Towards Understanding the Condensation of Neural Networks at Initial Training

Hanxu Zhou[1], **Qixuan Zhou**[1], **Tao Luo**[1], **Yaoyu Zhang**[1,2]*, **Zhi-Qin John Xu**[1]†
[1] School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC and Qing Yuan Research Institute, Shanghai Jiao Tong University
[2] Shanghai Center for Brain Science and Brain-Inspired Technology

Abstract

Empirical works show that for ReLU neural networks (NNs) with small initialization, input weights of hidden neurons (the input weight of a hidden neuron consists of the weight from its input layer to the hidden neuron and its bias term) condense on isolated orientations. The condensation dynamics implies that the training implicitly regularizes a NN towards one with a much smaller effective size. In this work, we illustrate the formation of the condensation in multi-layer fully connected NNs and show that the maximal number of condensed orientations in the initial training stage is twice the multiplicity of the activation function, where "multiplicity" indicates the multiple roots of activation function at origin. Our theoretical analysis confirms experiments for two cases, one is for the activation function of multiplicity one with arbitrary dimension input, which contains many common activation functions, and the other is for the layer with one-dimensional input and arbitrary multiplicity. This work makes a step towards understanding how small initialization leads NNs to condensation at the initial training stage.

1 Introduction

The question why over-parameterized neural networks (NNs) often show good generalization attracts much attention (Breiman, 1995; Zhang et al., 2021). Luo et al. (2021) found that when initialization is small, the input weights of hidden neurons in two-layer ReLU NNs (the input weight or the feature of a hidden neuron consists of the weight from its input layer to the hidden neuron and its bias term) condense on isolated orientations during the training. As illustrated in the cartoon example in Fig. 1, the condensation transforms a large network with one of only a few effective neurons, leading to an output function with low complexity. Since the complexity bounds the generalization error (Bartlett and Mendelson, 2002), the study of condensation could provide insight to how over-parameterized NNs are implicitly regularized to achieve good generalization performance in practice.

Small initialization leads NNs to rich non-linearity during the training (Mei et al., 2019; Rotskoff and Vanden-Eijnden, 2018; Chizat and Bach, 2018; Sirignano and Spiliopoulos, 2020). For example, in over-parameterized regime, small initialization can achieve low generalization error (Advani et al., 2020). Irrespective of network width, small initialization can make two-layer ReLU NNs converge to a solution with maximum margin (Phuong and Lampert, 2020). Small initialization also enables neural networks to learn features actively (Lyu et al., 2021; Luo et al., 2021). Condensation is an important phenomenon that reflects the feature learning process. Therefore, it is important to understand how condensation emerges during the training with small initialization.

*Corresponding author: zhyy.sjtu@sjtu.edu.cn.
†Corresponding author: xuzhiqin@sjtu.edu.cn.

Preprint. Under review.

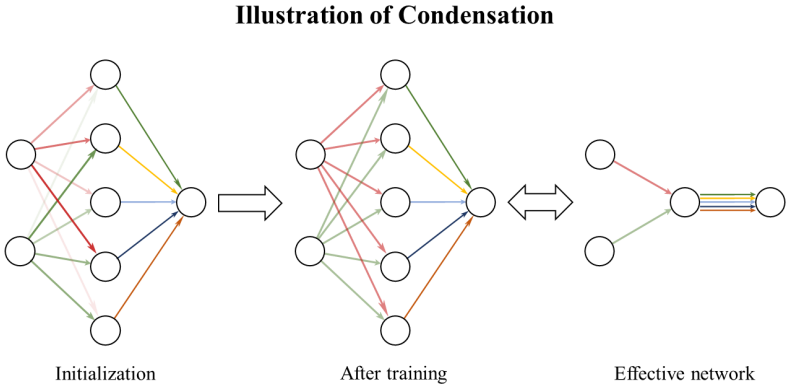


Figure 1: Illustration of condensation. The color and its intensity of a line indicate the strength of the weight. Initially, weights are random. After training, the weights from a input node to all hidden neurons are the same, i.e., condensation. Multiple hidden neurons can be replaced by an effective neuron, which has the same input weight as original hidden neurons and the output weight as the summation of all output weights of original hidden neurons.

The dynamic behavior of the training at the initial state is important for the whole training process, because it largely determines the training dynamics of a neural network and the region it ends up in (Fort et al., 2020; Hu et al., 2020), which impacts the characteristics of the neural networks in the final stage of training (Luo et al., 2021; Jiang et al., 2019; Li et al., 2018). For two-layer ReLU NNs, several works have studied the mechanism underlying the condensation at the initial training stage when the initialization of parameters goes to zero (Maennel et al., 2018; Pellegrini and Biroli, 2020). However, it still remains unclear that **for NNs of more general activation functions, how the condensation emerges at the initial training stage.**

In this work, we show that the condensation at the initial stage is closely related to the multiplicity p at $x = 0$, which means the derivative of activation at $x = 0$ is zero up to the $(p-1)th$-order and is non-zero for the p-th order. Many common activation functions, e.g., $\tanh(x)$, $\mathrm{sigmoid}(x)$, $\mathrm{softplus}(x)$, $\mathrm{Gelu}(x)$, $\mathrm{Swich}(x)$, etc, are all multiplicity $p = 1$, and $x\tanh(x)$ and $x^2\tanh(x)$ have multiplicity two and three, respectively. Our contribution is summarized as follows:

- Our extensive experiments suggest that the maximal number of condensed orientations in the initial training is twice the multiplicity of the activation function used in general NNs.
- We present a theory for the initial condensation with small initialization for two cases, one is for the activation function of multiplicity one with arbitrary dimension input, and the other is for the layer with one-dimensional input and arbitrary multiplicity. As many common activation functions are multiplicity $p = 1$, our theory would be of interest to general readers.

2 Related works

Luo et al. (2021) systematically study the effect of initialization for two-layer ReLU NN with infinite width by establishing a phase diagram, which shows three distinct regimes, i.e., linear regime (similar to the lazy regime) (Jacot et al., 2018; Arora et al., 2019; Zhang et al., 2020; E et al., 2020; Chizat and Bach, 2019), critical regime (Mei et al., 2019; Rotskoff and Vanden-Eijnden, 2018; Chizat and Bach, 2018; Sirignano and Spiliopoulos, 2020) and condensed regime (non-linear regime), based on the relative change of input weights as the width approaches infinity, which tends to 0, $O(1)$ and $+\infty$, respectively. As shown in Luo et al. (2021), two-layer ReLU NNs with infinite width do not condense in the neural tangent kernel (NTK) regime , slightly condense in the mean-field regime, and clearly condense in the non-linear regime. However, in Luo et al. (2021), it is not clear how general the condensation phenomenon is when other activation functions are used and why there is condensation.

Zhang et al. (2021a,b) propose a general Embedding Principle of loss landscape of DNNs that unravels a hierarchical structure of the loss landscape of NNs, i.e., loss landscape of an DNN contains all critical points of all the narrower DNNs. The embedding principle shows that a large DNN can experience critical points where the DNN condenses and its output is the same as that of a much smaller DNN. However, embedding principle does not explain how the training can take the DNN to such critical points.

The condensation is consistent with previous works that suggest that NNs may learn data from simple to complex patterns (Arpit et al., 2017; Xu et al., 2019; Rahaman et al., 2019; Xu et al., 2020; Jin et al., 2020; Kalimeris et al., 2019). For example, an implicit bias of frequency principle is widely observed that NNs often learn the target function from low to high frequency (Xu et al., 2019; Rahaman et al., 2019; Xu et al., 2020), which has been utilized to understand various phenomena (Ma et al., 2020; Xu and Zhou, 2021) and inspired algorithm design (Liu et al., 2020; Cai et al., 2020; Tancik et al., 2020; Li et al., 2020, 2021).

3 Preliminary: Neural networks and initial stage

A two-layer NN is

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{m} a_j \sigma(\boldsymbol{w}_j \cdot \boldsymbol{x}), \tag{1}$$

where $\sigma(\cdot)$ is the activation function, $\boldsymbol{w}_j = (\bar{\boldsymbol{w}}_j, \boldsymbol{b}_j) \in \mathbb{R}^{d+1}$ is the neuron feature including the input weight and bias terms, and $\boldsymbol{x} = (\bar{\boldsymbol{x}}, 1) \in \mathbb{R}^{d+1}$ is combination of the input sample and scalar 1, $\boldsymbol{\theta}$ is the set of all parameters, i.e., $\{a_j, \boldsymbol{w}_j\}_{j=1}^m$. For simplicity, **we call $\boldsymbol{w}_j$ as input weight or weight** and $\boldsymbol{x}$ as input sample.

A L-layer NN can be recursively defined by feeding the output of the previous layer as the input to the current hidden layer, i.e.,

$$\begin{gathered}\boldsymbol{x}^{[0]} = (\boldsymbol{x}, 1), \quad \boldsymbol{x}^{[1]} = (\sigma(\boldsymbol{W}^{[1]}.\boldsymbol{x}^{[0]}), 1), \quad \boldsymbol{x}^{[l]} = (\sigma(\boldsymbol{W}^{[l]} \cdot \boldsymbol{x}^{[l-1]}), 1), \text{ for } l \in \{2, 3, ..., L\} \\ f(\boldsymbol{\theta}, \boldsymbol{x}) = \boldsymbol{a}^\intercal \boldsymbol{x}^{[L]} \triangleq f_{\boldsymbol{\theta}}(\boldsymbol{x}),\end{gathered} \tag{2}$$

where $\boldsymbol{W}^{[l]} = (\bar{\boldsymbol{W}}^{[l]}, \boldsymbol{b}^{[l]}) \in \mathbb{R}^{m_l \times (m_{l-1}+1)}$, and m_l represents the dimension of the l-th hidden layer. For simplicity, **we also call each row of $\boldsymbol{W}^{[l]}$ as input weight or weight** and $\boldsymbol{x}^{[l-1]}$ as input to the l-th hidden layer. The target function is denoted as $f^*(\boldsymbol{x})$. The training loss function is mean squared error

$$R_S(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} (f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i))^2. \tag{3}$$

Without loss of generality, we assume that the output is one-dimensional for theoretical analysis, because, for high-dimensional cases, we only need to sum up the components directly. For summation, it does not affect the results of our theories. We consider the gradient flow training

$$\dot{\boldsymbol{\theta}} = -\nabla_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}). \tag{4}$$

To ensure that the training is close to the gradient flow, all the learning rates used in this paper are relatively small. We characterize the activation function by the following definition.

Definition 1 (multiplicity p). Suppose that $\sigma(x)$ satisfies the following condition, there exists a $p \in \mathbb{N}$ and $p \geq 1$, such that the s-th order derivative $\sigma^{(s)}(0) = 0$ for $s = 1, 2, \cdots, p-1$, and $\sigma^{(p)}(0) \neq 0$, then we say σ has multiplicity p.

Remark 3.1. Here are some examples. $\tanh(x)$, $\mathrm{sigmoid}(x)$ and $\mathrm{softplus}(x)$ have multiplicity $p = 1$. $x \tanh(x)$ has multiplicity $p = 2$.

We illustrate small initialization and initial stage as follows

Small initialization: $\boldsymbol{W}^{[l]} \sim o(1)$ and $\boldsymbol{W}^{[l]} \cdot \boldsymbol{x}^{[l-1]} \sim o(1)$ for all l's and $\boldsymbol{x}^{[l-1]}$.

We want to remark that it dose not make sense to define the initial stage by the number of training steps, because the training is affected by many factors, such as the learning rate. Therefore, in our

experiments, the epochs we use to show phenomena can across a wide range. Alternatively, we consider the initial stage as follows.

Initial stage: the period when the leading-order Taylor expansion w.r.t activated neurons is still valid for theoretical analysis of Sec. 5.

As the parameters of a neural network evolve, the loss value will also decay accordingly, which is easy to be directly observed. Therefore, we propose *an intuitive definition of the initial stage* of training by the size of loss in this article, that is the stage before the value of loss function decays to 70% of its initial value. Such a definition is reasonable, for generally a loss could decay to 1% of its initial value or even lower. The loss of the all experiments in the article can be found in Appendix A.3, and they do meet the definition of the initial stage here.

Cosine similarity: The cosine similarity for two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined as

$$D(\boldsymbol{u}, \boldsymbol{v}) = \frac{\boldsymbol{u}^\intercal \boldsymbol{v}}{(\boldsymbol{u}^\intercal \boldsymbol{u})^{1/2} (\boldsymbol{v}^\intercal \boldsymbol{v})^{1/2}}. \tag{5}$$

4 Initial condensation of input weights

In this section, we would empirically show how the condensation differs among NNs with activation function of different multiplicities in the order of a practical example, multidimensional synthetic data, and 1-d input synthetic data, followed by theoretical analysis in the next section.

4.1 Experimental setup

For Synthetic dataset: Throughout this work, we use fully-connected neural network with size, d-m-$\cdots$-m-d_{out}. The input dimension d is determined by the training data. The output dimension is $d_{out} = 1$. The number of hidden neurons m is specified in each experiment. All parameters are initialized by a Gaussian distribution $N(0, var)$. The total data size is n. The training method is Adam with full batch, learning rate lr and MSE loss. We sample the training data uniformly from a sub-domain of $\mathbb{R}^d$. *The sampling range and the target function are chosen randomly to show the generality of our experiments.*

For CIFAR10 dataset: We use Resnet18-like neural network, which has been described in Fig. 2, and the input dimension is $d = 32 * 32 * 3$. The output dimension is $d_{out} = 10$. All parameters are initialized by a Gaussian distribution $N(0, var)$. The total data size is n. The training method is Adam with batch size 128, learning rate lr and cross-entropy loss.

