Pathological spectra of the Fisher information metric and its variants in deep neural networks

Ryo Karakida; Shotaro Akaho; and Shun-ichi Amari

Abstract

The Fisher information matrix (FIM) plays an essential role in statistics and machine learning as a Riemannian metric tensor. Focusing on the FIM and its variants in deep neural networks (DNNs), we reveal their characteristic behavior when the network is sufficiently wide and has random weights and biases. Various FIMs asymptotically show pathological eigenvalue spectra in the sense that a small number of eigenvalues take on large values while most of them are close to zero. This implies that the local shape of the parameter space or loss landscape is very steep in a few specific directions and almost flat in the other directions. Similar pathological spectra appear in other variants of FIMs: one is the neural tangent kernel; another is a metric for the input signal and feature space that arises from feedforward signal propagation. The quantitative understanding of the FIM and its variants provided here offers important perspectives on learning and signal processing in large-scale DNNs.

1 Introduction

Deep neural networks (DNNs) have outperformed many standard machine-learning methods in practical applications [1]. Despite their practical success, many theoretical aspects of DNNs remain to be uncovered, and there are still many heuristics used in deep learning. We need a solid theoretical foundation for elucidating how and under what conditions DNNs and their learning algorithms work well.

The Fisher information matrix (FIM) is a fundamental metric tensor that appears in statistics and machine learning. The FIM determines the Cramér-Rao bound for parameter estimation. An empirical FIM is equivalent to the Hessian of the loss function around a certain global minimum, and it affects the performance of optimization in machine learning. In information geometry, the FIM defines the Riemannian metric tensor of the parameter manifold of a statistical model [2]. The natural gradient method is a first-order gradient method in the Riemannian space, and it is characterized by the FIM and invariance under parameter coordinate transformations [3–7]. It can be used to train various DNNs faster than conventional gradient methods can. The FIM also acts as a regularizer to prevent catastrophic forgetting [8]; a DNN trained on one dataset can learn another dataset without forgetting information if the parameter change is regularized with the diagonal of the FIM.

However, our understanding of the FIM for neural networks has so far been limited to empirical studies and theoretical analyses of simple networks. Numerical experiments confirmed that the eigenvalue spectra of the FIM and those of the Hessian are highly distorted; that is, most eigenvalues are close to zero, while others take on large values [9–11]. Focusing on shallow neural networks, Pennington and Worah [12] theoretically analyzed the FIM’s eigenvalue spectra by using random matrix theory, and Fukumizu [13] derived a condition under which the FIM becomes singular. Liang et al. [14] have connected FIMs to the generalization ability of DNNs by using model complexity, but their results are restricted to linear networks. Thus, theoretical evaluations of deeply nonlinear
cases seem to be difficult because of iterated nonlinear transformations. To go one step further, it would be helpful if a framework that is widely applicable to various DNNs could be constructed.

Investigating DNNs with random weights has given promising results. When such DNNs are sufficiently wide, we can formulate their behavior by using simpler analytical equations through coarse-graining of the model parameters, as is discussed in mean field theory \[15\]-\[21\] and random matrix theory \[22\]-\[24\]. For example, Schoenholz et al. \[18\] proposed a mean field theory for backpropagation in fully-connected DNNs. This theory characterizes the amplitudes of gradients by using specific quantities, i.e., order parameters in statistical physics, and enables us to quantitatively predict parameter regions that can avoid vanishing or explosive gradients. This theory is applicable to a wide class of DNNs with various non-linear activation functions and depths. Such DNNs with random weights are substantially connected to Gaussian process and kernel methods \[25\]-\[28\]. Furthermore, the theory of the neural tangent kernel (NTK) explains that even trained parameters are close enough to the random initialization in sufficiently wide DNNs and the performance of trained DNNs is determined by the NTK on the initialization \[28\]-\[30\].

Karakida et al. \[31\] recently focused on FIM corresponding to the least square loss and proposed a framework to express certain eigenvalue statistic by using order parameters. They revealed that when conventional fully-connected networks with random initialization are sufficiently wide, the FIM’s eigenvalue spectrum asymptotically becomes pathologically distorted. As the network width increases, most of the eigenvalues become asymptotically close to zero while a small number of them take on huge values and become outliers. The distorted shape of the eigenvalue spectrum is consistent with empirical reports \[9\]-\[10\]. While LeCun et al. \[32\] implied that such pathologically large eigenvalue might appear in multi-layered networks and affect the training dynamics, its theoretical elucidation has been limited to a data covariance matrix in a linear regression model. The results of \[31\] verify the large eigenvalues suggested in \[32\] and enables us to quantify them in wide and multi-layered networks. The obtained eigenvalue statistics are crucial in practice. As we make the network wider, the largest eigenvalue becomes larger and we have to make the learning rate smaller for the gradient methods to converge \[9\]-\[31\]. Using statistics obtained in \[31\], Sun and Nielsen \[33\] investigated a new formulation of the minimum description length in DNNs that showed improved generalization performance.

In this paper, we extend the framework of the previous work \[31\] and reveal that various FIMs and variants show pathological spectra. Our contributions are summarized as follows:

- **FIM with soft-max outputs**: While the FIM analyzed in the previous work \[31\] corresponds to a squared error loss for regression tasks, the typical loss function used in classification tasks is the cross-entropy loss with soft-max outputs. We analyze this FIM for classification tasks and reveal that its spectrum is pathological. There are at least \( C \) dominant eigenvalues, which is consistent with a recent experimental report \[11\].

