ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200221

Mean Field Theory for Deep Dropout Networks: Digging up Gradient Backpropagation Deeply

Wei Huang [1] and **Richard Yi Da Xu** [2] and **Weitao Du** [3] and **Yutian Zeng** [4] and **Yunce Zhao** [5]

Abstract. In recent years, the mean field theory has been applied to the study of neural networks and has achieved a great deal of success. The theory has been applied to various neural network structures, including CNNs, RNNs, Residual networks, and Batch normalization. Inevitably, recent work has also covered the use of dropout. The mean field theory shows that the existence of depth scales that limit the maximum depth of signal propagation and gradient backpropagation. However, the gradient backpropagation is derived under the *gradient independence assumption* that weights used during feed forward are drawn independently from the ones used in backpropagation. This is not how neural networks are trained in a real setting. Instead, the same weights used in a feed-forward step needs to be carried over to its corresponding backpropagation. Using this realistic condition, we perform theoretical computation on linear dropout networks and a series of experiments on dropout networks with different activation functions. Our empirical results show an interesting phenomenon that the length gradients can backpropagate for a *single* input and a *pair* of inputs are governed by the same depth scale. Besides, we study the relationship between variance and mean of statistical metrics of the gradient and shown an emergence of universality. Finally, we investigate the maximum trainable length for deep dropout networks through a series of experiments using MNIST and CIFAR10 and provide a more precise empirical formula that describes the trainable length than original work.

1 Introduction

Deep neural networks have achieved exceptional results in a range of fields since its inception [11]. Recent seminal innovations have been proposed to improve the performance of neural networks further. For example, residual networks [7] and batch normalization [8], which were introduced to overcome the gradient vanishing and exploding problem, enabled the trainable length to be very deep. Another technology is the dropout [22], which is a regularization technique for reducing the over-fitting problem. It is also the focus of this paper. In dropout, network units are randomly dropped during training, which can prevent complex co-adaptations [22].

More recently, we have witnessed several signs of progress made using mean field theory [20, 21, 17] in deep learning. The mean field considers networks after random initialization, whose weights and biases were i.i.d. Gaussian distributed, and the width of each layer tends to infinity. As a result of studying signal propagation under mean field theory, an order-to-chaos expressivity phase transition split by a critical line has been found [20]. Later, how parameter initialization may impact the gradient of backpropagation was studied, and the conclusion that the ordered and chaotic phases correspond to regions of vanishing and exploding gradient respectively was shown [21]. The results were also equivalently applied to networks with or without dropout.

The main contribution of the mean field theory for random networks is that it shows the existence of depth scales that limit the maximum depth of signal propagation and gradient backpropagation. Practically, the result is to show a hypothesis that random networks may be trained precisely when information can travel through them. Thus, the depth scales provide bounds on how deep a network may be trained for a specific choice of hyper-parameters [21]. This ansatz was tested and verified by practical experiments on MNIST and CIFAR10 dataset with wide width fully-connected networks [21], deep dropout networks [21], and residual networks [24].

However, the mean field calculation for the gradient is based on the so-called *gradient independence assumption*, which states that the weights used during feed forward are drawn independently from the ones used in backpropagation. This is in an effort to make the calculation of gradient feasible regardless of the choice of activation functions. This assumption was later formulated explicitly [24] for residual networks and was illustrated in a review [25]. While it enjoys the correct prediction of gradient dynamics in some cases, our experiments show that under the condition in which the weights in feed-forward are carried over to its backpropagation, the length that gradients can backpropagate for a *single* and a *pair* of inputs are governed by the same depth scale on deep dropout networks instead.

By further studying the mean and variance of gradient statistics metrics on deep dropout networks, we show an emergence of universality for the relationship between the mean and variance. This universality exists regardless of the choice of hyper-parameters, including dropout rate and activation function. After summarizing the theoretical results about the trainable length of deep dropout networks governed by maximum depth of signal propagation and gradient backpropagation, we perform a series of experiments to investigate it. Empirically, we find a more precise way to describe the maximum trainable length for deep dropout networks, compared with the original results [21].

2 Related Work

The mean field theory has been applied to different network architectures, including CNNs [10], RNNs [15], Residual networks [7], Batch normalization [8], LSTM [5], and GRUs [2]. These networks

[1] University of Technology Sydney, Australia, email:Wei.Huang-6@student.uts.edu.au
[2] University of Technology Sydney, Australia, email: YiDa.Xu@uts.edu.au
[3] Northwestern University, USA, email: weitao.du@northwestern.edu
[4] Xiamen University, China, email: 19020161152850@stu.xmu.edu.cn
[5] University of Technology Sydney, Australia, email: Yunce.Zhao@student.uts.edu.au

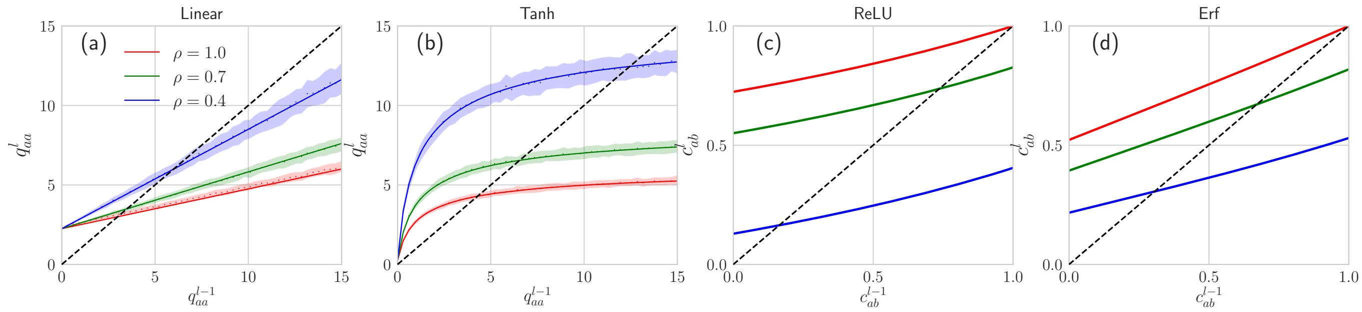

Figure 1. The iterative squared length mapping of Equation (2) and Equation (4) with different activations and dropout rates. (a) The iterative length map of q_{aa}^l on a Linear network at $\sigma_w = 0.5$ and $\sigma_b = 1.5$. Theoretical predictions (solid lines) match well with network simulations (dots) within a standard error (shadow). The intersection between map and unity line determine its fixed points q_{ab}^*. Different color correspond to different dropout rates: $\rho = 1$ is red, $\rho = 0.7$ is green, and $\rho = 0.4$ is blue. (b) The iterative length map of q_{aa}^l on a Tanh network at $\sigma_w = 2.5$ and $\sigma_b = 0.5$. (c) The iterative length map of c_{ab}^l on a ReLU network at $\sigma_w = 0.9$ and $\sigma_b = 0.5$. Only intersection of network at $\rho = 1$ (red) is $c_{ab}^* = 1$, the others are $c_{ab}^* < 1$. (d) The iterative length map of c_{ab}^l on a Erf network at $\sigma_w = 0.9$ and $\sigma_b = 0.5$. Again, $c_{ab}^* = 1$ only holds at $\rho = 1$.