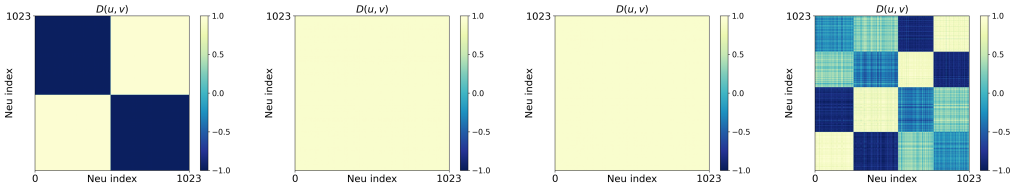
4.2 A practical example

The condensation of the weights of between the fully-connected (FC) layers of a Resnet18-like neural network on CIFAR10 is shown in Fig. 2, whose activation functions for FC layers are $\tanh(x)$, $\mathrm{sigmoid}(x)$, $\mathrm{softplus}(x)$ and $x\tanh(x)$, indicated by the corresponding sub-captions, respectively. As shown in Fig. 2(a), for activation function $\tanh(x)$, the color indicates cosine similarity $D(u, v)$ of two hidden neurons' weights, whose indexes are indicated by the abscissa and the ordinate, respectively. If the neurons are in the same beige block, $D(u, v) \sim 1$ (navy-blue block, $D(u, v) \sim -1$), their input weights have the same (opposite) direction. Input weights of hidden neurons in Fig. 2(a) condense at two opposite directions, i.e., one line. Similarly, weights of hidden neurons for NNs with $\mathrm{sigmoid}(x)$ and $\mathrm{softplus}(x)$ (Fig. 2(b, c)), which are frequently used and have multiplicity one, condense at one direction. As the multiplicity increases, NNs with $x\tanh x$ (Fig. 2(d)) condense at two different lines. These experiments suggest that the condensation is closely related to the multiplicity of the activation function.

In these experiments, we find that the performance of the Resnet18-like network with tanh FC layers with small initialization is similar to the one with common initialization in Appendix A.4.

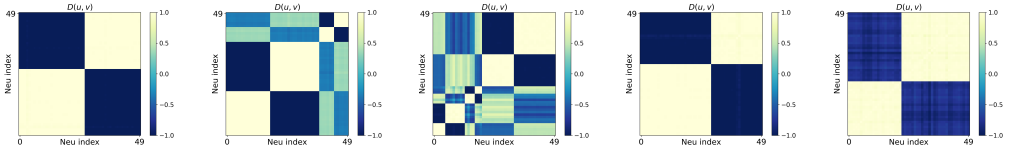
4.3 Multidimensional synthetic data

For convenience of experiments, we then use synthetic data to perform extensive experiments to study the relation between the condensation and the multiplicity of the activation function.



(a) $\tanh(x)$ (b) $\mathrm{sigmoid}(x)$ (c) $\mathrm{softplus}(x)$ (d) $x\tanh(x)$

Figure 2: Condensation of Resnet18-like neural networks on CIFAR10. Each network consists of the convolution part of resnet18 and fully-connected (FC) layers with size 1024-1024-10 and softmax. The color in figures indicates the cosine similarity of normalized input weights of two neurons in the first FC layer, whose indexes are indicated by the abscissa and the ordinate, respectively. The convolution part is equipped with ReLU activation and initialized by Glorot normal distribution (Glorot and Bengio, 2010). The activation functions are $\tanh(x)$, $\mathrm{sigmoid}(x)$, $\mathrm{softplus}(x)$ and $x\tanh(x)$ for FC layers in (a), (b), (c), and (d), separately. The numbers of steps selected in the sub-pictures are epoch 20, epoch 30, epoch 30 and epoch 61, respectively. The learning rate is 3×10^{-8}, 1×10^{-8}, 1×10^{-8} and 5×10^{-6}, separately .The FC layers are initialized by $N(0, \frac{1}{m_{\mathrm{out}}^3})$, and Adam optimizer with cross-entropy loss and batch size 128 are used for all experiments.



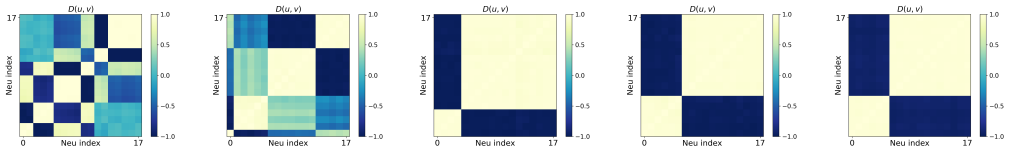
(a) $\tanh(x)$ (b) $x\tanh(x)$ (c) $x^2\tanh(x)$ (d) $\mathrm{sigmoid}(x)$ (e) $\mathrm{softplus}(x)$

Figure 3: Condensation of two-layer NNs. The color indicates $D(u,v)$ of two hidden neurons' input weights at epoch 100, whose indexes are indicated by the abscissa and the ordinate, respectively. If neurons are in the same beige block, $D(u,v)\sim 1$ (navy-blue block, $D(u,v)\sim -1$), their input weights have the same (opposite) direction. The activation functions are indicated by the sub-captions. The training data is 80 points sampled from $\sum_{k=1}^{5} 3.5\sin(5x_k+1)$, where each x_k is uniformly sampled from $[-4,2]$. $n=80$, $d=5$, $m=50$, $d_{out}=1$, $var=0.005^2$. $lr=10^{-3}, 8\times10^{-4}, 2.5\times10^{-4}$ for (a-c), (d) and (e), respectively.

We use two-layer fully-connected NNs with size 5-50-1 to fit $n=80$ training data sampled from a 5-dimensional function $\sum_{k=1}^{5} 3.5\sin(5x_k+1)$, where $\boldsymbol{x}=(x_1,x_2,\cdots,x_5)^{\intercal}\in\mathbb{R}^5$ and each x_k is uniformly sampled from $[-4,2]$. As shown in Fig. 3(a), for activation function $\tanh(x)$, input weights of hidden neurons condense at two opposite directions, i.e., one line. As the multiplicity increases, NNs with $x\tanh(x)$ (Fig. 3(b)) and $x^2\tanh x$ (Fig. 3(c)) condense at two and three different lines, respectively. For activation function $\mathrm{sigmoid}(x)$ in Fig. 3(d) and $\mathrm{softplus}(x)$ in Fig. 3(e), NNs also condense at two opposite directions.

For multi-layer NNs with different activation functions, we show that the condensation for all hidden layers is similar to the two-layer NNs. In deep networks, residual connection is often introduced to overcome the vanishing of gradient. To show the generality of condensation, we perform an experiment of six-layer NNs with residual connections. To show the difference of various activation functions, we set the activation functions for hidden layer 1 to hidden layer 5 as $x^2\tanh(x)$, $x\tanh(x)$, $\mathrm{sigmoid}(x)$, $\tanh(x)$ and $\mathrm{softplus}(x)$, respectively. The structure of the residual is $\boldsymbol{h}_{l+1}(\boldsymbol{x})=\sigma(\boldsymbol{W}_l\boldsymbol{h}_l(\boldsymbol{x})+b_l)+\boldsymbol{h}_l(\boldsymbol{x})$, where $\boldsymbol{h}_l(\boldsymbol{x})$ is the output of the l-th layer. As shown in Fig. 4, input weights condense at three, two, one, one and one lines for hidden layer 1 to hidden layer 5, respectively. Note that residual connections are not necessary. We show an experiment of the same structure as in Fig. 4 but without residual connections in Appendix A.5.

Through these experiments, we conjecture that the maximal number of condensed orientations at initial training is twice the multiplicity of the activation function used. To understand the mechanism



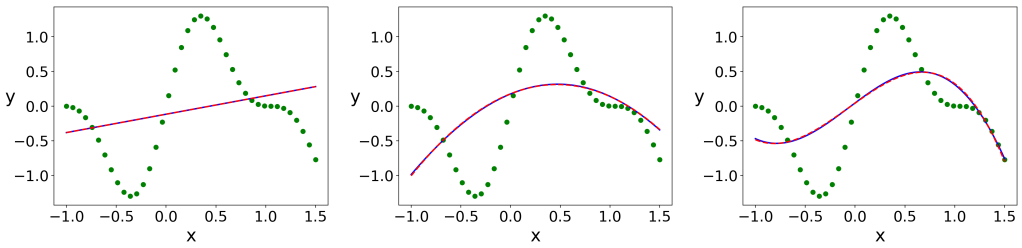
(a) layer 1 (b) layer 2 (c) layer 3 (d) layer 4 (e) layer 5

Figure 4: Condensation of six-layer NNs with residual connections. The activation functions for hidden layer 1 to hidden layer 5 are $x^2\tanh(x)$, $x\tanh(x)$, $\mathrm{sigmoid}(x)$, $\tanh(x)$ and $\mathrm{softplus}(x)$, respectively. The numbers of steps selected in the sub-pictures are epoch 1000, epoch 900, epoch 900, epoch 1400 and epoch 1400, respectively, while the NN is only trained once. The color indicates $D(u,v)$ of two hidden neurons' input weights, whose indexes are indicated by the abscissa and the ordinate, respectively. The training data is 80 points sampled from a 3-dimensional function $\sum_{k=1}^{3} 4\sin(12x_k+1)$, where each x_k is uniformly sampled from $[-4,2]$. $n=80$, $d=3$, $m=18$, $d_{out}=1$, $var=0.01^2$, $lr=4\times10^{-5}$.

of the initial condensation, we turn to experiments of 1-d input and two-layer NNs, which can be clearly visualized in the next subsection.

4.4 1-d input and two-layer NN

For 1-d data, we visualize the evolution of the two-layer NN output and each weight, which confirms the connection between the condensation and the multiplicity of the activation function.



(a) $\tanh(x)$ (b) $x\tanh(x)$ (c) $x^2\tanh(x)$

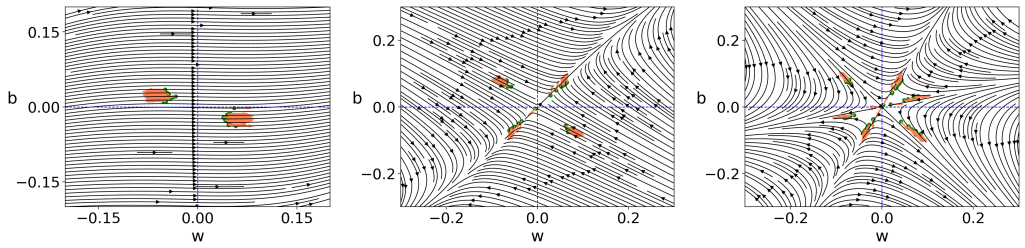
Figure 5: The outputs of two-layer NNs at epoch 1000 with activation function $\tanh(x)$, $x\tanh(x)$, and $x^2\tanh(x)$ are displayed, respectively. The training data is 40 points uniformly sampled from $\sin(3x)+\sin(6x)/2$ with $x\in[-1,1.5]$, illustrated by green dots. The blue solid lines are the NN outputs at test points, while the red dashed auxiliary lines are the first, second, third and first order polynomial fittings of the test points for (a, b, c), respectively. Parameters are $n=40$, $d=1$, $m=100$, $d_{out}=1$, $var=0.005^2$, $lr=5\times10^{-4}$.

We display the outputs at initial training in Fig. 5. Due to the small magnitude of parameters, an activation function with multiplicity p can be well approximated by a p-th order polynominal, thus, the NN output can also be approximated by a p-th order polynominal. As shown in Fig. 5, the NN outputs with activation function $\tanh(x)$, $x\tanh(x)$ and $x^2\tanh(x)$ overlap well with the auxiliary of a linear, a quadratic and a cubic polynominal curve, respectively in the beginning. This experiment, although simple, but convincingly shows that NN does not always learn a linear function at the initial training stage and the complexity of such learning depends on the activation function.

We visualize the direction field for input weight $\boldsymbol{w}_j := (w_j, b_j)$, following the gradient flow,

$$\dot{\boldsymbol{w}}_j = -\frac{a_j}{n}\sum_{i=1}^{n} e_i \sigma'(\boldsymbol{w}_j\cdot\boldsymbol{x}_i)\boldsymbol{x}_i,$$

where $e_i := f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)$. Since we only care about the direction of $\boldsymbol{w}_j$ and a_j is a scalar at each epoch, we can visualize $\dot{\boldsymbol{w}}_j$ by $\dot{\boldsymbol{w}}_j/a_j$. For simplicity, we do not distinguish $\dot{\boldsymbol{w}}_j/a_j$ and $\dot{\boldsymbol{w}}_j$ if there



(a) $\tanh(x)$ (b) $x\tanh(x)$ (c) $x^2\tanh(x)$

Figure 6: The direction field for input weight $\boldsymbol{w} := (w, b)$ of the dynamics in (4.4) at epoch 200. All settings are the same as Fig. 5. Around the original point, the field has one, two, three stables lines, on which an input weight would keep its direction, for $\tanh(x)$, $x\tanh(x)$, and $x^2\tanh(x)$, respectively. We also display the value of each weight by the green dots and the corresponding directions by the orange arrows.

is no ambiguity. When we compute $\dot{\boldsymbol{w}}_j$ for different j's, $e_i\boldsymbol{x}_i$ for $(i = 1, \cdots, n)$ is independent with j. Then, at each epoch, for a set of $\{e_i, \boldsymbol{x}_i\}_{i=1}^n$, we can consider the following direction field

$$\dot{\boldsymbol{\omega}} = -\frac{1}{n}\sum_{i=1}^{n} e_i\boldsymbol{x}_i\sigma'(\boldsymbol{\omega}\cdot\boldsymbol{x}_i).$$

When $\boldsymbol{\omega}$ is set as $\boldsymbol{w}_j$, we can obtain $\dot{\boldsymbol{w}}_j$. As shown in Fig. 6, around the original point, the field has one, two, three stables lines, on which a neuron would keep its direction, for $\tanh(x)$, $x\tanh(x)$, and $x^2\tanh(x)$, respectively. We also display the input weight of each neuron on the field by the green dots and their corresponding velocity directions by the orange arrows. Similarly to the high-dimensional cases, NNs with multiplicity p activation functions condense at p different lines for $p = 1, 2, 3$. Therefore, It is reasonable to conjecture that the maximal number of condensed orientations is twice the multiplicity of the activation function used.

