The Spectrum of Fisher Information of Deep Networks Achieving Dynamical Isometry

Tomohiro Hayase
Fujitsu Laboratories

Ryo Karakida
AIST

Abstract

The Fisher information matrix (FIM) is fundamental to understanding the trainability of deep neural nets (DNN), since it describes the parameter space's local metric. We investigate the spectral distribution of the conditional FIM, which is the FIM given a single sample, by focusing on fully-connected networks achieving dynamical isometry. Then, while dynamical isometry is known to keep specific backpropagated signals independent of the depth, we find that the parameter space's local metric linearly depends on the depth even under the dynamical isometry. More precisely, we reveal that the conditional FIM's spectrum concentrates around the maximum and the value grows linearly as the depth increases. To examine the spectrum, considering random initialization and the wide limit, we construct an algebraic methodology based on the free probability theory. As a byproduct, we provide an analysis of the solvable spectral distribution in two-hidden-layer cases. Lastly, experimental results verify that the appropriate learning rate for the online training of DNNs is in inverse proportional to depth, which is determined by the conditional FIM's spectrum.

1 Introduction

Deep neural networks (DNNs) have empirically succeeded in achieving high performances in various machine-learning tasks (LeCun et al., 2015; Goodfellow et al., 2016). Nevertheless, their theoretical understanding has been limited, and their success depends much on the heuristic search setting, such as architectures and hyper-parameters. In order to understand and improve the training of DNNs, researchers have developed some theories to investigate, for instance, vanishing/exploding gradient problems (Schoenholz et al., 2017), the shape of the loss landscape (Pennington and Worah, 2018; Karakida et al., 2019a), the global convergence of training and the generalization (Jacot et al., 2018).

The Fisher information matrix (FIM) has been a fundamental quantity for such theoretical understandings. The FIM describes the local metric of the loss surface concerning the KL-divergence function (Amari, 2016). In particular, the eigenvalue spectrum describes the efficiency of optimization methods. For instance, the maximum eigenvalue determines an appropriate size of the learning rate of the first-order gradient method for convergence (LeCun et al., 1991; Karakida et al., 2019a; Wu et al., 2018). In spite of its importance, the spectrum of FIMs in neural networks is not revealed enough from a theoretical perspective. The reason is that it has been limited to random matrix theory for shallow networks (Pennington and Worah, 2018) or mean-field theory for eigenvalue bounds, which may be loose in general (Karakida et al., 2019b). Thus, we need an alternative approach applicable to DNNs.

It is well known that it is difficult to reduce the training error in very deep models without careful prevention of the vanishing/exploding of the gradient. Naive settings (i.e., activation function and initialization) cause vanishing/exploding gradients, as long as the network is relatively deep. The dynamical isometry (Saxe et al., 2014; Pennington et al., 2018) was proposed to solve this problem. The dynamic isometry can facilitate training by setting the input-output Jacobian's singular values to be one, where the input-output Jacobian is the Jacobian matrix of the DNN at a given input. Experiments have shown that with initial values and models satisfying dynamical isometry, very deep models can be trained without gradient vanishing/exploding; (Pennington et al., 2018; Xiao et al., 2018; Sokol and Park, 2020) have found that DNNs achieve approxi-

Proceedings of the 24^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

mately dynamical isometry over random orthogonal weights, but they do not do so over random Gaussian weights.

We investigate the asymptotic spectrum of the FIM of multi-layer perceptrons satisfying dynamical isometry in the present work. In order to handle the mathematical difficulty to treat the spectrum of the full FIM, we focus on the *conditional FIM* given a single sample.

While analyzing the spectrum of DNN, we often face mathematical difficulties caused by the non-linearity of the activation function, the depth of the network, and random matrices. To handle such difficulty, we use the free probability theory (FPT). The FPT, which is invented by Voiculescu for understanding von Neuman algebras (Voiculescu, 1985), provides algebraic tools of random matrix theory (Voiculescu, 1991). Since DNN's FIM is a random matrix polynomial of the Jacobian, the FPT provides tools for understanding FIM's spectral distribution. We use these tools to obtain the propagation of spectral distributions through the layers and then obtain the critical recursive equation of the spectral distributions.

Our findings are the following.

The main finding is that the dynamical isometry makes the spectrum of the conditional FIM concentrates on the maximum spectrum, which value grows linearly as the depth increases (Theorem 4.3). An interesting phenomenon is that the FIM spectrum depends on the depth, unlike the spectrum of the input-output Jacobian, which is independent of the depth. It follows from this that as the depth increases, the parameter space's local geometry linearly depends on the depth and the first-order optimization also suffers from it. Now, our approach for the spectral analysis of DNNs consists of three steps. In the first step, we consider the dual of the conditional FIM in order to focus on the non-trivial non-zero eigenvalues. In the second step, the dual's eigenvalue distribution is decomposed to a free multiplicative convolution of two distributions. Subsequently, we introduce the recursive equations (15) of the dual's spectral distribution throughout the layers. In the last step, the induction on depth shows simultaneously that the maximal value of the limit spectrum of the dual is an atom with a large weight, and we get the recursive equation of the maximum (30). We emphasize that in our setting, the atom helps us simplify the analysis of the maximum.

Secondly, we discover a solvable case on the spectrum distribution of the dual conditional FIM of a non-trivial DNN. In more detail, we explicitly show the asymptotic spectrum of the dual conditional FIM of a two-hidden-layer (a total of three layers) DNN (Theorem 3.2). As long as we know, this is the first solvable case ever for the FIM's spectrum of the deep architecture and clarifies the connection to the universal law of random matrices in FPT.

Thirdly, we empirically confirm in Section 5.1 that the spectrum of the FIM at a small number of samples has the same property as the single-sample version. Section 4.3 describes the rationale for a part of the experiment.

Lastly, in Section 5.2, experimental results confirm that the FIM's dependence on depth determines the appropriate magnitude of the learning rate for the convergence of the first-order optimization at the initial phase of the online training of DNNs.

Our analysis is the first step towards a theoretical understanding of the FIM of DNN achieving dynamical isometry. By extending our framework, we expect to see the spectrum of other FIMs in more varied settings.