have been investigated by [23, 1, 24, 26, 6], respectively, which form a large family of the mean field theory for deep neural networks.

Following the mean field theory, [17] studied all singular values of the input-output Jacobian and found a strong connection between *dynamical isometry* and fast training speed. Later, the analysis of the spectrum of input-output Jacobian has been developed to provide a detailed analytic understanding [18] and a nonlinear random matrix theory for deep learning [19]. The study of the spectrum of input-output Jacobian is based on the mean field theory, which will not be addressed in this work since it is trivial to extend the analysis method by [17] to the dropout networks.

In contrast to the mean field theory view to the random networks, [3] studied the relationship between random networks and kernels while [12, 14] adopted another view of Gaussian processes (GPs) in the realm of Bayesian learning. The correspondence between single infinite neural networks and Gaussian process was first observed by [16]. Moreover, a study of the dynamics of networks in the infinite width limit, termed as the neural tangent kernel, has achieved great success [9, 13] recently.

Finally, dropout training in deep neural networks can be viewed as approximate Bayesian inference in deep Gaussian processes [4]. Further, dropout can be used in the Neural Network GP [12]. While this topic is interesting, we do not include the Bayesian learning of random dropout networks in our work.

3 Background

In this section, we review the mean field theory for deep dropout networks. We give the main definitions, setup, and notations, and introduce the results of theory for random networks at initialization, including signal feed-forward and gradient backpropagation, respectively.

3.1 Feed Forward

Consider a feed-forward, fully-connected, untrained, and dropout network of depth L with layer width N. We denote synaptic weight and bias for the l-th layer by W_{ij}^l and b_i^l; pre-activations and post-activations by z_i^l and y_i^l respectively. Finally, we take the input to be $y_i^0 = x_i$ and the dropout keep rate to be ρ. The information propagation in this network is governed by,

$$z_i^l = \frac{1}{\rho}\sum_j W_{ij}^l p_j^l y_j^{l-1} + b_i^l, \quad y_i^l = \phi(z_i^l), \tag{1}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is the activation function and $p \sim \text{Bernoulli}(\rho)$. We adopt the mean field theory assumption [20, 21], where $W_{ij}^l \sim \mathcal{N}(0, \frac{\sigma_w^2}{N})$, $b_i^l \sim \mathcal{N}(0, \sigma_b^2)$, and the width N tends to infinite. Since the weights and biases are randomly distributed, these equations define a probability distribution on the pre-activations over an ensemble of untrained neural networks. Under the mean field approximation, z_i^l can be replaced by a Gaussian distribution with zero mean.

Consider a *single* input $x_{i;a}$, where the subscript a refers to the index of input. We define the length quantities $q_{aa}^l = \frac{1}{N}\sum_{i=1}^N (z_{i;a}^l)^2$, which is the mean squared pre-activations. According to the mean field approximation, the length quantity is described by the recursion relation,

$$q_{aa}^l = \frac{\sigma_w^2}{\rho}\int \mathcal{D}z\phi^2(\sqrt{q_{aa}^{l-1}}z) + \sigma_b^2, \tag{2}$$

where $\int \mathcal{D}z = \frac{1}{\sqrt{2\pi}}\int dz e^{-\frac{1}{2}z^2}$ is the measure for a normal distribution. This equation describes how a single input evolves through a random neural network. To study the property of evolution, we investigate the fixed point at $q_{aa}^* \equiv \lim_{l\to\infty} q_{aa}^l$. One way to estimate the fixed point is to plot Equation (2) with the unity line, and the intersection is the fixed point. We show the result for Equation (2) with Linear dropout network and Tanh dropout network in Figure 1(a)(b). Note that the smaller the dropout rate ρ, the larger the fixed point value q_{aa}^*.

The propagation of a *pair* of inputs $x_{i;a}$ and $x_{i;b}$, where the subscript a and b refer to different inputs, can be studied by looking at the correlation between the two inputs after l layers. We definite this correlation quantity as $q_{ab}^l = \frac{1}{N}\sum_{i=1}^N (z_{i;a}^l z_{i;b}^l)$. Similarly, the correlation q_{ab}^l will be given by the recurrence relation,

$$q_{ab}^l = \sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1)\phi(u_2) + \sigma_b^2, \tag{3}$$

where $u_1 = \sqrt{q_{aa}^{l-1}} z_1$ and $u_2 = \sqrt{q_{bb}^{l-1}}(c_{ab}^{l-1} z_1 +$

$\sqrt{1-(c_{ab}^{l-1})^2}z_2)$, with

$$c_{ab}^l = q_{ab}^l / \sqrt{q_{aa}^l q_{bb}^l}. \quad (4)$$

This equation also have a fixed point at $c_{ab}^* \equiv \lim_{l\to\infty} c_{ab}^l$. It is known that $c_{ab}^* = 1$ when $\rho = 1$, while $c_{ab}^* < 1$ when $\rho < 1$ [21]. We show the result of Equation (4) on the ReLU and Erf dropout networks in Figure 1(c)(d), which demonstrate the main conclusion about fixed-point without ($\rho = 1$) and with ($\rho < 1$) dropout.