Taken together, we have empirically shown that the multiplicity of the activation function is a key factor that determines the complexity of the initial output and condensation. To facilitate the understanding of the evolution of condensation in the initial stage, we show several steps during the initial stage of each example in Appendix A.6.

5 Analysis of the initial condensation of input weights

In this section, we would present a preliminary analysis to understand how the multiplicity of the activation function affects the initial condensation. At each training step, we consider the velocity field of weights in each hidden layer of a neural networks.

Considering a network with L hidden layers, we use row vector $\boldsymbol{W}_j^{[k]}$ to represent the weight from the $(k$-1)-th layer to the j-th neuron in the k-th layer. Since condensation is always accompany with small initialization, together with the initial stage defined in Sec. 3, we make the following assumptions,

Assumption 1. Small initialization infers that $\boldsymbol{W}_j^{[k]}\cdot\boldsymbol{x}^{[k-1]} \sim o(1)$ applies for all k's and j's.

Assumption 2. During the initial stage of condensation, Taylor expansion of each corresponding activated neurons holds, i.e.,

$$\sigma'(\boldsymbol{W}_j^{[k]}\cdot\boldsymbol{x}^{[k-1]}) = \sigma'(0) + \cdots + \frac{\sigma^{(\gamma)}(0)}{(\gamma-1)!}(\boldsymbol{W}_j^{[k]}\cdot\boldsymbol{x}^{[k-1]})^{\gamma-1} + O((\boldsymbol{W}_j^{[k]}\cdot\boldsymbol{x}^{[k-1]})^{\gamma}) \tag{6}$$

Suppose the activation function has multiplicity p, i.e., $\sigma^{(s)}(0) = 0$ for $s = 1, 2, \cdots, p-1$, and $\sigma^{(p)}(0) \neq 0$. Then, together with Assumption 2, we have

$$\sigma'(\boldsymbol{W}_j^{[k]}\cdot\boldsymbol{x}^{[k-1]}) \approx \frac{\sigma^{(p)}(0)}{(p-1)!}(\boldsymbol{W}_j^{[k]}\cdot\boldsymbol{x}^{[k-1]})^{p-1}. \tag{7}$$

For each k and j, $\boldsymbol{W}_j^{[k]}$ satisfies the following dynamics, (see Appendix A.2)

$$\dot{r} = \boldsymbol{u} \cdot \dot{\boldsymbol{w}}, \quad \dot{\boldsymbol{u}} = \frac{\dot{\boldsymbol{w}} - (\dot{\boldsymbol{w}} \cdot \boldsymbol{u})\boldsymbol{u}}{r}. \tag{8}$$

where $\boldsymbol{w}$ can represent $\boldsymbol{W}_j^{[k]\intercal}$ for all k's and j's, $r = \|\boldsymbol{w}\|_2$ is the amplitude, and $\boldsymbol{u} = \boldsymbol{w}/r$.

For convenience, we define an operator $\mathcal{P}$ satisfying $\mathcal{P}\boldsymbol{w} := \dot{\boldsymbol{w}} - \boldsymbol{u}(\dot{\boldsymbol{w}} \cdot \boldsymbol{u})$. To specify the condensation for theoretical analysis, we make the following definition,

Condensation: the weight evolves towards a direction which will not change in the direction field and is defined as follows,

$$\dot{\boldsymbol{u}} = 0 \Leftrightarrow \mathcal{P}\boldsymbol{w} := \dot{\boldsymbol{w}} - \boldsymbol{u}(\dot{\boldsymbol{w}} \cdot \boldsymbol{u}) = 0. \tag{9}$$

Since $\dot{\boldsymbol{w}} \cdot \boldsymbol{u}$ is a scalar, $\dot{\boldsymbol{w}}$ is parallel with $\boldsymbol{u}$. $\boldsymbol{u}$ is a unit vector, therefore, we have $\boldsymbol{u} = \pm\dot{\boldsymbol{w}}/\|\dot{\boldsymbol{w}}\|_2$.

Remark 1 (An intuitive explanation for condensation). In this work, we consider NNs with sufficiently small parameters. Suppose $r = \|\boldsymbol{w}\|_2 \sim O(\epsilon)$, where ϵ is a small quantity, then dynamics (8) will show that $O(\dot{r}) \sim O(\dot{\boldsymbol{w}})$ and $O(\dot{\boldsymbol{u}}) \sim O(\dot{r})/O(\epsilon)$. Here $O(\dot{r}) \sim O(\dot{\boldsymbol{w}})$ refers that the evolution of $\dot{r}$ is at the same order as every component of $\dot{\boldsymbol{w}}$, and it is the same for $O(\dot{\boldsymbol{u}}) \sim O(\dot{r})/O(\epsilon)$. Therefore, the orientation $\boldsymbol{u}$ would move much more quickly than the amplitude r. By the dynamics for $\boldsymbol{w}$ (Equ. 10) and the Taylor approximation with multiplicity p (Equ. 7), it is easy to find that the solutions for Equ. 9 are finite. Then, taken together, the orientation $\boldsymbol{u}$ would converge rapidly into certain directions, leading to condensation.

In the following, we study the case of (i) $p = 1$ and (ii) $m_{k-1} = 1$ (the dimension of input of the k-th layer equals one), and reach the following theorem,

Theorem 5.1. *Under Assumption 1 and 2, suppose we only consider the leading-order Taylor expansion of Equ. 9, then the maximal number of roots for Equ. 9 is twice the multiplicity of the activation function used as initialization towards zero for two cases: (i) $p = 1$ and (ii) $m_{k-1} = 1$.*

Proof. **Case 1:** $p = 1$

By gradient flow, we can obtain the dynamics for $\boldsymbol{W}_j^{[k]}$ (see Appendix A.2),

$$\dot{\boldsymbol{w}}^\intercal = \dot{\boldsymbol{W}}_j^{[k]} = -\frac{1}{n}\sum_{i=1}^{n}\left(f(\boldsymbol{\theta}, \boldsymbol{x}_i) - y_i\right)\left[\mathrm{diag}\{\sigma'(\boldsymbol{W}^{[k]} \cdot \boldsymbol{x}_i^{[k-1]})\}(E^{[k+1:L]}\boldsymbol{a})\right]_j \boldsymbol{x}_i^{[k-1]\intercal}, \tag{10}$$

where we use $E^l = \boldsymbol{W}^{[l]\intercal}\,\mathrm{diag}\{\sigma'(\boldsymbol{W}^{[l]} \cdot \boldsymbol{x}^{[l-1]})\}$, for $l \in \{2, 3, ..., L\}$, $E^{[q:p]} = E^q E^{q+1} ... E^p$, and $\boldsymbol{x}_i^{[k]}$ represents the neurons of the k-th layer generated by the i-th sample.

For a fixed step, we only consider the gradient of loss w.r.t. $\boldsymbol{W}_j^{[k]}$. According to our assumption $p = 1$, we have $\sigma'(0) \neq 0$. Suppose that parameters are small and denote $e_i := (f(\boldsymbol{\theta}, \boldsymbol{x}_i) - y_i)$. By Taylor expansion,

$$\begin{aligned}\mathcal{P}\boldsymbol{w} \overset{\text{leading order}}{\approx} \mathcal{Q}\boldsymbol{w} := -\frac{1}{n}\{(\mathrm{diag}\{\sigma'(\boldsymbol{0})\}(E^{[k+1:L]}\boldsymbol{a}))_j \cdot \sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}\} \\ +\{(\frac{1}{n}(\mathrm{diag}\{\sigma'(\boldsymbol{0})\}(E^{[k+1:L]}\boldsymbol{a}))_j \cdot \sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]} \cdot \boldsymbol{u})\boldsymbol{u}\} = 0,\end{aligned}$$

where operator $\mathcal{Q}$ is the leading-order approximation of operator $\mathcal{P}$, and here $E^{[k+1:L]}$ is independent with i, because $\mathrm{diag}\{\sigma'(\boldsymbol{W}^{[l]}\boldsymbol{x}^{[l-1]})\} \approx \mathrm{diag}\{\sigma'(0)\}$. Since $\mathrm{diag}\{\sigma'(\boldsymbol{0})\} = c\boldsymbol{I}, c \neq 0$ by assumption, and, WLOG, we assume $\boldsymbol{a} \neq 0$, then

$$\mathcal{Q}\boldsymbol{w} = 0 \Leftrightarrow \sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]} = \left(\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]} \cdot \boldsymbol{u}\right)\boldsymbol{u}.$$

We have

$$\boldsymbol{u}=\frac{\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}}{\|\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}\|_2} \ \ or \ \ \boldsymbol{u}=-\frac{\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}}{\|\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}\|_2}.$$

This calculation shows that for layer k, the input weights for any hidden neuron j have the same two stable directions. Together with the analysis before, i.e., when parameters are sufficiently small, the orientation $\boldsymbol{u}$ would move much more quickly than the amplitude r, all input weights would move towards the same direction or the opposite direction, i.e., condensation on a line, under small initialization.

Case 2: the k-th layer with one-dimensional input, i.e., $m_{k-1}=1$

By the definition of the multiplicity p, we have

$$\sigma'(\boldsymbol{w}\cdot\boldsymbol{x}_i)=\frac{\sigma^{(p)}(0)}{(p-1)!}(\boldsymbol{w}\cdot\boldsymbol{x}_i)^{p-1}+o((\boldsymbol{w}\cdot\boldsymbol{x}_i)^{p-1}).$$

where $(\cdot)^{p-1}$ and $\sigma^{(p)}(\cdot)$ operate on component here. Then up to the leading order in terms of the magnitude of $\boldsymbol{\theta}$, we have (see Appendix A.2)

$$\mathcal{P}\boldsymbol{w}\overset{\text{leading order}}{\approx}\mathcal{Q}\boldsymbol{w}:=-\{(\frac{1}{n}\sum_{i=1}^{n} e_i\boldsymbol{x}_i^{[k-1]}(\boldsymbol{w}^\intercal\boldsymbol{x}_i^{[k-1]})^{p-1})\cdot[\mathrm{diag}\{\frac{\sigma^{(p)}(\boldsymbol{0})}{(p-1)!}\}(E^{[k+1:L]}\boldsymbol{a})]_j\}$$

$$+\{((\frac{1}{n}\sum_{i=1}^{n} e_i\boldsymbol{x}_i^{[k-1]}(\boldsymbol{w}^\intercal\boldsymbol{x}_i^{[k-1]})^{p-1})\cdot[\mathrm{diag}\{\frac{\sigma^{(p)}(\boldsymbol{0})}{(p-1)!}\}(E^{[k+1:L]}\boldsymbol{a})]_j\cdot\boldsymbol{u})\boldsymbol{u}\}.$$

WLOG, we also assume $\boldsymbol{a}\neq 0$. And by definition, $\boldsymbol{w}=r\boldsymbol{u}$, we have

$$\mathcal{Q}\boldsymbol{w}=0\ \Leftrightarrow\ \boldsymbol{u}=\frac{\frac{1}{n}\sum_{i=1}^{n} e_i\boldsymbol{x}_i^{[k-1]}(\boldsymbol{u}^\intercal\boldsymbol{x}_i^{[k-1]})^{p-1}}{\|\frac{1}{n}\sum_{i=1}^{n} e_i\boldsymbol{x}_i^{[k-1]}(\boldsymbol{u}^\intercal\boldsymbol{x}_i^{[k-1]})^{p-1}\|_2}\ \ or\ \ \boldsymbol{u}=-\frac{\frac{1}{n}\sum_{i=1}^{n} e_i\boldsymbol{x}_i^{[k-1]}(\boldsymbol{u}^\intercal\boldsymbol{x}_i^{[k-1]})^{p-1}}{\|\frac{1}{n}\sum_{i=1}^{n} e_i\boldsymbol{x}_i^{[k-1]}(\boldsymbol{u}^\intercal\boldsymbol{x}_i^{[k-1]})^{p-1}\|_2}.$$

Since $m_{k-1}+1=2$ (for $\boldsymbol{x}^{[k-1]}=(\sigma(\boldsymbol{W}^{[k-1]}\boldsymbol{x}^{[k-2]}),1)$ and $m_{k-1}=1$), we denote $\boldsymbol{u}=(u_1,u_2)^\intercal\in\mathbb{R}^2$ and $\boldsymbol{x}_i^{[k-1]}=((\boldsymbol{x}_i^{[k-1]})_1,(\boldsymbol{x}_i^{[k-1]})_2)^\intercal\in\mathbb{R}^2$, then,

$$\frac{\sum_{i=1}^{n}(u_1(\boldsymbol{x}_i^{[k-1]})_1+u_2(\boldsymbol{x}_i^{[k-1]})_2)^{p-1}e_i(\boldsymbol{x}_i^{[k-1]})_1}{\sum_{i=1}^{n}(u_1(\boldsymbol{x}_i^{[k-1]})_1+u_2(\boldsymbol{x}_i^{[k-1]})_2)^{p-1}e_i(\boldsymbol{x}_i^{[k-1]})_2}=\frac{u_1}{u_2}\triangleq\hat{u}.$$

We obtain the equation for $\hat{u}$,

$$\sum_{i=1}^{n}(\hat{u}(\boldsymbol{x}_i^{[k-1]})_1+(\boldsymbol{x}_i^{[k-1]})_2)^{p-1}e_i(\boldsymbol{x}_i^{[k-1]})_1=\hat{u}\sum_{i=1}^{n}(\hat{u}(\boldsymbol{x}_i^{[k-1]})_1+(\boldsymbol{x}_i^{[k-1]})_2)^{p-1}e_i(\boldsymbol{x}_i^{[k-1]})_2.$$

Since it is an univariate p-th order equation, $\hat{u}=\frac{u_1}{u_2}$ has at most p complex roots. Because $\boldsymbol{u}$ is a unit vector, $\boldsymbol{u}$ at most has p pairs of values, in which each pair are opposite. □

Taken together, our theoretical analysis is consistent with our experiments, that is, the maximal number of condensed orientations is twice the multiplicity of the activation function used when parameters are small. As many commonly used activation functions are either multiplicity $p=1$ or ReLU-like, our theoretical analysis is widely applied and sheds light on practical training.