1.1 Related Works

Dynamical Isometry and Edge of Chaos A DNN is said to be on the edge of chaos if it preserves the norm of the gradient and the mean squared singular value of the Jacobian throughout layers. However, vanishing or exploding of gradients in specific directions still occurs in the worst case. To prevent them, we need both orthogonal initialization and weight scale depending on activation functions, and that is the finding of the theory of dynamical isometry (Pennington et al., 2017, 2018). The theory of dynamical isometry has been extended to various architectures (Burkholz and Dubatovka, 2019; Gilboa et al., 2019; Tarnowski et al., 2019), and experiments on CNN showed that the dynamical isometry reduced the training error of 10,000 layers of models (Xiao et al., 2018). In contrast, even if the backpropagation signal is isotropic, our study shows that the parameter space's curvature essentially depends on the number of layers. In particular, our study shows that many eigenvalues concentrate on the number of the point of the depth. For example, it suggests that the learning rate, which has been implicitly set, needs to be set to a smaller and more appropriate value depending on the number of layers.

Fisher Information Matrix of DNN Several works have provided the spectral analysis of FIM in limited cases. (Pennington and Worah, 2018) has analyzed the spectrum of FIM via random matrix theory but limited to shallow networks and random Gaussian weight matrices. (Karakida et al., 2019b,a) obtained some bounds for the FIM's eigenvalues in DNNs, but their bounds are loose in general and also limited to Gaussian weights. Saxe et al. (2014) treats the loss's Hessian eigenvalues, but the work is restricted to a linear activation. On the contrary, we investigate the FIM spectrum of deep non-linear networks on random orthogonal weights, which satisfy the dynamical isometry.

Neural Tangent Kernel Additionally, let us remark that (Jacot et al., 2018) uses a version of the dual FIM Θ with Gaussian initialization as the kernel matrix and call it the neural tangent kernel (NTK). The FIM's eigenvalues also determine convergence characteristics of gradient descent in wide neural networks through the NTK. Settings of the last layer of DNNs are different between the theory of NTK and the theory of dynamical isometry. In the theory of NTK, the dimension of the final layer of the DNN is set to be lower-order than that of hidden layers. However, in the theory of dynamical isometry, the final layer's dimension is set to be of the same order as the hidden layers. Huang et al. (2020) treats the NTK regime of orthogonal initialization but does not analyze the case of dynamical isometry.

2 Preliminaries

2.1 Settings

Spectral Distribution For $M \times M$ symmetric matrix A with $M \in \mathbb{N}$, its spectral distribution μ_A is given by

$$\mu_A = \frac{1}{M}\sum_{k=1}^{M} \delta_{\lambda_k}, \tag{1}$$

where tr is the normalized trace, $\lambda_k (k = 1, \dots, M)$ are eigenvalues of A, and δ_λ is the discrete probability distribution whose support is $\{\lambda\} \subset \mathbb{R}$, In general, for a noncommutative probability space $(\mathcal{A}, \tau)$ (see supplementary material C), the spectral distribution μ of $A \in \mathcal{A}$ is a probability distribution μ on $\mathbb{R}$ satisfying $\tau(A^m) = \int t^m \mu(dt)$ for any $m \in \mathbb{N}$.

Dynamical Isometry We say that a feed-forward network achieves dynamical isometry if all singular values of the Jacobian of the network on an input are equal to one. In a later section, we introduce an approximate dynamical isometry in a similar setting as (Pennington et al., 2018).

Network Architecture We assume random weight matrices as is usual in the studies of FIM (Pennington and Worah, 2018; Karakida et al., 2019a) and dynamical isometry (Saxe et al., 2014; Pennington et al., 2018). Fix $L \in \mathbb{N}$. We consider an L-layer feed-forward neural network f_θ with $M \times M$ weight matrices $W_1, W_2, \dots, W_L$ and pointwise activation functions $\varphi^1, \dots \varphi^{L-1}$ on $\mathbb{R}$. Besides, we assume that φ^ℓ is continuous and differentiable except for finite points. Firstly, pick a single input $x \in \mathbb{R}^M$. Set $x^0 = x$. For $\ell = 1, \dots, L$, set

$$h^\ell = W_\ell x^{\ell-1} + b^\ell, \;\; x^\ell = \varphi^\ell(h^\ell). \tag{2}$$

We omit the bias parameters b^ℓ in (2) to simplify the analysis. Write $f_\theta(x) = h^L$. Write

$$D_\ell = \frac{\partial x^\ell}{\partial h^\ell}, \; \delta_{L\to\ell} = \frac{\partial h^L}{\partial h^\ell}. \tag{3}$$

Fisher Information Matrix We focus on the *the Fisher information matrix* (FIM) for supervised learning with a mean squared error (MSE) loss (Pennington and Worah, 2018; Karakida et al., 2019b; Pascanu and Bengio, 2014). Let us summarize its definition and basic properties. Given $x \in \mathbb{R}^M$ and θ, we consider a Gaussian probability model $p_\theta(y|x) = \exp\left(-\mathcal{L}\left(f_\theta(x) - y\right)\right)/\sqrt{2\pi}$ $(y \in \mathbb{R}^M)$. We define the MSE loss by

$$\mathcal{L}(u) = ||u||^2/2, (u \in \mathbb{R}^M), \tag{4}$$

where $||\cdot||$ is the Euclidean norm. In addition, consider a probability distribution $p(x)$ and a joint distribution $p_\theta(x, y) = p_\theta(y|x)p(x)$. Then, the FIM is defined by $\mathcal{I}(\theta) = \int [\nabla_\theta \log p_\theta(x, y)^\top \nabla_\theta \log p_\theta(x, y)] p_\theta(x, y) dx dy$, which is an $LM^2 \times LM^2$ matrix. Now, we denote by $\mathcal{I}(\theta|x)$ the *conditional FIM* (or pointwise FIM) given a single input x defined by

$$\mathcal{I}(\theta|x) = \int [\nabla_\theta \log p_\theta(y|x)^\top \nabla_\theta \log p_\theta(y|x)] p_\theta(y|x) dy. \tag{5}$$

Since $p_\theta(y|x)$ is Gaussian, the conditional FIM is equal to

$$\mathcal{I}(\theta|x) = \frac{\partial f_\theta(x)}{\partial \theta}^\top \frac{\partial f_\theta(x)}{\partial \theta}. \tag{6}$$