- **Diagonal Blocks of FIM**: We give a detailed analysis of the diagonal block parts of the FIM for regression tasks. Natural gradient algorithms often use a block diagonal approximation of the FIM \[34\]. We show that the diagonal blocks also suffer from pathological spectra.

- **Connection to NTK**: The NTK and FIM inherently share the same non-zero eigenvalues. Paying attention to a specific re-scaling of the parameters assumed in studies of NTK, we reveal that NTK’s eigenvalue statistics become independent of the width scale. Instead, the gap between the average and maximum eigenvalues increases with the sample size. This suggests that, as the sample size increases, the training dynamics converge non-uniformly and that calculations with the NTK become ill-conditioned. We also demonstrate a simple method to make eigenvalue statistics that are independent of both the width and the sample size.

- **Metric tensors for input and feature spaces**: We consider metric tensors for input and feature spaces spanned by neurons in input and hidden layers. These metric tensors potentially enable us to quantitatively evaluate the robustness of DNNs against perturbations in the input and feedforward propagated signals. We show that these metric tensors have the averages of eigenvalues that are asymptotically close to zero while their largest eigenvalues maintain constant values. In the sense that the outlier of the spectrum is much far from most of the eigenvalues, the spectrum is pathologically distorted, similar to FIMs.
In summary, this study presents a unified perspective on the asymptotical spectra common to various wide networks.

2 Preliminaries

2.1 Model

We investigated the fully-connected feedforward neural network shown in Fig. 1. The network consists of one input layer, \( L - 1 \) hidden layers \((l = 1, \ldots, L - 1)\), and one output layer. It includes shallow nets \((L = 2)\) and arbitrary deep nets \((L \geq 3)\). The network width is denoted by \(M_l\). The pre-activations \(u^l_i\) and activations of units \(h^l_i\) in the \(l\)-th layer are defined recursively in terms of the activations \(h^{l-1}_j\) of the previous layer:

\[
u^l_i = \sum_{j=1}^{M_{l-1}} W^l_{ij} h^{l-1}_j + b^l_i, \quad h^l_i = \phi(u^l_i),
\]

for \(l = 1, \ldots, L\), which will be explained in the following. The input signals are \(x_i = h^0_i\), which propagate layer by layer by Eq. (1). We define the weight matrices as \(W^l_{ij} \in \mathbb{R}^{M_l \times M_{l-1}}\) and the bias terms as \(b^l_i \in \mathbb{R}^{M_l}\). We will mainly focus on the case in which the activation function in the \(L\)-th layer (network output) is linear, i.e.,

\[
f_i := h^L_i = u^L_i.
\]

The softmax output is also discussed in Section 3.2.

FIM computations require the chain rule of backpropagated signals \(\delta^l_{k,i} := \partial f_k / \partial u^l_i\) and naturally appear in the derivatives of \(f_k\) with respect to the parameters:

\[
\frac{\partial f_k}{\partial W^l_{ij}} = \delta^l_{k,i} h^{l-1}_j, \quad \frac{\partial f_k}{\partial b^l_i} = \delta^l_{k,i},
\]

\[
\delta^l_{k,j} = \phi'(u^l_i) \sum_j \delta^{l+1}_{k,j} W^{l+1}_{ji},
\]

for \(l = 1, \ldots, L\). To avoid complicating the notation, we will omit the index \(k\) of the output unit, i.e., \(\delta^l_i = \delta^l_{k,i}\). To evaluate the above feedforward and backward signals, we assume the following conditions.

Random weights and biases: Suppose that the parameter set \(\{W^l_{ij}, b^l_i\}\) is an ensemble generated by

\[
W^l_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_\text{ws}/M_{l-1}), \quad b^l_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_\text{bi}),
\]

and thus is fixed, where \(\mathcal{N}(0, \sigma^2)\) denotes a Gaussian distribution with zero mean and variance \(\sigma^2\). Treating the case in which different layers have different variances is straightforward. Note that the variances of the weights are scaled in the order of \(1/M\). In practice, the learning of DNNs usually starts from random initialization with this scaling \cite{35,36}. Regarding the network width, we set

\[
M_l = \alpha_l M \quad (0 \leq l \leq L - 1), \quad M_L = C,
\]

and consider the limiting case of a sufficiently large \(M\) with constant coefficients \(\alpha_l > 0\). The number of output units is taken to be a constant \(C\), as is usually done in practice.

Input samples: We assume that there are \(T\) input samples \(x(t) \in \mathbb{R}^{M_0}\) \((t = 1, \ldots, T)\) generated identically and independently from the input distribution. We generate the samples by using a standard normal distribution, i.e.,

\[
x_j(t) \overset{i.i.d.}{\sim} \mathcal{N}(0, 1).
\]

Activation functions: Suppose the following two conditions: (i) the activation function \(\phi(x)\) has a polynomially bounded weak derivative. (ii) the network is non-centered, which means a DNN with bias terms \((\sigma_b \neq 0)\) or activation functions with a non-zero Gaussian mean. The definition of the non-zero Gaussian mean is \(\int Dz \phi(z) \neq 0\).
Condition (i) is used to obtain recurrence relations of backward order parameters \[21\]. Condition (ii) makes it easy to evaluate the FIM for regression tasks \[31, 37\]. The two conditions are valid in various realistic settings, because conventional networks include bias terms, and widely used activation functions, such as the sigmoid function and (leaky-) ReLUs, have bounded weak derivatives and non-zero Gaussian means. Different layers may have different activation functions.

