

# Network Dynamics-Based Framework for Understanding Deep Neural Networks

Yuchen Lin,<sup>\*</sup> Yong Zhang,<sup>\*</sup> Sihan Feng, and Hong Zhao<sup>†</sup>  
*Department of Physics, Xiamen University, Xiamen 361005, China*

(Dated:)

Advancements in artificial intelligence call for a deeper understanding of the fundamental mechanisms underlying deep learning. In this work, we propose a theoretical framework to analyze learning dynamics through the lens of dynamical systems theory. We redefine the notions of linearity and nonlinearity in neural networks by introducing two fundamental transformation units at the neuron level: order-preserving transformations and non-order-preserving transformations. Different transformation modes lead to distinct collective behaviors in weight vector organization, different modes of information extraction, and the emergence of qualitatively different learning phases. Transitions between these phases may occur during training, accounting for key phenomena such as grokking. To further characterize generalization and structural stability, we introduce the concept of attraction basins in both sample and weight spaces. The distribution of neurons with different transformation modes across layers, along with the structural characteristics of the two types of attraction basins, forms a set of core metrics for analyzing the performance of learning models. Hyperparameters such as depth, width, learning rate, and batch size act as control variables for fine-tuning these metrics. Our framework not only sheds light on the intrinsic advantages of deep learning, but also provides a novel perspective for optimizing network architectures and training strategies.

## I. INTRODUCTION

To understand deep neural networks (DNNs), several influential theories have been proposed, including the information bottleneck theory [1–4], the flat minima hypothesis [5–8], and geometric analyses [9]. These approaches primarily analyze the global behavior of the entire network. However, their high level of abstraction and computational complexity often necessitate simplifications, typically through model linearization [10–13] or shallow network approximations [14–16]. While valuable, it remains unclear how their conclusions can be fully extended to deep nonlinear architectures.

Despite neurons' foundational role as computational units—characterized by linear summation and nonlinear activation—in neural networks (NNs), existing theoretical approaches have yet to fully establish how their neuron-level properties impact global learning dynamics. This limitation is evident in their difficulty explicitly defining the nonlinearity of learning models at the neuron level and integrating these basic building blocks to explain local inter-neuron interactions leading to system-level learning phenomena. The conventional linear/nonlinear classification of learning models based solely on activation functions lacks precision, as demonstrated by the fact that networks with nonlinear activation functions often initially exhibit linear-like dynamics before gradually developing their full nonlinear characteristics during training [16, 17].

This paper introduces a neuron-level analytical framework for investigating learning dynamics, which is founded upon two novel concepts. Our first concept introduces fundamental transformation units, operating

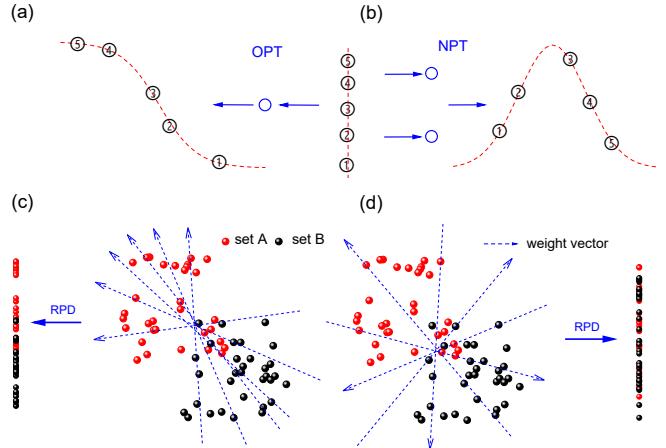


FIG. 1. Illustration of transformation modes and their effects. Circled numbers represent the local fields of five samples projected by a weight vector; hollow circles represent neurons. (a) The OPT mode preserves sample order and can be achieved by a single neuron. (b) The NPT mode disrupts the order: samples 4 and 5 output less than sample 3. This requires at least two cooperating neurons under typical monotonic nonlinear activations. (c) OPT-induced weight vectors concentrate to maximize outputs for sample set A, yielding higher projections than for set B. (d) NPT-induced weight vectors are isotropic. Together, (c) and (d) show how RPD reflects the transformation mode composition.

at the individual neuron level. For a training set of  $P$  samples, each neuron processes  $P$  local fields as an input sequence. This sequence undergoes transformation via two distinct modes: Order-preserving transformation (OPT), where the input sequence's ordering is maintained (Fig. 1(a)), and Non-order-preserving transformation (NPT), where the input sequence's ordering is altered (Fig. 1(b)). The 'ordering' here denotes the rank-

\* These authors contributed equally to this work.

† [zhaoh@xmu.edu.cn](mailto:zhaoh@xmu.edu.cn)

ing of local fields from largest to smallest. Crucially, OPT operations are effectively linear in terms of order preservation and can be implemented by most monotonic activation functions (including linear activation as a special case). In contrast, NPT operations, typically achieved through non-monotonic activation functions (e.g., Gaussian) or specific combinations of common activation functions like ReLU or tanh, produce localized peak-shaped responses via nonlinear folding operations. These distinct transformation modes profoundly influence weight dynamics and information extraction (Figs. 1(c)–1(d)). Consequently, the OPT/NPT ratio emerges as a quantitative measure of nonlinearity, simultaneously offering an interpretable design parameter for optimizing information processing.

Second, we propose defining attraction basins within both sample and weight spaces. The sample-space basin characterizes input-output sensitivity, whereas the weight-space basin quantifies parameter-space stability. These multi-level basins—including average, class-wise, and sample-specific forms—offer a means to delineate the internal structures formed during learning and to understand network performance at varying scales. Notably, the analysis of these basins and their inter-space coordination provides a powerful tool for unraveling how architectural parameters (depth, width) and training strategies (learning rate, batch size, dropout rate, etc.) impact the learning model’s performance.

Our proposed transformation modes establish the underlying operations of the learning process, while the basins of attraction characterize its resultant states. This unique interpretation allows us to view DNNs as layer-wise iterative dynamical mappings, thereby offering a powerful lens from dynamics theory to elucidate their learning dynamics.

An essential feature revealed by this framework is the emergence of distinct learning phases during training, each characterized by specific OPT/NPT state distributions across layers. In contemporary DNNs trained with standard optimization strategies (e.g., SGD or Adam with nonlinear activations), we observe a distinct progression in layer-wise behavior. During early training, shallow layers exhibit OPT-dominant configurations, where weight vectors maintain restricted orientations, limiting the diversity of extracted features. As training proceeds, these layers transition to an NPT-dominant regime, characterized by relaxed directional constraints and increased representational richness. In parallel, deep layers develop near-linear responses with high redundancy, suggesting opportunities for pruning. With extended training, all layers—except the output layer—converge toward similar OPT/NPT ratios, indicating a network-wide tendency toward overlearning.

We conduct a detailed case analysis based on the well-known grokking phenomenon—characterized by sudden improvements in test accuracy following prolonged periods of near-random performance [18–22]—to illustrate how phase transitions align with qualitative shifts in

model performance. Although both exhibit phase transition behavior, attraction basin analysis reveals that grokking in deep networks and in the reference shallow networks [23] occurs via distinct mechanisms.

The paper is organized as follows. Section II introduces transformation modes and attraction basins, along with methods for their quantitative characterization. Section III analyzes learning dynamics in shallow models, emphasizing phase emergence and transitions. Section IV examines DNNs, focusing on how architecture and training parameters shape key dynamical metrics. Section V explores the mechanisms underlying phase transitions in grokking. Section VI concludes with broader implications and future directions.