The main contribution of mean field theory for the fully-connected networks without dropout ($\rho = 1$) is that it presents a phase diagram, which is determined by a crucial quantity,

$$\chi_1 = \frac{\partial c_{ab}^l}{\partial c_{ab}^{l-1}} = \sigma_w^2 \int \mathcal{D}z[\phi'(\sqrt{q^*}z)]^2. \quad (5)$$

This quantity was firstly introduce by [20] to determine whether or not the $c_{ab}^* = 1$ is an attractive fixed point. When $\chi_1 > 1$, the fixed point is unstable. Conversely, when $\chi_1 < 1$, the fixed point is stable. Thus, the critical line $\chi_1 = 1$ separates two phases. One is the chaotic phase ($\chi_1 > 1$), where a pair of inputs end up asymptotically decorrelated, and the other is the ordered phase ($\chi_1 < 1$), in which a pair of inputs end up asymptotically correlated.

We give a comment on the difference between q_{aa}^l and c_{ab}^l here. The random networks in the infinite width limit can be viewed as the Gaussian processes, where q_{aa}^l and c_{ab}^l are the diagonal and non-diagonal elements of the compositional kernel[12], respectively. Intuitively, the non-diagonal element of the kernel measures the correlation between different data points while the diagonal component measures the information of one input itself.

The study of information propagation shows the existence of a depth-scales ξ_2, which represent the length of propagation of the following qualities:

$$|c_{ab}^l - c_{ab}^*| \sim e^{-l/\xi_2}. \quad (6)$$

where $\xi_2 = |1/\log \chi_2|$, with $\chi_2 = \sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi'(u_1^*)\phi'(u_2^*)$, where $u_1^* = \sqrt{q_{aa}^*}z_1$ and $u_2^* = \sqrt{q_{bb}^*}(c_{ab}^* z_1 + \sqrt{1-(c_{ab}^*)^2}z_2)$. Intuitively, the depth-scales ξ_2 measures how far can correlation between two different inputs survives through the network.

3.2 Back Propagation

There is a duality between the forward propagation of signals and the backpropagation of gradients. Given a loss E, we have

$$\frac{\partial E}{\partial W_{ij}^l} = \frac{p_j^l}{\rho}\phi(z_j^{l-1})\delta_i^l, \quad \delta_i^l = \phi'(z_i^l)\frac{p_i^{l+1}}{\rho}\sum_j \delta_j^{l+1} W_{ji}^{l+1}, \quad (7)$$

where $\delta_i^l = \frac{\partial E}{\partial z_i^l}$. We define the metric of gradient for both a *single* input and a *pair* of inputs cases:

$$g_{aa}^l \equiv \frac{1}{N^2}\sum_{ij}(\frac{\partial E_a}{\partial W_{ij}^l})^2, \quad g_{ab}^l \equiv \left|\frac{1}{N^2}\sum_{ij}\frac{\partial E_a}{W_{ij}^l}\frac{\partial E_b}{W_{ij}^l}\right|. \quad (8)$$

Within mean field theory, the scale of fluctuations of the gradient of weights in a layer will be proportional to $\tilde{q}_{aa}^l \equiv \mathbb{E}\left[\delta_{i;a}^l \delta_{i;a}^l\right]$, which can be written as, $g_{aa}^l \propto \tilde{q}_{aa}^l$ [21]. On the other hand, the correlation between gradients of a pair of inputs will be proportional to $\tilde{q}_{ab}^l \equiv \mathbb{E}\left[\delta_{i;a}^l \delta_{i;b}^l\right]$, namely, $g_{ab}^l \propto \tilde{q}_{ab}^l$.

In order to work out the recurrence relation for $\tilde{q}_{aa}^l$ and $\tilde{q}_{ab}^l$, an approximation was made [21], named *gradient independence assumption*, that the weights used during forward propagation are drawn independently from the weights used in backpropagation. In this way, the term $\phi'(z_i^l)$, δ_j^{l+1} and W_{ji}^{l+1} in Equation (7) can be addressed independently. Then, the recurrence behavior of $\tilde{q}_{aa}^l$ and $\tilde{q}_{ab}^l$ are achieved,

$$\tilde{q}_{aa}^l = \tilde{q}_{aa}^{l+1}\chi_1, \quad \tilde{q}_{ab}^l = \tilde{q}_{ab}^{l+1}\chi_2. \quad (9)$$

where we redefine the quantity χ_1 for the dropout networks,

$$\chi_1 = \frac{\sigma_w^2}{\rho}\int \mathcal{D}z[\phi'(\sqrt{q^*}z)]^2. \quad (10)$$

Equation (9) has an exponential solution with,

$$\tilde{q}_{aa}^l = \tilde{q}_{aa}^L e^{-(L-l)/\xi_1}, \quad \tilde{q}_{ab}^l = \tilde{q}_{ab}^L e^{-(L-l)/\xi_2}. \quad (11)$$

Similar to the signal propagation, gradient backpropagation can limit the trainable length in the way of gradient vanishing or gradient exploding, which is measured by the depth-scales ξ_1 and ξ_2.

4 Gradient Backpropagation

In this section, we first calculate the metrics of gradient g_{aa} and g_{ab} theoretically without the gradient independence assumption on linear dropout networks. We then conduct a series experiment for metrics of gradient on deep dropout networks, including non-linear cases. Finally, we show an emergence of a universal relationship between mean and variance of metrics of the gradient.

4.1 Breaking the gradient independence assumption

We follow the fact that weights used in a feed-forward are carried over to its back-propagation. We first provide a theoretical treatment to the linear networks in which we assume the output is the last layer of network $y_i^L = z_i^L$ without soft-max. The labels of data are set to be zeros, and the loss is the mean squared loss.

For space reason, we omit details of the calculation and present the primary analysis and final results here. The main problem is that we should expand δ_j^{l+1} when calculating δ_i^l in Equation (7), since δ_j^{l+1} can correlate with W_{ji}^{l+1} without the gradient independence assumption. Using g_{aa}^l as an example, we perform:

1. Starting from the last layer L, we compute $\delta_{i,a}^L = \frac{\partial E_a}{\partial z_{i,a}^L} = 2z_{i,a}^L$ and use this result to compute $g_{aa}^L = \mathbb{E}\left[(\frac{p_{j,a}^L}{\rho}z_{j,a}^{L-1}\delta_{i,a}^L)^2\right]$.
2. Then we compute $g_{aa}^{L-1} = \mathbb{E}\left[(\frac{p_{j,a}^{L-1}}{\rho}z_{j,a}^{L-2}\delta_{i,a}^{L-1})^2\right]$ with the result of $\delta_{i,a}^{L-1} = \frac{\partial E_a}{\partial z_{i,a}^L}\frac{\partial z_{i,a}^L}{\partial z_{i,a}^{L-1}} = \sum_j 2z_{j,a}^L \frac{p_{i,a}^L}{\rho}W_{ji}^L$ and $z_i^L = \frac{1}{\rho}\sum_j W_{ij}^L p_j^l z_j^{L-1} + b_i^L$.
3. By parity of reasoning, we obtained the results for the penultimate layer g_{aa}^{L-2}. The correlation between terms that contain W_{ij}^L and W_{ij}^{L-1} are considered.
4. As the index of the layer decreases, the amount of calculation becomes larger and larger. Thus we use the induction method to achieve the results for left layers.