2.2 Overview of metric tensors

We will analyze two types of metric tensors (metric matrices) that determine the responses of network outputs, i.e., the response to a local change in parameters and the response to a local change in the input and hidden neurons. One can systematically understand these tensors from the perspective of perturbations of variables.

We denote the set of network parameters as \( \theta := \{W^l_{ij}, b^l_i\} \) and its dimension as \( P \). Next suppose we choose one network output unit \( k \). If \( f_k \) is perturbed by an infinitesimal change \( d\theta \in \mathbb{R}^P \), its change is given by a quadratic form after performing a Taylor expansion, i.e.,

\[
E \left[ |f_k(x; \theta + d\theta) - f_k(x; \theta)|^2 \right] \sim d\theta^\top F_k d\theta,
\]

(7)

where \( \nabla_\theta \) is the derivative with respect to \( \theta \) and \( E[\cdot] \) denotes the expectation over an input distribution. The matrix \( F_k \) acts as a metric tensor for the parameter space. \( F_k \)'s eigenvalues determine the robustness of the network output \( f_k \) against the perturbation. As will be explained in Section 3.1, \( F_k = \sum_{k=1}^C F_k \)

has a special meaning because it is the Fisher information matrix (FIM).

Let us take a closer look at the structure of \( F_k \) by using block matrices. \( F_k \) can be partitioned into \( L^2 \) block matrices. We denote the \((l, l')\)-th block as \( F_{k}^{ll'} \) \((l, l' = 1, ..., L)\), whose weight part is given by \((F_{k}^{ll'})_{(ij)(i'j')} := E \left[ \nabla W_{ij} f_k \nabla W_{i'j'} f_k \right] \). One can represent it in matrix form:

\[
F_{k}^{ll'} = E[\delta_k^l (\delta_k^{l'})^\top \otimes h^{l-1}(h^{l-1})^\top],
\]

(10)

where \( \otimes \) represents the Kronecker product. The variables \( h \) and \( \delta \) are functions of \( x \), and the expectation is taken over \( x \).

In analogy with the FIM, one can introduce a metric tensor that measures the response to a change in the neural activities. Make a vector of all the activations in the input and hidden layers, i.e.,
We refer to $A_k$ as the **metric tensor for the input and feature spaces** because each $h^l$ acts as the input to the next layer and corresponds to the features realized in the network. We take the average over $T$ input samples; this operation includes the trivial case of one sample, i.e., $T = 1$.

One can partition $A_k$ into $L^2$ block matrices whose $(l, l')$-th block is expressed by an $M_l \times M_{l'}$ matrix:

$$A_k^{ll'} := (W^{ll'+1})^\top E[d_k^{l'}(\theta_{k}^{l'/1})^\top] W^{l'+1},$$

for $l, l' = 0, \ldots, L - 1$. In particular, the first diagonal block $A_k^{00}$ indicates the robustness of the network output against perturbation of the input:

$$E[|f_k(x + dx; \theta) - f_k(x; \theta)|^2] \sim dx^\top A_k^{00} dx.$$  

Similar (but different) quantities have been investigated in terms of the robustness against input noise, such as sensitivity [38], and robustness against adversarial examples [39].

### 2.3 Order parameters in mean field theory

We use the following four types of **order parameter**, i.e., $(\hat{q}_t^l, \hat{q}_st^l, \tilde{q}_t^l, \tilde{q}_st^l)$, which were used in various studies on wide DNNs [15, 17, 20, 26]. First, let us define the following variables for feedforward signal propagation:

$$\hat{q}_t^l := \frac{1}{M_l} \sum_{i=1}^{M_l} h_i^l(t)^2, \quad \hat{q}_st^l := \frac{1}{M_l} \sum_{i=1}^{M_l} h_i^l(s) h_i^l(t),$$

where $h_i^l(t)$ is the output of the $l$-th layer generated by the $t$-th input sample $x(t)$. The variable $\hat{q}_t^l$ describes the total activity in the $l$-th layer, and the variable $\hat{q}_st^l$ describes the overlap between the activities for different input samples $x(s)$ and $x(t)$. These variables have been utilized to describe the depth to which signals can propagate from the perspective of order-to-chaos phase transitions [17]. In the large $M$ limit, these variables can be recursively computed by integration over Gaussian distributions [15, 17]:

$$\hat{q}_t^{l+1} = \int Du \phi^2 \left( \sqrt{\hat{q}_t^{l+1}} u \right), \quad \hat{q}_st^{l+1} = I_\phi[\hat{q}_t^{l+1}, \hat{q}_st^{l+1}],$$

$$\tilde{q}_t^{l+1} := \sigma_w^2 \sigma_b^2, \quad \tilde{q}_st^{l+1} := \sigma_w^2 \sigma_b^2,$$