6 Discussion

In this work, we have shown that the characteristic of the activation function, i.e., multiplicity, is a key factor to understanding the complexity of NN output and the weight condensation at initial training. The condensation restricts the NN to be effectively low-capacity at the initial training stage, even for finite-width NNs. During the training, the NN increases its capacity to better fit the data,

leading to a potential explanation for their good generalization in practical problems. This work also serves as a starting point for further studying the condensation for multiple-layer neural networks throughout the training process.

For general multiplicity with high-dimensional input data, the theoretical analysis for the initial condensation is a very difficult problem, which is equivalent to counting the number of the roots of a high-order high-dimensional polynomial with a special structure originated from NNs. Training data can also affect the condensation but not the maximal number of condensed orientations. When data is simple, such as low frequency, the number of the condensed orientations can be less, some experiments of MNIST and CIFAR100 can be found in Appendix A.7.

References

L. Breiman, Reflections after refereeing papers for nips, The Mathematics of Generalization XX (1995) 11–15.

C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Communications of the ACM 64 (2021) 107–115.

T. Luo, Z.-Q. J. Xu, Z. Ma, Y. Zhang, Phase diagram for two-layer relu neural networks at infinite-width limit, Journal of Machine Learning Research 22 (2021) 1–47.

P. L. Bartlett, S. Mendelson, Rademacher and gaussian complexities: Risk bounds and structural results, Journal of Machine Learning Research 3 (2002) 463–482.

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

G. M. Rotskoff, E. Vanden-Eijnden, Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 7146–7155.

L. Chizat, F. Bach, On the global convergence of gradient descent for over-parameterized models using optimal transport, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 3040–3050.

J. Sirignano, K. Spiliopoulos, Mean field analysis of neural networks: A central limit theorem, Stochastic Processes and their Applications 130 (2020) 1820–1852.

M. S. Advani, A. M. Saxe, H. Sompolinsky, High-dimensional dynamics of generalization error in neural networks, Neural Networks 132 (2020) 428–446.

M. Phuong, C. H. Lampert, The inductive bias of relu networks on orthogonally separable data, in: International Conference on Learning Representations, 2020.

K. Lyu, Z. Li, R. Wang, S. Arora, Gradient descent on two-layer nets: Margin maximization and simplicity bias, Advances in Neural Information Processing Systems 34 (2021).

S. Fort, G. K. Dziugaite, M. Paul, S. Kharaghani, D. M. Roy, S. Ganguli, Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel, Advances in Neural Information Processing Systems 33 (2020) 5850–5861.

W. Hu, L. Xiao, B. Adlam, J. Pennington, The surprising simplicity of the early-time learning dynamics of neural networks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/c6dfc6b7c601ac2978357b7a81e2d7ae-Abstract.html.

Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, S. Bengio, Fantastic generalization measures and where to find them, arXiv preprint arXiv:1912.02178 (2019).

H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, Visualizing the loss landscape of neural nets, Advances in neural information processing systems 31 (2018).

H. Maennel, O. Bousquet, S. Gelly, Gradient descent quantizes relu network features, arXiv preprint arXiv:1803.08367 (2018).

F. Pellegrini, G. Biroli, An analytic theory of shallow networks dynamics for hinge loss classification, Advances in Neural Information Processing Systems 33 (2020).

A. Jacot, F. Gabriel, C. Hongler, Neural tangent kernel: convergence and generalization in neural networks, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 8580–8589.

S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, R. Wang, On exact computation with an infinitely wide neural net, Advances in Neural Information Processing Systems 32 (2019) 8141–8150.

Y. Zhang, Z.-Q. J. Xu, T. Luo, Z. Ma, A type of generalization error induced by initialization in deep neural networks, in: Mathematical and Scientific Machine Learning, PMLR, 2020, pp. 144–164.

W. E, C. Ma, L. Wu, A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics, Science China Mathematics (2020) 1–24.

L. Chizat, F. Bach, A note on lazy training in supervised differentiable programming, in: 32nd Conf. Neural Information Processing Systems (NeurIPS 2018), 2019.

Y. Zhang, Z. Zhang, T. Luo, Z.-Q. J. Xu, Embedding principle of loss landscape of deep neural networks, arXiv preprint arXiv:2105.14573 (2021a).

Y. Zhang, Y. Li, Z. Zhang, T. Luo, Z.-Q. J. Xu, Embedding principle: a hierarchical structure of loss landscape of deep neural networks, arXiv preprint arXiv:2111.15527 (2021b).

D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, et al., A closer look at memorization in deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 233–242.

Z.-Q. J. Xu, Y. Zhang, Y. Xiao, Training behavior of deep neural network in frequency domain, International Conference on Neural Information Processing (2019) 264–274.

N. Rahaman, D. Arpit, A. Baratin, F. Draxler, M. Lin, F. A. Hamprecht, Y. Bengio, A. Courville, On the spectral bias of deep neural networks, International Conference on Machine Learning (2019).

Z.-Q. J. Xu, Y. Zhang, T. Luo, Y. Xiao, Z. Ma, Frequency principle: Fourier analysis sheds light on deep neural networks, Communications in Computational Physics 28 (2020) 1746–1767.

P. Jin, L. Lu, Y. Tang, G. E. Karniadakis, Quantifying the generalization error in deep learning in terms of data distribution and neural network smoothness, Neural Networks 130 (2020) 85–99.

D. Kalimeris, G. Kaplun, P. Nakkiran, B. Edelman, T. Yang, B. Barak, H. Zhang, Sgd on neural networks learns functions of increasing complexity, Advances in Neural Information Processing Systems 32 (2019) 3496–3506.

C. Ma, L. Wu, E. Weinan, The slow deterioration of the generalization error of the random feature model, in: Mathematical and Scientific Machine Learning, PMLR, 2020, pp. 373–389.

Z.-Q. J. Xu, H. Zhou, Deep frequency principle towards understanding why deeper learning is faster, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021.

Z. Liu, W. Cai, Z.-Q. J. Xu, Multi-scale deep neural network (mscalednn) for solving poisson-boltzmann equation in complex domains, Communications in Computational Physics 28 (2020) 1970–2001.

W. Cai, X. Li, L. Liu, A phase shift deep neural network for high frequency approximation and wave problems, SIAM Journal on Scientific Computing 42 (2020) A3285–A3312.

M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, R. Ng, Fourier features let networks learn high frequency functions in low dimensional domains, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 7537–7547.

X.-A. Li, Z.-Q. J. Xu, L. Zhang, A multi-scale dnn algorithm for nonlinear elliptic equations with multiple scales, Communications in Computational Physics 28 (2020) 1886–1906.

X.-A. Li, Z.-Q. J. Xu, L. Zhang, Subspace decomposition based dnn algorithm for elliptic-type multi-scale pdes, arXiv preprint arXiv:2112.06660 (2021).

X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, 2010, pp. 249–256.

K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

A Appendix

A.1 Basic definitions

In this study, we first consider the neural network with 2 hidden layers,

A two-layer NN is

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{m} a_j \sigma(\boldsymbol{w}_j \cdot \boldsymbol{x}), \tag{11}$$

where $\sigma(\cdot)$ is the activation function, $\boldsymbol{w}_j = (\bar{\boldsymbol{w}}_j, \boldsymbol{b}_j) \in \mathbb{R}^{d+1}$ is the neuron feature including the input weight and bias terms, and $\boldsymbol{x} = (\bar{\boldsymbol{x}}, 1) \in \mathbb{R}^{d+1}$ is combination of the input sample and scalar 1, $\boldsymbol{\theta}$ is the set of all parameters, i.e., $\{a_j, \boldsymbol{w}_j\}_{j=1}^m$. For simplicity, **we call $\boldsymbol{w}_j$ as input weight or weight** and $\boldsymbol{x}$ as input sample.

Then, we consider the neural network with l hidden layers,

$$\begin{gathered} \boldsymbol{x}^{[0]} = (\boldsymbol{x}, 1), \quad \boldsymbol{x}^{[1]} = (\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x}^{[0]}), 1), \quad \boldsymbol{x}^{[l]} = (\sigma(\boldsymbol{W}^{[l]}\boldsymbol{x}^{[l-1]}), 1), \text{ for } l \in \{2, 3, ..., L\} \\ f(\boldsymbol{\theta}, \boldsymbol{x}) = \boldsymbol{a}^\intercal \boldsymbol{x}^{[L]} \triangleq f_{\boldsymbol{\theta}}(\boldsymbol{x}), \end{gathered} \tag{12}$$

where $\boldsymbol{W}^{[l]} = (\bar{\boldsymbol{W}}^{[l]}, \boldsymbol{b}^{[l]}) \in \mathbb{R}^{(m_l \times m_{l-1})}$, and m_l represents the dimension of the l-th hidden layer. The initialization of $\boldsymbol{W}_{k,k'}^{[l]}, l \in \{1, 2, 3, ..., L\}$ and $\boldsymbol{a}_k$ obey normal distribution $\boldsymbol{W}_{k,k'}^{[l]} \sim \mathcal{N}(0, \beta_l^2)$ for $l \in \{1, 2, 3, ..., L\}$ and $\boldsymbol{a}_k \sim \mathcal{N}(0, \beta_{L+1}^2)$.

The loss function is mean squared error given below,

$$R_s(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} (f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2. \tag{13}$$

For simplification, we denote $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ as f in following.

A.2 Derivations for concerned quantities

A.2.1 Neural networks with three hidden layers

In order to better understand the gradient of the parameter matrix of the multi-layer neural network, we first consider the case of the three-layer neural network,

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) := \boldsymbol{a}^\intercal \sigma(\boldsymbol{W}^{[2]} \sigma(\boldsymbol{W}^{[1]} \boldsymbol{x})), \tag{14}$$

with the mean squared error as the loss function,

$$R_s(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} (f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2. \tag{15}$$

We calculate $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{W}^{[2]}}$ and $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{W}^{[1]}}$ respectively, using differential form,

$$\mathrm{d}f = \mathrm{tr}((\frac{\partial f}{\partial \boldsymbol{x}})^\intercal \mathrm{d}f). \tag{16}$$

We consider $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{W}^{[2]}}$ first,

$$\begin{aligned} \mathrm{d}f &= \mathrm{tr}\{\mathrm{d}(\boldsymbol{a}^\intercal \sigma(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]}))\} \\ &= \mathrm{tr}\{\boldsymbol{a}^\intercal \mathrm{d}(\sigma(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]}))\} \\ &= \mathrm{tr}\{\boldsymbol{a}^\intercal \sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]}) \odot \mathrm{d}\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]}\} \\ &= \mathrm{tr}\{(\boldsymbol{a} \odot \sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]})^\intercal \mathrm{d}\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]}\} \\ &= \mathrm{tr}\{\boldsymbol{x}^{[1]}(\boldsymbol{a} \odot \sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]})^\intercal \mathrm{d}\boldsymbol{W}^{[2]}\} \\ &= \mathrm{tr}\{((\boldsymbol{a} \odot \sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]}))\boldsymbol{x}^{[1]\intercal})^\intercal \mathrm{d}\boldsymbol{W}^{[2]}\}, \end{aligned} \tag{17}$$

where $\odot$ is Hadamard Product, and it is the multiplication of matrix elements of the same position. Hence,

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{W}^{[2]}} &= (\boldsymbol{a}\odot\sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]}))\boldsymbol{x}^{[1]\intercal}\\ &= \mathrm{diag}\{\sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]})\}\boldsymbol{a}\boldsymbol{x}^{[1]\intercal}.\end{aligned} \tag{18}$$

Then, we consider $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{W}^{[1]}}$,

$$\begin{aligned}\mathrm{d}f &= \mathrm{tr}\{(\boldsymbol{a}\odot\sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]}))^{\intercal}\boldsymbol{W}^{[2]}\mathrm{d}\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x})\}\\ &= \mathrm{tr}\{(\boldsymbol{W}^{[2]\intercal}(\boldsymbol{a}\odot\sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]})))^{\intercal}\sigma^{'}(\boldsymbol{W}^{[1]}\boldsymbol{x})\odot\mathrm{d}(\boldsymbol{W}^{[1]}\boldsymbol{x})\}\\ &= \mathrm{tr}\{((\boldsymbol{W}^{[2]\intercal}(\boldsymbol{a}\odot\sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]}))\odot\sigma^{'}(\boldsymbol{W}^{[1]}\boldsymbol{x}))^{\intercal}\mathrm{d}(\boldsymbol{W}^{[1]}\boldsymbol{x})\}\\ &= \mathrm{tr}\{[((\boldsymbol{W}^{[2]\intercal}(\boldsymbol{a}\odot\sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]}))\odot\sigma^{'}(\boldsymbol{W}^{[1]}\boldsymbol{x}))\boldsymbol{x}^{\intercal}]^{\intercal}\mathrm{d}(\boldsymbol{W}^{[1]})\}.\end{aligned} \tag{19}$$

Hence, we have,

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{W}^{[1]}} &= ((\boldsymbol{W}^{[2]\intercal}(\boldsymbol{a}\odot\sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]})))\odot\sigma^{'}(\boldsymbol{W}^{[1]}\boldsymbol{x}))\boldsymbol{x}^{\intercal}\\ &= \mathrm{diag}\{\sigma^{'}(\boldsymbol{W}^{[1]}\boldsymbol{x})\}\boldsymbol{W}^{[2]\intercal}\mathrm{diag}\{\sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}^{[1]})\}\boldsymbol{a}\boldsymbol{x}^{\intercal}.\end{aligned} \tag{20}$$