## II. BASIC CONCEPTS AND METHODS

### A. Models and Algorithm

Without loss of generality, we consider a fully connected DNN architecture for classification, defined by the following equations:

$$\begin{aligned} x_{i_l}^{(l)} &= f(h_{i_l}^{(l)}), \\ h_{i_l}^{(l)} &= \sum_{i_{l-1}=1}^{N_{l-1}} w_{i_l i_{l-1}}^{(l)} x_{i_{l-1}}^{(l-1)}. \end{aligned} \quad (1)$$

Here,  $x_{i_l}^{(l)}$  denotes the output of the  $i_l$ -th neuron in layer  $l$ , and  $h_{i_l}^{(l)}$  is the corresponding local field. The weight  $w_{i_l i_{l-1}}^{(l)}$  connects the  $i_{l-1}$ -th neuron in layer  $l-1$  to the  $i_l$ -th neuron in layer  $l$ . The function  $f(\cdot)$  represents the activation function, and  $N_l$  denotes the number of neurons in layer  $l$ . All computations are implemented in PyTorch, a widely used deep learning framework.

### B. Fundamental transformation units of a learning model

Our approach is grounded in a fundamental postulate: classification relies on the mutual information between samples, requiring that each sample be simultaneously associated with one class and dissociated from others. The first hidden layer is thus expected to maximize information extraction, while subsequent layers should preserve or increase information throughput during processing.

Specifically, a weight vector  $\mathbf{w}$ —a row of the weight matrix  $\mathbf{W}$ —projects a sample vector  $\mathbf{x}^\mu$  onto a local field  $h(\mu) = \mathbf{w} \cdot \mathbf{x}^\mu$ , which serves as the input of the  $\mu$ -th sample to a neuron. The sequence  $h(\mu)$ , where  $\mu = 1, 2, \dots, P$ , represents the local fields induced by this projection over all  $P$  samples. It is not the raw

values of  $h(\mu)$  themselves, but rather their relative ordering—i.e., the ranks of  $h(\mu)$  sorted in descending order—and the spatial intervals between them, that encode the relational structure and mutual information among samples along this direction. These features define the neuron’s informational input. Different neurons extract distinct aspects of information via different weight vectors. To maximize information extraction, the configuration in Fig. 1(c) is preferable to that in Fig. 1(d), as it enables exploration of a broader range of projection directions.

The projection acts as a transformation of the sample space along the weight vector. Neurons operate in two distinct modes: OPT and NPT.

**1. The OPT mode** preserves the ordering of the input sequence  $h(\mu)$ , as illustrated in Fig. 1(a). This behavior can be realized even with strongly nonlinear but monotonic activation functions such as ReLU or tanh. The transformation is effectively linear in terms of preserving input ordering, with linear activation representing a special case of this mode.

**2. The NPT mode** alters the ordering of the input sequence. This can arise either from individual neuron with non-monotonic activations, such as the Gaussian-type function  $y = \exp(-h^2)$ , or from combinations of neurons. For monotonic activations like tanh, the minimal combination of two neurons, e.g.,  $y = w_1 \tanh(h_1) + w_2 \tanh(h_2)$ , is sufficient to produce a maximum at an arbitrary position within the input sequence. (For clarity, we assume identical input sequences across neurons.) ReLU activations exhibit a similar capability. Both cases can result in folding-like transformations, as illustrated in Fig. 1(b). More complex combinations of neurons can further enhance this effect. This mode is inherently nonlinear in its functional consequences, consistent with the notion of nonlinearity in dynamical systems, where the emergence of fundamentally nonlinear behavior requires both stretching and folding, rather than merely the presence of nonlinear terms in the governing equations.

Thus, we define fundamental local processing units that perform qualitatively distinct transformations—either linear (OPT) or nonlinear (NPT). These modes give rise to different collective behaviors of the weight vectors. In the OPT mode, weight vectors must align with specific directions to promote the local fields of target samples to higher ranks within the projected sequence, thereby maximizing output activations, as illustrated in Fig. 1(c). This behavior is effective for extracting linearly separable features. However, to minimize training loss for rare or difficult samples, these weight vectors may converge toward peculiar orientations tailored to those samples. While this may reduce loss locally, it risks overfitting and compromises generalization. Moreover, once the linearly separable structure is extracted, the loss can no longer be effectively reduced, limiting further optimization.

In contrast, the NPT mode imposes no directional constraints on the weight vectors, enabling information ex-

traction from a broader set of orientations, as illustrated in Fig. 1(d). This flexibility allows neurons to capture more complex input relationships. However, operating in a more nonlinear regime also makes them more sensitive to input variations. An effective learning model should thus integrate neurons operating in both modes to balance expressiveness and stability.

### C. Linear substitution map (L-Map) and rank probability distribution (RPD)

The core question that remains is how to characterize the distribution of OPT and NPT neurons in each hidden layer and thus quantify the linearity layer by layer. We propose the following metrics to address these issues. We first introduce them in the case of a single hidden-layer network, and then extend to DNNs.

Feeding all  $P$  samples into the network yields, for each hidden-layer neuron, a sequence of local fields  $h_i^{(2)}(\mu)$ , where  $\mu = 1, 2, \dots, P$ . For the  $i$ th neuron, we construct a signed projection sequence  $h_i^{(2)}(\mu) \cdot \text{sign}(W_{ki}^{(3)})$ , which is connected to the  $k$ th output neuron (corresponding to class  $k$ ). For inputs from class- $k$  samples, the output neuron corresponding to class  $k$  is expected to produce higher activations. Under the OPT mode, the ordering of the transformed sequence is preserved and neurons operate independently. Thus, to maximize the output activation, the values  $h_i^{(2)}(\mu) \cdot \text{sign}(W_{ki}^{(3)})$  for class- $k$  samples should achieve higher ranks within the sequence, resulting in a collective alignment of weight vectors. In contrast, under the NPT mode, ordering is not preserved, and maximization can be achieved through combinations of neurons without requiring collective alignment.

To characterize the above effects, we rank the samples belonging to class  $k$  among all  $P$  samples according to their values in the projection sequence. Without loss of generality, we sort the sequence in descending order. We then compute the RPD of this class for the given neuron. Applying the same methodology, we obtain RPDs for all neurons in the hidden layer. Since our primary interest lies in statistical behavior, we perform an ensemble average over all classes and neurons to produce a smooth RPD that characterizes the overall property of the layer. When needed, per-class RPDs within each hidden layer can also be examined individually. The RPD captures the collective alignment properties of weight vectors, as illustrated in Fig. 1(c) and Fig. 1(d). In other words, the steepness of the RPD provides a quantitative probe of the relative proportions of OPT and NPT neurons in a given hidden layer.

To extend the analysis to DNNs, we introduce the L-map as follows. By replacing all nonlinear activation functions beyond the  $l$ th layer with the identity function  $f(h) = h$ , we obtain

$$\mathbf{W}_{\text{L-map}}^{(l)} = \mathbf{W}^{(l)} \cdot \mathbf{W}^{(l+1)} \cdots \mathbf{W}^{(L)}, \quad (2)$$

where  $\mathbf{W}^{(l)}$  denotes the weight matrix of the  $l$ th layer.

