure their second moment as [see Fig. 3(b) for illustration]

$$\sigma^{(l),pq} = \frac{1}{N_l}\sum_{i=1}^{N_l} \mathbf{X}_i^{(l),p}\mathbf{X}_i^{(l),q}, \tag{11}$$

$$= \mathbb{E}\left\{\left[\sum_{j=1}^{N_{l-1}} \mathbf{W}_{ij}^{(l)}\psi\left(\mathbf{X}_j^{(l-1),p}\right) + \varepsilon_i^{(l)}\right]\left[\sum_{j=1}^{N_{l-1}} \mathbf{W}_{ij}^{(l)}\psi\left(\mathbf{X}_j^{(l-1),q}\right) + \varepsilon_i^{(l)}\right]\right\}, \tag{12}$$

$$= \sum_{j=1}^{N_{l-1}} \mathbb{E}\left[\left(\mathbf{W}_{ij}^{(l)}\right)^2\right]\mathbb{E}\left[\psi\left(\mathbf{X}_j^{(l-1),p}\right)\psi\left(\mathbf{X}_j^{(l-1),q}\right)\right] + \mathbb{E}\left[\left(\varepsilon_i^{(l)}\right)^2\right], \tag{13}$$

$$= \sigma_w^2 \int_{\mathbb{R}}\int_{\mathbb{R}} \psi(\sigma^{(l-1),p}x)\psi\left\{\sigma^{(l-1),q}\left[\mathcal{C}_{pq}^{(l-1)}x + \sqrt{1-\left(\mathcal{C}_{pq}^{(l-1)}\right)^2}y\right]\right\}\mathsf{D}x\mathsf{D}y + \sigma_b^2, \tag{14}$$

where $\sigma^{(l-1),p}$ and $\sigma^{(l-1),q}$ denotes the standard deviations of $\mathbf{X}^p$ and $\mathbf{X}^q$ in the $(l-1)$-layer defined following Eq. (9). Variables x and y are independent standard Gaussian variables. Notion $\mathcal{C}_{pq}^{(l-1)}$ denotes the correlation between $\mathbf{X}^p$ and $\mathbf{X}^q$ in the $(l-1)$ layer. Please note that subscript i in Eqs. (11)–(13) can be eventually dropped in Eq. (14) because there is independent and identical assumption on the components of $\mathbf{X}^p$ and $\mathbf{X}^q$.

C. Phase space of information propagation

Let us contextualize the above mathematical definitions with physics backgrounds. While studying information propagation, we expect to understand how the global extrinsic curvature of latent Riemannian geometry in inputs (i.e., the relation between two input signals $\mathbf{X}^p$ and $\mathbf{X}^q$), a key factor underlying the expressivity of neural networks [58], evolves across layers.

We wonder if the difference between $\mathbf{X}^p$ and $\mathbf{X}^q$ will be principally maintained, enlarged, or reduced during information propagation. To answer this question, we first explore the stable fixed point of $\sigma^{(l)}$ in Eq. (9) as a function of $(\sigma_w, \sigma_b) \in \mathbb{R}\times(0,\infty)$ because the length of each propagating signal in the downstream layer will rapidly converge to this stable fixed point. After confirming the stable fixed point, denoted by $\sigma^*(\sigma_w,\sigma_b)$, we set $\sigma^{(l),p} = \sigma^{(l),q} = \sigma^*(\sigma_w,\sigma_b)$ in Eq. (14) and divide Eq. (14) by $\sigma^*(\sigma_w,\sigma_b)$ to obtain the iterative dynamics of the correlation between $\mathbf{X}^p$ and $\mathbf{X}^q$ [see Fig. 3(b)],

$$\mathcal{C}_{pq}^{(l)} = \frac{1}{\sigma^*(\sigma_w,\sigma_b)^2}\left(\sigma_w^2\int_{\mathbb{R}}\int_{\mathbb{R}} \psi[\sigma^*(\sigma_w,\sigma_b)x]\psi\left\{\sigma^*(\sigma_w,\sigma_b)\left[\mathcal{C}_{pq}^{(l-1)}x + \sqrt{1-\left(\mathcal{C}_{pq}^{(l-1)}\right)^2}y\right]\right\}\mathsf{D}x\mathsf{D}y + \sigma_b^2\right), \tag{15}$$

whose fixed point can be readily found as $\mathcal{C}_{pq}^* = 1$ after direct calculation. The stability of this fixed point, however, cannot be directly confirmed. Therefore, we need to analyze $\zeta(\sigma_w,\sigma_b)$, the slope of $\mathcal{C}_{pq}^{(l)}$ around $\mathcal{C}_{pq}^* = 1$ given a setting (σ_w,σ_b) [see Fig. 3(d)],

$$\zeta(\sigma_w,\sigma_b) = \left.\frac{\partial\mathcal{C}_{pq}^{(l)}}{\partial\mathcal{C}_{pq}^{(l-1)}}\right|_{\mathcal{C}_{pq}^{(l)}=1}, \tag{16}$$

$$= \sigma_w^2\int_{\mathbb{R}} \psi'(\sigma^*(\sigma_w,\sigma_b)x)^2\mathsf{D}x, \tag{17}$$

$$= \frac{1}{N_l}\mathbb{E}\{\mathrm{tr}[(\mathbf{D}^{(l)}\mathbf{W}^{(l)})^T\mathbf{D}^{(l)}\mathbf{W}^{(l)}]\}. \tag{18}$$

In Eq. (17), notion $\psi'(\cdot)$ denotes the derivative function of $\psi(\cdot)$. In Eq. (18), notion $\mathbf{D}^{(l)}$ is a diagonal matrix $\mathbf{D}^{(l)} = \mathrm{diag}([\psi'(\mathbf{X}_1^{(l)}),\ldots,\psi'(\mathbf{X}_{N_l}^{(l)})])$ such that $\mathbf{J}^{(l)} = \mathbf{D}^{(l)}\mathbf{W}^{(l)}$ can be understood as the input-output Jacobian matrix of the lth layer [46,47]. Notion $\mathrm{tr}(\cdot)$ denotes matrix trace. The expectation is calculated by averaging across all possible configurations of $\mathbf{D}^{(l)}\mathbf{W}^{(l)}$ in Eq. (18). As suggested by Eq. (18), parameter $\zeta(\sigma_w,\sigma_b)$ can be understood as a stretch factor because any random perturbation η in the $(l-1)$th layer, $\mathbf{X}^{(l)} + \nu$, implies a subsequent perturbation in the lth layer, $\mathbf{X}^{(l+1)} + \mathbf{J}^{(l-1)}\nu$, with a stretch effect measured by $\zeta(\sigma_w,\sigma_b) = \mathbb{E}(\|\mathbf{J}^{(l-1)}\nu\|_2^2/\|\nu\|_2^2)$ [58]. The effect corresponds to growth if $\zeta(\sigma_w,\sigma_b) > 1$ and corresponds to shrinkage if $\zeta(\sigma_w,\sigma_b) < 1$ [see Fig. 3(d)].

The fixed point $\mathcal{C}^*_{pq} = 1$ is stable when $\zeta(\sigma_w, \sigma_b) < 1$ while it is unstable when $\zeta(\sigma_w, \sigma_b) > 1$ [46,47,58]. In the case where $\mathcal{C}^*_{pq} = 1$ is stable, all possible relations between $\mathbf{X}^p$ and $\mathbf{X}^q$ eventually converge to a strong correlation (i.e., $\mathbf{X}^{(l),p}$ and $\mathbf{X}^{(l),q}$ become increasingly similar as l increases). In the case where $\mathcal{C}^*_{pq} = 1$ is not stable, $\mathbf{X}^p$ and $\mathbf{X}^q$ become increasingly separable as they propagate. Consequently, the condition with $\zeta(\sigma_w, \sigma_b) = 1$ naturally defines a boundary separating between two phases on the plane of (σ_w, σ_b). The first phase, corresponding to the case where signals become separable during information propagation (i.e., $\zeta(\sigma_w, \sigma_b) > 1$), is referred to as the disordered phase. The second phase, corresponding to the case where signals converge to correlated states [i.e., $\zeta(\sigma_w, \sigma_b) < 1$], is the ordered phase. Please see Fig. 3(d) for illustrations of the phase space. As suggested by Refs. [46,49], the ordered and disordered phases correspond to vanishing and exploding gradient problems, respectively.

Apart from defining a phase transition boundary between ordered and disordered phases, we can further consider the dynamic isometry condition, a special point on this boundary [see Fig. 3(d)]. Specifically, we can understand $\zeta(\sigma_w, \sigma_b)$ in Eq. (18) as the second moment of the singular values of $\mathbf{J}^{(l)} = \mathbf{D}^{(l)}\mathbf{W}^{(l)}$ or, equivalently, the first moment of the singular values of $\mathbf{H}^{(l)} = (\mathbf{J}^{(l)})^T\mathbf{J}^{(l)}$,

$$\zeta(\sigma_w, \sigma_b) = \frac{1}{N_l}\sum_{i=1}^{N_l}\theta_i^2 = \frac{1}{N_l}\sum_{i=1}^{N_l}\lambda_i, \tag{19}$$

where $[\theta_1, \ldots, \theta_{N_l}]$ and $[\lambda_1, \ldots, \lambda_{N_l}]$ are the singular values of $\mathbf{J}^{(l)}$ and $\mathbf{H}^{(l)}$, respectively. Dynamic isometry is defined as a case where the singular values of $\mathbf{H}^{(l)}$ not only have a first moment of 1 but also satisfy $\lambda_i = 1$ for each $i \in \{1, \ldots, N_l\}$. To reach the dynamic isometry condition, we can consider a situation where the second moment of the singular values of $\mathbf{H}^{(l)}$ approaches to 0 while the first moment equals 1. As suggested by Ref. [49], free probability theory [59] can be applied to derive the probability distribution of the singular values of $\mathbf{H}^{(l)}$ to realize this objective. Detailed calculations of dynamic isometry point $(\sigma_w^\diamond, \sigma_b^\diamond)$ across different activation functions $\psi(\cdot)$ or network architectures have been provided by Refs. [47,49–51] and we summarize the general method in Appendix A. Based on the method, it has been demonstrated that orthogonal weights, i.e., $(\mathbf{W}^{(l)})^T\mathbf{W}^{(l)} = \mathbf{I}$ for each l (notion $\mathbf{I}$ denotes the identity matrix), and a non-ReLU-type activation function $\psi(\cdot)$, i.e., $\psi(r) \neq \max(0, r)$ for each $r \in \mathbb{R}$, can achieve dynamical isometry in neural networks [47,49].

Till now, we have reviewed and unified existing studies on mean-field behaviors and dynamic isometry of neural networks [46,47,49–51,56–58] to present a general framework for analyzing information propagation in infinite-width neural networks. This framework supports us to rethink the limitation of existing theories and explore undiscovered laws governing neural networks.

III. ON THE LIMITATION OF CLASSIC MEAN-FIELD APPROXIMATION

In this section, we point out the limitation of classic mean-field approximation in characterizing neural networks as information channels.

A. Rethinking the classic mean-field approximation: A strict perspective

Let us rethink the validity of the classic mean-field approximation summarized in Sec. II in defining the channel capacity of neural networks. At the first glance, the rethinking seems to be unnecessary because Sec. II has suggested how neural networks serve as channels where information propagates across layers in a mean-field manner. However, as we suggested below, the independent and identical assumption in classic mean-field theory may imply unexpected errors in correlation measurements.