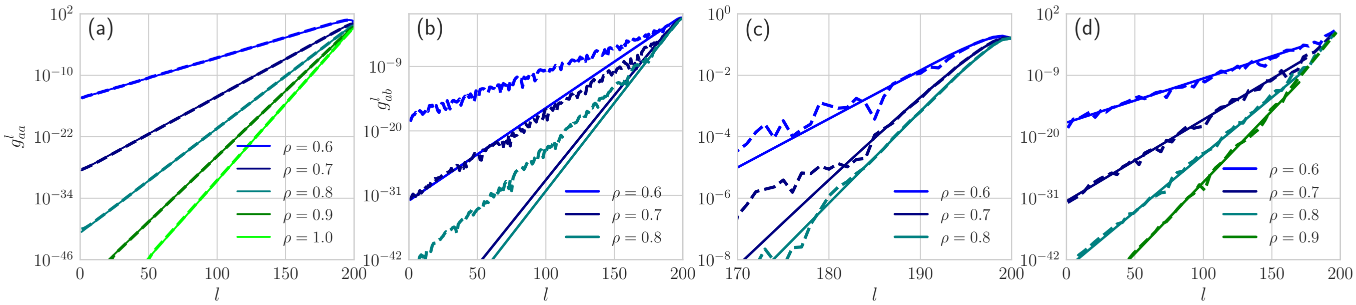

Figure 2. Theoretical calculations versus network simulations for metric of gradient. (a) g^l_{aa} as a function of layer l, for a 200 layers random linear network with $\sigma_w^2 = 0.5$ and $\sigma_b^2 = 0.1$. Excellent agreement is observed between empirical simulations of networks of width 1000 (dashed lines) and theoretical calculations (solid lines). (b) g^l_{ab} as a function of layer l. Theoretical calculations (solid lines) fail to predict empirical simulations (dashed lines). (c) g^l_{ab} as a function of layer l in the range of length $l = 170 - 200$. Theoretical calculations (solid lines) can predict empirical simulations (dashed lines) in the few last layers. (d) g^l_{ab} as a function of layer l. The solid lines are $g^l_{ab} \propto \chi_1^{L-l}$ for different ρ. Theoretical calculations failed to predict empirical simulations (dashed lines).

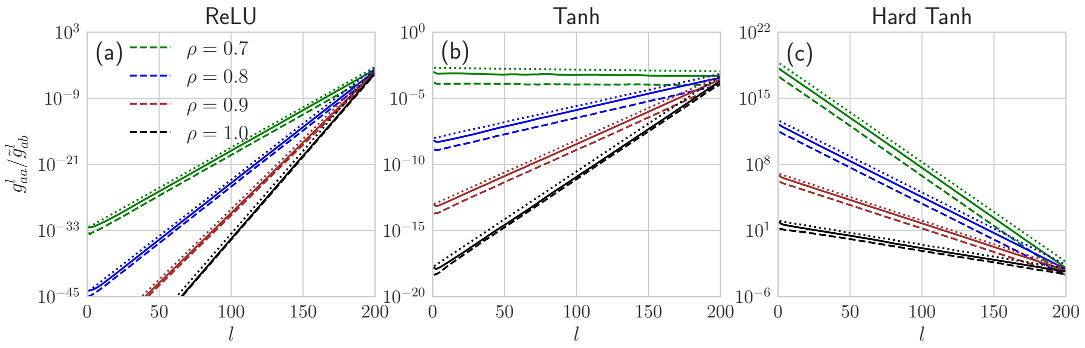

Figure 3. The metric of gradient with one and two different inputs, g^l_{aa} (solid lines), $\tilde{g}^l_{ab}$ (dashed lines), and $g^l \propto \chi_1^{L-l}$ (dotted lines) as a function of layer l with different activation. (a) ReLU network with $\sigma_w^2 = 1.0$ and $\sigma_b^2 = 0.1$. (b) Tanh network with $\sigma_w^2 = 1.4$ and $\sigma_b^2 = 0.1$. (c) Hard Tanh network with $\sigma_w^2 = 1.4$ and $\sigma_b^2 = 0.1$. Excellent agreement is observed between empirical simulations of g^l_{aa}, $\tilde{g}^l_{ab}$, and formula $g^l = g^{l+1}\chi_1$.

We use the same approach to derive the result for g^l_{ab}. As a result, we have,

$$\begin{aligned} g^l_{aa} &= 4(\frac{q^*_{aa}}{\rho})^2(\frac{\sigma_w^2}{\rho})^{L-l}[\rho + \sum_{j=1}^{L-l}(\frac{\sigma_w^2}{\rho})^j], \\ g^l_{ab} &= 4(q^*_{ab})^2(\sigma_w^2)^{L-l}[1 + \sum_{j=1}^{L-l}(\frac{\sigma_w^2}{\rho^2})^j]. \end{aligned} \tag{12}$$

By analyzing the first formula of Equation (12), we find that $g^l_{aa} = g^{l+1}_{aa}\chi_1$. This can be better observed by dividing the expression related to layer l into two factors: one is $(\frac{\sigma_w^2}{\rho})^{L-l}$, and the other is $\sum_{j=1}^{L-l}(\frac{\sigma_w^2}{\rho})^j$. The first factor accounts for $g^l_{aa} = g^{l+1}_{aa}\chi_1$, where $\chi_1 = \frac{\sigma_w^2}{\rho}$ for linear dropout networks. And second factor will be stable after several layers starting from the last layer L due to $\sigma_w^2 < \rho$. We show an excellent match between the theoretical calculation above with simulation using networks with width $N = 500$ and layer $L = 200$ over 100 different instantiations of the network in Figure 2(a).