for $l = 0, \ldots, L - 1$. Because the input samples generated by Eq. (6) yield $\hat{q}_t^0 = 1$ and $\hat{q}_st^0 = 0$ for all $s$ and $t$, $\hat{q}_st^l$ in each layer takes the same value for all $s \neq t$; so does $\tilde{q}_t^l$ for all $t$. The notation $Du = du \exp(-u^2/2)/\sqrt{2\pi}$ means integration over the standard Gaussian density. A two-dimensional Gaussian integral is given by

$$I_\phi[a, b] := \int Dy Dx \phi(\sqrt{a}x) \phi(\sqrt{a}(cx + \sqrt{1-c^2 y}))$$

with $c = b/a$. One can represent this integral in a bit simpler form, i.e., $I_\phi[a, b] = \int Dy (\int Dx \phi(\sqrt{a} - bx + \sqrt{b} y))^2$.

Next, let us define the following variables for backpropagated signals:

$$\hat{q}_t^l := \sum_{i=1}^{M_l} \delta_i^l(t)^2, \quad \hat{q}_st^l := \sum_{i=1}^{M_l} \delta_i^l(s) \delta_i^l(t).$$

Above, we omitted $k$, the index of the output $f_k$, because the symmetry in the layer makes the above variables independent of $k$ in the large $M$ limit. The backward variables are defined by the summations, while the feedforward ones in [15] are defined by the averages. Because we suppose
\( C = O(1) \), each \( \delta_i^l \) is of \( O(1/\sqrt{M}) \) and their sums are of \( O(1) \) in terms of the order notation \( O(\cdot) \). The variable \( q_i^l \) is the magnitude of the backward signals and \( \bar{q}_i^l \) is their overlap. Previous studies found that \( q_i^l \) and \( \bar{q}_i^l \), in the large \( M \) limit are easily computed using the following recurrence relations \([18, 21]\).

\[
\bar{q}_i^l = \sigma_w^2 \bar{q}_i^{l+1} \int Du \left[ \phi'(\sqrt{q_i^l} u) \right]^2, \quad q_i^l = \sigma_w^2 \bar{q}_i^{l+1} I_{\phi'[q_i^l, \bar{q}_i^l]},
\]

for \( l = 0, \ldots, L - 1 \). A linear network output \([2]\) leads to the following initialization of the recurrences: \( q_i^0 = \bar{q}_i^0 = 1 \). The previous studies showed excellent agreement between these backward order parameters and experimental results \([18, 20]\). Although those studies required the so-called \textit{gradient independence assumption} to derive these recurrences, Yang \([21]\) recently proved that such an assumption is unnecessary when condition (i) of the activation function is satisfied.

The order parameters \((q_i^l, \bar{q}_i^l, \bar{q}_i^l, \bar{q}_i^l)\) depend only on the type of activation function, depth, and the variance parameters \( \sigma_w^2 \) and \( \sigma_b^2 \). The recurrence relations for the order parameters require \( L \) iterations of one- and two-dimensional numerical integrals. Moreover, we can obtain explicit forms of the recurrence relations for some of the activation functions \([31]\).

### 3 Eigenvalue statistics of FIMs

This section shows the asymptotic eigenvalue statistics of the FIMs. When we have an \( P \times P \) metric tensor whose eigenvalues are \( \lambda_i \ (i = 1, \ldots, P) \), we compute the following quantities:

\[
m_\lambda := \frac{1}{P} \sum_{i=1}^{P} \lambda_i, \quad s_\lambda := \frac{1}{P} \sum_{i=1}^{P} \lambda_i^2, \quad \lambda_{\text{max}} := \max_i \lambda_i.
\]

The obtained results are universal for any sample size \( T \) and network ranging in size from shallow \((L = 2)\) to arbitrarily deep \((L \geq 3)\).

### 3.1 FIM for regression tasks

This subsection overviews the results obtained in the previous studies \([31, 57]\). The metric tensor \( F \) is equivalent to the Fisher information matrix (FIM) of neural network models \([3, 7]\), originally defined by

\[
F := \mathbb{E} \left[ \nabla_\theta \log p(x; y; \theta) \nabla_\theta \log p(x; y; \theta)^\top \right].
\]

The statistical model is given by \( p(x; y; \theta) = p(y|x; \theta) q(x) \), where \( p(y|x; \theta) \) is the conditional probability distribution of the DNN of output \( y \) given input \( x \), and \( q(x) \) is an input distribution. The expectation \( \mathbb{E}[\cdot] \) is taken over the input-output pairs \((x, y)\) of the joint distribution \( p(x; y; \theta) \). This FIM appears in the Kullback-Leibler divergence between a statistical model and an infinitesimal change to it: \( \text{KL}[p(x, y; \theta) : p(x, y; \theta + d\theta)] \sim d\theta^\top F d\theta \). The parameter space \( \theta \) forms a Riemannian manifold and the FIM acts as its Riemannian metric tensor \([2]\).

Basically, there are two types of FIM for supervised learning, depending on the definition of the statistical model. One type corresponds to the squared error loss for regression tasks; the other corresponds to the cross-entropy loss for classification tasks. The latter is discussed in Section \([3.3]\).