We mainly investigate this conditional FIM in the following analysis. Since the distribution $p(x)$ of the input does not depend on θ, the FIM is given by

$$\mathcal{I}(\theta) = \int \frac{\partial f_\theta(x)}{\partial \theta}^\top \frac{\partial f_\theta(x)}{\partial \theta} p(x) dx. \tag{7}$$

We regard $p(x)$ as an empirical distribution of input samples and the FIM (7) is usually referred to as the empirical FIM (Kunstner et al., 2019; Pennington and Worah, 2018; Karakida et al., 2019b). As is known in information geometry (Amari, 2016), the FIM works as a degenerate metric on the parameter space: the Kullback-Leibler divergence between the statistical model and itself perturbed by $d\theta$ is given by $D_{\mathrm{KL}}(p_\theta || p_{\theta+d\theta}) = d\theta^\top \mathcal{I}(\theta) d\theta$. More intuitive understanding is that we can write the Hessian of the loss as $\frac{\partial}{\partial\theta}^2\mathbb{E}_{x,y}[\mathcal{L}(f_\theta(x)-y)] = \mathcal{I}(\theta) + \mathbb{E}_{x,y}[(f_\theta(x)-y)^\top \frac{\partial}{\partial\theta}^2 f_\theta(x)]$. Hence the FIM also characterizes the local geometry of the loss surface around a global minimum with a zero training error.

Dual Fisher Information Matrix Now, in order to ignore $\mathcal{I}(\theta|x)$'s trivial eigenvalue zero, we introduce the *dual conditional FIM* given by

$$H_L(x,\theta) = \frac{1}{M}\frac{\partial f_\theta(x)}{\partial\theta}\frac{\partial f_\theta(x)}{\partial\theta}^\top, \tag{8}$$

which is an $M \times M$ matrix. If there is no confusion, we omit the arguments and denote it by H_L. Except for trivial zero eigenvalues, $\mathcal{I}(\theta|x)/M$ and $H_L(x,\theta)$ share the same eigenvalues as the following:

$$\mu_{I(\theta|x)/M} = \frac{LM^2 - M}{LM^2}\delta_0 + \frac{1}{L}\mu_{H_L(x,\theta)}, \tag{9}$$

where μ_A is the spectral distribution for a matrix A. Note that we multiplied the normalization factor $1/M$ in (8) because the loss $\mathcal{L}(y)$ is $O(M)$ as $M \to \infty$ when the output y has the constant order second moments.

2.2 Free Probability Theory

Freeness Free probability theory gives an asymptotic analysis of families of random matrices in the infinite-dimensional limit. The asymptotic freeness is a vital notion of free probability theory to separate the random matrices' spectral analysis into their respective spectral analysis. We refer readers to the supplemental material C and (Voiculescu et al., 1992; Mingo and Speicher, 2017) for more information on the asymptotic freeness.

S-transform Given probability distribution ν, set $G_\nu(z) = \int (z-t)^{-1}\nu(dt)$ and $h_\nu(z) = zG_\nu(z) - 1$. Then the S-transform (Voiculescu, 1987) of ν is defined as

$$S_\nu(z) = \frac{1+z}{z}\frac{1}{h_\nu^{-1}(z)}. \tag{10}$$

For example, given discrete distribution $\nu = \alpha\delta_0 + (1-\alpha)\delta_\gamma$ with $0 \le \alpha \le 1$ and $\gamma > 0$, we have $S_\nu(z) = \gamma^{-1}(z+\alpha)^{-1}(z+1)$. If two operators A and B are free and each spectral distribution is given by μ and ν respectively, then the spectral distribution of AB is given by the free multiplicative convolution, denoted by $\mu \boxtimes \nu$ (Voiculescu, 1987). Moreover, it holds that

$$S_{\mu\boxtimes\nu}(z) = S_\mu(z)S_\nu(z). \tag{11}$$

3 Propagation of Spectral Distributions

3.1 Recursive Equations

We use several assumptions in the mean-field theory of neural networks (Pennington et al., 2017, 2018; Karakida et al., 2019a) used in the analysis of dynamical isometry. Firstly, we assume that W_ℓ/σ_ℓ are independent and uniformly distributed on $M \times M$ orthogonal matrices, where $\sigma_1, \ldots, \sigma_L > 0$ are constant. Secondly, set

$$\hat{q}_\ell = ||x_\ell||^2/M. \tag{12}$$

Assume that $\hat{q}_0$ converges to $q_0 > 0$. With an appropriate choice of activation function, the empirical distribution of each hidden unit x_ℓ converges to the centered normal distribution (Pennington et al., 2018). Set $q_\ell = \lim_{M\to\infty}\hat{q}_\ell$. Lastly, we assume the following asymptotic freeness.

Assumption 3.1. We assume that $(D_\ell)_{\ell=1}^{L-1}$ is asymptotically free from $(W_\ell, W_\ell^\top)_{\ell=1}^{L}$ as $M \to \infty$ almost surely.

Note that Assumption 3.1 is weaker than the assumption of the forward-backward independence that researches of dynamical isometry assumed (Pennington et al., 2018, 2017; Karakida et al., 2019a). Several works prove or treat the asymptotic freeness with Gaussian initialization (Hanin and Nica, 2019; Yang, 2019; Pastur, 2020), and we expect that it will also hold with orthogonal initialization.

Now we have prepared to discuss the propagation of spectral distributions. It holds that

$$H_L = \sum_{\ell=1}^{L}\hat{q}_{\ell-1}\delta_{L\to\ell}\delta_{L\to\ell}^\top. \tag{13}$$

Since $\delta_{L\to\ell} = W_L D_{L-1}\delta_{L-1\to\ell}$ $(\ell < L)$, it holds that

$$H_{\ell+1} = \hat{q}_\ell I + W_{\ell+1}D_\ell H_\ell D_\ell W_{\ell+1}^\top, \tag{14}$$

where I is the identity matrix. Let μ_ℓ (resp. ν_ℓ) be the limit spectral distribution as $M \to \infty$ of H_ℓ (resp. D_ℓ^2). Note that $\mu_1 = \delta_{q_0}$. By Assumption 3.1, we have the following propagation equation of limit spectral distributions. For $\ell = 1, \ldots, L-1$, we have

$$\mu_{\ell+1} = (q_\ell + \sigma_{\ell+1}^2\,\cdot\,)_*(\nu_\ell \boxtimes \mu_\ell), \tag{15}$$

where the distribution $(b + a\,\cdot\,)_*\mu$ is the pushforward of μ with the map $x \mapsto b + ax$ for a given distribution μ.