Through the chain rule, we can get the evolution equation of $\boldsymbol{W}^{[1]}$ and $\boldsymbol{W}^{[2]}$,

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{W}^{[1]}}{\mathrm{d}t} &= -\frac{\mathrm{d}R_s(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{W}^{[1]}}\\ &= -\frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{\theta},\boldsymbol{x}_i)-y_i)\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{W}^{[1]}}\\ &= -\frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{\theta},\boldsymbol{x}_i)-y_i)\mathrm{diag}\{\sigma^{'}(\boldsymbol{W}^{[1]}\boldsymbol{x}_i)\}\boldsymbol{W}^{[2]\intercal}\mathrm{diag}\{\sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}_i^{[1]})\}\boldsymbol{a}\boldsymbol{x}_i^{\intercal},\end{aligned} \tag{21}$$

and

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{W}^{[2]}}{\mathrm{d}t} &= -\frac{\mathrm{d}R_s(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{W}^{[2]}}\\ &= -\frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{\theta},\boldsymbol{x}_i)-y_i)\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{W}^{[1]}}\\ &= -\frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{\theta},\boldsymbol{x}_i)-y_i)\mathrm{diag}\{\sigma^{'}(\boldsymbol{W}^{[2]}\boldsymbol{x}_i^{[1]})\}\boldsymbol{a}\boldsymbol{x}_i^{[1]\intercal}.\end{aligned} \tag{22}$$

A.2.2 L hidden layers condition

And, we consider the neural network with L hidden layers,

$$\begin{aligned}\mathrm{d}f &= \mathrm{tr}\{\mathrm{d}\boldsymbol{a}^{\intercal}\mathrm{d}\sigma(\boldsymbol{W}^{[L]}\boldsymbol{x}^{[L-1]})\}\\ &= \mathrm{tr}\{(\boldsymbol{a}\odot\sigma^{'}(\boldsymbol{W}^{[L]}\boldsymbol{x}^{[L-1]}))^{\intercal}\mathrm{d}\boldsymbol{W}^{[L]}\sigma(\boldsymbol{W}^{[L-1]}\boldsymbol{x}^{[L-2]})\}\\ &= \mathrm{tr}\{(\boldsymbol{W}^{[L]\intercal}\Lambda_L)^{\intercal}\sigma^{'}(\boldsymbol{W}^{[L-1]}\boldsymbol{x}^{[L-2]})\odot\mathrm{d}\boldsymbol{W}^{[L-1]}\sigma(\boldsymbol{W}^{[L-2]}\boldsymbol{x}^{[L-3]})\}\\ &= \mathrm{tr}\{((\boldsymbol{W}^{[L]\intercal}\Lambda_L)\odot\sigma^{'}(\boldsymbol{W}^{[L-1]}\boldsymbol{x}^{[L-2]}))^{\intercal}\boldsymbol{W}^{[L-1]}\mathrm{d}\sigma(\boldsymbol{W}^{[L-2]}\boldsymbol{x}^{[L-3]})\}\\ &= (\boldsymbol{W}^{[L-1]\intercal}\Lambda_{L-1})^{\intercal}\mathrm{d}\sigma(\boldsymbol{W}^{[L-2]}\boldsymbol{x}^{[L-3]})\\ &= \dots\\ &= \mathrm{tr}\{\Lambda_k^{\intercal}\mathrm{d}\boldsymbol{W}^{[k]}\boldsymbol{x}^{[k-1]}\}\\ &= \mathrm{tr}\{(\Lambda_k\boldsymbol{x}^{[k-1]\intercal})^{\intercal}\mathrm{d}\boldsymbol{W}^{[k]}\},\end{aligned} \tag{23}$$

where $\Lambda_l \triangleq (\boldsymbol{W}^{[l+1]^\intercal}\Lambda_{l+1}) \odot \sigma^{'}(\boldsymbol{W}^{[l]}\boldsymbol{x}^{[l-1]})$ for $l = k, k+1 \dots L-1$ and $\Lambda_L \triangleq \boldsymbol{a} \odot \sigma^{'}(\boldsymbol{W}^{[L]}\boldsymbol{x}^{[L-1]})$.

Hence, we get,

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{W}^{[k]}} = \Lambda_k \boldsymbol{x}^{[k-1]^\intercal}. \tag{24}$$

Through the chain rule, we can get the evolution equation of $\boldsymbol{W}^{[k]}$,

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{W}^{[k]}}{\mathrm{d}t} &= -\frac{\mathrm{d}R_s(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{W}^{[k]}} \\ &= -\frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{\theta},\boldsymbol{x}_i)-y_i)\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{W}^{[k]}} \\ &= -\frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{\theta},\boldsymbol{x}_i)-y_i)\Lambda_k \boldsymbol{x}_i^{[k-1]^\intercal}.\end{aligned} \tag{25}$$

Through $\boldsymbol{a} \odot \sigma'(\boldsymbol{W}\boldsymbol{x}) = \mathrm{diag}\{\sigma^{'}(\boldsymbol{W}\boldsymbol{x})\}\boldsymbol{a}$,

Finally, the dynamic system can be obtained:

$$\begin{aligned}\dot{\boldsymbol{a}} = \frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^{[L]}\left(f(\boldsymbol{\theta},\boldsymbol{x}_i)-y_i\right), \\ \dot{\boldsymbol{W}}^{[L]} = \frac{\mathrm{d}\boldsymbol{W}^{[L]}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n}\mathrm{diag}\{\sigma'(\boldsymbol{W}^{[L]}\boldsymbol{x}_i^{[L-1]})\}\boldsymbol{a}\boldsymbol{x}_i^{[L-1]^\intercal}\left(f(\boldsymbol{\theta},\boldsymbol{x}_i)-y_i\right), \\ \dot{\boldsymbol{W}}^{[k]} = \frac{\mathrm{d}\boldsymbol{W}^{[k]}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n}\mathrm{diag}\{\sigma'(\boldsymbol{W}^{[k]}\boldsymbol{x}_i^{[k-1]})\}E^{[k+1:L]}\boldsymbol{a}\boldsymbol{x}_i^{[k-1]^\intercal}\left(f(\boldsymbol{\theta},\boldsymbol{x}_i)-y_i\right)\ \forall i \in [1:L-1],\end{aligned} \tag{26}$$

where we use $E^l(\boldsymbol{x}) = \boldsymbol{W}^{[l]^\intercal}\,\mathrm{diag}\{\sigma'(\boldsymbol{W}^{[l]}\boldsymbol{x}^{[l-1]})\}$. And $E^{[q:p]} = E^q E^{q+1}...E^p$.

Let $r_{k,j} = \|\boldsymbol{W}_j^{[k]}\|_2$. We have

$$\frac{\mathrm{d}}{\mathrm{d}t}|r_{k,j}|^2 = \frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{W}_j^{[k]}\|^2. \tag{27}$$

Then we obtain

$$\dot{r}_{k,j}r_{k,j} = \dot{\boldsymbol{W}}_j^{[k]} \cdot \boldsymbol{W}_j^{[k]}. \tag{28}$$

Finally, we get

$$\begin{aligned}\dot{r}_{k,j} = \frac{\mathrm{d}r_{k,j}}{\mathrm{d}t} &= \dot{\boldsymbol{W}}_j^{[k]} \cdot \boldsymbol{W}_j^{[k]}/r_{k,j} \\ &= \dot{\boldsymbol{W}}_j^{[k]} \cdot \boldsymbol{u}_{k,j},\end{aligned} \tag{29}$$

where $\boldsymbol{u}_{k,j} = \frac{\boldsymbol{W}_j^{[k]}}{r_{k,j}}$ is a unit vector. Then we have,

$$\begin{aligned}\dot{\boldsymbol{u}}_{k,j} = \frac{\mathrm{d}\boldsymbol{u}_{k,j}}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\boldsymbol{W}_j^{[k]}}{r_{k,j}}\right) \\ &= \frac{\dot{\boldsymbol{W}}_j^{[k]}r_{k,j} - \boldsymbol{W}_j^{[k]}\dot{r}_{k,j}}{r_{k,j}^2} \\ &= \frac{\dot{\boldsymbol{W}}_j^{[k]}r_{k,j} - \boldsymbol{W}_j^{[k]}(\dot{\boldsymbol{W}}_j^{[k]} \cdot \boldsymbol{u}_{k,j})}{r_{k,j}^2} \\ &= \frac{\dot{\boldsymbol{W}}_j^{[k]} - \boldsymbol{u}_{k,j}(\dot{\boldsymbol{W}}_j^{[k]} \cdot \boldsymbol{u}_{k,j})}{r_{k,j}}.\end{aligned} \tag{30}$$

To conclude, the quantities we concern are summarized as follows,

$$
\begin{cases}
\dot{\boldsymbol{a}} = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^{[L]} \left(f(\boldsymbol{\theta}, \boldsymbol{x}_i) - y_i\right) & (31)\\
\dot{\boldsymbol{W}}^{[L]} = -\frac{1}{n}\sum_{i=1}^{n} \mathrm{diag}\{\sigma'(\boldsymbol{W}^{[L]}\boldsymbol{x}_i^{[L-1]})\}\boldsymbol{a}\boldsymbol{x}_i^{[L-1]\intercal} \left(f(\boldsymbol{\theta}, \boldsymbol{x}_i) - y_i\right), & (32)\\
\dot{\boldsymbol{W}}^{[k]} = -\frac{1}{n}\sum_{i=1}^{n} \mathrm{diag}\{\sigma'(\boldsymbol{W}^{[k]}\boldsymbol{x}_i^{[k-1]})\}E^{[k+1:L]}\boldsymbol{a}\boldsymbol{x}_i^{[k-1]\intercal} \left(f(\boldsymbol{\theta}, \boldsymbol{x}_i) - y_i\right)\ \forall k \in [1:L-1] & (33)\\
\dot{r}_{k,j} = \dot{\boldsymbol{W}}_j^{[k]} \cdot \boldsymbol{u}_{k,j} & (34)\\
\dot{\boldsymbol{u}}_{k,j} = \dfrac{\dot{\boldsymbol{W}}_j^{[k]} - \boldsymbol{u}_{k,j}(\dot{\boldsymbol{W}}_j^{[k]} \cdot \boldsymbol{u}_{k,j})}{r_{k,j}}, & (35)
\end{cases}
$$

where we use $E^l(\boldsymbol{x}) = \boldsymbol{W}^{[l]\intercal}\,\mathrm{diag}\{\sigma'(\boldsymbol{W}^{[l]}\boldsymbol{x}^{[l-1]})\}$. And $E^{[q:p]} = E^q E^{q+1}...E^p$.

A.2.3 Prove for $\mathcal{P}w$ in 5

We calculate $\mathcal{P}\boldsymbol{w} \overset{\text{leading order}}{\approx} \mathcal{Q}\boldsymbol{w}$ as following,

$$
\begin{aligned}
\mathcal{P}\boldsymbol{w}\approx\mathcal{Q}\boldsymbol{w} :=& -\frac{1}{n}\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}[\mathrm{diag}\{\sigma'(\boldsymbol{W}^{[k]}\boldsymbol{x}_i^{[k-1]})\}(E^{[k+1:L]}\boldsymbol{a})]_j\\
&+ (\frac{1}{n}\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}[\mathrm{diag}\{\sigma'(\boldsymbol{W}^{[k]}\boldsymbol{x}_i^{[k-1]})\}(E^{[k+1:L]}\boldsymbol{a})]_j \cdot \boldsymbol{u})\boldsymbol{u}\\
=& -\frac{1}{n}\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}[\mathrm{diag}\{\frac{\sigma^{(p)}(\boldsymbol{0})}{(p-1)!} \odot (\boldsymbol{W}^{[k]}\boldsymbol{x}_i^{[k-1]})^{p-1}\}(E^{[k+1:L]}\boldsymbol{a})]_j\\
&+ (\frac{1}{n}\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}[\mathrm{diag}\{\frac{\sigma^{(p)}(\boldsymbol{0})}{(p-1)!} \odot (\boldsymbol{W}^{[k]}\boldsymbol{x}_i^{[k-1]})^{p-1}\}(E^{[k+1:L]}\boldsymbol{a})]_j \cdot \boldsymbol{u})\boldsymbol{u}\\
=& -\frac{1}{n}\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}[\mathrm{diag}\{(\boldsymbol{W}^{[k]}\boldsymbol{x}_i^{[k-1]})^{p-1}\}\,\mathrm{diag}\{\frac{\sigma^{(p)}(\boldsymbol{0})}{(p-1)!}\}(E^{[k+1:L]}\boldsymbol{a})]_j\\
&+ (\frac{1}{n}\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}[\mathrm{diag}\{(\boldsymbol{W}^{[k]}\boldsymbol{x}_i^{[k-1]})^{p-1}\}\,\mathrm{diag}\{\frac{\sigma^{(p)}(\boldsymbol{0})}{(p-1)!}\}(E^{[k+1:L]}\boldsymbol{a})]_j \cdot \boldsymbol{u})\boldsymbol{u}\\
=& -\frac{1}{n}\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}[\mathrm{diag}\{(\boldsymbol{W}^{[k]}\boldsymbol{x}_i^{[k-1]})^{p-1}\}]_j\,\mathrm{diag}\{\frac{\sigma^{(p)}(\boldsymbol{0})}{(p-1)!}\}(E^{[k+1:L]}\boldsymbol{a})\\
&+ (\frac{1}{n}\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}[\mathrm{diag}\{(\boldsymbol{W}^{[k]}\boldsymbol{x}_i^{[k-1]})^{p-1}\}]_j\,\mathrm{diag}\{\frac{\sigma^{(p)}(\boldsymbol{0})}{(p-1)!}\}(E^{[k+1:L]}\boldsymbol{a}) \cdot \boldsymbol{u})\boldsymbol{u}\\
=& -(\frac{1}{n}\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}(\boldsymbol{W}_j^{[k]}\boldsymbol{x}_i^{[k-1]})^{p-1})[\mathrm{diag}\{\frac{\sigma^{(p)}(\boldsymbol{0})}{(p-1)!}\}(E^{[k+1:L]}\boldsymbol{a})]_j\\
&+ ((\frac{1}{n}\sum_{i=1}^{n} e_i \boldsymbol{x}_i^{[k-1]}(\boldsymbol{W}_j^{[k]}\boldsymbol{x}_i^{[k-1]})^{p-1})[\mathrm{diag}\{\frac{\sigma^{(p)}(\boldsymbol{0})}{(p-1)!}\}(E^{[k+1:L]}\boldsymbol{a})]_j \cdot \boldsymbol{u})\boldsymbol{u},
\end{aligned}
\tag{36}
$$

where $(\cdot)^{p-1}$ and $\sigma^p(\cdot)$ operate on component here.