In practice, however, additional components such as batch normalization (BN) [24] are commonly employed during training to stabilize learning and accelerate convergence. BN standardizes activations within each mini-batch and applies a feature-wise affine transformation. These operations modify the network's L-map and must be properly accounted for in the analysis.

The BN operates differently in training and evaluation modes [25]. During training, it normalizes each layer using the mean and variance computed from the current mini-batch. In evaluation mode, it instead applies a moving average of the mean and variance accumulated throughout training. The updates follow the formulas:

$$\begin{aligned}\hat{\mu}_t &= (1 - \alpha)\hat{\mu}_{t-1} + \alpha\mu_t, \\ \hat{\sigma}_t^2 &= (1 - \alpha)\hat{\sigma}_{t-1}^2 + \alpha\sigma_t^2,\end{aligned}\quad (3)$$

where  $\mu_t$  and  $\sigma_t^2$  are the mean and variance of the current mini-batch, and  $\hat{\mu}_t$ ,  $\hat{\sigma}_t^2$  are the accumulated estimates used in evaluation. The momentum parameter is typically set as  $\alpha = 0.1$ . Setting  $\alpha = 1$  effectively corresponds to using only the current batch statistics, i.e., the behavior during training.

After normalization, BN applies a learnable scaling factor  $\gamma_i^{(l)}$  to each feature  $i$  in layer  $l$ , enabling the network to recover suitable activation magnitudes. The combined normalization and rescaling can be expressed as a diagonal matrix  $\mathbf{D}^{(l)} = \text{diag}(\gamma_i^{(l)}/\hat{\sigma}_i^{(l)})$ , where  $\hat{\sigma}_i^{(l)}$  is the estimated standard deviation for feature  $i$ . This matrix captures the feature-wise transformation introduced by BN at layer  $l$  during inference.

Consequently, the L-map is updated to

$$\begin{aligned}\mathbf{W}_{\text{L-map}}^{(l)} &= (\mathbf{D}^{(l)}\mathbf{W}^{(l)})(\mathbf{D}^{(l+1)}\mathbf{W}^{(l+1)}) \\ &\dots (\mathbf{D}^{(L)}\mathbf{W}^{(L)}).\end{aligned}\quad (4)$$

When the latter part of a DNN consists entirely of linear neurons, the L-map becomes mathematically equivalent to a linear perceptron and can effectively substitute for the original nonlinear portion. In this case, the connection  $W_{ki}^{(2)}$  can be replaced with  $W_{ki}^{\text{L-map}}$ , allowing the RPD of the  $l$ -th layer to be computed accordingly.

In DNNs with nonlinear activation functions, we also apply the RPD to estimate the proportion of neurons operating in the two modes, i.e., continue to apply the L-map to estimate the connections. This constitutes an approximation, as nonlinear effects from subsequent hidden layers may affect the accuracy of the estimation. However, when monotonic activation functions such as tanh or ReLU types are used, the accuracy of this approximation remains relatively high. Since the RPD depends only on the ranking of sample projections, preserving the sign of  $W_{ki}^{\text{L-map}}$ , rather than its absolute value, is sufficient. Replacing monotonic activation functions with linear ones maintains the input-output monotonicity, thereby enhancing the robustness of the sign preservation.

## D. Attraction Basins

Our second key analytical tool is the concept of attraction basins, a fundamental notion in nonlinear dynamical systems. The analysis of attraction basins has been applied to study the dynamics of asymmetric Hopfield neural networks [26–28], where a sharp transition from a chaotic phase to a memory phase emerges as the basins expand [27, 28]. We extend the concept of attraction basins to the context of DNNs. Here we define the attraction basin of a training sample in two distinct spaces: the sample space and the weight space.

First, we apply random perturbations to a training sample and evaluate whether the model retains its original prediction. Specifically, if the trained network still classifies a perturbed version  $\mathbf{x}^\mu + \delta\mathbf{x}$  of the  $\mu$ th sample into the same class, then  $\delta\mathbf{x}$  is considered to lie within the attraction basin of that sample. Each original input  $\mathbf{x}^\mu$  is first normalized to the range  $[0, 1]$ . Gaussian noise with a specified standard deviation is added, and the resulting sample is then rescaled to the original data range. By plotting the classification accuracy—averaged over multiple perturbation trials for each noise amplitude—we observe a gradual decline from noise-free accuracy to the level expected from random guessing. The noise amplitude at which the accuracy falls to 50% defines the size of the sample's attraction basin. This metric directly reflects the model's generalization capacity to input variations.

Another type of attraction basin is defined by perturbing the network weights and assessing whether a training sample remains correctly classified. Specifically, if  $\mathbf{x}^\mu$  is still recognized as belonging to the same class under a perturbed weight configuration  $\mathbf{w} + \delta\mathbf{w}$ , then  $\delta\mathbf{w}$  is regarded as lying within the attraction basin in weight space for that sample. To ensure consistency in the analysis, weights are first standardized using mean-variance normalization. Gaussian noise of controlled magnitude is then added, after which the weights are inverse-transformed back to their original scale. The updated weights are used to evaluate classification accuracy. This type of attraction basin characterizes the network's structural stability and its robustness to perturbations in the weight space.

This type of attraction basin provides a direct and intuitive measure of the robustness and breadth of the solution region in parameter space. Unlike traditional notions of flat minima [29, 30], which assess the sensitivity of the loss function to weight perturbations and often rely on computationally expensive Hessian-based analyses, we evaluate robustness through the direct impact of weight variations on training accuracy. The concept of flat minima aims to establish a correspondence between weight variations and sample variations. In contrast, we demonstrate that the attraction basins in sample space and weight space exhibit independence, and provide complementary metrics for characterizing the learning model's behavior.

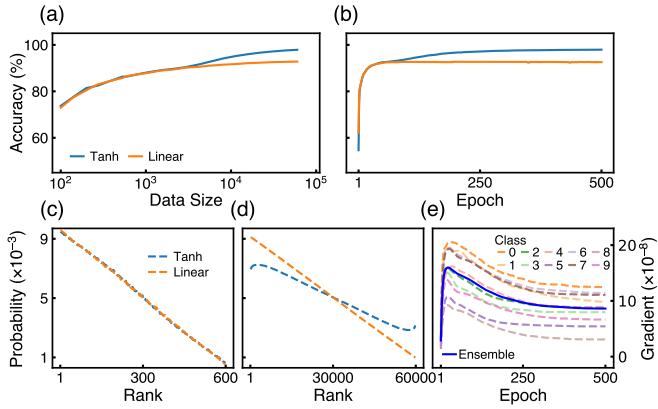


FIG. 2. Learning dynamics of a shallow network. (a) Accuracy as a function of training sample size. From top to bottom, the curves represent the performance of the NN with tanh activation, the LNN, and the L-map, respectively. (b) Accuracy vs. training epochs for a training set of 60,000 samples, with the same ordering of lines as in (a). (c) RPDs of the NN with tanh activation and the LNN trained on 600 samples. (d) RPDs of the NN with tanh activation and the LNN trained on 60,000 samples. (e) Evolution of RPDs for the NN with tanh activation at different training epochs.

### III. APPLICATIONS TO SHALLOW NETWORKS

In this section, we demonstrate how the above concepts can be applied to analyze a shallow neural network with the architecture 784–2048–10, trained on the MNIST dataset for handwritten digit recognition. The MNIST dataset [31] is a widely used benchmark for image classification, consisting of 60,000 training samples and 10,000 test samples, each represented as a  $28 \times 28$  grayscale image.