3.2 An Example: The Two-Hidden-Layer Case

To show a nontrivial example, we examine the solvable asymptotic spectrum of the conditional FIM in the case of a two-hidden-layer network (i.e. $L = 3$). Assume that

$$\nu_\ell = (1-\alpha_\ell)\delta_0 + \alpha_\ell \delta_{\gamma_\ell}, \tag{16}$$

where $0 < \alpha_\ell < 1$ and $\gamma_\ell > 0$. We get the distribution (16) if we choose activation as the shifted-ReLU ($\varphi(x) = ax$ if $x > b$ otherwise ab with $a, b > 0$) or the hard tanh given by

$$\varphi_{s,g}(x) = \begin{cases} gx, & \text{if } sg|x| < 1, \\ g \cdot \mathrm{sgn}(x), & \text{otherwise}, \end{cases} \tag{17}$$

where $s, g > 0$. These activation functions appear in (Pennington et al., 2018) for dynamical isometry. Then we have the following explicit representation of the H_3's asymptotic spectral distribution μ_3.

Theorem 3.2. *We have* $\mu_3(dx) = \mu_{\mathrm{atoms}}(dx) + \rho(x)dx$, *where*

$$\begin{aligned} \mu_{\mathrm{atoms}} = (1-\alpha_2)\delta_{\lambda_{\min}} + (\alpha_2 - \alpha_1)^+ \delta_{\lambda_{\mathrm{mid}}} \\ + (\alpha_1 + \alpha_2 - 1)^+ \delta_{\lambda_{\max}}(dx), \end{aligned} \tag{18}$$

$$\rho(x) = \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi(x - \lambda_{\mathrm{mid}})(\lambda_{\max} - x)} \mathbf{1}_{[\lambda_-, \lambda_+]}(x), \tag{19}$$

with the following notations: $a^+ = \max(a, 0)$ *for* $a \in \mathbb{R}$, $\mathbf{1}_X$ *is the indicator function for* $X \subset \mathbb{R}$,

$$\lambda_{\min} = q_2, \tag{20}$$

$$\lambda_{\mathrm{mid}} = q_2 + \sigma_3^2 \gamma_2 q_1, \tag{21}$$

$$\begin{aligned} \lambda_\pm = q_2 + \\ \sigma_3^2 \gamma_2 \{ q_1 + \\ \sigma_2^2 \gamma_1 q_0 [\sqrt{\alpha_1(1-\alpha_2)} \pm \sqrt{\alpha_2(1-\alpha_1)}]^2 \}, \end{aligned} \tag{22}$$

$$\lambda_{\max} = q_2 + \sigma_3^2 \gamma_2 (q_1 + \sigma_2^2 \gamma_1 q_0). \tag{23}$$

Proof. The proof is based on (11) and (15), and is postponed to the supplemental material A. □

Fig. 1 shows the agreement of the predicted distribution μ_3 by Theorem 3.2 and the empirical spectral distribution of H_3 .

Theorem 3.2 (23) reveals that the maximal eigenvalue is close to three, which is the number of layers, when (25) is satisifed and $q_* = 1$. Further, if (26) is satisfied, then $(\alpha_1 + \alpha_2 - 1)^+$ is close to one and the eigenvalues concentrate on the maximal eigenvalue. We observed in Fig. 1 (left) that most of the eigenvalues concentrated at the value of depth when a DNN achieved dynamical isometry approximately, but there were other peaks. Later, we show in Theorem 4.3 that the eigenvalues concentrate on the value of depth at the large depth.

Additionally, Theorem 3.2 (18) reveals that the weight of the minimum eigenvalue depends on how much the last activation's derivation vanishes.

Theorem 3.2 also gives insight into the spectrum of the DNN out of the dynamical isometry. Although Fig. 1 (right) is also out of dynamical isometry, the spectrum obeys the arcsin law known in FPT (Voiculescu et al., 1992) and is interesting its own right. Once the hyperparameters are standardized and each D_ℓ is a projection, the spectrum is attributed to the product of free two projections well examined in FPT.

4 Analysis through Approximate Dynamical Isometry

4.1 Assumptions

Let us review on how to achieve the dynamical isometry. For the sake of the prospect of the theory, let J be the Jacobian of the network with ignoring the last layer $h^L \mapsto x^L = W_L h^L$. We say that the network achieves dynamical isometry if all eigenvalues of JJ^T are one. Consider the limit spectral distribution ν_ℓ of D_ℓ^2 given by (16). At the deep limit, we consider the situation such that forward and backward signal propagation are stable. Hence we adopt the following assumption,

Assumption 4.1. (Pennington et al., 2018) For each $L \in \mathbb{N}$, sequences $q_\ell, \alpha_\ell, \gamma_\ell$ and $\sigma_{\ell+1}$ ($\ell = 0, 1, \dots, L-1$) are constant for every ℓ, but depend on L. In addition, each constant converges to a finite value as $L \to \infty$.

Now the limit spectral distribution $\mu_{J^\top J}$ of $J^\top J$ as $M \to \infty$ is given by $[(\sigma_{L-1}^2 \cdot)_* \nu_{L-1}]^{\boxtimes (L-1)}$. By (11),

$$S_{\mu_{J^\top J}}(z) = \left(\frac{1}{\sigma_{L-1}^2 \gamma_{L-1}} \left(1 + \frac{1 - \alpha_{L-1}}{z + \alpha_{L-1}}\right)\right)^{L-1}. \tag{24}$$

In order to achieve dynamical isometry, we need that $[(\sigma_{L-1}^2 \cdot)_* \nu_{L-1}]^{\boxtimes (L-1)}$ converges to a compactly supported distribution as $L \to \infty$, and $S_{JJ^T}(z)$ converges to a non-zero function of z. Since the right hand side is approximated by $\exp(L(1-\alpha_L)(z+\alpha_L)^{-1} - L \log \sigma_L^2 \gamma_L))$ as $L \to \infty$, we need

$$\log \sigma_L^2 \gamma_L = O(L^{-1}) \tag{25}$$

$$1 - \alpha_L = O(L^{-1}) \tag{26}$$

and as $L \to \infty$.