A.3 The Verification of the initial stage

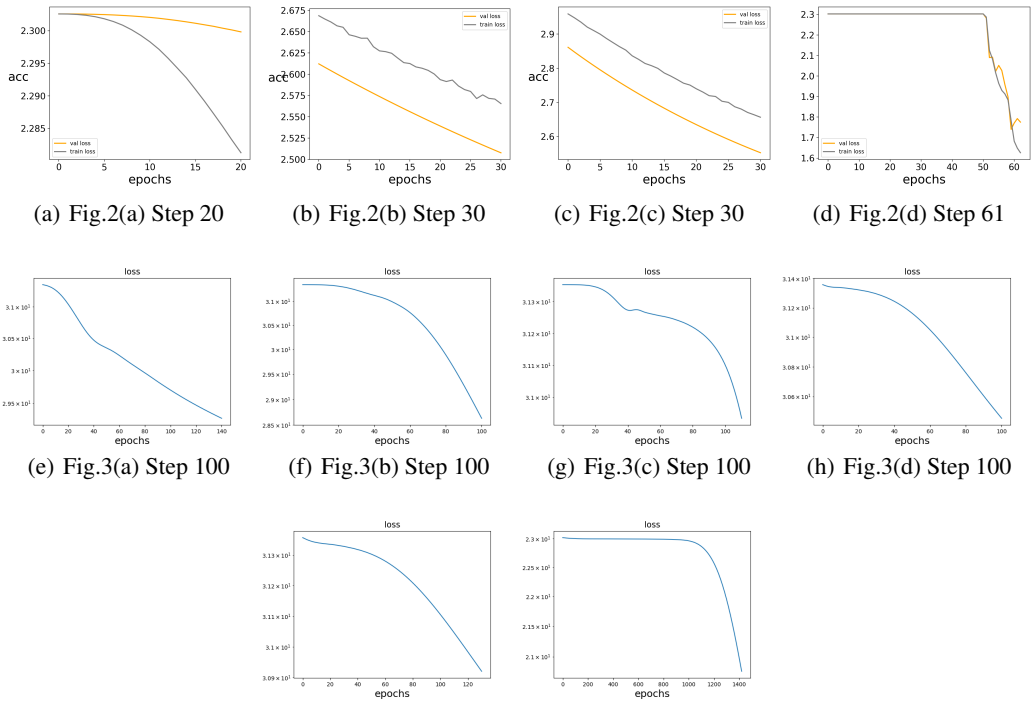
We put the loss of the experiments in the main text here to show that they are indeed in the initial stage of training by the definition.

As is shown in Fig.7 and Fig.8, at the steps demonstrated in the article, loss satisfies the definition of the initial stage, so we consider that they are in the initial stage of training.

Learning rate is not a sensitive to the appearance of condensation. However, a small learning rate could enable us to observe the condensation process in the initial stage under a gradient flow training, more clearly. For example, when the learning rate is relatively small, the initial stage of training may be relatively long, while when the learning rate is relatively large, the initial stage of training may be relatively small.

We empirically find that to ensure the training process follows a gradient follow, where the loss decays monotonically, we have to select a smaller learning rate for large multiplicity p. Therefore, it looks like we have a longer training in our experiments with large p. Note that for a small learning rate in the experiments of small p, we can observe similar phenomena.

In all subsequent experiments in the Appendix, we will no longer show the loss graph of each experiment one by one, but we make sure that they are indeed in the initial stage of training.

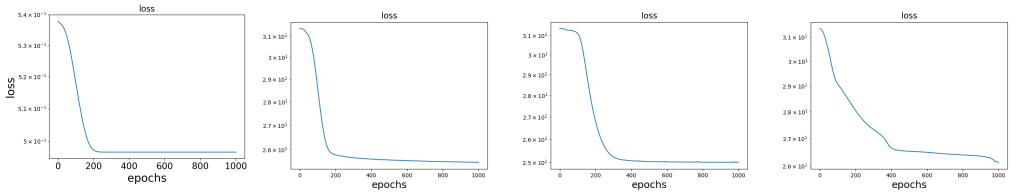


(a) Fig.2(a) Step 20 (b) Fig.2(b) Step 30 (c) Fig.2(c) Step 30 (d) Fig.2(d) Step 61

(e) Fig.3(a) Step 100 (f) Fig.3(b) Step 100 (g) Fig.3(c) Step 100 (h) Fig.3(d) Step 100

(i) Fig.3(e) Step 100 (j) Fig.4 Step 900 to 1400

Figure 7: Losses from Fig. 2 to Fig.4. The original figures and the numbers of steps corresponding to each sub-picture are written in the sub-captions.



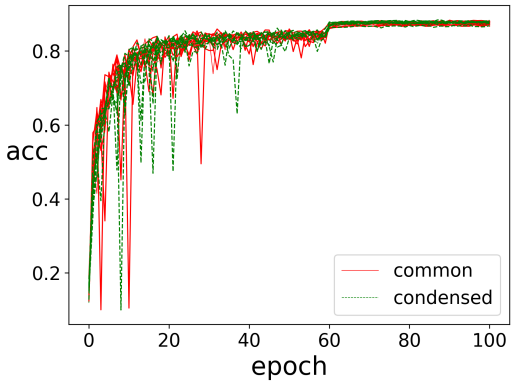
(a) Fig.5(a) Step 1000 (b) Fig.5(b) Step 1000 (c) Fig.5(c) Step 1000 (d) Fig.5(d) Step 1000

(e) Fig.6(a) Step 200 (f) Fig.6(b) Step 200 (g) Fig.6(c) Step 200 (h) Fig.6(d) Step 200

Figure 8: Losses from Fig. 5 to Fig.6. The original figures and the numbers of steps corresponding to each sub-picture are written in the sub-captions.

A.4 Performance of tanh activation function in condensed regime

For practical networks, such as resnet18-like (He et al., 2016) in learning CIFAR10, as shown in Fig. 9 and Table 1, we find that the performance of networks with initialization in the condensed regime is vary similar to the common initialization methods. For both initialization methods, the test accuracy is about 86.5% to 88.5%, where the highest test accuracy and lowest test accuracy for common methods are 88.07% and 86.73%, respectively, while the highest one and lowest one for condensed methods are 88.26% and 86.88%, respectively. This implies that performance of common and condensed initialization is similar.



(a) test accuracy

Figure 9: The test accuracy of Resnet18-like networks with different initialization methods. Each network consists of the convolution part of resnet18 and fully-connected (FC) layers with size 1024-1024-10 and softmax. The convolution part is equipped with ReLU activation and initialized by Glorot normal distribution (Glorot and Bengio, 2010). For FC layers, the activation is $\tanh(x)$ and they are initialized by three common methods (red) and three condensed ones (green) as indicated in Table 1. The learning rate is 10^{-3} for epoch 1-60 and 10^{-4} for epoch 61-100. Adam optimizer with cross-entropy loss and batch size 128 are used for all experiments.

Table 1: Comparison of test accuracy of resnet18 in learning CIFAR10 with common (Glorot and Bengio, 2010) and condensed Gaussian initializations. $\bar{m} = (m_{\rm in} + m_{\rm out})/2$. $m_{\rm in}$: in-layer width. $m_{\rm out}$: out-layer width. Each line is a trial.

	common			condensed		
	Glorot_uniform	Glorot_normal	$N(0, \frac{1}{\bar{m}})$	$N(0, \frac{1}{m_{\rm out}^4})$	$N(0, \frac{1}{m_{\rm out}^3})$	$N(0, (\frac{1}{\bar{m}})^2)$
Test 1	0.8747	0.8759	0.8807	0.8749	0.8744	0.8765
Test 2	0.8715	0.8673	0.8733	0.8763	0.8799	**0.8826**
Test 3	0.8772	0.8794	0.8788	0.8688	0.8780	0.8771

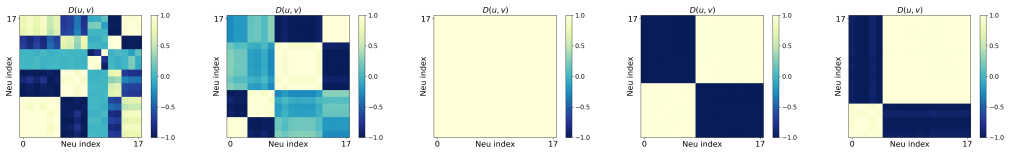
A.5 multi-layer experimental

The condensation of the six layer without residual connections is shown in 10, whose activation functions for hidden layer 1 to hidden layer 5 are $x^2\tanh(x)$, $x\tanh(x)$, $\mathrm{sigmoid}(x)$, $\tanh(x)$ and $\mathrm{softplus}(x)$, respectively.

The condensation of the three layer without residual connections is shown in 11, whose activation functions are same for each layer indicated by the corresponding sub-captions.

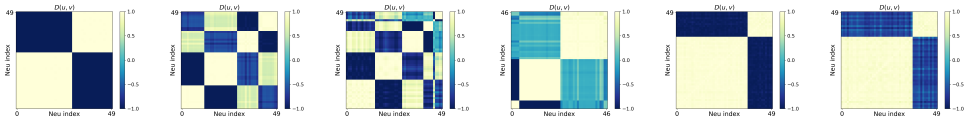
The condensation of the five layer without residual connections is shown in 12, whose activation functions are same for each layer indicated by the corresponding sub-captions.

The condensation of the five layer with residual connections is shown in 13, whose activation functions are same for each layer indicated by the corresponding sub-captions.

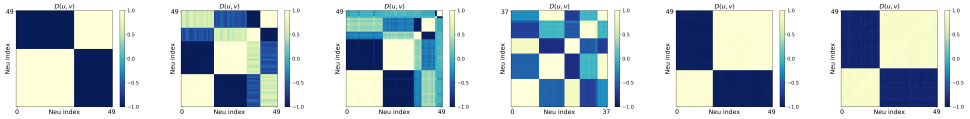


(a) layer 1 (b) layer 2 (c) layer 3 (d) layer 4 (e) layer 5

Figure 10: Condensation of six-layer NNs without residual connections. The activation functions for hidden layer 1 to hidden layer 5 are $x^2\tanh(x)$, $x\tanh(x)$, $\mathrm{sigmoid}(x)$, $\tanh(x)$ and $\mathrm{softplus}(x)$, respectively.The numbers of steps selected in the sub-pictures are epoch 6800, epoch 6800, epoch 6800, epoch 6800 and epoch 6300, respectively, while the NN is only trained once. The color indicates $D(u,v)$ of two hidden neurons' input weights, whose indexes are indicated by the abscissa and the ordinate, respectively. The training data is 80 points sampled from a 3-dimensional function $\sum_{k=1}^{3} 4\sin(12x_k+1)$, where each x_k is uniformly sampled from $[-4,2]$. $n=80$, $d=3$, $m=18$, $d_{out}=1$, $var=0.008^2$, $lr=5\times10^{-5}$.

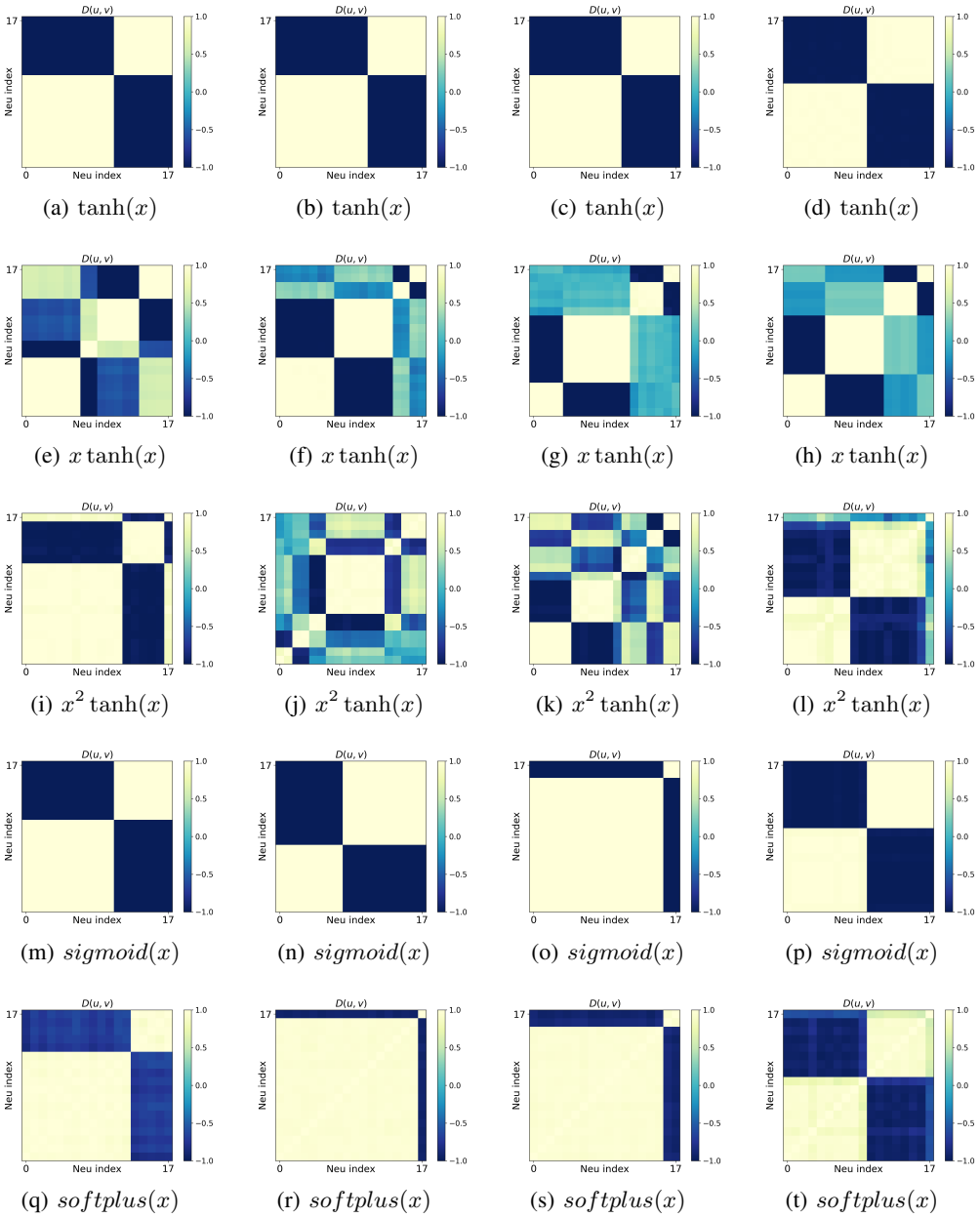


(a) $\tanh(x)$ (b) $x\tanh(x)$ (c) $x^2\tanh(x)$ (d) $\mathrm{ReLU}(x)$ (e) $\mathrm{sigmoid}(x)$ (f) $\mathrm{softplus}(x)$