Fig. 2(a) shows the test accuracy as a function of sample size for the network with tanh activation and for the linear neural network (LNN) with activation  $f(h) = h$ . For small training sets, the two accuracy curves converge, indicating that the nonlinear network behaves similarly to the LNN. Fig. 2(c) presents the RPDs of the hidden layer for both networks, trained on 600 samples. The RPDs exhibit high probability density in the high-ranking region and low density in the low-ranking region, clearly demonstrating the OPT-induced alignment effect of weight vectors. The close match between the two RPD curves suggests that learning is driven purely by OPT neurons (noting that the LNN can perform only the OPT operation).

As the training set size increases, the accuracy curves begin to diverge. Fig. 2(d) shows that at a training set size of 60,000, the two RPDs differ substantially, with the LNN exhibiting a steeper gradient. These observations indicate that a significant number of NPT-mode neurons are activated in the nonlinear network.

To understand how the linearity of the nonlinear network evolves during training, we track the test accuracy

and the RPD gradient over training steps using the full training set of 60,000 samples. As shown in Fig. 2(b), the accuracy curves for the nonlinear network and the LNN initially overlap but gradually diverge as training progresses. Fig. 2(e) shows that the RPD gradient of the nonlinear network increases during early training, peaks, and then declines. These results suggest that, in the early stage, the OPT mode drives weight vectors to align along specific directions, transforming the RPD from flat to steep. (Due to random initialization, the weight vectors are initially isotropic, resulting in a flat RPD.) At this stage, the nonlinear network effectively behaves as a linear network, as learning is governed by OPT-mode neurons. As training continues, NPT-mode neurons become activated, causing the weight vectors to spread over a broader range of directions, thereby reducing the RPD gradient. We also compute class-specific RPDs (Fig. 2(e)), which reveal distinct evolution patterns across digit classes, reflecting class-dependent degrees of linearity.

Therefore, the training dynamics may exhibit distinct learning phases. In the initial stage, the nonlinear network typically learns through the OPT mode. When trained on a small, linearly separable dataset, the learning objective can often be fully achieved by the OPT mode alone, and the weight vectors remain in the OPT phase. For larger datasets, however, the OPT mode may become insufficient to further reduce the loss, triggering the activation of NPT neurons. At this point, the learning transitions from a pure OPT phase to a mixed phase involving neurons operating in both OPT and NPT modes.

### IV. APPLICATIONS TO DNNs

In this section, we adopt the ReLU activation function for nonlinear DNNs and the identity function  $f(h) = h$  for linear DNNs. All models have an input dimension of 784 and an output dimension of 10, and are trained on the full MNIST training set comprising 60,000 samples. This investigation aims to reveal how optimization algorithms (SGD vs. Adam), hyperparameters (learning rate and batch size), and network architectures (depth and width) influence the layer-wise RPD distribution and the evolution of both attraction basins, thereby elucidating their underlying mechanisms.

#### A. RPD analysis

Figs. 3(a)–3(c) show the evolution of training accuracy, test accuracy, and the RPD gradient for each hidden layer over training steps in a 10-layer DNN with a width of 512, trained using SGD. ReLU activation is used in the first and second plots, with learning rates of 0.03 and 0.37, respectively. The third plot corresponds to training with a linear activation function and a learning rate of 0.03.

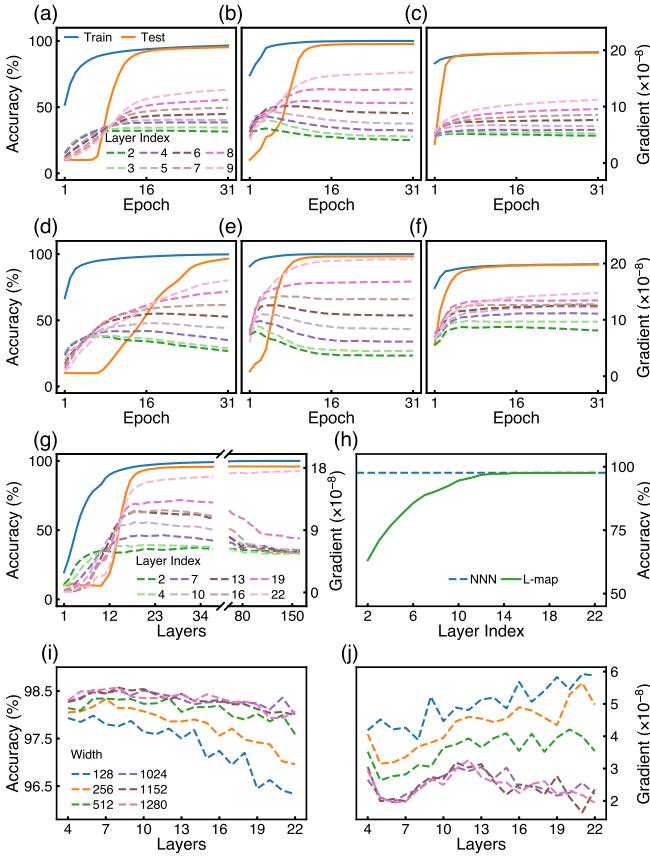


FIG. 3. Training dynamics and RPD analysis. (a)–(c) Evolution of training/test accuracy and RPD gradients in a 10-layer DNN (width 512) trained with SGD. (a), (b): ReLU activation with learning rates 0.03 and 0.37; (c): linear activation with learning rate 0.03. (d)–(f) Corresponding results using the Adam optimizer. Batch size is 60,000 in (d), (f), and 20,000 in (e). (g) Results for a 23-layer DNN (width 128) trained with Adam. (h) L-map pruning accuracy as a function of the starting layer for the 23-layer network. (i) Test accuracy as a function of depth for various widths. (j) First-layer RPD gradient as a function of depth and width.

Figs. 3(d)–3(f) present the results for DNNs trained using the Adam optimizer. The network architecture and activation functions exactly match those used in Figs. 3(a)–3(c). A batch size of 60,000 is used in the first and third plots, while the second uses a batch size of 20,000.

Figs. 3(a) and 3(d) show a clear phase transition: at the early stage of training, the RPD gradient is higher in the earlier layers and lower in the deeper layers (denoted as phase I), but this pattern later reverses (phase II). In the other plots, the learning process maintains a phase II-like gradient distribution from the outset.

The distribution of RPD gradients elucidates the specific mechanism of information processing. An overall increase in RPD gradients during the early training stage reflects the directional alignment of weight vectors induced by the OPT mode, which drives the transition from an initially isotropic distribution to more specific

orientations. The stabilized phase ultimately determines the network's performance, as elaborated in the following.

Phase II represents an ideal distribution pattern: a low RPD gradient in the first hidden layer enables information extraction from a broad range of weight vector directions, while the subsequently increasing gradients across deeper layers progressively transform sample representations into linearly separable forms.

The activation of NPT neurons enhances information extraction. The final RPD gradient distributions in Figs. 3(a) and 3(c) are almost identical, indicating that the two models extract largely similar information in the OPT mode. However, the former achieves a test accuracy of 97.14%, whereas the latter reaches only 92.37%. This discrepancy demonstrates that NPT neurons acquire additional information features by constructing and optimizing neuronal combinations based on the weight vector distribution established by OPT.

The RPD gradient in the first hidden layer plays a pivotal role in maximizing information extraction. A smaller gradient at this layer indicates a broader distribution of weight vector orientations, thereby enabling more effective information extraction from the dataset. As shown in Fig. 3(b), the model exhibits a lower first-layer RPD gradient than its two SGD-trained counterparts and ultimately achieves a superior test accuracy of 97.89%.