Now, the exact dynamical isometry, which means that the input-output Jacobian is an orthogonal matrix, is too strong. Thus we consider the following approximate condition.

Assumption 4.2. We assume that the following limits exist and $|\varepsilon_1|, |\varepsilon_2| < 1$:

$$\varepsilon_1 = \lim_{L \to \infty} L(1 - \alpha_L), \tag{27}$$

$$\varepsilon_2 = -\lim_{L \to \infty} L \log \sigma_L^2 \gamma_L. \tag{28}$$

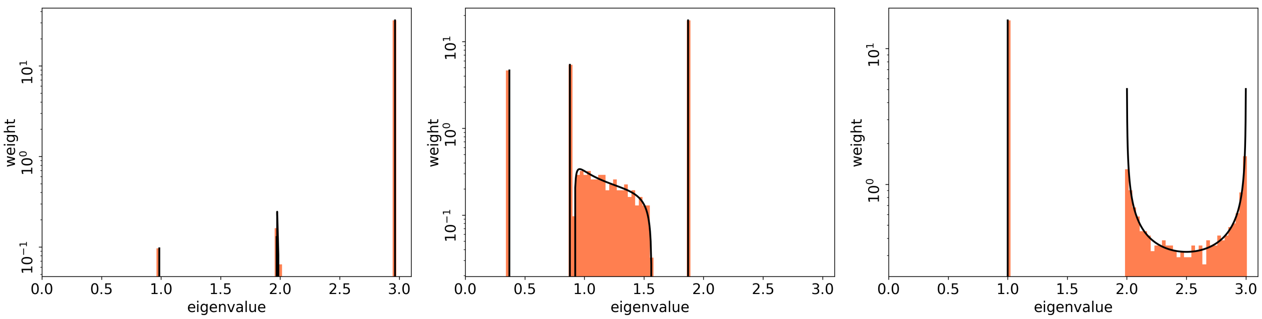

Figure 1: Normalized histograms of eigenvalues of H_3 (orange) and predicted μ_3 by Theorem 3.2 (black lines). The y-axis in each figure is logarithmic. We set $M = 1000$, $\hat{q}_0 = \sigma_\ell = 1$, the width of bins 0.31, and used the following setting. (Left): We used the hard tanh activation (17) with $s = 0.125$ and $g = 1.0013$ to achieve dynamical isometry. (Center): We used the hard tanh with $s = g = 1$. (Right): We constructed H_L based on (14) by replacing each Jacobian D_ℓ with an independent matrix whose spectral distribution is $(1/2)\delta_0 + (1/2)\delta_1$.

To achieve exact dynamical isometry, we need $\varepsilon_1 = \varepsilon_2 = 0$. Assumption 4.2 is implicitly assumed in Pennington et al. (2018).

4.2 Concentration around the Maximal Eigenvalue of the Conditional FIM

We investigate $||\mu_L||_\infty$, where we denote by $||\nu||_\infty$ the maximum of the support of ν for a compactly supported probability distribution ν on $\mathbb{R}_+$. We show that $||\mu_L||_\infty$ is $O(L)$ as $L \to \infty$. We discuss the intuitive reason for this ahead of time. Now, ν_ℓ is close to a delta measure at a point to achieve dynamical isometry. Then the recursive equation (15) looks like an affine transform. Therefore, $||\mu_L||_\infty$ is the result of L-times affine transforms, hence it is $O(L)$ as $L \to \infty$.

A key of the proof is to show that $||\mu_L||_\infty$ is an atom of μ_L under Assumption 4.2, which is an assumption to achieve approximate dynamical isometry. Recall that $x \in \mathbb{R}$ is an atom of a probability distribution ν if and only if $\nu(\{x\}) > 0$. Here we sketch the proof of the phenomenon. Firstly, the support of ν_ℓ, which is the limit spectral distribution of D_ℓ^2, consists of two atoms 0 and γ_ℓ, and the weight α_ℓ of γ_ℓ is close to one for sufficiently large ℓ. Secondly, we recall the following fact of the free multiplicative convolution: if a (resp. b) is an atom of a probability distribution μ (resp. ν) and $\mu(\{a\}) + \nu(\{b\}) - 1 > 0$, then ab is an atom of $\mu \boxtimes \nu$ and

$$\mu \boxtimes \nu(\{ab\}) = \mu(\{a\}) + \nu(\{b\}) - 1. \tag{29}$$

Thus, if $||\mu_\ell||_\infty$ is an atom of μ_ℓ with sufficiently large weight, then $||\nu_\ell||_\infty ||\mu_\ell||_\infty$ is an atom of $\nu_\ell \boxtimes \mu_\ell$. Note that (29) is different from that in the computation rule of the classical convolution. Additionally, by the recursive equation (15), we have the following recursive equation:

$$||\mu_{\ell+1}||_\infty = q_\ell + \sigma_{\ell+1}^2 ||\nu_\ell||_\infty ||\mu_\ell||_\infty. \tag{30}$$

This recursive equation allows us to use induction. Then we have our theorem.

Theorem 4.3. *Consider Assumption 4.1 and 4.2. Then for sufficiently larger L, it holds that $||\mu_L||_\infty$ is an atom of μ_L with weight $1-(L-1)(1-\alpha_{L-1})$, and*

$$\lim_{L\to\infty} L^{-1}||\mu_L||_\infty = q\frac{1-\exp(-\varepsilon_2)}{\varepsilon_2}. \tag{31}$$

In particular, the limit has the expansion $q(1-\varepsilon_2/2) + O(\varepsilon_2^2)$ as the further limit $\varepsilon_2 \to 0$.

The proof is postpone to Appendix D.

Recall that μ_L is equal to the limit distribution of eigenvalues of the conditional FIM $I(\theta|x)$ except for $LM^2 - M$ zeros. Theorem 4.3 shows that the maximum of the support is $O(L)$ as $L \to \infty$. Furthermore, we emphasize that the weight $1-(L-1)(1-\alpha_{L-1}) \sim 1-\varepsilon_1$ of the maximal eigenvalue $||\mu_L||$ is close to 1. Therefore, non-zero eigenvalues of the conditional FIM asymptotically concentrates around $qL(1-\varepsilon_2/2)$. In particular, the non-degenerate-part of the conditional FIM is approximated by a scaled identity operator.