Let us consider the following statistical model for the regression task:

\[
p(y|x; \theta) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} ||y - f(x; \theta)||^2 \right),
\]

where we denote the Euclidean norm as \( || \cdot || \). The squared error loss is given by the log-likelihood of this model. Substituting \( p(y|x; \theta) \) into the original definition of FIM \([20]\) and taking the integral over \( y \), one can easily confirm that it is equivalent to the metric tensor \([9]\) introduced by the perturbation.

When \( T \) input samples \( x(t) \ (t = 1, \ldots, T) \) are available, we can replace the expectation \( \mathbb{E}[\cdot] \) of the FIM with the empirical mean:

\[
F = \mathbb{E} \left[ \nabla_\theta f_k \nabla_\theta f_k^\top \right] = \frac{1}{T} \sum_{t=1}^{T} \nabla_\theta f_k(t) \nabla_\theta f_k(t)^\top,
\]

\( \Theta = 0 \)
Figure 2: Matrix representations of metric tensors. (a) Metric for parameter space, also known as (empirical) Fisher information matrix (FIM). In particular, $Q = I$ corresponds to the FIM for squared error loss. (b) Dual of FIM. Under a specific parameter transformation, this is equivalent to the neural tangent kernel. (c) Metric for input and feature spaces. Note that the figures omit the scalar factors of the metrics.

where we have abbreviated the network outputs as $f_k(t) = f_k(x(t); \theta)$ to avoid complicating the notation. This is an empirical FIM in the sense that the average is computed over empirical input samples. We can express it in the matrix form shown in Fig. 2. Let us investigate this type of empirical metric tensor for arbitrary $T$. One can set $T$ as a constant value or make it increase depending on $M$. The empirical FIM (22) converges to the expected FIM as $T \rightarrow \infty$. Note that in the context of natural gradient algorithms, one may approximate the FIM (20) by taking an average over empirical input-output pairs $(x(t), y(t))$ where $y(t)$ is a training label. Recently, Kunstner et al. [40] emphasized that in natural gradient algorithms, the FIM (22) performs better than that of the empirical pairs $(x(t), y(t))$.

The previous studies [31, 37] uncovered the following eigenvalue statistics of the FIM (22):

**Theorem 3.1** ([31], [37]). When $M$ is sufficiently large, the eigenvalue statistics of $F$ can be asymptotically evaluated as

$$m_\lambda \sim \kappa_1 \frac{C}{M}, \quad s_\lambda \sim \alpha \left( \frac{T - 1}{T} \kappa_2^2 + \frac{\kappa_1^2}{T} \right),$$

$$\lambda_{\max} \sim \alpha \left( \frac{T - 1}{T} \kappa_2 + \frac{\kappa_1}{T} \right),$$

where $\alpha := \sum_{l=1}^{L-1} \alpha_l \alpha_{l-1}$, and positive constants $\kappa_1$ and $\kappa_2$ are obtained using order parameters,

$$\kappa_1 := \sum_{l=1}^{L} \frac{\alpha_{l-1}}{\alpha} q_l^1 q_{l-1}^1, \quad \kappa_2 := \sum_{l=1}^{L} \frac{\alpha_{l-1} \alpha^{1/2}}{\alpha} q_{l+1}^1 q_{l-1}^1.$$

The eigenspace corresponding to $C$ largest eigenvalues is spanned by $C$ eigenvectors,

$$E[\nabla_\theta f_k] \quad (k = 1, …, C).$$

The mean of the eigenvalue spectrum asymptotically decreases in the order of $1/M$ in the large $M$ limit, while the variance takes a value of $O(1)$ and the largest eigenvalue takes a huge value of $O(M)$. Note that $\kappa_1$ is positive by definition and $\kappa_2$ is positive under the condition (ii) of activation functions.
Theorem 1 has the following implication. Since the eigenvalues are non-negative by definition, the obtained statistics means that most of the eigenvalues are asymptotically close to zero, while the other eigenvalues are widely distributed; this behavior has been empirically known for decades \[9][11]. Thus, when the network is sufficiently wide, one can see that the shape of the eigenvalue spectrum asymptotically becomes pathologically distorted. This implies that the parameter space of the DNNs is locally almost flat in most directions but highly distorted in a few specific directions. The following remarks are helpful for understanding further implications of the theorem.

**Remark 1: Dependence on the Depth.** As the depth \(L\) increases, \(\lambda_{\text{max}}\) linearly increases in the sense that it is proportional to the sum over \(L\) terms. As previous studies have reported \[17][18], a large \(L\) limit causes a qualitative change in the network state, known as a phase transition. Schoenholz et al. \[18\] recommends setting \((\sigma_w^2, \sigma_b^2)\) on the critical line of the phase transition. In such a case, \(\tilde{q}_l^l\) and \(\tilde{q}_{st}^l\) take finite values and \(\lambda_{\text{max}}\) scales linearly to the depth. In contrast, the means of the eigenvalues remain unchanged. Therefore, deeper networks have more distorted parameter spaces.