The results obtained using the Adam optimizer are qualitatively similar. Figs. 3(d)–3(f) achieve classification accuracies of 98.26%, 98.29%, and 92.46%, respectively, which strongly correlate with the level of the first-layer RPD gradient. These results demonstrate the general superiority of Adam over SGD, as it consistently yields a lower RPD gradient in the first hidden layer, thereby enabling more comprehensive sample information extraction.

Deeper networks may exhibit more distinct learning phases. Fig. 3(g) shows the results for a 23-layer DNN with a width of 128, trained using the Adam optimizer and a batch size of 60,000. In addition to phases I and II, we observe a third phase characterized by the convergence of RPD gradients across all layers (except the final layer) to approximately the same value. These results suggest that the learning process may transition through multiple phases with distinct layer-wise RPD gradient distributions, depending on the network architecture, training strategy, and hyperparameter configuration.

The degree of nonlinearity across DNN layers can be estimated via L-map pruning. We assess this by computing the pruning accuracy, which involves replacing layers from the  $l$ -th hidden layer to the output layer with their corresponding L-map. Fig. 3(h) shows the change in accuracy as a function of the starting layer for the 23-layer DNN, evaluated at the epoch corresponding to peak generalization performance.

The pruning accuracy differs from that of the original DNN when pruning begins at earlier layers, indicating

the presence of NPT effects and inherent nonlinearity in those layers. As the starting layer moves deeper, the pruning accuracy increases, suggesting increasing linearity with depth. Beyond a critical layer, the accuracy stabilizes and becomes comparable to that of the original DNN, implying that the later layers are functionally equivalent to a linear perceptron and can be safely pruned. This behavior is desirable for reducing redundancy and conserving computational resources [32].

An intriguing observation is that grokking consistently appears in the scenarios shown in Figs. 3(a), 3(d), and 3(g), all of which exhibit phase transition characteristics. The term *grokking* refers to the phenomenon where test accuracy remains at the level of random guessing even after training accuracy becomes high, before suddenly improving. Notably, the presence of grokking does not necessarily imply higher test accuracy. The underlying mechanism of grokking will be discussed in Section V.

We then extend the RPD analysis to examine the effects of both network depth and width. Fig. 3(i) presents test accuracy as a function of depth for DNNs with widths ranging from 128 to 1280. For a fixed width, accuracy initially increases with depth and then declines, with the reduction being more pronounced in narrower networks and more gradual in wider ones. The maximum accuracy is attained by networks with approximately 6 to 8 layers, a result that appears largely independent of width. For a fixed depth, accuracy generally increases with width and eventually saturates.

Fig. 3(j) shows the RPD gradient of the first hidden layer across DNNs with varying depths and widths, evaluated at the epoch corresponding to the highest test accuracy. The key findings are as follows. First, as width increases, the RPD gradient decreases. This trend correlates with the increase in test accuracy, suggesting that a higher proportion of NPT neurons in wider networks facilitates early-stage information extraction. Beyond a certain width, both the gradient and accuracy saturate. Second, the depth dependence of the RPD gradient reveals a pronounced minimum around 6 layers, indicating that network depth serves as a critical degree of freedom for minimizing the first-layer gradient and thereby maximizing information extraction.

## B. Attraction basin analysis

Fig. 4(a) shows that increasing depth enlarges the average attraction basin in the sample space. This highlights a key advantage of depth: layer-by-layer iteration progressively expands the attraction basin, thereby improving generalization ability. Fig. 4(b) shows that increasing depth reduces the attraction basin in the weight space, indicating that increasing the depth negatively affect structural stability. Both attraction basins exhibit a saturation trend with increasing depth, reflecting a trade-off between these two counteracting factors. These plots confirm that while deep networks offer superior general-

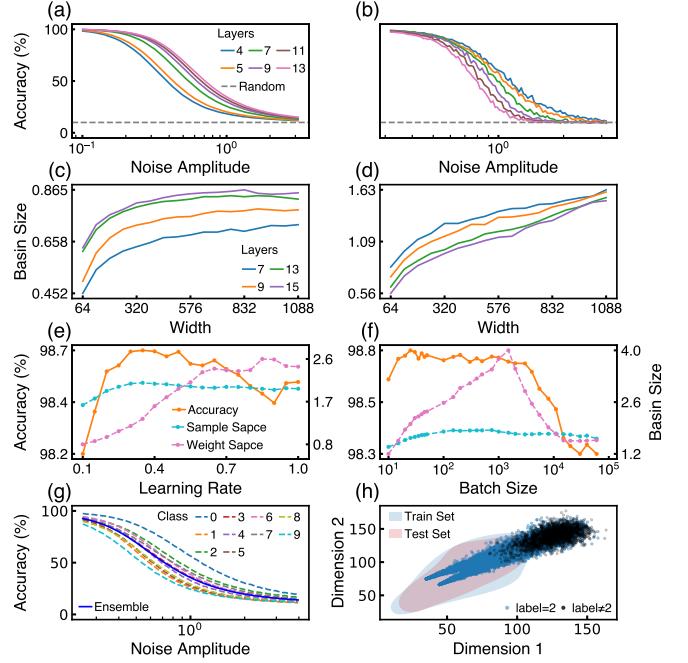


FIG. 4. Attraction basin analysis. (a) Accuracy of noised training samples vs. noise amplitude. (b) Accuracy of training samples under weight perturbations vs. noise amplitude. (c), (d) Average attraction basin sizes in sample space and weight space, respectively, as a function of network width for various depths. (e), (f) Basin sizes in sample and weight space vs. learning rate and batch size, respectively. (g) Accuracy of noisy samples from ten classes vs. noise amplitude in the 23-layer DNN. (h) 2D projection of attraction basins for digit "2" and its 3-pixel shifted variant.

ization capabilities, their structural sensitivity increases, suggesting the existence of an optimal depth for DNNs.

Figs. 4(c) and 4(d) show the average basin sizes in the sample and weight spaces as a function of width for DNNs with varying depths. For fixed width, we observe that sample-space basin sizes increase with depth, whereas weight-space basins shrink, consistent with the trends shown in Figs. 4(a) and 4(b). For fixed depth, both types of basins expand with width. This behavior stems from a common design strategy in DNNs, where the initial weight range is reduced as width increases. Without this initialization scheme, the observed correlation between width and weight-space basin size may no longer hold (see Supplementary Materials).

To validate this interpretation, we further examine the effect of initialization scale. Reducing the initial weight range enlarges the attraction basins in weight space but simultaneously reduces those in sample space, revealing a fundamental trade-off. These findings suggest that simultaneously enlarging both basins of attraction is crucial for maximizing network performance, which typically requires coordinated architectural and initialization strategies. Empirically, the initialization scales adopted in mainstream deep learning frameworks strike an effec-

tive balance between these competing factors. Supporting data are provided in the Supplementary Materials (SM).

Strategies such as tuning the learning rate and the batch size can be interpreted as efforts to simultaneously enlarge both types of attraction basins. Figs. 4(e) (using SGD) and 4(f) (using the Adam optimizer) show the basin sizes in the sample and weight spaces as functions of learning rate and batch size, respectively, for DNNs with 12 layers and width 1024. Test accuracy is plotted alongside for reference. In both cases, the highest accuracy is achieved in regions where both attraction basins expand simultaneously.