Correlation $\mathcal{C}$, irrespective of being measured between any pair of variables, should be located within the interval of $[-1, 1]$ [see Fig. 4(a)]. An approximation framework is invalid if it enables a correlation to be larger than 1 or smaller than -1 with a nonzero probability [see Fig. 4(a)]. Different from the correlation between two inputs, $\mathbf{X}^p$ and $\mathbf{X}^q$, in the lth layer defined by Eq. (15), here we consider $\mathcal{C}_{ij}^{(0,l)}$, the correlation between the ith component of an input signal, $\mathbf{X}_i$, and the jth component of propagated signal in the lth layer, $\mathbf{X}_j^{(l)}$, in Eq. (20). Please see Fig. 3(b) for illustration. This correlation reflects how an input evolves during its propagation from the 1-st layer to the lth layer. Specifically, we have

$$\mathcal{C}_{ij}^{(0,l)} = \frac{\mathbb{E}\left\{\left[\mathbf{X}_j^{(l)} - \mathbb{E}\left(\mathbf{X}_j^{(l)}\right)\right][\mathbf{X}_i - \mathbb{E}(\mathbf{X}_i)]\right\}}{\sigma_j^{(l)}\sigma_i^{(0)}}, \tag{20}$$

$$= \frac{1}{\sigma_j^{(l)}\sigma_i^{(0)}}\sum_{k=1}^{N_{l-1}}\left\{\mathbf{W}_{jk}^{(l)}\int_{\mathbb{R}}\int_{\mathbb{R}}\psi\left(\sigma_k^{(l-1)}x_k\right)\sigma_i^{(0)}\left[\mathcal{C}_{ik}^{(0,l-1)}x_k + \sqrt{1-\left(\mathcal{C}_{ik}^{(0,l-1)}\right)^2}y_k\right]\mathsf{D}x_k\mathsf{D}y_k\right\}, \tag{21}$$

$$= \frac{1}{\sigma_j^{(l)}}\sum_{k=1}^{N_{l-1}}\mathbf{W}_{jk}^{(l)}\mathcal{C}_{ik}^{(0,l-1)}\left[\int_{\mathbb{R}}\psi\left(\sigma_k^{(l-1)}z_k\right)z_k\mathsf{D}z_k\right], \tag{22}$$

where $\sigma_i^{(0)}$ denotes the second moment of $\mathbf{X}_i$, the ith component of input signal $\mathbf{X}$, and $\sigma_j^{(l)}$ denotes the second moment of $\mathbf{X}_j^{(l)}$, the jth component of propagated signal $\mathbf{X}^{(l)}$ [see Fig. 3(b) for illustration]. In Eq. (21), every $\mathsf{D}x_k$ and $\mathsf{D}y_k$

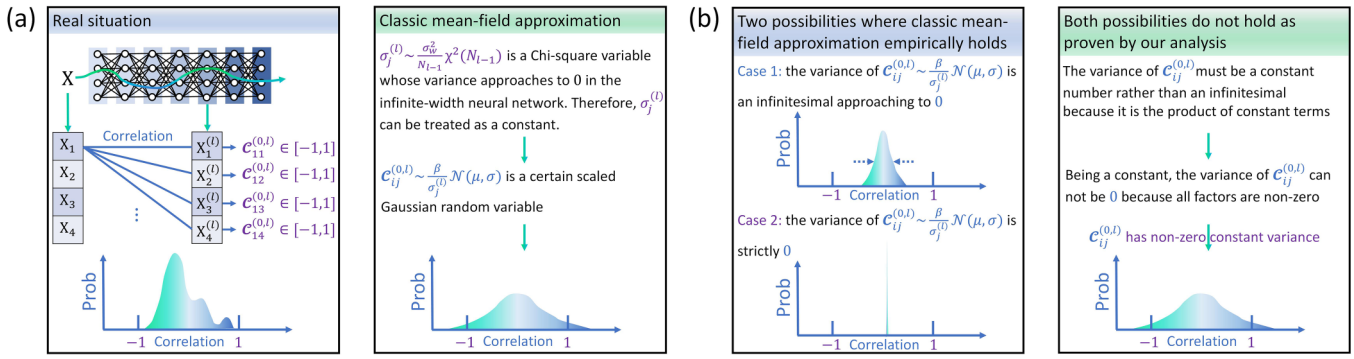

FIG. 4. Conceptual illustrations of the limitation of classic mean-field approximation in correlation measurement. (a) In real cases, the ith component of an input signal, $\mathbf{X}_i$, and the jth component of propagated signal in the lth layer, $\mathbf{X}_j^{(l)}$ should always be located within $[-1, 1]$. However, the independent and identical assumption held by classic mean-field assumption implies a nonzero probability for the measured correlation, $\mathcal{C}_{ij}^{(0,l)}$, to be larger than 1 or small than -1 because $\mathcal{C}_{ij}^{(0,l)}$ follows a specific Gaussian distribution. (b) Although a Gaussian variable with an infinitesimal variance approaching to 0 or a strictly zero variance can be empirically treated as a constant and may be valid to serve as a correlation (e.g., when the constant is located within $[-1, 1]$), the measured correlation $\mathcal{C}_{ij}^{(0,l)}$ is proven to dissatisfy both cases and has a constant nonzero variance. Consequently, it is invalid even from an empirical perspective.

are standard Gaussian measures. Because these measures are identically and independently distributed, their integration over $\mathbb{R}$ can be uniformly represented by the integration of a standard Gaussian measure, $\mathsf{D}z_k$, over $\mathbb{R}$ in Eq. (22).

To verify the possibility for $\mathcal{C}_{ij}^{(0,l)}$ to be larger than 1 or smaller than -1, we can analyze its support set (i.e., if $[-1, 1]$ is a proper subset of the support of $\mathcal{C}_{ij}^{(0,l)}$, then $\mathcal{C}_{ij}^{(0,l)}$ has a nonzero probability to reach an invalid value). As shown below, our analysis is implemented based on two main steps.

First, the second moment term in Eq. (22) is formally measured as [see Fig. 3(b) for illustration]

$$\left(\sigma_j^{(l)}\right)^2 = \mathbb{E}\left\{\left[\sum_{k=1}^{N_{l-1}} \mathbf{W}_{jk}^{(l)} \psi\left(\mathbf{X}_k^{(l-1)}\right) + \varepsilon_j^{(l)}\right]^2\right\} - \mathbb{E}\left[\sum_{k=1}^{N_{l-1}} \mathbf{W}_{jk}^{(l)} \psi\left(\mathbf{X}_k^{(l-1)}\right) + \varepsilon_j^{(l)}\right]^2, \tag{23}$$

which can be reformulated as

$$\left(\sigma_j^{(l)}\right)^2 = \left(\varepsilon_j^{(l)}\right)^2 + \sum_{k=1}^{N_{l-1}} \left(\mathbf{W}_{jk}^{(l)}\right)^2 \mathbb{E}\left[\psi\left(\mathbf{X}_k^{(l-1)}\right)^2\right] - \left(\varepsilon_j^{(l)}\right)^2, \tag{24}$$

$$= \sum_{k=1}^{N_{l-1}} \left(\mathbf{W}_{jk}^{(l)}\right)^2 \mathbb{E}\left[\psi\left(\mathbf{X}_k^{(l-1)}\right)^2\right]. \tag{25}$$

The reformulation holds because of $\mathbb{E}[\psi(\mathbf{X}_k^{(l-1)})] = 0$ (hint: all common non-ReLU-type activation functions in deep learning are odd functions while the probability density of $\mathbf{X}_k^{(l-1)}$ is an even function) and $\mathbb{E}[\psi(\mathbf{X}_a^{(l-1)})\psi(\mathbf{X}_b^{(l-1)})] = \delta_{a,b}$ where $\delta_{\cdot,\cdot}$ denotes the Kronecker δ function (hint: the independent and identical assumption). It is trivial that Eq. (25) is equivalent to

$$\left(\sigma_j^{(l)}\right)^2 = \alpha \sum_{k=1}^{N_{l-1}} \left(\mathbf{W}_{jk}^{(l)}\right)^2 \sim \frac{\sigma_w^2}{N_{l-1}} \chi^2(N_{l-1}), \tag{26}$$

where we define $\alpha = \mathbb{E}[\psi(\mathbf{X}_k^{(l-1)})^2]$ for each k under the independent and identical assumption (i.e., α is same for every k in the $(l-1)$th layer), notion $\chi^2(\cdot)$ denotes the Chi-square random variable [see Fig. 4(a)]. Equation (26) is derived from the independent and identical assumption that $\mathbf{W}_{jk}^{(l)}$ i.i.d. $\sim \mathcal{N}(0, \frac{\sigma_w^2}{N_{l-1}})$. In the infinite-width limit, we discover that the variance of such a Chi-square variable vanishes

$$\lim_{N_{l-1}\to\infty} 2N_{l-1} \frac{\sigma_w^4}{N_{l-1}^2} = 0. \tag{27}$$

In other words, this Chi-square variable can be reasonably treated as a constant in an infinite-width neural network [Fig. 4(a)].

Second, we can define $\beta = \int_{\mathbb{R}} \psi(\sigma_k^{(l-1)} z_k) z_k \mathsf{D}z_k$ for each k under the independent and identical assumption [i.e., β is same for every k in the $(l-1)$th layer], which supports us to rewrite Eq. (22) as

$$\mathcal{C}_{ij}^{(0,l)} = \frac{\beta}{\sigma_j^{(l)}} \sum_{k=1}^{N_{l-1}} \mathbf{W}_{jk}^{(l)} \mathcal{C}_{ik}^{(0,l-1)} \xrightarrow{d} \frac{\beta}{\sigma_j^{(l)}} \mathcal{N}(\mu, \sigma), \tag{28}$$

where the right part is derived based on the central limit theorem (i.e., $\sum_{k=1}^{N_{l-1}} \mathbf{W}_{jk}^{(l)} \mathcal{C}_{ik}^{(0,l-1)} \xrightarrow{d} \mathcal{N}(\mu, \sigma)$). As suggested by Eq. (28), $\mathcal{C}_{ij}^{(0,l)}$ is a Gaussian random variable. Therefore, interval $[-1, 1]$ is always a proper subset of the support of $\mathcal{C}_{ij}^{(0,l)}$, suggesting the limitation of the independent and identical assumption held by classic mean-field approximation [see Fig. 4(a) for illustration].

B. Rethinking the classic mean-field approximation: An empirical perspective

Certainly, one can still treat the independent and identical assumption as reasonably valid if the variance of $\mathcal{C}_{ij}^{(0,l)}$ equals 0 or becomes an infinitesimal approaching to 0. In these cases, correlation $\mathcal{C}_{ij}^{(0,l)}$ can be generally analyzed as a constant from an empirical perspective [see Fig. 4(b)].

However, as we suggest below, correlation $\mathcal{C}_{ij}^{(0,l)}$ intrinsically features a nonzero and finite (i.e., not being infinitesimal) variance in deep neural networks. Before our

formal explanations, we first note that any input $\mathbf{X}$ (e.g., data) of deep neural networks should have a finite dimension even though we apply classic mean-field approximation to consider infinite-width neural networks. This is because the dimensionality of $\mathbf{X}$ is determined by the learning task itself as *a priori* and should not be tampered. Given this property, let us consider a case where each component of $\mathbf{X}$ is independently and identically distributed (i.e., uncorrelated),

$$\mathcal{C}(\mathbf{X}_i, \mathbf{X}_j) := \mathcal{C}_{ij}^{(0,0)} = \delta_{i,j}. \tag{29}$$

By simple calculation based on Eqs. (20)–(22) and Eq. (29), we can derive the correlation between the ith component of an input signal, $\mathbf{X}_i$, and the jth component of propagated signal in the 1th layer,

$$\mathcal{C}_{ij}^{(0,1)} = \frac{1}{\sigma_j^{(1)}} \mathbf{W}_{ij}^{(1)} \left[\int_{\mathbb{R}} \psi(\sigma_i z_i) z_i \mathsf{D}z_i \right], \tag{30}$$

where $\sigma_j^{(1)}$ can be further calculated following Eq. (25)

$$\mathcal{C}_{ij}^{(0,1)} = \frac{\mathbf{W}_{ij}^{(1)} \left[\int_{\mathbb{R}} \psi(\sigma_i z_i) z_i \mathsf{D}z_i \right]}{\sum_{k=1}^{N_0} \left(\mathbf{W}_{jk}^{(1)} \right)^2 \mathbb{E}[\psi(\mathbf{X}_k)^2]}. \tag{31}$$

Notion N_0 measures the dimension of input $\mathbf{X}$. Based on Eq. (31) and Eqs. (20)–(22), it is trivial to derive the following variance items:

$$\mathbb{V}\left(\mathcal{C}_{ij}^{(0,1)}\right) = \frac{1}{\mathbb{E}[\psi(\mathbf{X}_k)^2]^2} \mathbb{V}\left[\frac{\mathbf{W}_{ij}^{(1)}}{\sum_{k=1}^{N_0} \left(\mathbf{W}_{jk}^{(1)} \right)^2} \right] \times \left[\int_{\mathbb{R}} \psi(\sigma_i z_i) z_i \mathsf{D}z_i \right]^2 \tag{32}$$

and

$$\mathbb{V}\left(\mathcal{C}_{ij}^{(0,l)}\right) = \frac{\sigma_w^2 \left[\int_{\mathbb{R}} \psi(\sigma_i z_i) z_i \mathsf{D}z_i \right]^2}{\left(\sigma_j^{(l)} \right)^2} \mathbb{V}\left(\mathcal{C}_{ij}^{(0,l-1)}\right), \tag{33}$$

where we momentarily use $\mathbb{V}(\cdot)$ to denote the variance to avoid confusions on mathematical symbols.

Let us primarily prove that $\mathbb{V}(\mathcal{C}_{ij}^{(0,l)})$ in Eq. (33) cannot be an infinitesimal approaching to 0 (i.e., $\mathbb{V}(\mathcal{C}_{ij}^{(0,l)})$ is a finite real number). Because $N_0 \in \mathbb{N}^+$ is finite, we know that each $\mathbf{W}_{ij}^{(1)}$ i.i.d. $\sim \mathcal{N}(0, \frac{\sigma_w^2}{N_0})$ has a finite variance, which further implies that $\mathbb{V}(\frac{\mathbf{W}_{ij}^{(1)}}{\sum_{k=1}^{N_0}(\mathbf{W}_{jk}^{(1)})^2}) = \mathbb{E}(\frac{(\mathbf{W}_{ij}^{(1)})^2}{[\sum_{k=1}^{N_0}(\mathbf{W}_{jk}^{(1)})^2]^2})$ is finite. Because the remaining part of the terms in Eq. (32) equal certain finite real numbers, a finite value of $\mathbb{V}(\frac{\mathbf{W}_{ij}^{(1)}}{\sum_{k=1}^{N_0}(\mathbf{W}_{jk}^{(1)})^2})$ makes $\mathcal{C}_{ij}^{(0,1)}$ intrinsically have a finite variance. According to Eq. (33), this property further makes $\mathbb{V}(\mathcal{C}_{ij}^{(0,l)})$ finite for any $l \in \mathbb{N}^+$. Please see Fig. 4(b) for a summary.