(g) $\tanh(x)$ (h) $x\tanh(x)$ (i) $x^2\tanh(x)$ (j) $\mathrm{ReLU}(x)$ (k) $\mathrm{sigmoid}(x)$ (l) $\mathrm{softplus}(x)$

Figure 11: Three-layer NN at epoch 700. (a-f) are for the input weights of the first hidden layer and (g-l) are for the input weights of the second hidden layer. The color indicates $D(u,v)$ of two hidden neurons' input weights, whose indexes are indicated by the abscissa and the ordinate, respectively. The training data is 80 points sampled from a 5-dimensional function $\sum_{k=1}^{5} 3\sin(8x_k+1)$, where each x_k is uniformly sampled from $[-4,2]$. $n=80$, $d=5$, $m=50$, $d_{out}=1$, $var=0.005^2$. $lr=10^{-4}, 2\times10^{-5}, 1.4\times10^{-5}$ for (a-d), (e) and (f), respectively. For (d) and (j), we discard hidden neurons, whose L_2-norm of its input weight is smaller than 0.1.



(a) $\tanh(x)$ (b) $\tanh(x)$ (c) $\tanh(x)$ (d) $\tanh(x)$

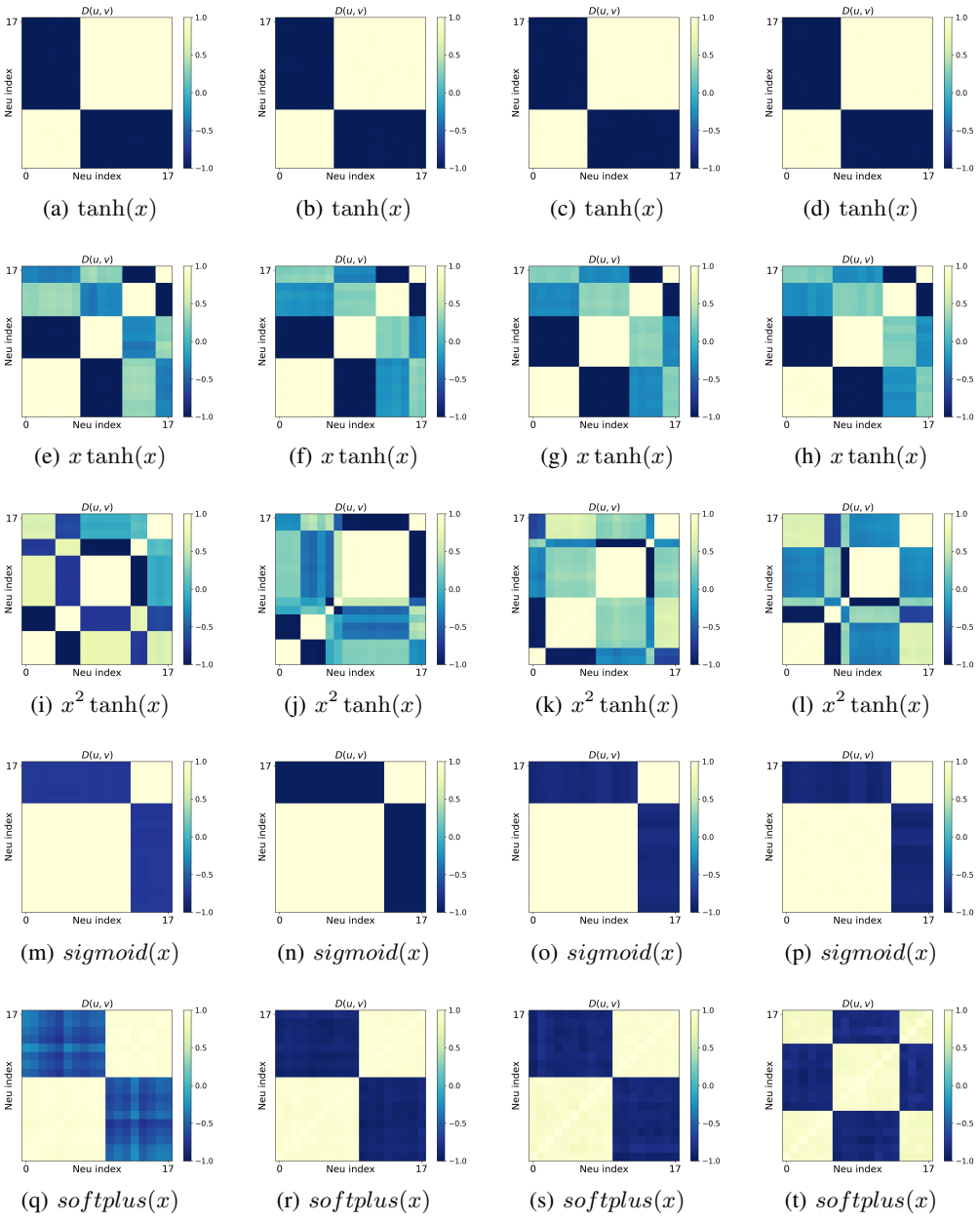
(e) $x\tanh(x)$ (f) $x\tanh(x)$ (g) $x\tanh(x)$ (h) $x\tanh(x)$

(i) $x^2\tanh(x)$ (j) $x^2\tanh(x)$ (k) $x^2\tanh(x)$ (l) $x^2\tanh(x)$

(m) $sigmoid(x)$ (n) $sigmoid(x)$ (o) $sigmoid(x)$ (p) $sigmoid(x)$

(q) $softplus(x)$ (r) $softplus(x)$ (s) $softplus(x)$ (t) $softplus(x)$

Figure 12: Five-layer NN. The first to fourth columns of each row are for the input weights of neurons from the first to the fourth hidden layers, respectively. The color indicates $D(u, v)$ of two hidden neurons' input weights, whose indexes are indicated by the abscissa and the ordinate, respectively. The training data is 80 points sampled from a 5-dimensional function $\sum_{k=1}^{3} 3\sin(10x_k + 1)$, where each x_k is uniformly sampled from $[-4, 2]$. $n = 80$, $d = 5$, $m = 18$, $d_{out} = 1$, $var = 0.008^2$. $lr = 1.5 \times 10^{-5}$, 1.5×10^{-5}, 1.5×10^{-5}, 1.5×10^{-5}, 1.5×10^{-6} and epoch is 10000, 10000, 26000, 10000, 20000 for $tanh(x)$, $xtanh(x)$, $x^2tanh(x)$, $sigmoid(x)$, $softplus(x)$, respectively.



(a) $\tanh(x)$ (b) $\tanh(x)$ (c) $\tanh(x)$ (d) $\tanh(x)$

(e) $x\tanh(x)$ (f) $x\tanh(x)$ (g) $x\tanh(x)$ (h) $x\tanh(x)$

(i) $x^2\tanh(x)$ (j) $x^2\tanh(x)$ (k) $x^2\tanh(x)$ (l) $x^2\tanh(x)$

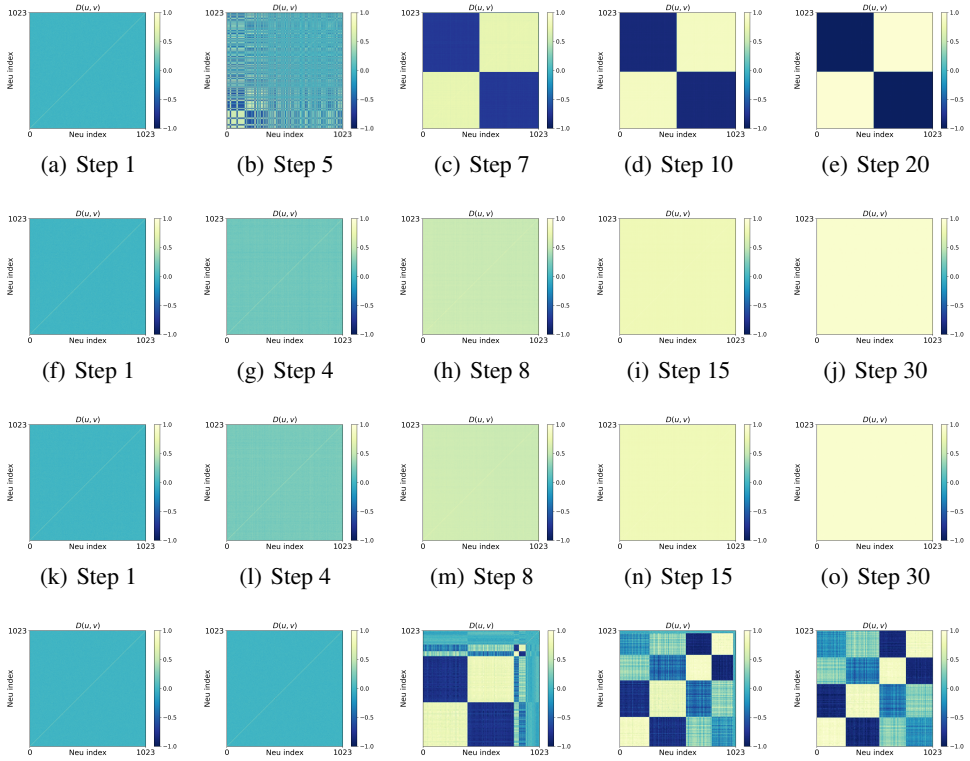
(m) $sigmoid(x)$ (n) $sigmoid(x)$ (o) $sigmoid(x)$ (p) $sigmoid(x)$

(q) $softplus(x)$ (r) $softplus(x)$ (s) $softplus(x)$ (t) $softplus(x)$

Figure 13: Five-layer NN. The first to fourth columns of each row are for the input weights of neurons from the first to the fourth hidden layers, respectively. The color indicates $D(u, v)$ of two hidden neurons' input weights, whose indexes are indicated by the abscissa and the ordinate, respectively. The training data is 80 points sampled from a 5-dimensional function $\sum_{k=1}^{3} 3\sin(10x_k + 1)$, where each x_k is uniformly sampled from $[-4, 2]$. $n = 80$, $d = 5$, $m = 18$, $d_{out} = 1$, $var = 0.008^2$. $lr = 1\times10^{-4}$, 1×10^{-4}, 1×10^{-4}, 5×10^{-5}, 5×10^{-5} and epoch is 400, 400, 400, 3000, 360, 400 for $tanh(x)$, $xtanh(x)$, $x^2tanh(x)$, $x^2tanh(x)$, $sigmoid(x)$, $softplus(x)$, respectively.

A.6 Several steps during the evolution of condensation at the initial stage

In the article, we only give the results of the last step of each condense, while the details of the evolution of condensation are lacking, which may provide a better understanding. Therefore, we show these details in Fig. 14, Fig. 15, Fig. 16 and Fig. 17, which also further illustrate the rationality of the experimental results and facilitate the understanding of the evolution of condensation in the initial stage.