Analyzing attraction basins at both the class and individual-sample levels provides multi-resolution insights into DNN dynamics. Fig. 4(g) shows the variation in classification accuracy across the ten digit classes as a function of noise amplitude. The results reveal substantial differences in the attraction basins of different classes.

Fig. 4(h) presents a two-dimensional projection of the attraction basins for a digit "2" and a variant shifted by 3 pixels along both axes. Here, the  $x$  and  $y$  coordinates correspond to the sums of the first and last 392 dimensions of the input vector, respectively. The plot also shows the distribution regions of both training and test samples in this 2D projection space. The original sample is located at the apex of the resulting conical structure. Notably, the effective attraction basins—defined as the regions overlapping with the sample distribution—remain distinctly separated for individual samples.

To better visualize this separability, we plot the attraction basins in a three-dimensional PCA projection in Fig. 5. The figure shows the attraction basins of 20 samples, comprising two samples per digit class. Most samples exhibit distinct effective attraction basins, which remain non-contiguous even among samples of the same class.

These findings suggest that the training set size has little impact on the average size of the attraction basin. To verify this, we constructed a dataset 17 times larger by applying  $\pm 1$  and  $\pm 2$  pixel translations along the  $x$ ,  $y$ , and diagonal ( $xy$ ) directions to the original 60,000 samples and retrained the network. The resulting average attraction basin size remained virtually unchanged (see Supplementary Material).

Furthermore, we observe that the bases of the conical attraction basins tend to converge toward common regions. As the noise amplitude increases, perturbed samples are increasingly drawn into the domains of non-target classes, leading to interwoven and overlapping attraction basins across samples. The extent of this overlap varies substantially among individual samples, providing a theoretical basis for the construction of adversarial examples. For instance, the digit "2" located in the lower-left corner exhibits significant overlap under minimal perturbation, indicating a high susceptibility to adversarial attacks. The figure shows a representative case in which

a slightly perturbed "2" is misclassified due to its attraction to the domain of digit "3".

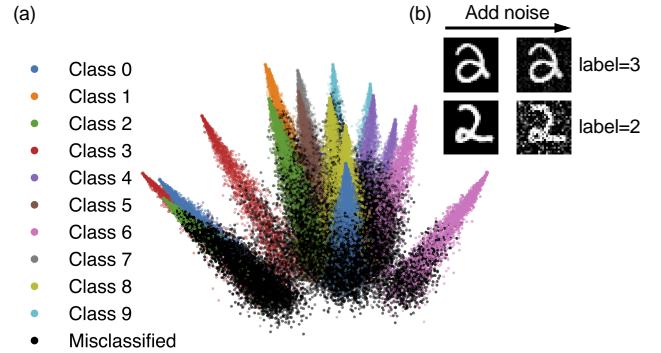


FIG. 5. Single sample attraction basins. (a)The attraction basins in a 3D PCA projection of 20 digits. (b) Digits with small attraction basins are easily attacked by perturbations.

## V. PHASE TRANSITION AND GROKKING

RPD and attraction basin analyses provide a systematic framework for understanding neural network behavior, offering insights into the underlying learning dynamics. In this section, we investigate the mechanisms behind the well-known grokking phenomenon. Grokking is often described as a delayed transition from memorization to generalization after extended training [23]. While this interpretation captures the behavior in shallow networks, we show that it does not accurately explain the dynamics in DNNs. Although multiple factors may contribute to grokking, a key prerequisite is a sharp shift in learning dynamics.

### A. Grokking in DNNs

We demonstrate that the grokking effect observed in DNNs is triggered by the phase transition in conjunction with the training strategy involving batch normalization (BN; see Eq. (3)). To visualize this effect, we project the sample representations from the hidden layer adjacent to the output layer into two dimensions. Specifically, we insert a bottleneck layer with two neurons before the output layer. Rather than retraining the model, we perform singular value decomposition on the weight matrix of the original final layer and extract the two leading right-singular vectors as principal directions. These directions are then used to construct the weights of the bottleneck layer. The resulting local fields  $h_1$  and  $h_2$  of the two bottleneck neurons are used as the coordinates in the two-dimensional representation.

For the 23-layer DNN, we visualize the 2D projections of training and test samples before and after grokking (Figs. 6(a) and 6(c) vs. 6(b) and 6(d)). Prior to grokking,

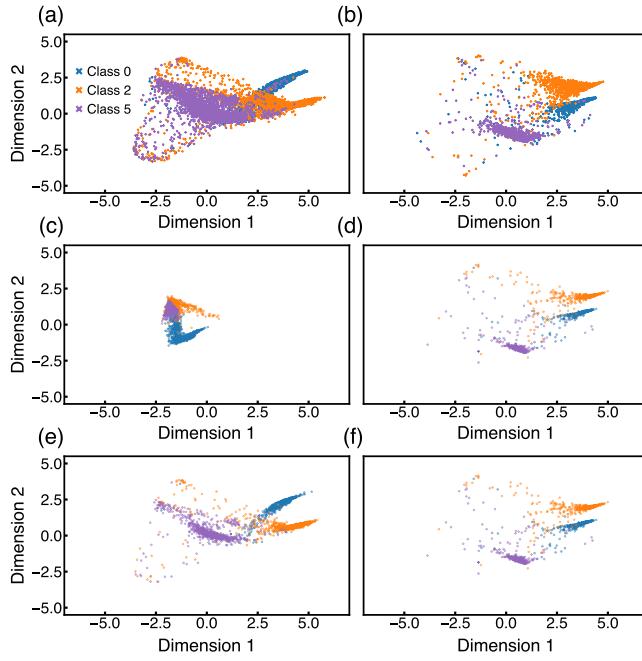


FIG. 6. Grokking mechanism in the 23-layer DNN. 2D representations of digits "0", "2", and "5" are shown before (left column) and after grokking (right column). (a) and (b) correspond to the training set in training mode. (c) and (d) show the test set in evaluation mode, while (e) and (f) show the test set in training mode. A shift in attraction basins occurs only in the evaluation mode after grokking.

the test samples are projected outside the region occupied by the training samples. After grokking, the projection regions of the test and training samples overlap substantially, indicating that the test samples now fall within the attraction basins established by the training data. This overlap accounts for the abrupt rise in test accuracy.

This grokking scenario—where test sample projections collectively shift from outside into the attraction basins of training samples—can be attributed to the perturbation introduced by BN in evaluation mode, under the condition of phase transition in RPD gradient distribution. As described by Eq. (3), BN relies on running estimates of mean and variance accumulated during earlier training stages. Phase transitions in learning dynamics (see Fig. 3(g)) can cause these historical estimates to diverge from the current true statistics, introducing a systematic bias in the inference of test samples. As training continues and the network settles into a stable learning phase, this discrepancy gradually diminishes. Consequently, test samples increasingly fall within the attraction basins of training samples, giving rise to the grokking phenomenon.

Indeed, when the training mode is applied during testing—by setting  $\alpha = 1$  in Eq. (3)—the grokking effect disappears. Under this setting, the projected regions of test samples (Fig. 6(e)) already overlap with those of the training samples (Fig. 6(a)) before grokking occurs. Af-

ter grokking, the test sample projections (Fig. 6(f)) remain similar to those in Fig. 6(d) and continue to coincide with the training sample projections (Fig. 6(e)). In this case, maintaining consistent attraction basins for both test and training samples eliminates the conditions necessary for grokking to emerge.