Given that $\mathbb{V}(\mathcal{C}_{ij}^{(0,l)})$ in Eq. (33) is a finite number, we turn to proving $\mathbb{V}(\mathcal{C}_{ij}^{(0,l)}) \neq 0$. The proof can be readily derived from the following facts. First, term $\frac{1}{\mathbb{E}[\psi(\mathbf{X}_k)^2]^2}$ in Eq. (32) cannot be 0 because the squared output of an appropriate nonlinear activation function $\psi(\cdot)$ cannot be infinite (otherwise deep neural networks inevitably involve with numerical issues). Second, term $\int_{\mathbb{R}} \psi(\sigma_i z_i) z_i \mathsf{D}z_i$ cannot be 0 because the integral of the product of $\psi(\cdot)$ and z_i, two odd functions, over $\mathbb{R}$ must be nonzero. Third, term $\frac{\mathbf{W}_{ij}^{(1)}}{\sum_{k=1}^{N_0}(\mathbf{W}_{jk}^{(1)})^2}$ cannot have a zero variance otherwise all weights in the 1-st layer become the same and immutable. Taken together, we know that $\mathbb{V}(\mathcal{C}_{ij}^{(0,1)}) \neq 0$ always holds in Eq. (32). Based on the iterative dynamics defined in Eq. (33), it is not difficult to see that $\mathbb{V}(\mathcal{C}_{ij}^{(0,l)}) \neq 0$ holds for any $l \in \mathbb{N}^+$ because the coefficient term of $\mathbb{V}(\mathcal{C}_{ij}^{(0,l-1)})$ can be trivially proven as nonzero following a similar idea. For more details, one can see Fig. 4(b) for a summary.

In sum, even from an empirical perspective, the independent and identical assumption is invalid in correlation measurement because $\mathcal{C}_{ij}^{(0,l)}$ is a Gaussian variable with nonzero and finite variance, whose support may include illogical values (i.e., smaller than -1 or larger than 1).

IV. MUTUAL INFORMATION MAXIMIZATION AT DYNAMIC ISOMETRY

In this section, we present our theory on the possibility for neural networks to be initialized toward optimal information channels. To overcome the limitation of classic mean-field approximation, we propose a restricted mean-field approximation framework that does not imply a Gaussian distribution of correlation with nonzero and finite variance. Then, we introduce Gaussian information bottleneck [60,61] into our analysis, which relates our objective of optimizing $\mathcal{I}(\phi(\mathbf{X}); \mathbf{X})$ with maximizing a lower bound of mutual information and supports analytic derivations. Finally, we analytically prove that mutual information between input and propagated signals is maximized at dynamic isometry.

A. Restricted mean-field approximation and Gaussian information bottleneck

As suggested in Sec. III A, the key limitation of classic mean-field approximation arises from the independent and identical assumption applied on all variables without constraints. Although some variables, such as weight and bias, can be assumed as independently and identically distributed [i.e., each $\mathbf{W}_{ij}^{(l)}$ i.i.d. $\sim \mathcal{N}(0, \frac{\sigma_w^2}{N_{l-1}})$ and each $\varepsilon_i^{(l)}$ i.i.d. $\sim \mathcal{N}(0, \sigma_b^2)$], it is less reasonable to apply independent and identical assumption on the components of input $\mathbf{X}$ (i.e., each $\mathbf{X}_i$ i.i.d. $\sim$ a certain distribution). This is because the joint distribution of the components of $\mathbf{X}$ has been defined by the learning task as *a priori* and should not be modified. Meanwhile, this unreasonable independent and identical assumption also makes $\mathbf{X}_i^{(l)}$ i.i.d. $\sim$ a certain distribution for each $l \in \mathbb{N}^+$, which eventually leads to the central limit theorem in Eq. (28) and implies a Gaussian distribution of correlation $\mathcal{C}_{ij}^{(0,l)}$.

Given the above analysis, a natural idea for developing a restricted mean-field approximation is to exclude the independent and identical assumption on the components of $\mathbf{X}$ [see Fig. 5(a)]. In the restricted approximation, there is no constraint on the joint distribution of $\mathbf{X}_i$. Therefore, $\mathbf{X}_i^{(l)}$ in the propagated signal may not be independently and identically distributed as well. A direct consequence of this

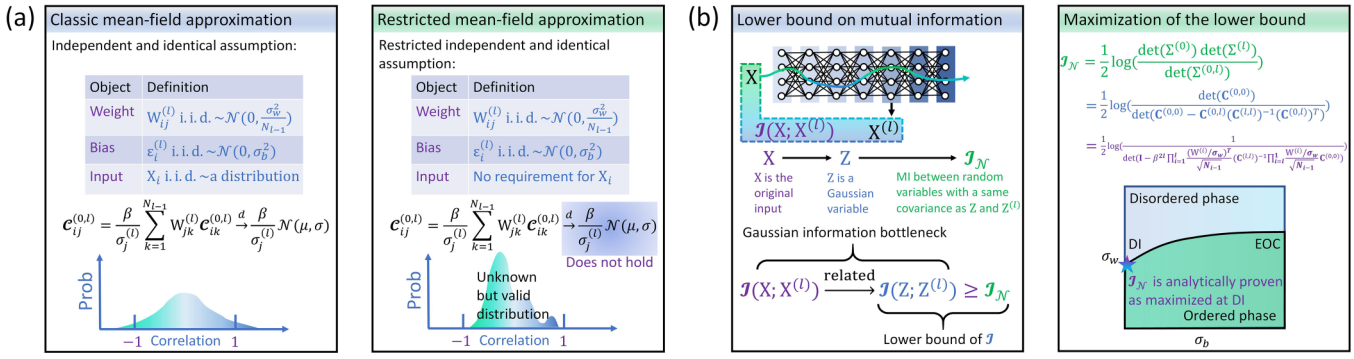

FIG. 5. Conceptual illustrations of the restricted mean-field approximation and the maximization of a lower bound of mutual information in the phase space of information propagation. (a) The restricted mean-field approximation differs from the classic one by excluding the independent and identical assumption on the components of input signal $\mathbf{X}$. This difference makes central limit theorem do not hold and imply a non-Gaussian distribution of correlation $\mathcal{C}_{ij}^{(0,l)}$. (b) Although the restricted mean-field approximation overcomes the limitation of the classic one, its loose constraints make the analytic characterization of information channel properties nearly impossible. To create the possibility of analytic derivations, Gaussian information bottleneck is introduced into our analysis to relate the arbitrarily distributed input signal $\mathbf{X}$ with a multivariate Gaussian variable $\mathbf{Z}$. Maximizing $\mathcal{I}_{\mathcal{N}}$, the lower bound of $\mathcal{I}(\mathbf{Z};\mathbf{Z}^{(l)})$, is closely related to optimizing $\mathcal{I}(\mathbf{X};\mathbf{X}^{(l)})$ according to Gaussian information bottleneck. Our theory analytically proves that $\mathcal{I}_{\mathcal{N}}$ is maximized at dynamic isometry, which is validated by computational experiments on real neural networks as well.

correction lies in that the central limit theorem in Eq. (28) no longer holds and the empirical distribution of $\mathcal{C}_{ij}^{(0,l)}$ does not converges to a Gaussian distribution [see Fig. 5(a) for illustration]. Consequently, the restricted mean-field approximation does not suffer from the limitation of the classic one while characterizing neural networks as information channels. Certainly, this correction also proposes critical challenges to analytic derivations because the distributions of $\mathbf{X}_i^{(l)}$ and $\mathcal{C}_{ij}^{(0,l)} = \frac{\beta}{\sigma_j^{(l)}} \sum_{k=1}^{N_{l-1}} \mathbf{W}_{jk}^{(l)} \mathcal{C}_{ik}^{(0,l-1)}$ lack close-form expressions in most general cases.

To create a possibility for analytic derivations, we suggest to include Gaussian information bottleneck [60,61] into our analysis. In general, we can consider a transform that maps $\mathbf{X}$ to an arbitrary Gaussian variable $\mathbf{Z}$ (there is no constraint on the first and second moments of $\mathbf{Z}$). With an appropriate transform, we can principally treat $\mathbf{Z}$ as the "Gaussian part" of $\mathbf{X}$. As suggested by Ref. [60], optimizing the information bottleneck or mutual information associated with $\mathbf{Z}$ will also reflect the corresponding optimization associated with $\mathbf{X}$ [see Fig. 5(b)]. The benefit of such a connection lies in that Gaussian information bottleneck and Gaussian mutual information have analytic expressions and clear mathematical properties to support our analysis [60,61]. Following the idea of Gaussian information bottleneck [60,61], we suggest to empirically consider a Gaussian counterpart $\mathbf{Z}$ of a given $\mathbf{X}$ where $\mathbf{Z}$ is derived following the approach introduced in Ref. [60]. The algorithm proposed by Ref. [60] ensures that the derived $\mathbf{Z}$ is a "Gaussian part" of $\mathbf{X}$ with $\mathcal{I}(\phi(\mathbf{Z});\mathbf{Z}) \leqslant \mathcal{I}(\phi(\mathbf{X});\mathbf{X})$. There is no other constraint on $\mathbf{Z}$. Meanwhile, signal $\mathbf{Z}^{(l)}$ is not required to maintain its Gaussian properties during information propagation in a neural network [i.e., $\phi(\mathbf{Z})$, the output of a neural network, can be an arbitrary variable].

In our work, we do not need to explicitly consider the actual form of $\mathbf{Z}$ derived following Ref. [60]. This is because the following inequality holds irrespective of what detailed properties that $\mathbf{Z}$ features as long as $\mathbf{Z}$ is a Gaussian variable [see Fig. 5(b)],

$$\mathcal{I}(\mathbf{Z};\mathbf{Z}^{(l)}) \geqslant \mathcal{I}_{\mathcal{N}}(\mathbf{Z};\mathbf{Z}^{(l)}), \tag{34}$$

$$= \frac{1}{2}\log\left[\frac{\det(\mathbf{\Sigma}^{(0)})\det(\mathbf{\Sigma}^{(l)})}{\det(\mathbf{\Sigma}^{(0,l)})}\right], \tag{35}$$

where $\Sigma^{(0)}$ and $\Sigma^{(l)}$ denote the covariance matrix of $\mathbf{Z}$ and $\mathbf{Z}^{(l)}$, respectively (i.e., each element in the matrix denotes the covariance between a pair of components of signals). Matrix $\Sigma^{(0,l)}$ is the covariance matrix measured between the components of $\mathbf{Z}$ and $\mathbf{Z}^{(l)}$, which is a direct generalization of $\Sigma^{(0)}$ and $\Sigma^{(l)}$. Please see Fig. 3(c) for illustrations. Notion $\mathcal{I}_{\mathcal{N}}(\mathbf{Z};\mathbf{Z}^{(l)})$ denotes a special case of mutual information where $\mathbf{Z}$ and $\mathbf{Z}^{(l)}$ are jointly Gaussian while they maintain the original moment properties (i.e., expectation and covariance remain the same). Please see detailed proofs of Eqs. (34) and (35) in Appendix A. Meanwhile, one can find an equivalent version of Eqs. (34) and (35) in Ref. [62].

Let us think about the above process reversely. We can analyze the case where $\mathbf{Z}$ is an arbitrary Gaussian variable to measure mutual information $\mathcal{I}(\phi(\mathbf{Z});\mathbf{Z})$. There always exists a certain input $\mathbf{X}$ that satisfies $\mathcal{I}(\phi(\mathbf{Z});\mathbf{Z}) \leqslant \mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ if we reversely solve the transform problem in Ref. [60]. As suggested by Eqs. (34) and (35), we can primarily focus on $\mathcal{I}_{\mathcal{N}}(\mathbf{Z};\mathbf{Z}^{(l)})$, a lower bound of $\mathcal{I}(\phi(\mathbf{Z});\mathbf{Z})$ in analysis because it has a closed-form expression. Once $\mathcal{I}(\phi(\mathbf{Z});\mathbf{Z})$ is maximized under specific condition, mutual information terms $\mathcal{I}(\phi(\mathbf{Z});\mathbf{Z})$ and $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ are both maximized because of their lower bound relations [see Fig. 5(b) for illustration].