Despite the successful prediction of theoretical calculation for g^l_{aa}, our theoretical results for g^l_{ab} only hold on the case of $\rho = 1$ while fail to predict the experimental behavior except for last few layers when $\rho < 1$, as shown in Figure 2(b)(c). After a few layers from L, the variances began to increase dramatically as shown in Figure 2(c). We noticed that unlike the case of computing q^l_{ab}, using χ_2 is prohibitive for computing g^l_{ab}. On the other hand, we try a function regarding χ_1 to fit g^l_{ab}, and find an interesting observations that χ_1 is a much more compatible term for g^l_{ab}, i.e, $g^l_{ab} = g^{l+1}_{ab}\chi_1$. This is demonstrated in Figure 2(d).

The incompatible phenomenon between theoretical calculation and experimental results for g^l_{ab} begins with the emergence of variance, as shown in Figure 2(c). One possible explanation is that the emergence of variance is caused by limited network length. Thus, we can reduce this variance by increasing network length only. To check if this explanation works, we further investigate the relationship between variance and mean of g^l_{ab} with different network widths N. The answer is that $g^l_{ab} = g^{l+1}_{ab}\chi_1$ holds regardless of the finite width. We will demonstrate it in the next section.

After studying the gradient behavior at the linear networks, we conduct a series of experiments on the nonlinear case since the theoretical formulation for nonlinear activation or with the soft-max layer is intractable. We firstly use g^l_{ab} as the metric of gradient and find it has a huge variance when $\rho < 1$. This is because the element of the gradient matrix with a pair of inputs can be either negative or

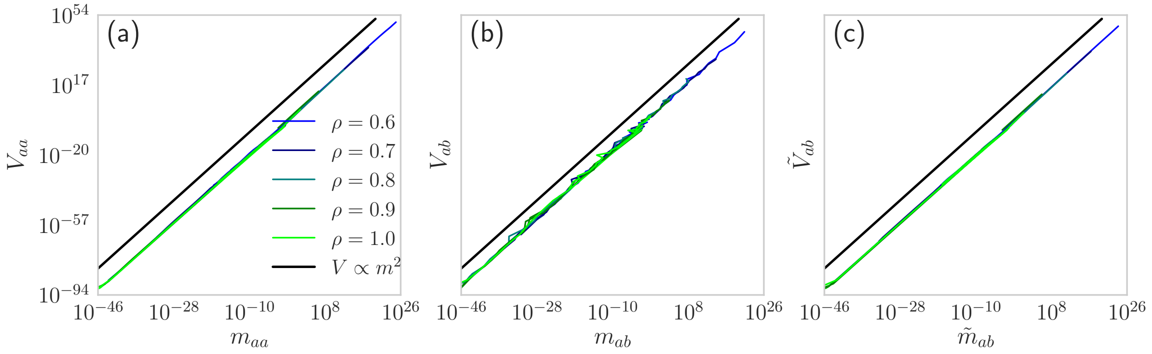

Figure 4. Universal relationship between variance and mean of g_{aa}^l, g_{ab}^l, and $\tilde{g}_{ab}^l$, on the 200 layers and width $N = 500$ random dropout networks. Different color represents a different dropout rate. The black line is the function of $V \propto m^2$. (a) V_{aa}^l as a function m_{aa}^l. (b) V_{ab}^l as a function of m_{ab}^l. (c) $\tilde{V}_{ab}^l$ as a function of $\tilde{m}_{ab}^l$. All the curves regarding different activations collapse to a line, and the power coefficient of all curves is consistent with 2.

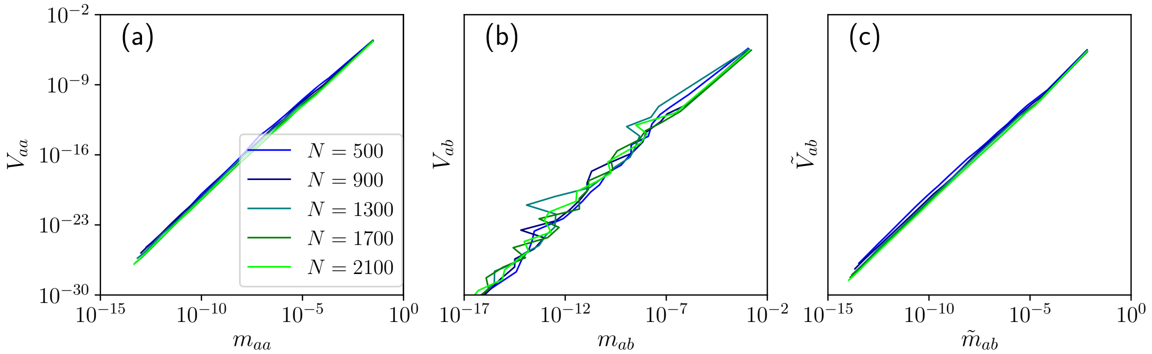

Figure 5. Universal relationship between variance and mean of g_{aa}^l, g_{ab}^l, and $\tilde{g}_{ab}^l$, on the 200 layers, Tanh random dropout networks with $\rho = 0.9$. All the curves regarding different width collapse to a line. Different color represents a different network width. (a) V_{aa}^l as a function m_{aa}^l. (b) V_{ab}^l as a function of m_{ab}^l. (c) $\tilde{V}_{ab}^l$ as a function of $\tilde{m}_{ab}^l$.

positive. To find a metric with low variance, we consider the metric $\tilde{g}_{ab}^l \equiv \frac{1}{N^2}\sum_{ij}|\frac{\partial E_a}{W_{ij}^l}\frac{\partial E_b}{W_{ij}^l}|$ whose elements are all positive. Besides, it is the ℓ_1 norm of the gradient matrix.

We plot g_{aa}^l and $\tilde{g}_{ab}^l$ as a function of l in Figure 3. Interestingly, our simulations show that both g_{ab}^l and $\tilde{g}_{ab}$ are governed by χ_1 in a range of activations. Thus we make a conjecture that the relation,

$$g_{aa}^l = g_{aa}^{l+1}\chi_1, \quad g_{ab}^l = g_{ab}^{l+1}\chi_1, \tag{13}$$

holds on deep dropout networks.