4.3 Expected vs Conditional FIM

While we have investigated the conditional FIM (6) so far, it will be curious to show how the obtained eigenvalue statistics could be related to those of the FIM (7). We investigate the analysis of mean eigenvalues of both the conditional FIM and the usual FIM. Analysis of the mean is easier than analysis of the maximum value, but is done to observe the effect of the dynamic isometry. Now, we denote by $m_1(\mu)$ the mean of a distribution μ given by $m_1(\mu) = \int x\mu(dx)$. We show that the mean of spectrum is $O(L)$ as $L \to \infty$.

Applying the decomposition formula of the free multiplicative convolution $m_1(\mu \boxtimes \nu) = m_1(\mu)m_1(\nu)$ to Eq. (15), we have the following recurrence formula of the mean:

$$m_1(\mu_{\ell+1}) = q_\ell + \sigma_{\ell+1}^2 m_1(\nu_\ell) m_1(\mu_\ell). \tag{32}$$

Then we have the followings.

Proposition 4.4. *Under Assumption 4.1 and 4.2, it holds that*

$$\lim_{L\to\infty} L^{-1} m_1(\mu_L) = q \frac{1 - \exp(-\varepsilon_1 - \varepsilon_2)}{\varepsilon_1 + \varepsilon_2}, \tag{33}$$

where $q = \lim_{L\to\infty} q_L$. In particular, the limit has the expansion $q(1 - (\varepsilon_1 + \varepsilon_2)/2) + O((\varepsilon_1 + \varepsilon_2)^2)$ as $\varepsilon_1, \varepsilon_2 \to 0$.

Proof. We have $m_1(\mu_{L+1}) = \sum_{\ell=0}^{L} q_\ell (\sigma_{\ell+1}^2 \alpha_\ell \gamma_\ell)^\ell = q_L \sum_{\ell=0}^{L} (\sigma_{L+1}^2 \alpha_L \gamma_L)^\ell$ by Assumption 4.1. Set $x_L = L(1 - \sigma_{L+1}^2 \alpha_L \gamma_L)$. We have $\lim_{L\to\infty} x_L = \varepsilon_1 + \varepsilon_2$. Then $L^{-1} \sum_{\ell=0}^{L} (\sigma_{L+1}^2 \alpha_L \gamma_L)^\ell = x_L^{-1}[1 - (1 - x_L/L)^L] \to (\varepsilon_1 + \varepsilon_2)^{-1}(1 - \exp(-\varepsilon_1 - \varepsilon_2))$. Then the assertion has been proven. $\square$

Next, consider the usual FIM. Fix $N \in \mathbb{N}$ and consider input vectors $x(1), \dots, x(N) \in \mathbb{R}^M$. Set $x^0(n) = x(n)$. Since $\mathcal{I}(\theta) = N^{-1} \sum_{n=1}^{N} \mathcal{I}(\theta | x(n))$, the FIM $\mathcal{I}(\theta)$ shares non-zero eigenvalues with the dual Θ, which is the $N \times N$ matrix whose (m, n)-entry is given by the following $M \times M$ matrix:

$$\Theta(m, n) = \frac{1}{N} \frac{\partial f_\theta(x(m))}{\partial \theta} \frac{\partial f_\theta(x(n))}{\partial \theta}^\top, \tag{34}$$

where $m, n = 1, \dots, N$. For $\ell = 1, \dots, L$, set recursively $h^\ell(n) = W_\ell x^{\ell-1}(n)$, $x^\ell(n) = \varphi^\ell(h^\ell(n))$, and define $\delta_{L\to\ell}(n)$ in the same way. Then $\Theta(m, n) = MN^{-1} \sum_{\ell=1}^{L} \Sigma_\ell(m, n) \delta_{L\to\ell}(m) \delta_{L\to\ell}(n)^\top$, where $\Sigma_\ell(m, n) = M^{-1} \sum_{i=1}^{M} x_{\ell,i}(m) x_{\ell,i}(n)$, and we have the following block-matrix representation:

$$\Theta = \frac{M}{N} \begin{bmatrix} H_L(x(1)) & * & \cdots & * \\ * & H_L(x(2)) & \cdots & * \\ \vdots & \cdots & \ddots & \vdots \\ * & * & \cdots & H_L(x(N)) \end{bmatrix} \tag{35}$$

Hence for the N-sample, considering the collection of eigenvalues of the dual conditional FIMs $(H_L(x(n)))_{n=1}^N$ is equivalent to considering block-diagonal approximation of the (scaled) dual FIM $MN^{-1}\Theta$. Even if it is not clear yet that the block-diagonal approximation behaves well, the mean of eigenvalues of full matrix is exactly determined by the diagonal part as the following assertion.

Corollary 4.5. *Denote by $m_{L,N}$ the wide limit $M \to \infty$ of the mean of eigenvalues of Θ/M. Then under the limit $L, N \to \infty$ with $L/N \to \alpha < \infty$, it holds that*

$$m_{L,N} \to \alpha q \frac{1 - \exp(-\varepsilon_1 - \varepsilon_2)}{\varepsilon_1 + \varepsilon_2}. \tag{36}$$

Proof. Fix L, N. The mean of eigenvalues of Θ/M is equal to $\mathrm{tr}(\Theta/M)$ and $\mathrm{tr}(\Theta/M) = \sum_{n=1}^{N} \mathrm{tr}(H_L(x(n)))/N^2 \to m_1(\mu_L)/N$. By Proposition 4.4, the assertion follows. $\square$

Corollary 4.5 implies that the mean eigenvalue of Θ/M is close to qLN^{-1} when L and N are of the same magnitude. In the next section, we empirically examine the block-diagonal approximation.