Nonetheless, evaluation mode alone does not necessarily induce grokking. In cases where no phase transition occurs (Figs. 3(b), 3(c), 3(e), and 3(f)), no notable grokking behavior is observed. This is because BN does not introduce significant distributional shifts within the same learning phase. Indeed, the 2D projection regions of test and training samples remain largely overlapping in these cases. These results underscore that grokking requires a transition in learning dynamics.

These findings suggest that the prevalent grokking behavior in deep neural networks cannot be fully accounted for by the conventional “memorization-to-generalization” narrative. Instead, it arises from phase transitions that cause batch normalization to introduce statistical mismatches between training and test data.

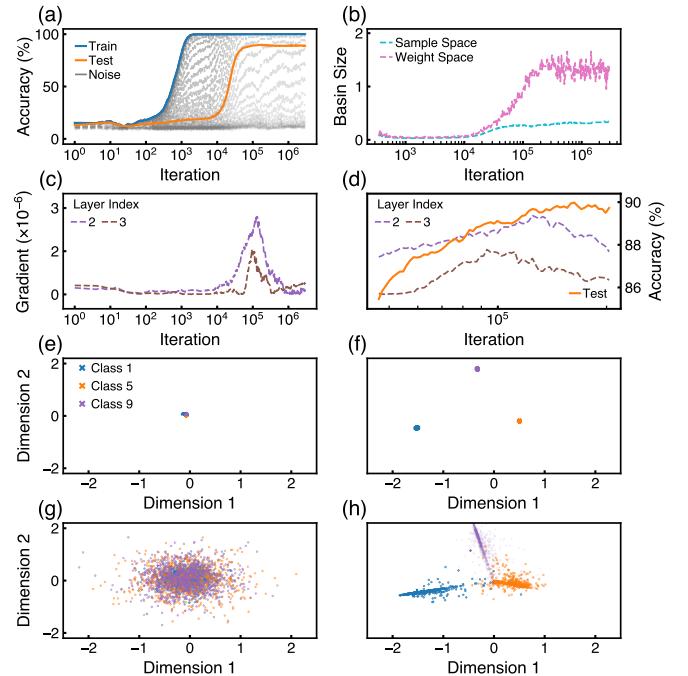


FIG. 7. Grokking mechanism in a shallow network. (a) Accuracy trajectories for training and test samples, including variants perturbed with different noise amplitudes. (b) Evolution of attraction basin sizes in both sample and weight spaces, with a clear expansion coinciding with the onset of grokking. (c) RPD gradients of the two hidden layers. (d) Continued training leads to improved test accuracy as RPD gradients decrease again, indicating enhanced information extraction. (e) and (f) show 2D projections of training samples for digits "1", "5", and "9" before and after grokking, respectively. (g) and (h) show the corresponding projections for test samples of the same digits before and after grokking.

## B. Grokking in shallow networks

Under standard design settings, grokking behavior is generally not observed in shallow networks. However, as demonstrated in [23], grokking can emerge when specific strategies are employed. In their setup, a 784-200-200-10 network was trained on 1,000 samples with large initial weights and relatively small weight decay. We show that these conditions indeed give rise to pronounced grokking behavior.

In Fig. 7(a), we reproduce the grokking phenomenon, characterized by a significant delay in test accuracy relative to training accuracy. The plot also shows the accuracy trajectories for samples with different noise levels. The evolution of both types of attraction basins is presented in Fig. 7(b) (not shown during the initial phase, as the basins have not yet formed). Fig. 7(c) displays the evolution of RPD gradients across the two hidden layers.

As shown in Fig. 7(a), when the noise amplitude remains below a certain threshold, the accuracy of noisy samples eventually approaches 100%, indicating the formation of stable attraction basins. In contrast, when the noise amplitude exceeds this threshold, accuracy declines, suggesting that the samples fall outside the effective attraction regions.

Fig. 7(b) shows a clear transition—an abrupt increase in both types of attraction basins coinciding with the onset of grokking. Prior to grokking, attraction basins are already formed for all training samples, as the training accuracy has reached nearly 100%. However, the average basin size remains on the order of  $10^{-2}$ , whereas the deviation between test and training samples is approximately  $10^{-1}$ . After grokking, the attraction basin size increases beyond  $10^{-1}$ . These findings provide an intuitive explanation for grokking: the phenomenon emerges when test samples shift from outside to inside the attraction basins of training samples.

To further illustrate that grokking entails a sharp transition in network dynamics, we visualize the 2D projections of training and test samples for digits "1", "5", and "9" before and after grokking (Figs. 7(e)–7(h)). Prior to grokking, the training samples form tightly clustered groups, while the test samples appear scattered and overlapping across classes, suggesting they fall outside the attraction basins. After grokking, the test projections become well-aligned with those of the training samples, indicating that they have entered the attraction basins. These patterns reveal a dynamical shift reminiscent of asymmetric Hopfield networks [26], where a sharp transition from a chaotic phase to a memory phase occurs as the attraction basins expand [27, 28].

The RPD analysis reveals deeper mechanisms underlying grokking and the associated transitions in learning dynamics. Fig. 7(c) shows that, prior to grokking, the RPDs in both hidden layers exhibit consistently small gradients. This observation suggests that the attraction basins at this stage are formed predominantly via the NPT mode. The activation of this mode arises from the

specific architectural and training choices in the four-layer model. In particular, to prominently elicit the grokking phenomenon, the model is initialized with large weights and trained with a small weight decay rate [23]. This small decay rate, together with the correspondingly small learning rate, prevents significant weight vector concentration and thus suppresses the activation of the OPT mode. Furthermore, the small learning rate contributes to the formation of narrow attraction basins. Consequently, prior to grokking, the network remains in an NPT-dominated learning phase.

Fig. 7(c) also shows that, around the grokking point, the RPD gradients peak simultaneously across both hidden layers. This marks the onset of a new learning phase characterized by a high density of OPT neurons in both layers. The underlying reason is as follows. During the NPT-dominated phase, the amplitude of the weight vectors gradually decreases (see SM). Once the weight amplitudes fall below a critical threshold, it becomes easier—despite the small learning rate—to induce changes in the directions of the weight vectors, thereby activating a large population of OPT neurons. These neurons, in turn, lead to a rapid expansion of the attraction basins, ultimately triggering the grokking transition.

Continued training leads to a further decrease in the RPD gradients over time, signaling a sustained increase in NPT neurons and a return to the NPT-dominated phase. This likely occurs because, at this stage, NPT neurons become more effective for further reducing the loss. As a result, test accuracy improves (Fig. 7(d)), since consistently small RPD gradients allow the network to explore broader directions in weight space for information extraction.

During this NPT-dominated phase, ongoing training expands the attraction basin in sample space while simultaneously shrinking the attraction basin in weight space. In particular, the attraction basin within the weight space exhibits increasingly severe fluctuations, gradually destabilizing the model. The competition between these opposing effects leads to a steady decline in test accuracy. This observation implies that training should be halted before excessive activation of NPT neurons occurs, thereby providing a theoretical justification for the widely used early stopping strategy [33–35].

Therefore, the grokking phenomenon observed in this shallow network aligns with the conventional definition of grokking as a shift from memorization to generalization. The underlying mechanism also stems from an abrupt transition in the network’s learning dynamics. Specifically, the network transitions from an NPT-dominated phase to a high-OPT phase, and then reverts back to an NPT-dominated phase during training. We attribute these transitions to the specific initialization scheme and the use of L2 regularization.