4.2 Emergence of Universality

We have studied three statistical metrics of the gradient, i.e. g_{aa}, g_{ab}, and $\tilde{g}_{ab}$ using their mean value. Inevitably, the variance of these metrics can give us essential information about the gradient. To do this, we performed a series of experiments to obtain the mean and variance of g_{aa}, g_{ab} and $\tilde{g}_{ab}$ with different activation and different network width N.

First, we show the relationship between variance and mean of the metric of gradient with different activations, including Linear, ReLU, Tanh, and Hard Tanh. We denote the mean of g_{aa}, g_{ab} and $\tilde{g}_{ab}$ as m_{aa}^l, m_{ab}^l, and $\tilde{m}_{ab}^l$, while naming the variance as V_{aa}^l, V_{ab}^l, and $\tilde{V}_{ab}^l$ respectively. We show the variance as a function of mean in Figure 4, and find the emergence of universality between the variance and mean regardless of dropout rate and choice of activation for g_{aa}, g_{ab}, and $\tilde{g}_{ab}^l$.

The plot of variance as a function of mean shows a power-law between them since it is like a straight line in the log-log plot. To estimate the power, we use a simple equation $V \propto m^2$ to compare with the experiment results. Surprisingly, all three curves are consistent with $V \propto m^2$. Thus we make a conjecture that the universal power coefficient between the variance and mean is 2.

Then, we investigate the relationship between variance and mean with different network width N and show the results in Figure 5. This time, we perform experiments on the $\rho = 0.9$ Tanh networks with different network width N. Again, the relationship between variance and mean satisfies universality, which means the Equation (13) does not depend on the network width of N.

We want to point out that we have performed the same investigation on q_{aa}^l and c_{ab}^l. However, we did not observe a similar universal relationship between variance and mean of q_{aa}^l and c_{ab}^l. This may occur due to the different behavior of q_{aa}^l (q_{ab}^l) and g_{aa}^l (g_{ab}^l). As Equation (6) shows, the mean of c_{ab}^l will converge to a fixed point after several layers, which means that the mean of c_{ab}^l will be stable in deeper layers. So, we won't expect a universal relation between

Table 1. Summary of depth-scale for theoretical results i.e. signal propagation and gradient backpropagation, and empirical results under different condition or assumption.

Summary	feed-forward propagation		gradient backpropagation		empirical results
metric	q_{aa}	q_{ab}	g_{aa}	g_{ab}	
realistic condition (our work)	-	ξ_2	ξ_1	ξ_1	$\min\{12\xi_1, 12\xi_2\}$
independent assumption [21]	-	ξ_2	ξ_1	ξ_2	$6\xi_2$

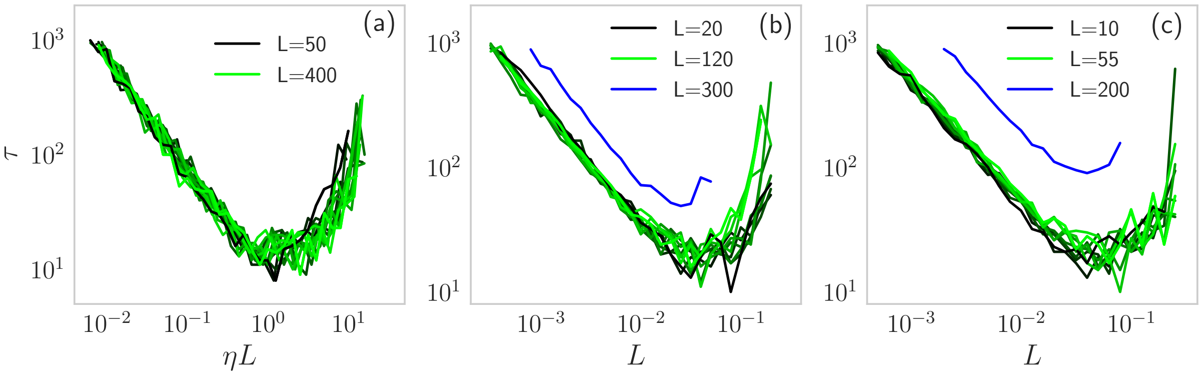

Figure 6. The number of steps τ to reach test accuracy $p \approx 0.25$ as a function learning rate η. (a) Network without dropout, colors reflect different network depth L from 50 (black) to 400 (green). They all collapse to a single universal curve when the learning rate η is re-scaled by L. (b) Network with dropout $\rho = 0.99$, colors reflect different network depth L from 20 (black) to 120 (green), additional $L = 300$ is colored blue for comparison. Curves with $L \leq 120$ collapse to a universal curve without any re-scale. (c) Network with dropout $\rho = 0.98$, colors reflect different network depth L from 10 (black) to 55 (green), additional $L = 200$ is colored blue for comparison. Curves with $L \leq 55$ collapse to a universal curve without any re-scale.

the mean and the variance in this case.

In summary, we have tried all the freedom of parameters that we can tune, the universal power coefficient between the variance and mean remains the same. We conclude that once the topological structure of the neural network is set, the power coefficient is universal.

5 Experiments

According to the theoretical results, during feed-forward, we expect that length-scale ξ_2 controls the propagation of c^l_{ab}, while ξ_1 measures the number of layers that gradient metrics g^l_{aa} and g^l_{ab} can survive during backpropagation. However, [21] claimed that both networks with or without dropout networks have a limited trainable length, which is governed by the depth-scale ξ_2. As our experimental results show, which be demonstrated later, this statement is not exactly right. To summarize, we present the comparison for the length-scale between [21] and our work in Table 1.

5.1 Training speed

Before investigating this problem, we study the relationship between training speed and choice of hyper-parameters. We confine the hyper-parameters at the critical line $\chi_1 = 1$ for the network with and without dropout and train networks of a range of length with width $N = 400$ for 10^3 steps with a batch size of 10^3 on the standard CIFAR10 dataset. Strictly speaking, $\chi_1 = 1$ is not the critical line when $\rho < 1$, since $\chi_2 < 1$. For learning rates of each network, we consider logarithmically spaced in steps 10^1. To search the optimal learning rate, we select a threshold accuracy of $p = 0.25$ and measure the first step τ when performance exceeds p. We show the steps τ as a function of learning rate η on the networks of dropout rate $\rho = 1.0, 0.99$, and 0.98 in Figure 6.