**Remark 2: Loss landscape and gradient methods.** The empirical FIM \((22)\) is equivalent to the Hessian of the loss around the global minimum with zero training loss \[31\]. This means that the local shape of the loss landscape is also asymptotically almost flat in most directions but very steep in a few specific directions. Karakida et al. \[37\] referred to the steep shape caused by \(\lambda_{\text{max}}\) as pathological sharpness. The sharpness of the loss landscape is connected to the learning rate of gradient methods for convergence. Experiments conducted by \[31\] confirmed that a learning rate \(\eta\) satisfying \(\eta < 2/\lambda_{\text{max}}\) is necessary for the steepest gradient method to converge. Because \(\lambda_{\text{max}}\) diverges as the width increases, we need to carefully choose an appropriately scaled learning rate to train the DNNs. Furthermore, because \(\lambda_{\text{max}}\) increases as the depth increases, a deeper network has a steeper loss landscape around the minimum, which requires a smaller learning rate.

**Remark 3: C largest eigenvalues.** The eigenspace corresponding to \(\lambda_{\text{max}}\) has the dimension of \(C\). \[37\] Fig. 3 (left) shows a typical spectrum of the FIM. We computed the eigenvalues of the FIM by using random Gaussian weights, biases, and inputs. We used deep Tanh networks with \(L = 3\), \(M = 200\), \(C = 10\), \(\alpha_l = 1\) and \((\sigma_w^2, \sigma_b^2) = (3, 0.64)\). Number of input samples was \(T = 100\). The red histogram was made from eigenvalues over 100 different networks with different random seeds. The histogram had two populations. The blue dashed histogram was made by eliminating the largest \(C\) eigenvalues. It coincides with the smaller population. Thus, one can see that the larger population corresponds to the \(C\) largest eigenvalues. The larger population in experiments can be distributed around \(\lambda_{\text{max}}\) because of finite \(M\).

The eigenvalue statistics of the smaller population were investigated in the previous work \[37\] in the context of batch normalization in the last layer. Such normalization includes mean subtraction, i.e., \(f_k := f_k - E[f_k]\). The previous work analyzed the corresponding FIM;

\[
\bar{F} := \sum_k E[\nabla_\theta f_k \nabla_\theta f_k^\top] = \sum_k E[\nabla_\theta f_k \nabla_\theta f_k^\top] - \sum_k E[\nabla_\theta f_k] E[\nabla_\theta f_k]^\top. \tag{23}
\]

The subtraction \((23)\) means eliminating the \(C\) largest eigenvalues from \(F\) because we asymptotically have \(\|E[\nabla_\theta f_k]\|^2 \sim \lambda_{\text{max}}\). Thus, the \(\bar{F}\)'s eigenvalues correspond to the smaller population in the figure. The previous work \[37\] theoretically confirmed that, under the condition \(T = O(M)\), the mean of \(\bar{F}\)'s eigenvalues is of \(O(1/M)\) and constructed the lower and upper bounds of the largest eigenvalue. Numerical experiments confirmed that the largest eigenvalue of \(\bar{F}\) is of \(O(1)\). Note that when \(T = O(M)\), the sample size is sufficiently large but the network satisfies \(P \gg T\) and keeps overparameterized.

### 3.2 Diagonal blocks of FIM

In the same way as Theorem 3.1, one can easily obtain eigenvalue statistics for diagonal blocks, that is, \(F^{ii}\).
Theorem 3.2. When $M$ is sufficiently large, the eigenvalue statistics of $F^{ll}$ are asymptotically evaluated as

$$m^l_{\lambda} \sim \frac{\bar{q}_l q_l^{-1}}{\alpha_l} C M,$$

$$s^l_{\lambda} \sim \frac{\alpha_l - 1}{\alpha_l} \left( \frac{T - 1}{T} (\bar{q}_{ll} q_{ll}^{-1})^2 + \frac{(\bar{q}_l q_l^{-1})^2}{T} \right) C,$$

$$\lambda^l_{\max} \sim \alpha_l - 1 \left( \frac{T - 1}{T} \bar{q}_{ll} q_{ll}^{-1} + \frac{\bar{q}_l q_l^{-1}}{T} \right) M.$$

The eigenspace of $F^{ll}$ corresponding to the $C$ largest eigenvalues is spanned by $C$ eigenvectors,

$$E[\nabla_{\theta} f_k] \ (k = 1, ..., C).$$

The theorem is proved in Appendix A. We have a mean of $O(1/M)$ and second moment of $O(1)$ in each hidden layer. The largest eigenvalue is of $O(M)$. Regarding the pathological largest eigenvalues, we have, asymptotically,

$$\lambda_{\max} = \sum_{i=1}^{L} \lambda^l_{\max}. \tag{24}$$

Fig. 3 (right) empirically confirms that $F^{22}$ has a similar pathological spectrum to that of $F$. Its experimental setting was the same as in the case of $F$.

It is helpful to investigate the relation between $F$ and its diagonal blocks when one considers the diagonal block approximation of $F$. Use of a diagonal block approximation can decrease the computational cost of natural gradient algorithms [6, 34]. When a matrix is composed only of diagonal blocks, its eigenvalues are given by those of each diagonal block. Therefore, $F$ approximated in this fashion has the same mean of the eigenvalues as the original $F$ and the largest eigenvalue $\lambda^l_{\max}$, which is of $O(M)$. Thus, the diagonal block approximation also suffers from a pathological spectrum. Eigenvalues that are close to zero can make the inversion of the FIM in the natural gradient methods unstable, whereas using a damping term seems to be an effective way of dealing with this instability [7].