B. Simultaneous maximization of mutual information and β

Given the importance of the lower bound of mutual information defined in Eqs. (34) and (35), we begin our formal analysis on it at first. To relate our calculations of Eq. (35) with the phase space of information propagation, we can first consider the mathematical connection between covariance matrix

and correlation matrix. In the limit of infinite-width, we can treat the variance of each component $\mathbf{Z}_i^{(l)}$ of the propagated signal $\mathbf{Z}^{(l)}$ as generally similar (i.e., the fluctuation of variance is sufficiently small compared with network width such that $\sigma_k^{(l)}$ principally maintains the same across different k in the lth layer). This simplification supports us to follow the idea underlying Eq. (22) to derive a correlation matrix,

$$\mathbf{C}^{(0,l)} = \beta \mathbf{C}^{(0,l-1)} (\mathbf{W}^{(l)}/\sigma_w)^T, \tag{36}$$

where the (i, j)th element in the matrix measures the correlation between $\mathbf{Z}_i$ and $\mathbf{Z}_j^{(l)}$ [see Fig. 3(c)]. Term β remains the same as its definition in Sec III A,

$$\beta = \sigma_w \int_{\mathbb{R}} \frac{\psi\left(\sigma_k^{(l-1)} z_k\right)}{\sigma_j^{(l)}} z_k \mathsf{D} z_k, \tag{37}$$

$$= \sigma_w \int_{\mathbb{R}} \frac{\psi[\sigma^*(\sigma_w, \sigma_b) z)]}{\sigma^*(\sigma_w, \sigma_b)} z \mathsf{D} z, \tag{38}$$

where Eq. (38) is derived by replacing $\sigma_j^{(l)}$ and $\sigma_k^{(l-1)}$ with the stable fixed point $\sigma^*(\sigma_w, \sigma_b)$ mentioned in Sec. II C. As suggested later in Sec. V, this replacement is reasonable because the convergence rates of $\sigma_j^{(l)}$ and $\sigma_k^{(l-1)}$ to $\sigma^*(\sigma_w, \sigma_b)$ are high. Please note that the subscript k in Eq. (37) can be eventually dropped in Eq. (38), because β is same across different k in the $(l-1)$th layer. Based on Eqs. (36) and (38), we can derive the recursion equation of $\mathbf{C}^{(0,l)}$,

$$\mathbf{C}^{(0,l)} = \beta^l \mathbf{C}^{(0,0)} \prod_{i=1}^{l} (\mathbf{W}^{(i)}/\sigma_w)^T. \tag{39}$$

Our next step is to relate Eq. (39) with the lower bound of mutual information in Eq. (35). Applying the property of the determinant of block matrix, we can reformulate term $\det(\mathbf{\Sigma}^{(0,l)})$ in Eq. (35) as

$$\det\left(\mathbf{\Sigma}^{(0,l)}\right) = \det\left(\mathbf{\Sigma}^{(l)}\right) \det\left[\mathbf{\Sigma}^{(0)} - \mathbf{D}^{(0)} \mathbf{C}^{(0,l)} \times \mathbf{D}^{(l)} \left(\mathbf{\Sigma}^{(l)}\right)^{-1} \mathbf{D}^{(l)} \left(\mathbf{C}^{(0,l)}\right)^T \mathbf{D}^{(0)}\right], \tag{40}$$

where $D^{(l)} = \mathrm{diag}(\mathbf{\Sigma}^{(l)})^{\frac{1}{2}}$ and $D^{(0)} = \mathrm{diag}(\mathbf{\Sigma}^{(0)})^{\frac{1}{2}}$, i.e., their diagonal elements are the standard deviations of signals. Please note that $\mathbf{Z}$ and $\mathbf{Z}^{(l)}$ are jointly Gaussian while calculating $\mathcal{I}_{\mathcal{N}}(\mathbf{Z}; \mathbf{Z}^{(l)})$. Therefore, their covariance matrices, $\mathbf{\Sigma}^{(0)}$ and $\mathbf{\Sigma}^{(l)}$, are invertible in Eq. (40). Then, we can notice that $\mathbf{\Sigma}^{(0)} = \mathbf{D}^{(0)} \mathbf{C}^{(0,0)} \mathbf{D}^{(0)}$ and $\mathbf{\Sigma}^{(l)} = \mathbf{D}^{(l)} \mathbf{C}^{(l,l)} \mathbf{D}^{(l)}$. Therefore, we can reformulate the lower bound of mutual information as

$$\mathcal{I}_{\mathcal{N}}(\mathbf{Z}; \mathbf{Z}^{(l)}) = \frac{1}{2} \log\left(\frac{\det(\mathbf{\Sigma}^{(0)}) \det(\mathbf{\Sigma}^{(l)})}{\det\left(\mathbf{\Sigma}^{(0,l)}\right)}\right), \tag{41}$$

$$= \frac{1}{2} \log\left(\frac{\det(\mathbf{\Sigma}^{(0)})}{\det[\mathbf{\Sigma}^{(0)} - \mathbf{D}^{(0)} \mathbf{C}^{(0,l)} (\mathbf{C}^{(l,l)})^{-1} (\mathbf{C}^{(0,l)})^T \mathbf{D}^{(0)}]}\right), \tag{42}$$

$$= \frac{1}{2} \log\left(\frac{\det(\mathbf{C}^{(0,0)})}{\det[\mathbf{C}^{(0,0)} - \mathbf{C}^{(0,l)} (\mathbf{C}^{(l,l)})^{-1} (\mathbf{C}^{(0,l)})^T]}\right). \tag{43}$$

In Eq. (42), we have replaced $\mathbf{D}^{(l)} (\mathbf{\Sigma}^{(l)})^{-1} \mathbf{D}^{(l)}$ with $(\mathbf{C}^{(l,l)})^{-1}$ for simplification. Equation (43) is obtained by dividing the numerator and denominator of Eq. (42) by $[\det(\mathbf{D}^{(0)})]^2$ simultaneously. In Eq. (43), matrix $\mathbf{C}^{(l,l)}$ is a constant matrix across different layers when signals arrive at their stable states shown in Eq. (38). Matrix $\mathbf{C}^{(0,0)}$ is fully determined by input $\mathbf{Z}$ [see Fig. 6(a)].

After substituting Eq. (39) into Eq. (43) and dividing the numerator and denominator of the derived result by $\mathbf{C}^{(0,0)}$, we can obtain the following equation:

$$\mathcal{I}_{\mathcal{N}}\left(\mathbf{X}; \mathbf{X}^{(l)}\right) = \frac{1}{2} \log\left(\frac{1}{\det\left[\mathbf{I} - \beta^{2l} \prod_{i=1}^{l} (\mathbf{W}^{(i)}/\sigma_w)^T (\mathbf{C}^{(l,l)})^{-1} \prod_{i=l}^{1} \mathbf{W}^{(i)}/\sigma_w \mathbf{C}^{(0,0)}\right]}\right). \tag{44}$$

Please note that we have replaced $(\mathbf{C}^{(0,0)})^T$ with $\mathbf{C}^{(0,0)}$ to improve the readability of Eq. (44) because $\mathbf{C}^{(0,0)}$ is a symmetric matrix. In Eq. (44), term $\prod_{i=1}^{l} \mathbf{W}^{(i)}/\sigma_w$ is completely determined by the type of weight distribution used in neural network initialization. Meanwhile, matrices $\mathbf{C}^{(0,0)}$ and $\mathbf{C}^{(l,l)}$ have been suggested as fully deterministic. Therefore, term $\beta = \beta(\sigma_w, \sigma_b)$ is the only one nontrivial variable in Eq. (44) remaining for analysis.

The above analysis hints us to focus on the possibility that $\mathcal{I}_{\mathcal{N}}(\mathbf{Z}; \mathbf{Z}^{(l)})$ can monotonically increase with $\beta = \beta(\sigma_w, \sigma_b)$ in Eq. (44) [see Fig. 6(a) for illustration]. Below, we present our detailed derivations. For convenience, we define

$$\widetilde{\mathbf{M}} = \prod_{i=1}^{l} (\mathbf{W}^{(i)}/\sigma_w)^T (\mathbf{C}^{(l,l)})^{-1} \prod_{i=l}^{1} \mathbf{W}^{(i)}/\sigma_w. \tag{45}$$

as a shorthand. We notice that we can equivalently verify whether $\det(\mathbf{I} - \beta^{2l} \widetilde{\mathbf{M}} \mathbf{C}^{(0,0)})$ monotonically decreases with β since $\mathcal{I}_{\mathcal{N}}(\mathbf{X}; \mathbf{X}^{(l)})$ monotonically decreases with $\det(\mathbf{I} - \beta^{2l} \widetilde{\mathbf{M}} \mathbf{C}^{(0,0)})$. Because $\mathbf{C}^{(0,0)}$ is a positive definite matrix, we can apply Cholesky factorization on it, i.e., $\mathbf{C}^{(0,0)} = LL^T$ where L is a lower triangular matrix whose diagonal elements are positive. Then, we have

$$\det(\mathbf{I} - \beta^{2l} \widetilde{\mathbf{M}} \mathbf{C}^{(0,0)}) = \det(\mathbf{I} - \beta^{2l} \widetilde{\mathbf{M}} L L^T), \tag{46}$$

$$= \det(L^T (L^T)^{-1} - \beta^{2l} L^T \widetilde{\mathbf{M}} L L^T (L^T)^{-1}), \tag{47}$$

$$= \det(\mathbf{I} - \beta^{2l} L^T \widetilde{\mathbf{M}} L). \tag{48}$$

Given that $\widetilde{\mathbf{M}}$ is a positive semi-definite matrix, we know that $L^T \widetilde{\mathbf{M}} L$ is also positive semi-definite. Let $\{\omega_i | i = 1, \ldots, \dim(\widetilde{\mathbf{M}}), \omega_i > 0\}$ be a set of eigenvalues of matrix

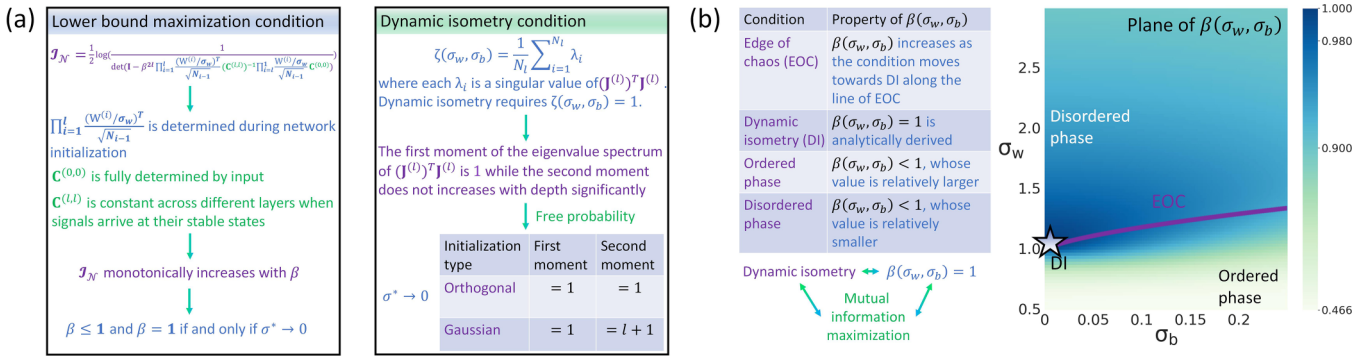

FIG. 6. Conceptual illustrations of the maximization of the lower bound of mutual information at dynamic isometry. (a) Because $\mathcal{I}_{\mathcal{N}}(\mathbf{Z};\mathbf{Z}^{(l)})$, the lower bound of mutual information, monotonically increases with $\beta \leqslant 1$, maximizing $\mathcal{I}_{\mathcal{N}}(\mathbf{Z};\mathbf{Z}^{(l)})$ is equivalent to maximizing $\beta \leqslant 1$, whose condition is analytically derived as $\sigma^* \to 0$. Meanwhile, the condition of dynamic isometry can also be proven as $\sigma^* \to 0$ based on the free probability theory. (b) Consequently, there are equivalent relations among dynamic isometry, lower bound maximization, and $\beta = 1$. As shown in the analytically calculated plane of $\beta(\sigma_w, \sigma_b)$ (the edge of chaos is represented by a purple line while dynamic isometry point is marked by a star), there is $\beta = 1$ at dynamic isometry (DI), suggesting the maximization of $\mathcal{I}_{\mathcal{N}}(\mathbf{Z};\mathbf{Z}^{(l)})$. Here the plane of $\beta(\sigma_w, \sigma_b)$ is calculated based on a commonly used nonlinear activation function, $\tanh(\cdot)$, in deep learning.