5 Empirical Analysis

5.1 Expected vs Conditional FIM

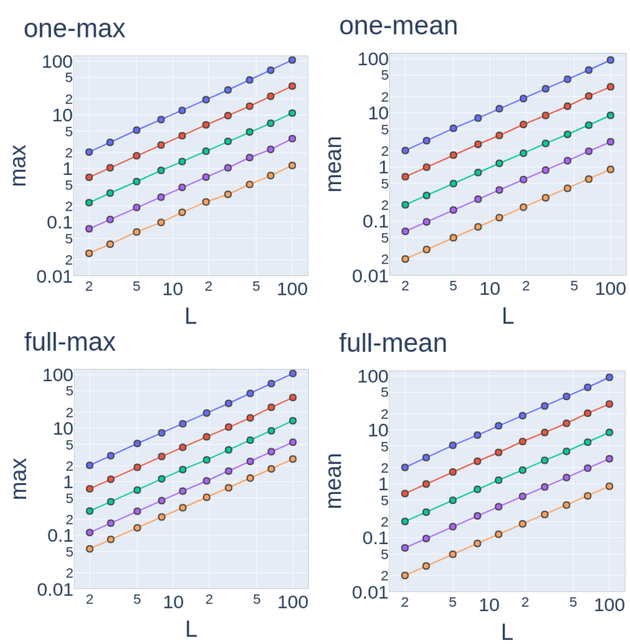

Figure 2: Eigenvalues of Θ/M (full matrix) and the collection of eigenvalues of $(H_L(x(n))/N)_{n=1}^N$ (one block); We show maximum and mean in each case. We set $M = 100$ and each axis logarithmic.

In order to investigate the difference between the usual FIM (7) for multi-samples and the conditional FIM (6), we numerically computed eigenvalues of their dual matrices with $\hat{q}_0 = 1$. Fig. 2 shows the statistics of

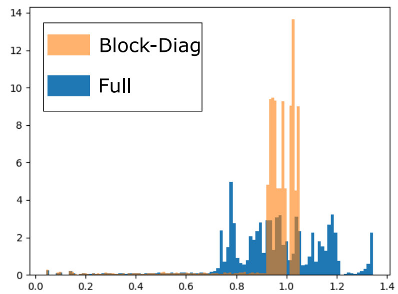

Figure 3: Histograms of eigenvalues.

their eigenvalues of Θ/M (Full) and $(H_L(x(n))/N)_{n=1}^N$ (Block-Diag) with $L = N = 10$. We observed that the maximum and the mean eigenvalue of each matrix are $O(L/N)$, except for the maximum eigenvalue of Θ/M with a large N.

Fig. 3 shows an example of eigenvalues of Θ/M (Full) and $(H_L(x(n))/N)_{n=1}^N$ (Block-Diag) concentrated around L/N for a small N. For the conditional FIM, we observed theoretical predictions Proposition 4.4 and Theorem 4.3 agree well with experimental results.

5.2 Training Dynamics

We investigate how our FIM's spectrum affects training dynamics. Consider the online gradient descent method:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta [M^{-1}\mathcal{L}(f_{\theta_t}(x(t) - y(t))]. \tag{37}$$

Under the first order Tayler approximation of f_θ around the initial parameter θ_0, we have the following approximation:

$$\begin{aligned}\theta_{t+1} &\sim (I - \eta H_L(x(t), \theta_0))\theta_t \\ &+ M^{-1} f_{\theta_0}(x)^\top (y - f_{\theta_0}(x) + \nabla_\theta f_{\theta_0}(x(t))^\top \theta_0). \end{aligned} \tag{38}$$

Hence at the initial phase of the training, the condition $||I - \eta H_L|| < 1$ is necessary to avoid the explosion of parameters. In particular,

$$\eta < \frac{2}{\lambda_{\max}(H_L)}. \tag{39}$$

By Theorem 4.3, we expect that the boundary $2/\lambda_{\max}(H_L)$ will be close to $2/(qL)$.

To confirm the boundary, we exhaustively searched training and testing accuracy while changing L and η with normalizing inputs with $\hat{q}_0 = 1$. In Fig. 4, we observed the theoretical prediction $\eta = 2/L$ by (39) coincided nicely with the boundary of the collapse of the accuracy obtained in experiments.

We normalized each input so that $\hat{q}_0 = 1$ and converted class labels to an orthonormal system in $\mathbb{R}^M$. In whole experiments, we commonly use the hard-tanh activation with $s^2 = 0.125$ and $g = 1.0013$ to archive dynamical isometry. After training, we computed the MSE loss $\mathcal{L}/M$ and accuracy on the dataset of 10000 samples separated from the dataset for training.

We trained the network on benchmark datasets Fashion-MNIST (Xiao et al., 2017) and CIFAR10 (Krizhevsky, 2009). The Fashion-MNIST (resp. CIFAR10) consists of 28×28 (resp. $3 \times 32 \times 32$) dimensional images and 10 class labels, We applied the online gradient descent on 500 data, which is uniformly sampled from whole data and fixed, for an epoch (resp. ten epochs) in the Fashion-MNIST (resp. CIFAR10). In Fig. 4, the difference between training and testing accuracy on the CIFAR10 was larger than that on the Fashion-MNIST. The reason for this is because of the overfitting of DNNs to the small dataset consists of only 500 data. However, agreement with the theoretical line was also visible in the testing accuracy on the CIFAR10.

Additionally, we observed that the testing loss slightly violated the boundary at large L (i.e., $\eta > 2/L$ and the test loss was not large). (See the supplemental material B for the detail.) In this region, we found that the spectral distribution of H_L during training was far different from that in the initial state. Thus, although the parameters did not explode, we had a qualitative change of the optimization in the seeping region. In this sense, the theoretical boundary explains well the state of training.

6 Conclusion

Our study establishes a springboard for a new way to examine Fisher information of neural networks with a powerful methodology provided by the free probability theory. In particular, we have shown that the dual conditional FIM's eigenvalues get concentrated at the maximum when DNNs achieve approximately dynamical isometry. Furthermore, we have explicitly solved a propagation equation and shown the spectrum distribution's exact form in the case depth is three.

As the study's evidence indicates, the empirical Fisher information matrix is in L/N growth rate. Interestingly, a depth-dependent learning rate has been empirically observed in DNNs achieving dynamical isometry (Pennington et al., 2017) and required for the convergence of training in deep linear networks Hu et al. (2020). Our theoretical results support their empirical observation and clarify the intrinsic distortion of the parameter space.