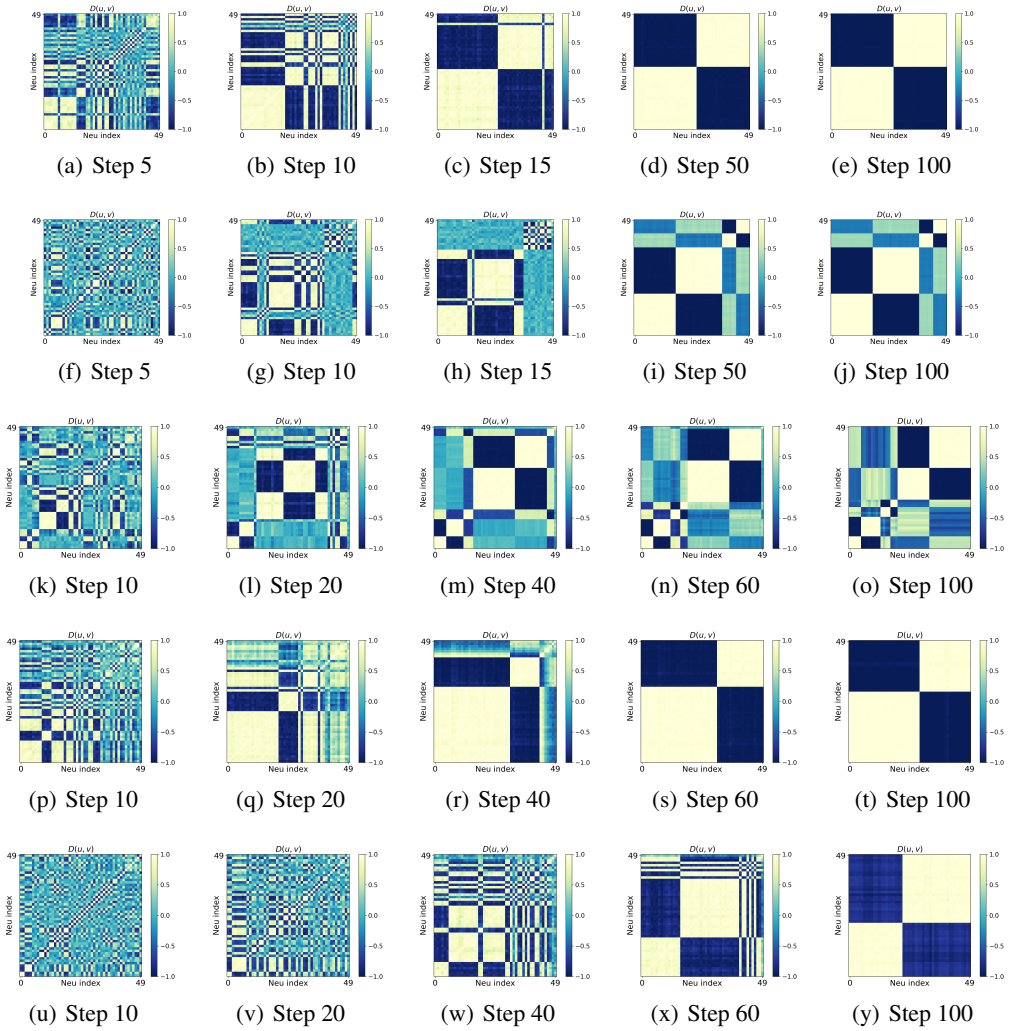
(a) Step 1 (b) Step 5 (c) Step 7 (d) Step 10 (e) Step 20

(f) Step 1 (g) Step 4 (h) Step 8 (i) Step 15 (j) Step 30

(k) Step 1 (l) Step 4 (m) Step 8 (n) Step 15 (o) Step 30

(p) Step 30 (q) Step 51 (r) Step 54 (s) Step 58 (t) Step 61

Figure 14: Evolution of condensation of Fig. 2(a), Fig. 2(b), Fig. 2(c), and Fig. 2(d). The evolution from the first row to the fourth row are corresponding to the Fig. 2(a), Fig. 2(b), Fig. 2(c), and Fig. 2(d). The numbers of evolutionary steps are shown in the sub-captions, where sub-figures in the last row are the epochs in the article.



(a) Step 5 (b) Step 10 (c) Step 15 (d) Step 50 (e) Step 100

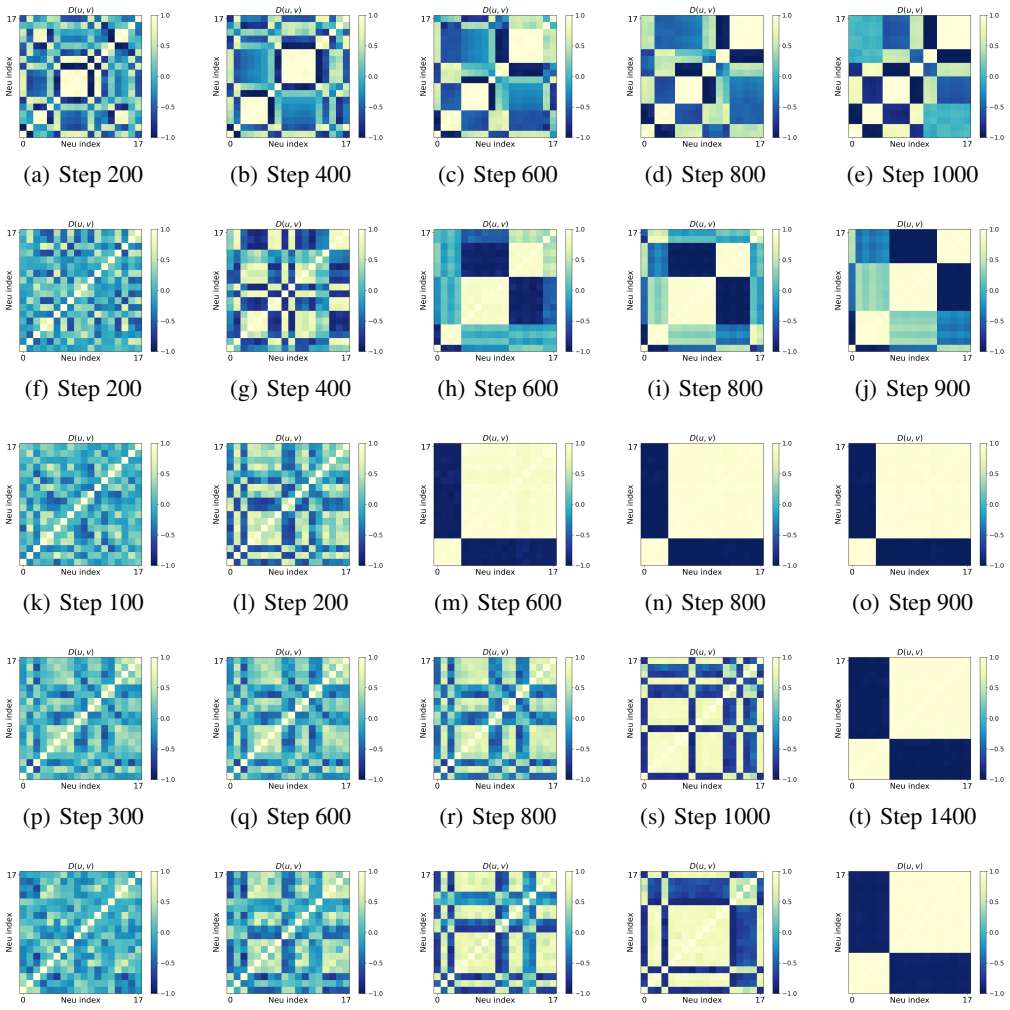
(f) Step 5 (g) Step 10 (h) Step 15 (i) Step 50 (j) Step 100

(k) Step 10 (l) Step 20 (m) Step 40 (n) Step 60 (o) Step 100

(p) Step 10 (q) Step 20 (r) Step 40 (s) Step 60 (t) Step 100

(u) Step 10 (v) Step 20 (w) Step 40 (x) Step 60 (y) Step 100

Figure 15: Evolution of condensation from Fig. 3(a) to 3(e). The evolution from the first row to the fifth row are corresponding to the Fig. 3(a), Fig. 3(b), Fig. 3(c), Fig. 3(d), Fig. 3(e). The numbers of evolutionary steps are shown in the sub-captions, where sub-figures in the last row are the epochs in the article.



(a) Step 200 (b) Step 400 (c) Step 600 (d) Step 800 (e) Step 1000

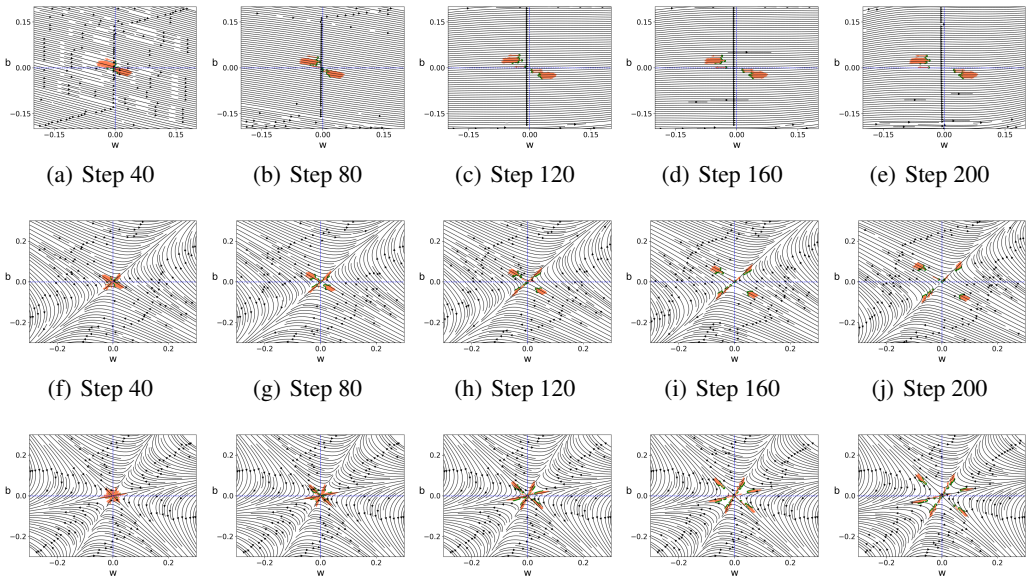
(f) Step 200 (g) Step 400 (h) Step 600 (i) Step 800 (j) Step 900

(k) Step 100 (l) Step 200 (m) Step 600 (n) Step 800 (o) Step 900

(p) Step 300 (q) Step 600 (r) Step 800 (s) Step 1000 (t) Step 1400

(u) Step 100 (v) Step 500 (w) Step 900 (x) Step 1000 (y) Step 1400

Figure 16: Evolution of condensation from Fig. 4(a) to 4(e). The evolution from the first row to the fifth row are corresponding to the Fig. 4(a), Fig. 4(b), Fig. 4(c), Fig. 4(d), Fig. 4(e). The numbers of evolutionary steps are shown in the sub-captions, where sub-figures in the last row are the epochs in the article.



(a) Step 40 (b) Step 80 (c) Step 120 (d) Step 160 (e) Step 200

(f) Step 40 (g) Step 80 (h) Step 120 (i) Step 160 (j) Step 200

(k) Step 40 (l) Step 80 (m) Step 120 (n) Step 160 (o) Step 200

Figure 17: Evolution of condensation from Fig. 6(a) to 6(c). The evolution from the first row to the fifth row are corresponding to the Fig. 6(a), Fig. 6(b), and Fig. 6(c). The numbers of evolutionary steps are shown in the sub-captions, where sub-figures in the last row are the epochs in the article.

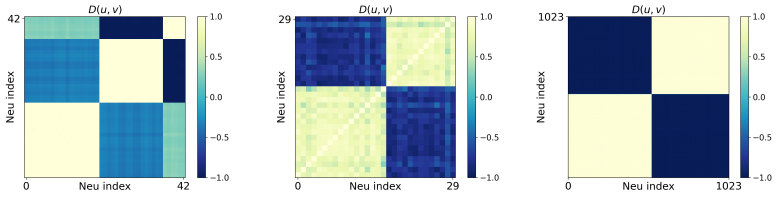
A.7 The influence of training data on condensation

We also find that when the training data is less oscillated, the NN may condense at fewer directions. For example, as shown in Fig. 18(a), compared with the high frequency function in Fig. 3, we only change the target function to be a lower-frequency function, i.e., $\sum_{k=1}^{5} 3.5\sin(2x_k+1)$. In this case, the NN with $x^2\tanh(x)$ only condenses at three directions, in which two are opposite. For MNIST data in Fig. 18(b), we find that, the NN with $x^2\tanh(x)$ condenses at one line, which may suggest that the function for fitting MNIST dataset is a low-frequency function. For CIFAR100 data in Fig. 18(c), we find that input weights of the first FC layer with $x\tanh(x)$ condense at only one line, which implies that features extracted by the convolution part of the NN may own low complexity.

These experiments does not contradict to our results, which claim that the *maximal* number of condensed orientations in the initial training is twice the multiplicity of the activation function used in general NNs.

For CIFAR100 dataset, we use Resnet18-like neural network, which has been described in Fig. 2. Besides, the input dimension is $d = 32 * 32 * 3$, the output dimension is $d_{out} = 100$, and all parameters are initialized by a Gaussian distribution $N(0, var)$. The total data size is n. The training method is Adam with batch size 128, learning rate lr and cross-entropy loss.

For MNIST dataset, we use fully-connected neural network with size, d-m-$\cdots$-m-d_{out}. The input dimension is $d = 784$, and the output dimension is $d_{out} = 10$. The number of hidden neurons m is specified in Fig. 18. All parameters are initialized by a Gaussian distribution $N(0, var)$. The total data size is n. The training method is Adam with full batch, learning rate lr and MSE loss.



(a) $\sum_{k=1}^{5} 3.5\sin(2x_k+1)$ (b) MNIST (c) CIFAR100

Figure 18: Condensation of low-frequency functions with two-layer NNs in (a,b) and condensation of the first FC layer of the Resnet18-like network on CIFAR100 in (c). The color indicates $D(u, v)$ of two hidden neurons' input weights, whose indexes are indicated by the abscissa and the ordinate. For (a,b), two-layer NN at epoch: 100 with activation function: $x^2\tanh(x)$. For (a), we discard about 15% of hidden neurons, in which the L_2-norm of each input weight is smaller than 0.04, while remaining those bigger than 0.4. The mean magnitude here for each parameter is $(0.4^2/785)^{0.5}$ $\sim$0.01, which should also be quite small. All settings in (a) are the same as Fig. 3, except for the lower frequency target function. Parameters for (b) are $n = 60000$, $d = 784$, $m = 30$, $d_{out} = 10$, $var = 0.001^2$. $lr = 5 \times 10^{-5}$. The structure and parameters of the Resnet18-like neural network for (c) is the same as Fig. 2, except for the data set CIFAR100 and learning rate $lr = 1 \times 10^{-6}$.