$L^T\widetilde{\mathbf{M}}L$, we have

$$\det(\mathbf{I} - \beta^{2l}\widetilde{\mathbf{M}}\mathbf{C}^{(0,0)}) = \prod_{i=1}^{\dim(\widetilde{\mathbf{M}})}(1 - \beta^{2l}\omega_i). \tag{49}$$

Let us assume that the range of β^2 is $[0, \widehat{\beta}^2)$, where 0 stands for zero correlation. To ensure the nonnegativity of $\mathcal{I}_{\mathcal{N}}(\mathbf{X};\mathbf{X}^{(l)})$ (i.e., mutual information cannot be negative), we know that $\det(\mathbf{I} - \beta^{2l}\widetilde{\mathbf{M}}\mathbf{C}^{(0,0)})$, a continuous function with respect to β^{2l}, should be in an interval of $(0, 1]$. According to Eq. (49), this nonnegativity requires that $\omega_i \in (0, \widehat{\beta}^{-2l})$. Meanwhile, to ensure that $\det(\mathbf{I} - \beta^{2l}\widetilde{\mathbf{M}}\mathbf{C}^{(0,0)})$ can monotonically decreases with $\beta = \beta(\sigma_w, \sigma_b)$, we can derive $\omega_i \in (0, \widehat{\beta}^{-2l})$ based on the continuity and nonnegativity. We can immediately find that these two requirements of ω_i are consistent with each other. Therefore, we can prove that $\mathcal{I}_{\mathcal{N}}(\mathbf{Z};\mathbf{Z}^{(l)})$ monotonically increases with $\beta(\sigma_w, \sigma_b)$ [see Fig. 6(a) for a summary].

Given a specific weight distribution and a certain network depth defined for neural network initialization, terms $\mathbf{C}^{(0,0)}$, $\prod_{i=1}^{l}\mathbf{W}^{(i)}/\sigma_w$, and $\mathbf{C}^{(l,l)}$ are principally fixed. Therefore, initializing neural networks for mutual information maximization essentially requires us to maximize $\beta(\sigma_w, \sigma_b)$, whose condition is analyzed in Sec. IV C.

C. Mutual information maximization at dynamic isometry

As we have proved in Sec. IV B, the maximization of the lower bound of mutual information is equivalent to the maximization of $\beta(\sigma_w, \sigma_b)$. Below, we prove that $\beta(\sigma_w, \sigma_b)$ is maximized at dynamic isometry. For convenience, we use β, ζ, and σ^* as the shorthands of $\beta(\sigma_w, \sigma_b)$, $\zeta(\sigma_w, \sigma_b)$, and $\sigma^*(\sigma_w, \sigma_b)$ in our derivations.

Reformulating the one-dimensional integral as two-dimensional integral by Cauchy–Schwarz inequality, we can derive

$$\left[\int_{\mathbb{R}}\psi(\sigma^* z)z\mathsf{D}z\right]^2 = \int_{\mathbb{R}}\int_{\mathbb{R}}\psi(\sigma^* x)\psi(\sigma^* y)xy\mathsf{D}x\mathsf{D}y, \tag{50}$$

$$\leqslant \int_{\mathbb{R}}\int_{\mathbb{R}}\psi(\sigma^* x)^2 y^2\mathsf{D}x\mathsf{D}y, \tag{51}$$

where equality holds only if $\psi(\sigma^* x)^2 \propto x^2$. Because σ^* satisfies

$$(\sigma^*)^2 = \sigma_w^2\int_{\mathbb{R}}\psi(\sigma^* z)^2\mathsf{D}z + \sigma_b^2, \tag{52}$$

we can combine Eq. (52) with Eqs. (38) and (51) to prove $\beta < 1$ when $\sigma^* \neq 0$,

$$\left[\int_{\mathbb{R}}\psi(\sigma^* z)z\mathsf{D}z\right]^2 < \int_{\mathbb{R}}\int_{\mathbb{R}}\psi(\sigma^* x)^2 y^2\mathsf{D}x\mathsf{D}y, \tag{53}$$

$$\sigma_w^2\left[\int_{\mathbb{R}}\psi(\sigma^* z)z\mathsf{D}z\right]^2 < (\sigma^*)^2, \tag{54}$$

$$\beta^2 < 1. \tag{55}$$

As for the situation where $\sigma^* \to 0$, we can prove that $\mathcal{I}_{\mathcal{N}}(\mathbf{X};\mathbf{X}^{(l)})$ is maximized at dynamic isometry by proving $\beta = 1$ under the corresponding condition. As *a priori* knowledge, we can know $\lim_{z\to+\infty}\psi(z) = c \in \mathbb{R}$ (i.e., the integral result is a constant) and $|\psi'(0)| \geqslant |\psi'(z)|$ because $\psi(\cdot)$, an activation function of neural networks, is usually an odd function that is convex in $[0, \infty]$ and satisfies $\lim_{z\to+\infty}\psi(z) = c \in \mathbb{R}$. These properties support us to derive the following proof. First, based on the convex property of $\psi(\cdot)$ and $1 = \int_{\mathbb{R}} z^2\mathsf{D}z$ (i.e., $\mathsf{D}z$ is a Gaussian measure), we can derive

$$(\sigma^{(l)})^2 = \int_{\mathbb{R}}(\sigma^{(l)})^2 z^2\mathsf{D}z \geqslant \sigma_w^2\int_{\mathbb{R}}\psi(\sigma^{(l)}z)^2\mathsf{D}z \tag{56}$$

when $\sigma_w \leqslant \frac{1}{\psi'(0)}$. Here the equality holds only if $\sigma^{(l)} = 0$. Based on Eq. (56), we can see that any $(\sigma^{(l)})^2$ will decrease until it arrives at $(\sigma^*)^2 = 0$ when $\sigma_w \leqslant \frac{1}{\psi'(0)}$ and $\sigma_b = 0$. In other words, point $\sigma^* \to 0$ is a stable fixed point only if $\sigma_w \leqslant$

$\frac{1}{\psi'(0)}$ and $\sigma_b = 0$. Second, given the condition for $\sigma^* \to 0$ to become a stable fixed point, we can further prove that $\beta = 1$ may emerge under this condition. Our derivation utilizes an equality obtained in Appendix B,

$$\psi'(0) = \lim_{\sigma^* \to 0} \int_{\mathbb{R}} \frac{\psi(\sigma^* z)}{\sigma^*} z \mathsf{D}z. \tag{57}$$

Substituting Eq. (57) into Eq. (38), we can see the desired combination of (σ_w, σ_b) for $\beta = 1$ when $\sigma^* \to 0$,

$$\sigma_w = \frac{\beta}{\lim_{\sigma^* \to 0} \int_{\mathbb{R}} \frac{\psi(\sigma^* z)}{\sigma^*(\sigma_w, \sigma_b)} z \mathsf{D}z} = \frac{1}{\psi'(0)}, \tag{58}$$

$$\sigma_b = 0. \tag{59}$$

To this point, we have derived Eqs. (58) and (59) as the sufficient and necessary condition for β to take its maximum, i.e., $\mathcal{I}_{\mathcal{N}}(\mathbf{X}; \mathbf{X}^{(l)})$ takes its maximum, which will be shown as exactly the condition of dynamical isometry in our subsequent analysis.

Although previous studies have addressed the condition of dynamical isometry in linear [63] and nonlinear neural networks [47,49] as suggested in Sec. II C, it remains unclear if it is possible to relate the condition of dynamical isometry with our theory. Below, we present our detailed derivations of dynamic isometry point $(\sigma_w^\diamond, \sigma_b^\diamond)$ based on free probability theory [59] to suggest such a possibility.

The key idea in our derviations is to take advantage of the property of $\mathcal{S}$-transform concerning matrix multiplication [64,65]

$$\mathcal{S}_{UV}(z) = \mathcal{S}_U(z) \mathcal{S}_V(z), \tag{60}$$

where U and V are two freely independent random matrices. Because the Jacobian matrix of the neural network can be defined as

$$\mathbf{J} = \frac{\partial \mathbf{X}^{(l)}}{\partial \mathbf{X}} = \prod_{i=1}^{l} \widetilde{\mathbf{D}}^{(i)} \mathbf{W}^{(i)}, \tag{61}$$

we can derive the $\mathcal{S}$-transform of the Jacobian matrix

$$\mathcal{S}_{\mathbf{J}\mathbf{J}^T}(z) = \mathcal{S}^L_{(\widetilde{\mathbf{D}}^{(i)})^2} \mathcal{S}^L_{(\mathbf{W}^{(i)})^2 \mathbf{W}^{(i)}}(z), \tag{62}$$

where $\widetilde{\mathbf{D}}^{(i)}$ is a diagonal matrix whose diagonal elements are $\widetilde{\mathbf{D}}^{(i)}_{jj} = \psi'(\mathbf{X}^{(i)}_j)$ (please note that the definition is different from the matrix $\mathbf{D}^{(i)}$ analyzed before). In the derivation of Eq. (62), we have applied that $\widetilde{\mathbf{D}}^{(a)} = \widetilde{\mathbf{D}}^{(b)}$ and $\mathbf{W}^{(a)} = \mathbf{W}^{(b)}$ for any pair of (a, b). This property holds because every layer in the neural network shares the same network initialization settings and signals in every layer share the same marginal distribution if they are at the stable fixed point.

Following the idea in Ref. [49], we can derive an implicit equation for eigenvalue spectrum of $\mathbf{H}^{(l)} = (\mathbf{J}^{(l)})^T \mathbf{J}^{(l)}$ based on Eq. (62),

$$M_{\mathbf{H}^{(l)}}(z) = M_{\widetilde{\mathbf{D}}^2} \left\{ z^{\frac{1}{l}} \mathcal{S}_{\mathbf{W}^T \mathbf{W}}[M_{\mathbf{H}^{(l)}}(z)] \left[1 + \frac{1}{M_{\mathbf{H}^{(l)}}(z)} \right]^{1 - \frac{1}{l}} \right\}, \tag{63}$$

where M is a moment generating function. After expanding Eq. (63) in the powers of z^{-1}, the expression of the first two moments of the eigenvalue spectrum of $\mathbf{H}^{(l)}$ can be obtained as

$$m_1 = \left(\sigma_w^2 \tilde{\mu}_1\right)^l, \tag{64}$$

$$m_2 = \left(\sigma_w^2 \tilde{\mu}_1\right)^{2l} \left(\frac{\tilde{\mu}_2}{\tilde{\mu}_1^2} + \frac{1}{l} - 1 - s_1 \right) l, \tag{65}$$

where we define $\tilde{\mu}_k = \int \psi'(\sigma^* z)^{2k} \mathsf{D}z$. Meanwhile, one can notice that m_1 is exactly equivalent to ζ defined in Eqs. (16) and (17). For scaled orthogonal initialization (i.e., weight matrices are initialized as orthogonal random matrices), we have $s_1 = 0$. For scaled Gaussian initialization (i.e., weight matrices are initialized as Gaussian random matrices), we have $s_1 = -1$. More calculation details of s_1 can be seen in Ref. [49].

As suggested in Sec. II C, dynamic isometry requires that the first moment of the eigenvalue spectrum of $\mathbf{H}^{(l)}$ equals 1 while the second moment approaches to 0. In deep neural networks, we reasonably relax the restriction on the second moment and require that the second moment does not increases with network depth l significantly. Applying the Cauchy-Schwarz inequality, we can derive

$$\tilde{\mu}_1^2 = \left(\int \psi'(\sigma^* z)^2 \mathsf{D}z \right)^2 \leqslant \int \psi'(\sigma^* z)^4 \mathsf{D}z = \tilde{\mu}_2, \tag{66}$$

where the equality holds (i.e., $\tilde{\mu}_1^2 = \tilde{\mu}_2$) if

$$\psi'(\sigma^* z)^2 \propto \psi'(\sigma^* z)^4 \Leftrightarrow \sigma^* \to 0. \tag{67}$$

In the case where $m_1 = 1$ and $\tilde{\mu}_1^2 = \tilde{\mu}_2$, we can readily obtain $m_2 = 1$ (i.e., the second moment is a small constant) for orthogonal initialization and $m_2 = l + 1$ (i.e., the second moment increases with network depth linearly instead of exhibiting explosive growth) for Gaussian initialization. Moreover, we can know that $\sigma^* \to 0$ enables $m_1 = 1$ (or equivalently $\zeta = 1$) to imply $\sigma_w = \frac{1}{\psi'(0)}$, which is exactly the condition of the maximization of β (or the maximization of the lower bound of mutual information) defined in Eqs. (58) and (59) [see Fig. 6(a) for a summary].

In sum, we have proven that the maximization of the lower bound of mutual information and dynamic isometry shares the same condition [i.e., Eqs. (58) and (59)]. In other words, mutual information $\mathcal{I}(\mathbf{Z}; \mathbf{Z}^{(l)})$ and its lower bound $\mathcal{I}_{\mathcal{N}}(\mathbf{Z}; \mathbf{Z}^{(l)})$ are maximized at dynamic isometry [see Fig. 6(b)]. Because $\mathbf{Z}$ is a "Gaussian part" of $\mathbf{X}$ with $\mathcal{I}(\mathbf{Z}^{(l)}; \mathbf{Z}) \leqslant \mathcal{I}(\mathbf{X}^{(l)}; \mathbf{X})$ (or equivalently $\mathcal{I}(\phi(\mathbf{Z}); \mathbf{Z}) \leqslant \mathcal{I}(\phi(\mathbf{X}); \mathbf{X})$) as suggested in Sec. IV A, we know that $\mathcal{I}(\mathbf{X}^{(l)}; \mathbf{X})$ is maximized at dynamic isometry in more general cases.