3.3 FIM for multi-label classification tasks

The cross-entropy loss is typically used in multi-label classification tasks. We define the $C$-dimensional softmax function by

$$g_k(t) := \frac{\exp(f_k(t))}{\sum_k \exp(f_k(t))}, \tag{25}$$

Figure 3: Eigenvalue spectra of FIMs in experiments with deep Tanh networks: (left) case of $F$, (right) case of diagonal block $F^{22}$. The vertical axis is the cumulative number of eigenvalues over 100 different networks. The black histograms show the original spectra, while the red dashed ones show the spectra without the $C$ largest eigenvalues. The blue lines represent the theoretical values of the largest eigenvalues.
for $i = 1, \ldots, C$. The cross-entropy loss comes from the log-likelihood of the following statistical model:

$$p(y|x; \theta) = \prod_i^C g_i(t)^{y_i}, \quad (26)$$

where $y$ is a $C$-dimensional one-hot vector. The cross-entropy loss is given by $-\mathbb{E}[\sum_i g_i(t) \log g_i(t)]$. Substituting the statistical model into the definition of the FIM [20] and taking the summation over $y$, we find that

$$F_{\text{cross}} = \frac{1}{T} \sum_{s,t=1}^C Q_{st}(k,k') \nabla_\theta f_t(s) \nabla_\theta f_t(k')^T. \quad (27)$$

$$Q_{st}(k,k') := \{g_k(t)\delta_{kk'} - g_k(t)g_{k'}(t)\} \delta_{st}. \quad (28)$$

This $F_{\text{cross}}$ is also derived in [4, 5]. One can also characterize the $F_{\text{cross}}$ by the robustness of the softmax function against the perturbation,

$$\mathbb{E} \left[ ||g(x; \theta + \delta \theta) - g(x; \theta)||^2 \right] \sim \delta \theta^T F_{\text{cross}} \delta \theta. \quad (29)$$

$F_{\text{cross}}$ is linked to $F$ through the matrix representation shown in Fig. 2 (a). One can view $F$ as a matrix representation with $Q = I$, that is, the identity matrix. In contrast, $F_{\text{cross}}$ corresponds to the $Q$ defined in [28]. This matrix representation is useful for deriving the eigenvalue statistics (see Appendix B) and the following theorem:

**Theorem 3.3.** When $M$ is sufficiently large, the eigenvalue statistics of $F_{\text{cross}}$ are asymptotically evaluated as

$$m_\lambda \sim \beta_1 C \frac{k_1}{M}, \quad s_\lambda \sim \alpha \left( \beta_2 \kappa_2^2 + \beta_3 \frac{k_1^2}{T} \right),$$

$$\beta_4 \alpha \left( \frac{T - 1}{T} \kappa_2 + \frac{k_1}{T} \right) M \leq \lambda_{\text{max}} \leq \sqrt{\alpha s_\lambda M},$$

where the coefficients are given by

$$\beta_1 := 1 - \frac{1}{T} \sum_i^C g_i(t)^2,$$

$$\beta_2 := \frac{1}{T^2} \left\{ \sum_{s \neq t}^C g_i(t)g_i(s) - 2 \sum_i^C g_i(t)^2 g_i(s) + \left( \sum_i^C g_i(t)g_i(s) \right)^2 \right\},$$

$$\beta_3 := \frac{1}{T} \left\{ \sum_{i}^C (1 - 2g_i(t))g_i(t)^2 + \left( \sum_i^C g_i(t)^2 \right)^2 \right\},$$

$$\beta_4 := \max_{1 \leq k \leq C} \frac{1}{T} \sum_{t}^L g_k(t)(1 - g_k(t)).$$

We find that the eigenvalue spectrum shows the same width dependence as the FIM for regression tasks. Although the evaluation of $\lambda_{\text{max}}$ in Theorem 3.3 is based on inequalities, one can see that $\lambda_{\text{max}}$ is of $O(M)$ and it linearly increases as the depth $L$ increases. The soft-max functions appear in the coefficients $\beta_k$. It should be noted that the values of $\beta_k$ generally depend on the index $i$ of each soft-max output. This is because the values of the softmax functions depend on the specific configuration of $W^L$ and $b^L$.

Fig. 4 shows that our theory predicts experimental results rather well for artificial data. We computed the eigenvalues of $F_{\text{cross}}$ with random Gaussian weights, biases, and inputs. We set $L = 3, M = 100$, $C = 10, \alpha_l = 1$ and $(\sigma_{w_l}^2, \sigma_{b_l}^2) = (3, 0.64)$ in the tanh case, $(2, 0.1)$ in the ReLU case, and $(1, 0.1)$ in the linear case. The number of input samples was set to $T = 100$. The predictions of Theorem 3.3 coincided with the experimental results for sufficiently large widths.
Figure 4: $F_{\text{cross}}$’s eigenvalue statistics: means (left), second moments (center), and maximum (right). Black points and error bars show means and standard deviations of the experimental results over 100 different networks with different random seeds. The blue lines represent the theoretical results obtained in the large $M$ limit. For $\lambda_{\text{max}}$, the dashed lines show the theoretical upper bound, while the solid ones show the lower bound.