We find that for networks without dropout, there is a universal scaling $\tau = f_1(\eta L)$ between the steps and learning rate, where f_1 is a scaling function, as shown in Figure 6(a). Note that it is different to the result that $\tau/\sqrt{L} = f'_1(\eta L)$ in [17] where they use the standard CIFAR10 dataset augmented with random flips, crops, and so on. The difference may be caused by the pretreatment of the dataset in [17]. Besides, we study the networks with $\rho = 0.99$ and $\rho = 0.98$, and find that the scaling $\tau = f_2(\eta)$ can be kept under a limited length $L = 120$ for $\rho = 0.99$ and $L = 55$ for $\rho = 0.98$, as shown in Figure 6(b) and (c) respectively.

5.2 Trainable length

Now we study the problem of trainable length. We consider random networks of depth $10 \leq L \leq 250$, and $1 \leq \sigma^2_w \leq 4$ with $\sigma^2_b = 0.05$. We train these networks using Stochastic Gradient Descent (SGD) and RMSProp on MNIST and CIFAR10 with Gaussian and Orthogonal weights, which can be seen as another variant of weight initialization in the mean field theory [17]. We perform four experiments on the network without dropout ($\rho = 1$) with different datasets, optimizer, and learning rate to conduct a comprehensive study, and plot the results in Figure 7(a)-(d). Besides, four experiments are conducted on the dropout networks ($\rho < 1$), and results are shown in Figure 7(e)-(h). We color in bright yellow the training accuracy that networks achieved as a function of σ^2_w and L for different dropout rates. From the heatmap, we can observe a boundary in which accuracy began to drop. We noticed that there are two boundaries, left

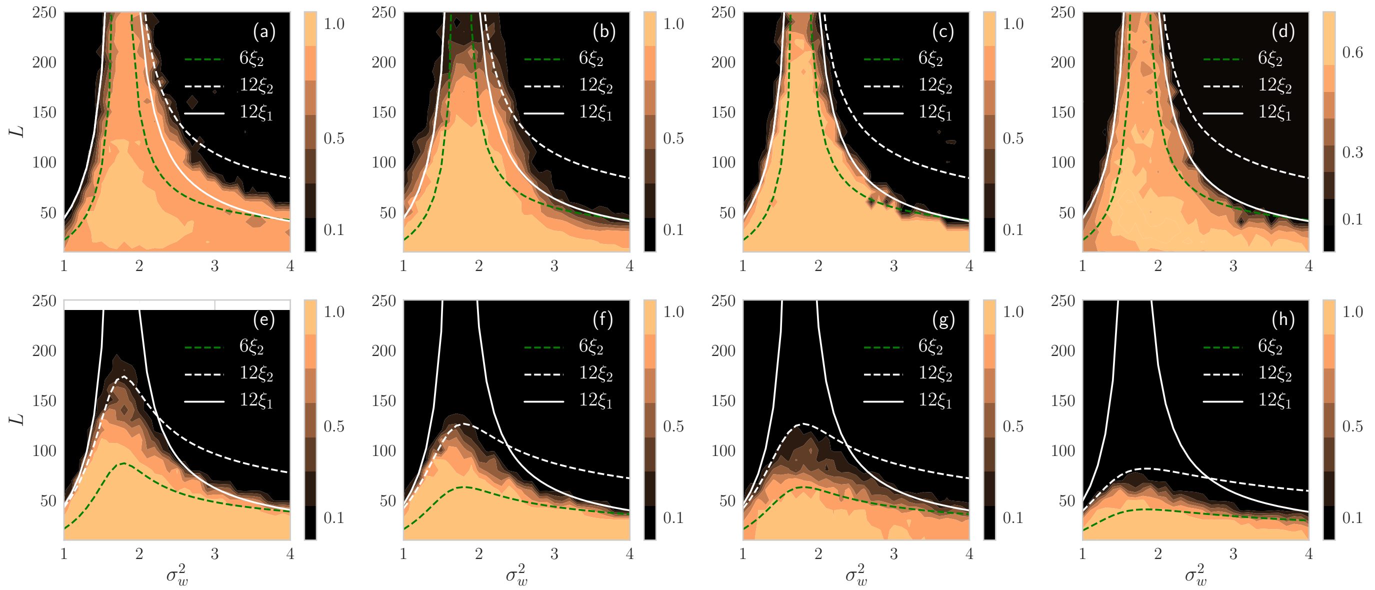

Figure 7. The training accuracy for neural networks as a function of the depth L and initial weight variance σ_w^2 from a high accuracy (bright yellow) to low accuracy (black). Comparison is made by plotting $12\xi_1$ (white solid line), $6\xi_2$ (green dashed line), and $12\xi_2$ (white dashed line). (a) 2000 training steps of $\rho = 1$ network with Gaussian weights on the MNIST using SGD. (b) 1000 training steps of $\rho = 1$ network with Gaussian weights on the MNIST using RMSProp. (c) 2000 training steps of $\rho = 1$ network with Orthogonal weights on the MNIST. (d) 3000 training steps of $\rho = 1$ network with Orthogonal weights on CIFAR10. (e) 3000 training steps of $\rho = 0.99$ network with Orthogonal weights on the MNIST. (f) 3000 training steps of $\rho = 0.98$ network with Orthogonal weights on the MNIST using SGD. (g) 10000 training steps of $\rho = 0.98$ network with Gaussian weights on the MNIST. (h) 3000 training steps of $\rho = 0.95$ network with Orthogonal weights on the MNIST using SGD.

and right. In order to show its relationship with ξ_1 and ξ_2, we superimpose them onto the heatmap.

In figure 7(a), we use the same learning rate and optimizer as those in Figure 5(a)-(c) of [21]. We use a learning rate of 10^{-3} for SGD when $L \leq 200$, and 10^{-4} for larger L. From the plot, we find the $6\xi_2$ underestimates the scope of train-ability in the σ_w^2-L plane, while $12\xi_1$ is more compatible with the experimental result. We note the phenomenon that $6\xi_2$ underestimates the scope of train-ability also happened in Figure 5(b)(c) of [21]. In figure 7(b), we adopt the same learning rate and optimizer as those in Figure 5(d) of [21], where we use a learning of 10^{-5} and RMSProp optimizer. Here, the only difference is that we use 1000 training steps instead of 300 training steps in [21]. According to the simulation result, $12\xi_1$ (solid line) and ξ_2 (dashed line) are identical on the left boundary, while they differ on the right side. We make a comparison between $12\xi_1$ and $12\xi_2$, and find that $12\xi_1$ has a much better argument with the trainable length while $12\xi_2$ overrates the trainable length on the right side.