V. EXPERIMENTAL VALIDATIONS

In the previous sections, we have presented our main theory on the equivalence of dynamic isometry and mutual information maximization, which is developed on infinite-width neural networks. It is reasonable to question whether our theory is valid on real finite-width neural networks in deep learning or not. Below, we validate our theory on real neural networks with various settings (e.g., different widths, depths, inputs, and initialization conditions).

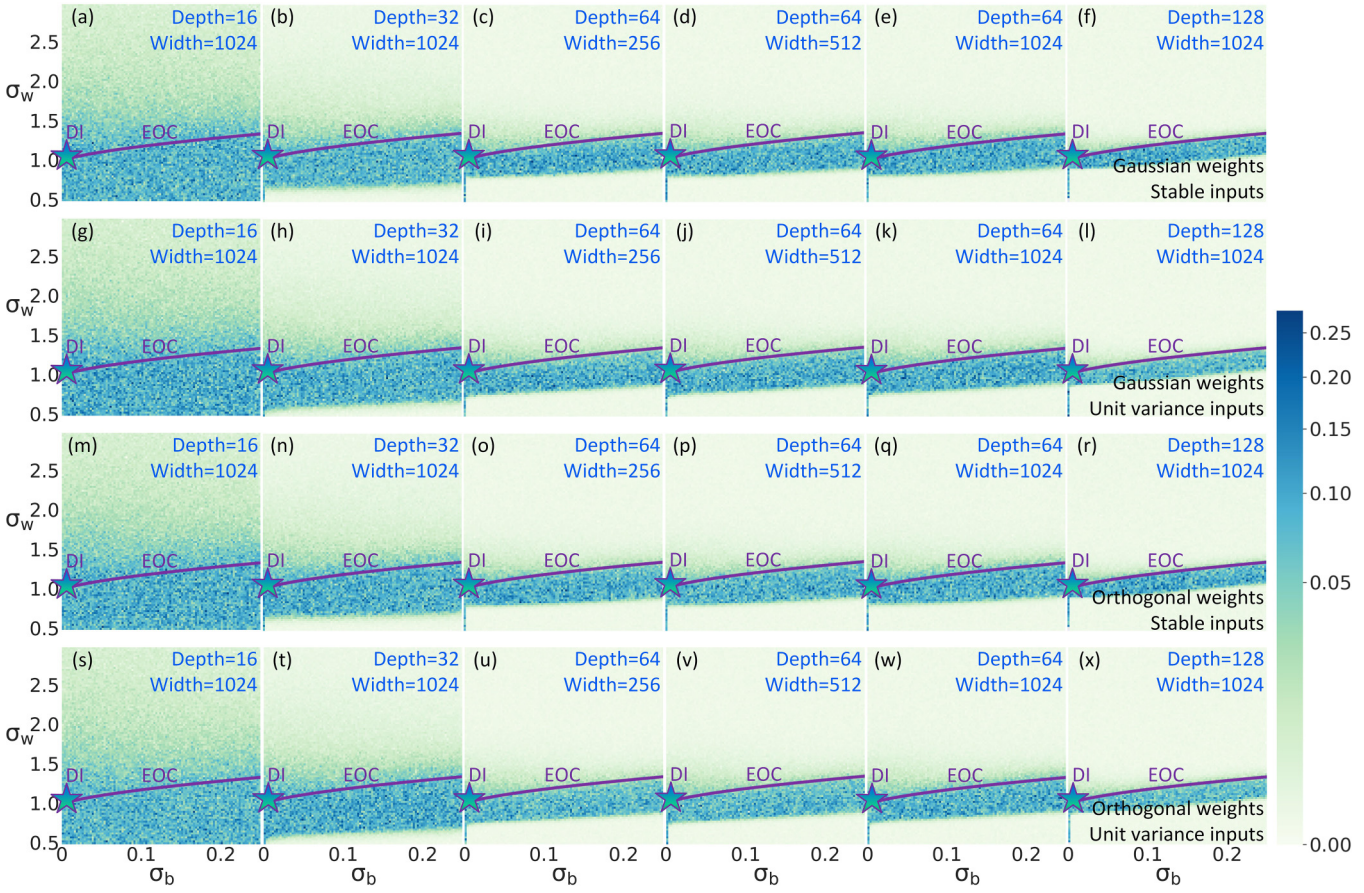

FIG. 7. The plane of the lower bound of mutual information computationally derived on real neural networks. Experiment settings, such as width, depth, input type, and initialization type, are presented along corresponding planes. Same as Fig. 6(b), the edge of chaos (EOC) is marked by a purple line while dynamic isometry is marked by a star.

In Fig. 7, we implement our experiments on finite-width neural networks with a widely applied nonlinear activation function, $\tanh(\cdot)$. To comprehensively verify the robustness of our theory against finite size effects, we design these neural networks with different widths and depths. To suggest the applicability of our theory on more general cases where input $\mathbf{Z}$ may not necessarily propagate at its stable state, we distinguish between stable inputs (i.e., propagating at the stable state as our theoretical derivations require) and unit variance inputs (i.e., with a unit covariance matrix that have not been considered in our previous derivations). To show the capacity of our theory to characterize orthogonal and Gaussian initialization, we conduct experiments under both initialization conditions. In our experiments, we measure the lower bound of mutual information between input $\mathbf{Z}$ and output $\phi(\mathbf{Z})$ (i.e., $\mathbf{Z}^{(l)}$ when l stands for the last layer of the neural network) based on Eq. (35), in which $\mathbf{\Sigma}^{(0)}$, $\mathbf{\Sigma}^{(l)}$, and $\mathbf{\Sigma}^{(0,l)}$ are computationally estimated from the data. Given a combination of width, depth, input type, and initialization type, we repeat our measurement under each condition of (σ_w, σ_b) to obtain a plane of the lower bound of mutual information. To offer a clear vision, we also illustrate the measured lower bound of mutual information along the edge of chaos given orthogonal or Gaussian initialization and stable inputs as two instances (see Fig. 8). As shown in Figs. 7 and 8, our experiment results are consistent with theoretical prediction that the lower bound of mutual information is maximized at the dynamic isometry point $(\sigma_w^\diamond, \sigma_b^\diamond) = (1, 0)$ (note that this point is confirmed by whether $\sigma^* \to 0$). Moreover, the distribution of the lower bound of mutual information corroborates the distribution of β analytically calculated in Fig. 6(b), where both β and the lower bound of mutual information on the edge of chaos increase as the condition moves toward dynamic isometry.

In sum, we have observed consistency between our theory and experiments, which suggests the applicability of our theory on real neural networks in deep learning.

VI. ANALYSIS WITH INFORMATION BOTTLENECK

To this point, we have analytically developed and computationally validated our theory about mutual information maximization at dynamic isometry. As an interdisciplinary attempt, our work not only focuses on the statistical physics of neural networks but also aims at offering insights on deep learning techniques. When $\mathcal{I}_\mathcal{N}(\mathbf{Z}; \mathbf{Z}^{(l)})$, the lower bound of mutual information $\mathcal{I}(\mathbf{Z}; \mathbf{Z}^{(l)})$ is maximized because the neural network is initialized at dynamic isometry, we can know $\mathcal{I}(\mathbf{X}; \mathbf{X}^{(l)})$ is also maximized according to $\mathcal{I}(\mathbf{Z}^{(l)}; \mathbf{Z}) \leqslant \mathcal{I}(\mathbf{X}^{(l)}; \mathbf{X})$ in Sec. IV A. However, we have suggested that maximizing $\mathcal{I}(\mathbf{X}; \mathbf{X}^{(l)})$ is not equivalent to making the neural network an optimal channel in deep learning. Below, we attempt to present a comprehen-

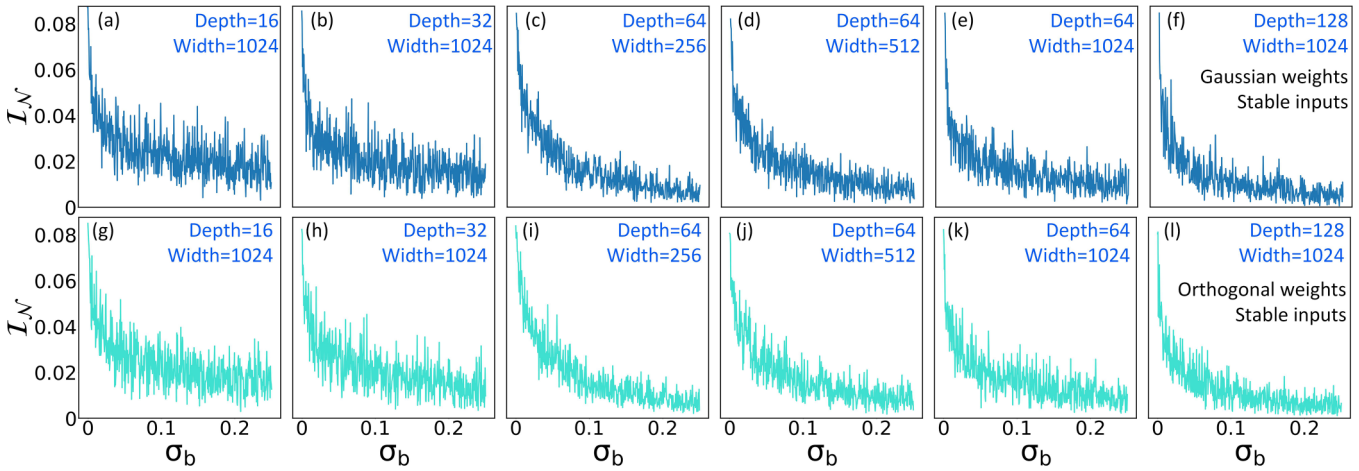

FIG. 8. Two instances of the computationally measured lower bound of mutual information on the edge of chaos. Experiment settings, such as width, depth, input type, and initialization type, are presented. One can see that $\mathcal{I}_{\mathcal{N}}$, the lower bound of mutual information, is maximized near $\sigma_b = 0$, which corresponds to the dynamic isometry point.

sive analysis on the precise relation between maximizing $\mathcal{I}(\mathbf{X};\mathbf{X}^{(l)})$ and driving neural networks toward optimal channels.

Given the difference between unsupervised and supervised learning, we subdivide our analysis into two cases:

(1) In supervised learning, both sample, $\mathbf{X}$, and target, $\mathbf{Y}$, are accessible for the neural network. Therefore, the optimization objective reduces to the classic information bottleneck [35–37], which is actually a special case of rate distortion theory [52] and sufficient statistics theory [53]

$$\max_{\phi} \mathcal{L}_s(\phi) := \underbrace{\mathcal{I}(\phi(\mathbf{X});\mathbf{Y})}_{\text{Encoding}} \underbrace{-\tau\mathcal{I}(\phi(\mathbf{X});\mathbf{X})}_{\text{Compression}}. \tag{68}$$

In general, Eq. (68) defines an objective that the neural network maximizes $\mathcal{I}(\phi(\mathbf{X});\mathbf{Y})$, the capacity to learn $\mathbf{Y}$, during encoding and minimizes the complexity of representation, $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$, during compression. During initialization, we suggest to maximize $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ by initializing the neural network at dynamic isometry because we know $\mathcal{I}(\phi(\mathbf{X});\mathbf{Y}) \leqslant \mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ according to the Markov chain in Eq. (1). This approach can avoid that $\mathcal{I}(\phi(\mathbf{X});\mathbf{Y})$ is bounded by a small value of $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ and cannot be thoroughly optimized during encoding. During compression, we suggest to minimize $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ while controlling the loss of $\mathcal{I}(\phi(\mathbf{X});\mathbf{Y})$ (i.e., ensuring $\Delta|\mathcal{I}(\phi(\mathbf{X});\mathbf{Y})| < \kappa$ where $\kappa \to 0$) to avoid neural network overfitting.

(2) In unsupervised learning, the only information accessible to the neural network is sample $\mathbf{X}$. Therefore, the optimization of neural network toward optimal channel can be implemented following

$$\max_{\phi} \mathcal{L}_u(\phi) := \underbrace{\mathcal{I}(\phi(\mathbf{X});\mathbf{X})}_{\text{Encoding}} \underbrace{-\tau\mathcal{I}(\phi(\mathbf{X});\mathbf{A})}_{\text{Compression}}, \tag{69}$$

where $\mathbf{A} \subset \mathbb{N}^+$ denotes the index set of samples. Parameter $\tau \in (0,\infty)$ denotes a Lagrange multiplier. The objective in Eq. (69) requires the neural network to maximize the encoded information in its representation, $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$, during encoding and reduce the dependence of neural network representation on sample index, $\mathcal{I}(\phi(\mathbf{X});\mathbf{A})$, during compression. This optimization enables the neural network to learn $\mathbf{Y} = \gamma(\mathbf{X})$ under the assumption that sample distribution matches target distribution [e.g., mapping γ is invertible such that $\mathcal{I}(\phi(\mathbf{X});\mathbf{Y})$ can be indirectly optimized through maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$]. An intrinsic difference between Eq. (69) and classic information bottleneck [35–37] lies that there is no strict Markov chain among $\mathbf{A}$, $\mathbf{X}$, $\phi(\mathbf{X})$, and $\mathbf{Y}$ because the definition of index set $\mathbf{A}$ is rather flexible in practice. During initialization, we suggest to maximize $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ by initializing the neural network at dynamic isometry. During encoding, we suggest to continue to maximize $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ and apply random data shuffling, a standard trick in real training processes [54,55], to make the neural network learn samples rather than overfit sample index. During compression, we suggest to minimize $\mathcal{I}(\phi(\mathbf{X});\mathbf{A})$ and ensure $\Delta|\mathcal{I}(\phi(\mathbf{X});\mathbf{X})| < \kappa$ for $\kappa \to 0$.