## VI. CONCLUSION AND DISCUSSIONS

The fundamental transformation modes and attraction basins offer sensitive and operational tools for probing learning dynamics, providing an intuitive framework for analyzing machine learning systems. The distinct information extraction mechanisms of OPT and NPT neurons give rise to different collective behaviors of weight vectors, which in turn shape the distribution of OPT and NPT neurons across hidden layers and result in distinct learning phases. The ratio of OPT to NPT neurons characterizes the degree of nonlinearity in each hidden layer, thereby resolving prior ambiguities in defining network linearity versus neuron's nonlinearity. Accordingly, the roles of network depth and width, as well as various training strategies, can be interpreted within this unified framework.

Attraction basins in sample space and weight space respectively characterize a learning model's generalization capability and parametric robustness. The average basin of attraction, class-wise basins, and single-sample basins capture the dynamical properties of learning models at different levels. The synchronous expansion of both attraction basins, together with the optimal distribution of OPT and NPT neuron ratios across hidden layers, forms a set of key metrics for achieving peak network performance.

Learning phases—shaped by training time, initialization conditions, hyperparameters, dataset size, activation

functions, and different training strategies —are closely correlated with network performance. In particular, phase transitions give rise to significant phenomena such as grokking. While such transitions are a prerequisite for observing grokking, our analysis reveals a fundamental difference between grokking in DNNs and shallow networks. In DNNs, grokking emerges from the evaluation mode of the BN strategy: test samples initially lie outside the attraction basins of the training data and are progressively drawn into them. By contrast, grokking in shallow networks arises from abrupt transitions in the structure of the attraction basin in sample space.

Finally, we emphasize that our analysis suggests that neither linear networks nor shallow networks can capture the essential properties of nonlinear DNNs. The former fails to embody the functional role of NPT neurons, while the latter lacks the architectural flexibility to adjust the distribution of OPT and NPT neurons across layers.

## ACKNOWLEDGMENTS

Specifically, Y.C.L. was responsible for performing the calculations and preparing the figures presented in this paper. Y.Z. conducted an extensive literature review and provided valuable insights during the analysis and discussion stages. S.H.F. carried out the initial exploratory research. Lastly, H.Z. proposed the concept of classifying neuron modes and introduced the idea of utilizing the basin of attraction for analyzing deep mechanisms. Additionally, H.Z. wrote the manuscript.

- 
- [1] N. Tishby and N. Zaslavsky, in *2015 IEEE Information Theory Workshop (ITW)* (2015) pp. 1–5.
  - [2] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, in *International Conference on Learning Representations* (2017).
  - [3] R. Shwartz-Ziv, *Information flow in deep neural networks* (2022), arXiv:2202.06749 [cs.LG].
  - [4] K. A. Murphy and D. S. Bassett, *Phys. Rev. Lett.* **132**, 197201 (2024).
  - [5] S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey, arXiv preprint arXiv:1711.04623 (2017).
  - [6] C. Ma and L. Ying, in *Advances in Neural Information Processing Systems*, edited by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (2021).
  - [7] Y. Feng and Y. Tu, *Proceedings of the National Academy of Sciences* **118**, e2015617118 (2021), <https://www.pnas.org/doi/pdf/10.1073/pnas.2015617118>.
  - [8] N. Yang, C. Tang, and Y. Tu, *Phys. Rev. Lett.* **130**, 237101 (2023).
  - [9] G. Naitzat, A. Zhitnikov, and L.-H. Lim, *Journal of Machine Learning Research* **21**, 1 (2020).
  - [10] A. Jacot, F. Gabriel, and C. Hongler, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18 (Curran Associates Inc., Red Hook, NY, USA, 2018) p. 8580–8589.
  - [11] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang, On exact computation with an infinitely wide neural net, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2019).
  - [12] A. Saxe, J. McClelland, and S. Ganguli, in *International Conference on Learning Representations 2014* (2014).
  - [13] Z. Ji and M. Telgarsky, in *International Conference on Learning Representations* (2019).
  - [14] Y. Dandi, F. Krzakala, B. Loureiro, L. Pesce, and L. Stephan, How two-layer neural networks learn, one (giant) step at a time (2023), arXiv:2305.18270 [stat.ML].
  - [15] T. Luo, Z.-Q. J. Xu, Z. Ma, and Y. Zhang, *J. Mach. Learn. Res.* **22** (2021).
  - [16] Y. Xu and L. Ziyin, Three mechanisms of feature learning in the exact solution of a latent variable model (2024), arXiv:2401.07085 [cs.LG].
  - [17] M. Geiger, L. Petrini, and M. Wyart, *Physics Reports* **924**, 1 (2021), landscape and training regimes in deep learning.
  - [18] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra, Grokking: Generalization beyond overfitting on small algorithmic datasets (2022), arXiv:2201.02177 [cs.LG].

- [19] M. Belkin, D. Hsu, S. Ma, and S. Mandal, *Proceedings of the National Academy of Sciences* **116**, 15849 (2019), <https://www.pnas.org/doi/pdf/10.1073/pnas.1903070116>.
- [20] T. Kumar, B. Bordelon, S. J. Gershman, and C. Pehlevan, in *The Twelfth International Conference on Learning Representations* (2024).
- [21] R. Schaeffer, Z. Robertson, A. Boopathy, M. Khona, K. Pistunova, J. W. Rocks, I. R. Fiete, A. Gromov, and S. Koyejo, in *The Third Blogpost Track at ICLR 2024* (2024).
- [22] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, in *International Conference on Learning Representations* (2020).
- [23] Z. Liu, E. J. Michaud, and M. Tegmark, *Omnigrok: Grokking beyond algorithmic data* (2023), arXiv:2210.01117 [cs.LG].
- [24] S. Ioffe and C. Szegedy, in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15 (JMLR.org, 2015) p. 448–456.
- [25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, in *Advances in Neural Information Processing Systems*, Vol. 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019).
- [26] H. Zhao, *Phys. Rev. E* **70**, 066137 (2004).
- [27] Q. Zhou, T. Jin, and H. Zhao, *Neural Computation* **21**, 2931 (2009).
- [28] T. Jin and H. Zhao, *Phys. Rev. E* **72**, 066111 (2005).
- [29] S. Hochreiter and J. Schmidhuber, *Neural Computation* **9**, 1 (1997), <https://direct.mit.edu/neco/article-pdf/9/1/1/813385/neco.1997.9.1.1.pdf>.
- [30] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, in *International Conference on Learning Representations* (2017).
- [31] L. Deng, *IEEE Signal Processing Magazine* **29**, 141 (2012).
- [32] H. Cheng, M. Zhang, and J. Q. Shi, *IEEE Transactions on Pattern Analysis & Machine Intelligence* **46**, 10558 (2024).
- [33] L. Prechelt, Early stopping - but when?, in *Neural Networks: Tricks of the Trade*, edited by G. B. Orr and K.-R. Müller (Springer Berlin Heidelberg, Berlin, Heidelberg, 1998) pp. 55–69.
- [34] Y. Yao, L. Rosasco, and A. Caponnetto, *Constructive approximation* **26**, 289 (2007).
- [35] R. Caruana, S. Lawrence, and C. Giles, in *Advances in Neural Information Processing Systems*, Vol. 13, edited by T. Leen, T. Dietterich, and V. Tresp (MIT Press, 2000).