Based on the analysis of Figure 7(a)(b), we may conclude that $12\xi_1$ can be used to measure the maximum trainable length of the network without dropout. We further reinforce this conclusion by performing experiments on different learning rates, weight initialization, and datasets. In figure(c), we use orthogonal weight initialization. In figure(d), we perform experiment on CIFAR10 dataset and adopt a learning rate of $\eta = c/L$, where c is constant. These learning rates were selected for the reason that each learning rate can lead to the fast step to a certain test accuracy at $\chi_1 = 1$, as shown in Figure 6. In a word, we attribute the maximum trainable length to $L \leq \min\{12\xi_1, 12\xi_2\} = 12\xi_1$, where the relation $\xi_1 \leq \xi_2$ holds on the network without dropout.

Furthermore, we consider the dropout case in Figure 7(e)-(h). We have studied three different dropout rate: $\rho = 0.99$ (Figure 7(e)), $\rho = 0.98$ (Figure 7(f)(g)), and $\rho = 0.95$ (Figure 7(h)). We find that both ξ_1 and ξ_2 have connections to the trainable length: the networks appear to be trainable when $L \leq \min\{12\xi_1, 12\xi_2\}$. Networks on the left side are influenced by $12\xi_2$ while they are constrained by the $12\xi_1$ on the right size. Note that the formula $L \leq \min\{12\xi_1, 12\xi_2\}$ is valid in the no dropout case as discussed above. To conclude, we show an improved relationship between maximum trainable length and length scale ξ_1 and ξ_2 than [21]. This conclusion that both ξ_1 and ξ_2 have connections to the trainable length instead of only ξ_2 [21] is more compatible with the theoretical results.

6 Discussion

In this paper, we have investigated the dropout networks by calculating its statistical metrics of gradient during the backpropagation at initialization and conjecture that both gradients metric with a single input and a pair of inputs are governed by the same quantity χ_1. We further investigate the relationship between variance and mean of statistical metrics empirically and find an emergence of universality. Our finding of a universal relationship between variance and mean of statistical metrics of gradient backpropagation suggests a deeper mechanism behind it. This mechanism may be comprehended better by studying more different network structures such as Resnet. Finally, for networks with or without dropout, we attribute the maximum trainable length to the formula $L \leq \min\{12\xi_1, 12\xi_2\}$, which is novel and important.

REFERENCES

[1] Minmin Chen, Jeffrey Pennington, and Samuel S Schoenholz, 'Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks', *arXiv preprint arXiv:1806.05394*, (2018).

[2] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, 'Empirical evaluation of gated recurrent neural networks on sequence modeling', *arXiv preprint arXiv:1412.3555*, (2014).

[3] Amit Daniely, Roy Frostig, and Yoram Singer, 'Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity', in *Advances In Neural Information Processing Systems*, pp. 2253–2261, (2016).

[4] Yarin Gal and Zoubin Ghahramani, 'Dropout as a Bayesian approximation: Representing model uncertainty in deep learning', in *international conference on machine learning*, pp. 1050–1059, (2016).

[5] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins, 'Learning to forget: Continual prediction with LSTM', (1999).

[6] Dar Gilboa, Bo Chang, Minmin Chen, Greg Yang, Samuel S Schoenholz, Ed H Chi, and Jeffrey Pennington, 'Dynamical isometry and a mean field theory of LSTMs and GRUs', *arXiv preprint arXiv:1901.08987*, (2019).

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, (2016).

[8] Sergey Ioffe and Christian Szegedy, 'Batch normalization: Accelerating deep network training by reducing internal covariate shift', *arXiv preprint arXiv:1502.03167*, (2015).

[9] Arthur Jacot, Franck Gabriel, and Clément Hongler, 'Neural tangent kernel: Convergence and generalization in neural networks', in *Advances in neural information processing systems*, pp. 8580–8589, (2018).

[10] Yann LeCun, Yoshua Bengio, et al., 'Convolutional networks for images, speech, and time series', *The handbook of brain theory and neural networks*, **3361**(10), 1995, (1995).

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, 'Deep learning', *nature*, **521**(7553), 436, (2015).

[12] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein, 'Deep neural networks as Gaussian processes', *arXiv preprint arXiv:1711.00165*, (2017).

[13] Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington, 'Wide neural networks of any depth evolve as linear models under gradient descent', *arXiv preprint arXiv:1902.06720*, (2019).

[14] Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani, 'Gaussian process behaviour in wide deep neural networks', *arXiv preprint arXiv:1804.11271*, (2018).

[15] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur, 'Recurrent neural network based language model', in *Eleventh annual conference of the international speech communication association*, (2010).

[16] Radford M Neal, 'Priors for infinite networks', in *Bayesian Learning for Neural Networks*, 29–53, Springer, (1996).

[17] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli, 'Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice', in *Advances in neural information processing systems*, pp. 4785–4795, (2017).

[18] Jeffrey Pennington, Samuel S Schoenholz, and Surya Ganguli, 'The emergence of spectral universality in deep networks', *arXiv preprint arXiv:1802.09979*, (2018).

[19] Jeffrey Pennington and Pratik Worah, 'Nonlinear random matrix theory for deep learning', in *Advances in Neural Information Processing Systems*, pp. 2637–2646, (2017).

[20] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli, 'Exponential expressivity in deep neural networks through transient chaos', in *Advances in neural information processing systems*, pp. 3360–3368, (2016).

[21] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein, 'Deep information propagation', *arXiv preprint arXiv:1611.01232*, (2016).

[22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, 'Dropout: a simple way to prevent neural networks from overfitting', *The Journal of Machine Learning Research*, **15**(1), 1929–1958, (2014).

[23] Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel S Schoenholz, and Jeffrey Pennington, 'Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks', *arXiv preprint arXiv:1806.05393*, (2018).

[24] Ge Yang and Samuel Schoenholz, 'Mean field residual networks: On the edge of chaos', in *Advances in neural information processing systems*, pp. 7103–7114, (2017).

[25] Greg Yang, 'Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation', *arXiv preprint arXiv:1902.04760*, (2019).

[26] Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, and Samuel S Schoenholz, 'A mean field theory of batch normalization', *arXiv preprint arXiv:1902.08129*, (2019).