In sum, although neural network initialization cannot completely determine the performance in subsequent learning tasks, maximizing $\mathcal{I}(\phi(\mathbf{X});\mathbf{X})$ based on initialization at dynamic isometry is beneficial for neural network optimization during the encoding phase. The above analysis may be closely related to the empirically observed benefits of dynamic isometry for neural network training (e.g., for convolutional neural networks [51], feed-forward neural networks [46], and recurrent neural networks [66]).

VII. CONCLUSION

In this research, we have explored a frequently neglected possibility that neural networks can be initialized toward optimal information channels in deep learning.

Compared with prior studies on related topics (e.g., studies on mean-field dynamics [46,47,49–51,56–58] and mutual information maximization [62] in neural networks), our work may contribute to both physics and deep learning in the following aspects (see a summary in Fig. 2). First, we present a unified framework to summarize existing works

concerning the classic mean-field approximation of information propagation in neural networks. Second, we indicate the limitation of classic mean-field approximation in characterizing neural networks as information channels (i.e., the implied unreasonable distribution of the correlation measured between inputs and propagated signals). Third, we propose a restricted mean-field approximation of infinite-width neural networks to overcome the limitation of the classic one. Based on the proposed approximation framework and the mechanism underlying Gaussian information bottleneck, we analytically prove that neural networks can realize mutual information maximization between inputs and outputs when they are initialized at dynamic isometry, a case where neural networks serve as norm-preserving random mappings during information propagation. Although initially proposed for infinite-width neural networks, our theory is successfully validated on real finite-width neural networks. Fourth, we have explored an in-depth analysis on the relation between the mutual information maximization emerged at dynamic isometry and driving neural networks toward optimal channels in deep learning tasks. These contributions may help researchers to study the dynamics (i.e., dynamic isometry and information propagation dynamics) and information (i.e., mutual information and channel optimality) of neural networks jointly rather than separately.

As a preliminary research, there are diverse intriguing details in our work remained for future exploration. Below, we suggest two potential directions.

First, our presented analyses are limited to non-ReLU-type activation functions (e.g., $\tanh(\cdot)$ or the sigmoidal function) while ReLU function and its variants are provisionally excluded from our framework. This is because the ReLU function family has been empirically demonstrated as incapable of dynamic isometry, irrespective of being equipped with Gaussian or orthogonal initialization [47]. However, there exist numerous influential neural network applications built on the ReLU function family (e.g., see instances in Refs. [67–71]), which are nonnegligible for theoretical analyses. Therefore, an important direction for improving our theory is to study the statistical physics underlying the optimal initialization of ReLU neural networks given that they cannot be driven toward dynamic isometry [47] and may not become optimal channels following our present theory.

Second, it may be important to accurately quantify finite-size effects on our results under more general conditions. In our present experiment, we demonstrate the robustness of our theory against finite-size effects by verifying it on real finite-width neural networks with different widths, depths, inputs, and initialization conditions. Among these experiment settings, inputs are chosen as the synthetic signals that either propagate at the stable state (as required by theoretical derivations) or are unstable and defined with unit variances. The reason why we do not use real deep learning data sets (e.g., the ImageNet [72] and the MNIST [73] data sets) as inputs lies in the difficulties of high-dimensional covariance [74–76] and mutual information [77–79] estimations. Estimating probability densities and every concept built on them in high-dimensional spaces can be extremely nontrivial and error-prone [80,81]. Therefore, a direct verification of our theory on the real deep learning data sets whose probability spaces are high-dimensional and sparse may inevitably suffer from the distractions caused by inaccurate estimators (e.g., it is difficult to tell whether the deviations between observed results and theoretical predictions arise from finite size effects, error-prone estimations, or the invalidation of our theory). In the future, developing reliable covariance or mutual information estimators and verifying our theory during real deep learning tasks serve as necessary steps to generalize our theory. We are positive for such a direction because numerous deep learning experiments have offered indirect verification of our theory [46,51,66]. Specifically, on real deep learning data sets (e.g., the MNIST [73] and the CIFAR-10 [71] data sets), the convolutional neural networks [51], feed-forward neural networks [46], and recurrent neural networks [66] initialized at dynamic isometry generally outperform the alternatives initialized under other conditions in image classification and sequence classification tasks (i.e., achieve higher accuracy). Although these empirical studies have not estimated covariance and mutual information due to the same difficulty met by us, their results suggest that initialization at dynamic isometry does help neural networks become optimal during subsequent learning processes.

To conclude, the suggested connection between statistical physics and deep learning in this work may be considered as a starting point for more comprehensive interdisciplinary studies bridging between these two fields.

ACKNOWLEDGMENTS

The authors appreciate Wenqing Wei, who studies at the School of Science and Engineering, Chinese University of Hong Kong (Shenzhen), for discussions about how $\mathcal{I}_{\mathcal{N}}(\mathbf{X};\mathbf{X}^{(l)})$ monotonically increases with $\beta(\sigma_w,\sigma_b)$. Sirui Huang, who studies at the Qiuzhen College, Tsinghua University, is acknowledged for applying the dominated convergence theorem to derive Eq. (B3). This project is supported by the Huawei Innovation Research Program (Grant No. TC20221109044) as well as the Artificial and General Intelligence Research Program of Guo Qiang Research Institute at Tsinghua University (Grant No. 2020GQG1017).

APPENDIX A: CALCULATION OF DYNAMIC ISOMETRY

In this section, we present our proof of Eqs. (34) and (35). One can also see similar derivations in Ref. [62].

Let us consider $f(\mathbf{Z})$, the probability density function of an arbitrary variable, and $g(\mathbf{Z})$, the probability density function of a multivariate Gaussian variable

$$g(\mathbf{Z}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{\Sigma})}} \exp\left\{-\frac{1}{2}[\mathbf{Z}-\mathbb{E}(\mathbf{Z})]^T \mathbf{\Sigma}^{(-1)}[\mathbf{Z}-\mathbb{E}(\mathbf{Z})]\right\}, \tag{A1}$$

where $\mathbf{\Sigma}$ and $\mathbb{E}(\mathbf{Z})$ denote the variance matrix and the mean vector shared by $f(\mathbf{Z})$ and $g(\mathbf{Z})$. Then, we derive

$$\int f(\mathbf{Z}) \log [g(\mathbf{Z})] d\mathbf{Z} = -\frac{1}{2} \log [(2\pi)^n \det(\mathbf{\Sigma})] - \frac{1}{2} \int f(\mathbf{Z}) \log\{[\mathbf{Z} - \mathbb{E}(\mathbf{Z})]^T \mathbf{\Sigma}^{-1} [\mathbf{Z} - \mathbb{E}(\mathbf{Z})]\} d\mathbf{Z}, \tag{A2}$$

$$= -\frac{1}{2} \log [(2\pi)^n \det(\mathbf{\Sigma})] - \frac{\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma})}{2}, \tag{A3}$$

$$= \int g(\mathbf{Z}) \log [g(\mathbf{X})] d\mathbf{Z}. \tag{A4}$$

After replacing $g(\mathbf{Z})$ and $f(\mathbf{Z})$ by $g(\mathbf{Z}; \mathbf{Z}^{(l)})$ and $f(\mathbf{Z}; \mathbf{Z}^{(l)})$, constraining the marginal distribution as $f_{\mathbf{Z}}(\mathbf{Z}) = g_{\mathbf{Z}}(\mathbf{Z})$, and using $f(\mathbf{Z}; \mathbf{Z}^{(l)})$ as the joint distribution of $(\mathbf{Z}; \mathbf{Z}^{(l)})$, we can obtain

$$\mathcal{I}(\mathbf{Z}; \mathbf{Z}^{(l)}) - \mathcal{I}_{\mathcal{N}}(\mathbf{Z}; \mathbf{Z}^{(l)}) = \int f(\mathbf{Z}; \mathbf{Z}^{(l)}) \log \left[\frac{f(\mathbf{Z}; \mathbf{Z}^{(l)})}{f_{\mathbf{Z}}(\mathbf{Z}) f_{\mathbf{Z}^{(l)}}(\mathbf{Z}^{(l)})} \right] d\mathbf{Z} d\mathbf{Z}^{(l)} - \int g(\mathbf{Z}; \mathbf{Z}^{(l)}) \log \left[\frac{g(\mathbf{Z}; \mathbf{Z}^{(l)})}{g_{\mathbf{Z}}(\mathbf{Z}) g_{\mathbf{Z}^{(l)}}(\mathbf{Z}^{(l)})} \right] d\mathbf{Z} d\mathbf{Z}^{(l)}. \tag{A5}$$

Using $f_{\mathbf{Z}}(\mathbf{Z}) = g_{\mathbf{Z}}(\mathbf{Z})$, we can further derive

$$\mathcal{I}(\mathbf{Z}; \mathbf{Z}^{(l)}) - \mathcal{I}_{\mathcal{N}}(\mathbf{Z}; \mathbf{Z}^{(l)}) = -\int f_{\mathbf{Z}^{(l)}}(\mathbf{Z}^{(l)}) \log[f_{\mathbf{Z}^{(l)}}(\mathbf{Z}^{(l)})] d\mathbf{Z}^{(l)} + \int g_{\mathbf{Z}^{(l)}}(\mathbf{Z}^{(l)}) \log[g_{\mathbf{Z}^{(l)}}(\mathbf{Z}^{(l)})] d\mathbf{Z}^{(l)}$$
$$+ \int f(\mathbf{Z}; \mathbf{Z}^{(l)}) \log[f(\mathbf{Z}; \mathbf{Z}^{(l)})] d\mathbf{Z} d\mathbf{Z}^{(l)} - \int g(\mathbf{Z}; \mathbf{Z}^{(l)}) \log[g(\mathbf{Z}; \mathbf{Z}^{(l)})] d\mathbf{Z} d\mathbf{Z}^{(l)}. \tag{A6}$$

Based on Eq. (A4), we can reformulate Eq. (A6) as

$$\mathcal{I}(\mathbf{Z}; \mathbf{Z}^{(l)}) - \mathcal{I}_{\mathcal{N}}(\mathbf{Z}; \mathbf{Z}^{(l)}) = \int f_{\mathbf{Z}^{(l)}}(\mathbf{Z}^{(l)}) \log \left[\frac{g_{\mathbf{Z}^{(l)}}(\mathbf{Z}^{(l)})}{f_{\mathbf{Z}^{(l)}}(\mathbf{Z}^{(l)})} \right] d\mathbf{Z}^{(l)} - \int f(\mathbf{Z}; \mathbf{Z}^{(l)}) \log \left[\frac{g(\mathbf{Z}; \mathbf{Z}^{(l)})}{f(\mathbf{Z}; \mathbf{Z}^{(l)})} \right] d\mathbf{Z} d\mathbf{Z}^{(l)}, \tag{A7}$$

$$= \int f(\mathbf{Z}; \mathbf{Z}^{(l)}) \log \left[\frac{g_{\mathbf{Z}^{(l)}}(\mathbf{Z}^{(l)}) f(\mathbf{Z}; \mathbf{Z}^{(l)})}{f_{\mathbf{Z}^{(l)}}(\mathbf{Z}^{(l)}) g(\mathbf{Z}; \mathbf{Z}^{(l)})} \right] d\mathbf{Z} d\mathbf{Z}^{(l)}, \tag{A8}$$

$$= \int f_{\mathbf{Z}^{(l)}}(\mathbf{Z}^{(l)}) D_{KL}(f_{\mathbf{Z}|\mathbf{Z}^{(l)}} || g_{\mathbf{Z}|\mathbf{Z}^{(l)}}) d\mathbf{Z}^{(l)}, \tag{A9}$$

$$\geqslant 0. \tag{A10}$$

Based on Eq. (A10), Eqs. (34) and (35) in the main text can be proven.

APPENDIX B: NECESSARY DERIVATIONS OF EQ. (57)

In this section, we present our derivations of Eq. (57). Let us reformulate the right side of Eq. (57) as

$$\int_{\mathbb{R}} \frac{\psi(\sigma^* z)}{\sigma^*} z \mathsf{D}z = \int_{\mathbb{R}} \frac{\psi(\sigma^* z) - \psi(0)}{\sigma^* z} z^2 \mathsf{D}z, \tag{B1}$$

where we have used the fact that $\psi(0) = 0$. Because $\psi(\cdot)$ is a odd function that is convex in $[0, +\infty]$, we can know $|\psi'(0)| > |\psi'(z)|$ for any $z \neq 0$. Then, we have

$$\lim_{\sigma^* \to 0} \frac{\psi(\sigma^* z) - \psi(0)}{\sigma^* z} z^2 \mathsf{D}z = \psi'(0). \tag{B2}$$

Based on the dominated convergence theorem [82], we can obtain

$$\lim_{\sigma^* \to 0} \int_{\mathbb{R}} \frac{\psi(\sigma^* z) - \psi(0)}{\sigma^* z} z^2 \mathsf{D}z = \psi'(0) \int_{\mathbb{R}} z^2 \mathsf{D}z = \psi'(0) \tag{B3}$$

based on Eq. (B2), which finishes our derivations on Eq. (57) in the main text.