We are aware that our spectral analysis of the FIM may have a few limitations. A limitation is that our analysis is based on the asymptotic freeness of Jacobi matrices

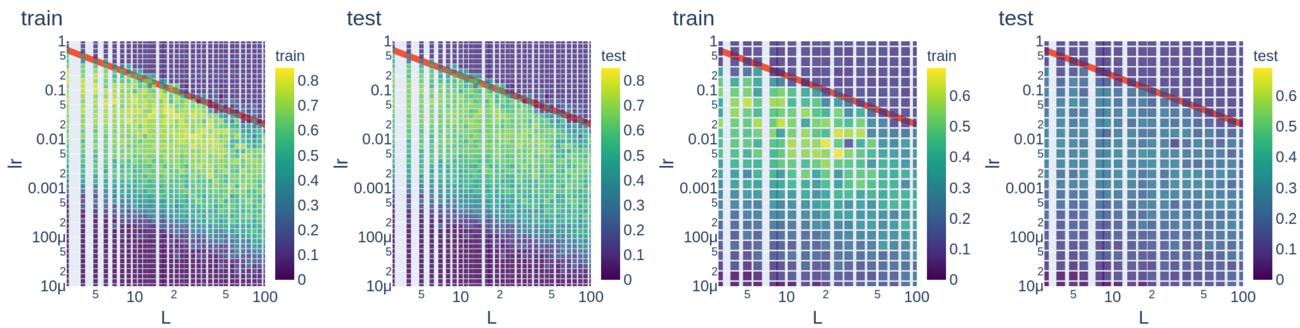

Figure 4: Accuracy heatmaps for different value of L (x-axis) and lr ($= \eta$) (y-axis) after online training. In each figure, the line is $\eta = 2/L$. Each axis is logarithmic. Each network was trained on Fashion-MNIST or CIFAR10. In each experiment, the training dataset consists of 500 data points sampled uniformly from a whole data set, and the testing dataset consists of 10000 samples separated from the training dataset. For the Fashion-MNIST (resp. CIFAR10), the network was trained for a single epoch (resp. ten epoch) training. The leftmost figure (resp. the second from left) is the heatmap of the training (resp. testing) accuracy evaluated after training on Fashion-MNIST. The third figure from left (resp. the rightmost) is the heatmap of the training (resp. testing) accuracy evaluated after training on CIFAR10.

(Assumption 3.1). We expect that the works (Hanin and Nica, 2019; Yang, 2019; Pastur, 2020), which prove or treat the asymptotic freeness with Gaussian initialization, will help us to prove it with the orthogonal initialization. Another limitation is that the batch size is limited to be small in our theory. Future analysis of block random matrices via free probability theory will investigate the full Fisher information matrix spectrum.

7 Acknowledgement

The authors gratefully acknowledge valuable and profound comments of Roland Speicher. The authors would like to thank Hiroaki Yoshida and Noriyoshi Sakuma for constructive discussion. TH acknowledges the funding support from JST ACT-X Grant number JPMJAX190N. RK acknowledges the funding support from JST ACT-X Grant Number JPMJAX190A.

References

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL `https://doi.org/10.1038/nature14539`.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *ICLR 2017, arXiv:1611.01232*, 2017.

Jeffrey Pennington and Pratik Worah. The spectrum of the Fisher information matrix of a single-hidden-layer neural network. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 5410–5419, 2018.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of Fisher information in deep neural networks: Mean field approach. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS); (arXiv1806.01316)*, pages 1032–1041, 2019a.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems (NeurIPS)*, pages 8571–8580, 2018.

Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016.

Yann LeCun, Ido Kanter, and Sara A. Solla. Eigenvalues of covariance matrices: Application to neural-network learning. *Physical Review Letters*, 66(18): 2396–2399, 1991. ISSN 00319007. doi: 10.1103/PhysRevLett.66.2396.

Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8279–8288. Curran Associates, Inc., 2018.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. The normalization method for alleviating pathological sharpness in wide neural networks. In *Advances in neural information processing systems (NeurIPS)*, 2019b.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *ICLR 2014, arXiv:1312.6120*, 2014.

Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1924–1932, 2018.

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. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 5393–5402, 2018.

Piotr A Sokol and Il Memming Park. Information geometry of orthogonal initializations and training. *ICLR 2020, arXiv:1810.03785*, 2020.

Dan V Voiculescu. Symmetries of some reduced free product C*-algebras. In *Operator Algebras and their Connections with Topology and Ergodic Theory*, volume 1132 of *Lecture Notes in Math.*, pages 556–588. Springer, Berlin, 1985.

Dan V Voiculescu. Limit laws for random matrices and free products. *Invention Math.*, 104:201–220, 1991.

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 (NeurIPS)*, pages 4785–4795, 2017.

Rebekka Burkholz and Alina Dubatovka. Initialization of relus for dynamical isometry. In *Advances in Neural Information Processing Systems*, pages 2385–2395, 2019.

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.

Wojciech Tarnowski, Piotr Warchoł, Stanisław Jastrzębski, Jacek Tabor, and Maciej Nowak. Dynamical isometry is achieved in residual networks in a universal way for any activation function. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89, pages 2221–2230. PMLR, 16–18 Apr 2019. URL `http://proceedings.mlr.press/v89/tarnowski19a.html`.

Wei Huang, Weitao Du, and Richard Yi Da Xu. On the neural tangent kernel of deep networks with orthogonal initialization. *arXiv preprint, arXiv:2004.05867*, 2020.

Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. *ICLR 2014, arXiv:1301.3584*, 2014.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical Fisher approximation. In *Advances in neural information processing systems*, 2019.

Dan V Voiculescu, Ken J Dykema, and Alexandru Nica. *Free random variables*. Number 1 in CRM Monograph Series. American Mathematical Soc., 1992.

James A Mingo and Roland Speicher. *Free probability and random matrices*, volume 35 of *Fields Institute Monograph*. Springer-Verlag New York, 2017.

Dan V Voiculescu. Multiplication of certain noncommuting random variables. *J. Operator Theory*, 18:223–235, 1987.

Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376:1–36, 2019.

Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv:1902.04760*, 2019.

Leonid Pastur. On random matrices arising in deep neural networks: Gaussian case. arXiv preprint, arXiv:2001.06188, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint, arXiv:1708.07747, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Tech Report*, 2009.

Wei Hu, Lechao Xiao, and Jeffrey Pennington. Provable benefit of orthogonal initialization in optimizing deep linear networks. *ICLR 2020, arXiv:2001.05992*, 2020.