[1] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798 (2013).

[2] T. Lesort, N. Díaz-Rodríguez, J.-F. Goudou, and D. Filliat, State representation learning for control: An overview, Neural Netw. **108**, 379 (2018).

[3] W. L. Hamilton, R. Ying, and J. Leskovec, Representation learning on graphs: Methods and applications, IEEE Data Eng. Bull. **40**, 52 (2017).

[4] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, Toward causal representation learning, Proc. IEEE **109**, 612 (2021).

[5] H. U. Dike, Y. Zhou, K. K. Deveerasetty, and Q. Wu, Unsupervised learning based on artificial neural network: A review, in *Proceedings of the IEEE International Conference on Cyborg and Bionic Systems (CBS)* (IEEE, Piscataway, NJ, 2018), pp. 322–327.

[6] R. Xu and D. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw. **16**, 645 (2005).

[7] D. Xu and Y. Tian, A comprehensive survey of clustering algorithms, Ann. Data Sci. **2**, 165 (2015).

[8] A. Radford, L. Metz, and S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv:1511.06434 (2015).

[9] N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar, A theoretical analysis of contrastive unsupervised representation learning, in *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California*, edited by K. Chaudhuri and R. Salakhutdinov, Proceedings of Machine Learning Research, Vol. 97 (PMLR, 2019), pp. 5628–5637.

[10] S. B. Kotsiantis, I. Zaharakis, P. Pintelas *et al.*, Supervised machine learning: A review of classification techniques, Emerging Artificial Intelligence Applications in Computer Engineering, **160**, 3 (2007).

[11] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, Berlin, 2009), Vol. 2.

[12] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for Machine Learning* (Cambridge University Press, Cambridge, UK, 2020).

[13] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, Learning from noisy labels with deep neural networks: A survey, IEEE Trans. Neural Netw. Learn. Syst., 1 (2022).

[14] B. Frénay and M. Verleysen, Classification in the presence of label noise: a survey, IEEE Trans. Neural Netw. Learn. Syst. **25**, 845 (2013).

[15] T. M. Cover, *Elements of Information Theory* (John Wiley & Sons, New York, NY, 1999).

[16] O. Shamir, S. Sabato, and N. Tishby, Learning and generalization with the information bottleneck, Theor. Comput. Sci. **411**, 2696 (2010).

[17] W. Kang and S. Ulukus, A new data processing inequality and its applications in distributed source and channel coding, IEEE Trans. Inf. Theory **57**, 56 (2010).

[18] C. Zhou, Q. Zhuang, M. Mattina, and P. N. Whatmough, Strong data processing inequality in neural networks with noisy neurons and its implications, in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)* (IEEE, Piscataway, NJ, 2021), pp. 1170–1175.

[19] Aaron van den Oord, Y. Li, and O. Vinyals, Representation learning with contrastive predictive coding, arXiv:1807.03748 (2018).

[20] O. Henaff, Data-efficient image recognition with contrastive predictive coding, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2020), pp. 4182–4192.

[21] Y. Tian, D. Krishnan, and P. Isola, Contrastive multiview coding, in *Proceedings of the European Conference on Computer Vision* (Springer, Berlin, 2020), pp. 776–794.

[22] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, Learning deep representations by mutual information estimation and maximization, *International Conference on Learning Representations* (OpenReview.net, 2019).

[23] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, On variational bounds of mutual information, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2019), pp. 5171–5180.

[24] D. McAllester and K. Stratos, Formal limitations on the measurement of mutual information, in *Proceedings of the International Conference on Artificial Intelligence and Statistics* (PMLR, 2020), pp. 875–884.

[25] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic, On mutual information maximization for representation learning, *International Conference on Learning Representations* (OpenReview.net, 2020).

[26] S. Becker and G. E. Hinton, Self-organizing neural network that discovers surfaces in random-dot stereograms, Nature **355**, 161 (1992).

[27] R. Linsker, Self-organization in a perceptual network, Computer **21**, 105 (1988).

[28] W. Ge, Deep metric learning with hierarchical triplet loss, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 269–285.

[29] B. Yu and D. Tao, Deep metric learning with tuplet margin loss, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2019), pp. 6490–6499.

[30] K. Sohn, Improved deep metric learning with multiclass n-pair loss objective, in *Proceedings of the 30th Conference on Neural Information Processing Systems, Barcelona, Spain*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Vol. 29 (Curran Associates, Inc., 2016).

[31] Q. Li and H. Sompolinsky, Statistical Mechanics of Deep Linear Neural Networks: The Backpropagating Kernel Renormalization, Phys. Rev. X **11**, 031059 (2021).

[32] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, Statistical mechanics of deep learning, Annu. Rev. Condens. Matter Phys. **11**, 501 (2020).

[33] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Proceedings of the 32nd Conference on Neural Information Processing Systems, Montréal, Canada*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Vol. 31 (Curran Associates, Inc., 2018).

[34] E. Golikov, E. Pokonechnyy, and V. Korviakov, Neural tangent kernel: A survey, arXiv:2208.13614 (2022).

[35] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, in *Proceedings of the 37th Allerton Conference on Communication and Computation* (IEEE, 1999), pp. 368–377.

[36] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, Deep variational information bottleneck, *International Conference on Learning Representations* (OpenReview.net, 2017).

[37] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, beta-VAE: Learning basic visual concepts with a constrained variational framework, in *Proceedings of the International Conference on Learning Representations* (OpenReview.net, 2017).

[38] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, Deep neural networks as Gaussian processes, *International Conference on Learning Representations* (OpenReview.net, 2018).

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

[40] A. Garriga-Alonso, C. E. Rasmussen, and L. Aitchison, Deep convolutional networks as shallow Gaussian processes, *International Conference on Learning Representations* (OpenReview.net, 2019).

[41] 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 *Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, Canada*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Vol. 32 (Curran Associates, Inc., 2019).

[42] L. Chizat, E. Oyallon, and F. Bach, On lazy training in differentiable programming, in *Proceddings of the 33rd Conference on Neural Information Processing Systems, Vancouver, Canada*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Vol. 32 (Curran Associates, Inc., 2019).

[43] R. Novak, L. Xiao, J. Lee, Y. Bahri, G. Yang, J. Hron, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein, Bayesian deep convolutional networks with many channels are Gaussian processes, *International Conference on Learning Representations* (OpenReview.net, 2019).

[44] G. Yang, Wide feedforward or recurrent neural networks of any architecture are Gaussian processes, in *Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, Canada*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Vol. 32 (Curran Associates, Inc., 2019).

[45] J. Sirignano and K. Spiliopoulos, Mean field analysis of deep neural networks, Math. Oper. Res. **47**, 120 (2022).

[46] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, Deep information propagation, *International Conference on Learning Representations* (OpenReview.net, 2017).

[47] J. Pennington, S. Schoenholz, and S. Ganguli, Resurrecting the sigmoid in deep learning through dynamical isometry: Theory and practice, in *Proceedings of the 31st Conferenceon Neural Information Processing Systems, Long Beach, CA, USA*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Vol. 30 (Curran Associates, Inc., 2017).

[48] G. Yang and S. Schoenholz, Mean field residual networks: On the edge of chaos, in *31st Conference on Neural Information Processing Systems,Long Beach, CA, USA*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Vol. 30 (Curran Associates, Inc., 2017).

[49] J. Pennington, S. Schoenholz, and S. Ganguli, The emergence of spectral universality in deep networks, in *Proceedings of the International Conference on Artificial Intelligence and Statistics* (PMLR, 2018), pp. 1924–1932.

[50] M. Chen, J. Pennington, and S. Schoenholz, Dynamical isometry and a mean field theory of rnns: Gating enables signal propagation in recurrent neural networks, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2018), pp. 873–882.

[51] 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 *International Conference on Machine Learning* (PMLR, 2018), pp. 5393–5402.

[52] T. Berger, Rate-distortion theory, *Wiley Encyclopedia of Telecommunications* (Wiley-Interscience, Hoboken, NJ, 2003).

[53] H. J. Kleven, Sufficient statistics revisited, Annu. Rev. Econ. **13**, 515 (2021).

[54] T. T. Nguyen, F. Trahay, J. Domke, A. Drozd, E. Vatai, J. Liao, M. Wahib, and B. Gerofi, Why globally re-shuffle? revisiting data shuffling in large scale deep learning, in *Proceedings of the IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (IEEE, Piscataway, NJ, 2022), pp. 1085–1096.

[55] C. Summers and M. J. Dinneen, Nondeterminism and instability in neural network optimization, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2021), pp. 9913–9922.

[56] S. Mei, T. Misiakiewicz, and A. Montanari, Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit, in *Proceedings of the Conference on Learning Theory* (PMLR, 2019), pp. 2388–2464.

[57] P.-M. Nguyen, Mean field limit of the learning dynamics of multilayer neural networks, arXiv:1902.02880 (2019).

[58] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, in *Proceedings of the 30th Conference on Neural Information Processing Systems, Barcelona, Spain*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Vol. 29 (Curran Associates, Inc., 2016).

[59] J. A. Mingo and R. Speicher, *Free Probability and Random Matrices* (Springer, Berlin, 2017), Vol. 35.

[60] A. Painsky and N. Tishby, Gaussian lower bound for the information bottleneck limit, J. Mach. Learn. Res. **18**, 1 (2018).

[61] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, Information bottleneck for Gaussian variables, in *Advances in Neural Information Processing Systems*, edited by S. Thrun, L. Saul, and B. Schölkopf, Vol. 16 (MIT Press, 2003).

[62] G. Ughi, Studies on neural networks: Information propagation at initialisation and robustness to adversarial examples, Ph.D. thesis, University of Oxford (2022).

[63] A. M. Saxe, J. L. McClelland, and S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, *International Conference on Learning Representations* (OpenReview.net, 2014).

[64] R. Speicher, Multiplicative functions on the lattice of noncrossing partitions and free convolution, Math. Ann. **298**, 611 (1994).

[65] D. V. Voiculescu, K. J. Dykema, and A. Nica, *Free Random Variables*, 1 (American Mathematical Society, Providence, RI, 1992).

[66] D. Gilboa, B. Chang, M. Chen, G. Yang, S. S. Schoenholz, E. H. Chi, and J. Pennington, Dynamical isometry and a mean field theory of lstms and grus, arXiv:1901.08987 (2019).

[67] D. Zou, Y. Cao, D. Zhou, and Q. Gu, Gradient descent optimizes overparameterized deep ReLU networks, Mach. Learn. **109**, 467 (2020).

[68] M. Chen, H. Jiang, W. Liao, and T. Zhao, Efficient approximation of deep ReLU networks for functions on low dimensional manifolds, in *Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, Canada*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Vol. 32 (Curran Associates, Inc., 2019).

[69] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, Mobilenets: Efficient

convolutional neural networks for mobile vision applications, arXiv:1704.04861 (2017).

[70] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, Transformers: State-of-the-art natural language processing, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP* (Association for Computational Linguistics, 2020), pp. 38–45.

[71] 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* (2016), pp. 770–778.

[72] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2009), pp. 248–255.

[73] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE **86**, 2278 (1998).

[74] M. Pourahmadi, *High-dimensional Covariance Estimation: With High-dimensional Data* (John Wiley & Sons, New York, NY, 2013), Vol. 882.

[75] S. Zhou, P. Rütimann, M. Xu, and P. Bühlmann, High-dimensional covariance estimation based on gaussian graphical models, J. Mach. Learn. Res. **12**, 2975 (2011).

[76] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence, Electron. J. Stat. **5**, 935 (2011).

[77] J. Song and S. Ermon, Understanding the limitations of variational mutual information estimators, *International Conference on Learning Representations* (OpenReview.net, 2020).

[78] W. Gao, S. Kannan, S. Oh, and P. Viswanath, Estimating mutual information for discrete-continuous mixtures, in *Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Vol. 30 (Curran Associates, Inc., 2017).

[79] J. Walters-Williams and Y. Li, Estimation of mutual information: A survey, in *Proceedings of the 4th International Conference on Rough Sets and Knowledge Technology (RSKT 2009), Gold Coast, Australia, July 14–16* (Springer, Berlin, 2009), pp. 389–396.

[80] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, Estimating the support of a high-dimensional distribution, Neural Comput. **13**, 1443 (2001).

[81] Z. Wang and D. W. Scott, Nonparametric density estimation for high-dimensional data–algorithms and applications, WIREs Comput. Stat. **11**, e1461 (2019).

[82] E. M. Stein and R. Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces* (Princeton University Press, Princeton, NJ, 2009).