Exhaustive experiments on the cross entropy loss have recently confirmed that there are $C$ dominant large eigenvalues (so-called outliers) \[11\]. Consistent with the results of this experimental study, we found that there are at least $C$ eigenvalues depending on the width scale:

**Theorem 3.4.** $F_{\text{cross}}$ has the first $C$ largest eigenvalues of $O(M)$.

The theorem is proved in Appendix C. These $C$ large eigenvalues are reminiscent of the $C$ largest eigenvalues of $F$ shown in Theorem 3.1. Informally speaking, the $Q$ matrix in $F_{\text{cross}}$ scatters the $C$ largest eigenvalues of $F$. It would be interesting to extend the above theorem and theoretically quantify the $C$ dominant eigenvalues precisely.

4 Connection to Neural Tangent Kernel

4.1 Scale-dependent eigenvalue statistics

The empirical FIM (\[22\]) is essentially connected to a recently proposed Gram matrix, i.e., the Neural Tangent Kernel (NTK). Jacot et al. [28] formulated the NTK as

$$\Theta := \nabla_\theta f \nabla_\theta f^\top$$

with $\theta = \{\omega_{l,j}, \beta_l\}$.

\[\text{(30)}\]

where they assumed a special scaling,

$$W_{l,j} = \frac{\sigma_w}{\sqrt{M_{l-1}}} \omega_{l,j}^l, \quad b_l^l = \sigma_b \beta_l^l, \quad \omega_{l,j}^l, \beta_l^l \sim N(0, 1).$$

\[\text{(31)}\]

This scaling is called NTK parameterization [29]. Under certain conditions with sufficiently large $M$, the NTK is known to govern the training dynamics of the wide network:

$$\frac{df}{dt} = \frac{\eta}{T} \Theta(y - f),$$

\[\text{(32)}\]
Figure 5: Spectra of the NTK ($\Theta$) in experiments with deep ReLU networks: (left) Spectra with $M = 1000$ and various $T$. The eigenvalues are normalized by $1/T$ for comparison. (right) Spectra under mean subtraction in the last layer and the condition $T = M$. The vertical axes represent the probability density obtained from the cumulative number of eigenvalues over 400 different networks.

where the notation $t$ corresponds to the time step of the parameter update and $\eta$ represents the learning rate. Surprisingly, the NTK with random initialization, $\Theta$, determines the training process, and we can analytically solve for the dynamics of $f$ at $t$. Specifically, NTK’s eigenvalues determine the speed of convergence of the training dynamics. Moreover, one can predict the network output on the test samples by using the Gaussian process with the NTK [28, 29].

The NTK and empirical FIM share essentially the same eigenvalues. Consider the left-to-right reversal of $F$ denoted by $F^*$ in Fig. 2 (b). we call it as the dual of $F$. We have $F^* = \frac{1}{T} \Theta$. Karakida et al. [31] analyzed $F^*$ under the usual parameterization to derive Theorem 3.1 because $F$ and $F^*$ have the same non-zero eigenvalues. We can obtain the eigenvalue statistics of $\Theta$ by taking the NTK parameterization and leveraging the proof of Theorem 3.1.

**Theorem 4.1.** When $M$ is sufficiently large, the eigenvalue statistics of $\Theta$ are asymptotically evaluated as

$$m_\lambda \sim \alpha \kappa'_1 C, \quad s_\lambda \sim \alpha^2 (T - 1) \kappa'_2 + \kappa'_1 C,$$

$$\lambda_{max} \sim \alpha ((T - 1) \kappa'_2 + \kappa'_1).$$

The positive constants $\kappa_1$ and $\kappa_2$ are obtained using order parameters,

$$\kappa'_1 := \frac{1}{\alpha} \sum_{l=1}^{L} (\sigma^2 w_{l} q_{l}^{-1} + \sigma^2 b_{l} q_{l}),$$

$$\kappa'_2 := \frac{1}{\alpha} \sum_{l=1}^{L} (\sigma^2 w_{l} q_{l}^{2} q_{l}^{-1} + \sigma^2 b_{l} q_{l}).$$

The proof is given in Appendix D. The NTK parameterization makes the eigenvalue statistics independent of the width scale. This is because the NTK parameterization maintains the scale of the weights but changes the scale of the gradients with respect to the weights. It also makes $(\kappa_1, \kappa_2)$ shift to $(\kappa'_1, \kappa'_2)$. This shift occurs because the NTK parameterization makes the order of the weight gradients $\nabla_W f$ comparable to that of the bias gradients $\nabla_b f$. The second terms in $\kappa'_1$ and $\kappa'_2$ correspond to a non-negligible contribution from $\nabla_b f$. The coefficients $(\kappa'_1, \kappa'_2)$ are equivalent to $\Theta^\infty_{\infty}(x(s), x(t))$ in Theorem 1 of [28].

While $m_\lambda$ is independent of the sample size $T$, $\lambda_{max}$ depends on it. This means that the NTK dynamics converge non-uniformly. Most of the eigenvalues are relatively small and the NTK dynamics converge more slowly in the corresponding eigenspace. In addition, a prediction made with the NTK requires the inverse of the NTK to be computed [28, 29]. When the sample size is large, the condition number of the NTK, i.e., $\lambda_{max}/\lambda_{min}$, is also large and the computation with the inverse NTK is expected to be numerically inaccurate.
4.2 Scale-independent NTK

A natural question is under what condition do NTK’s eigenvalue statistics become independent of both the width and the sample size? As indicated in Eq. (23), the mean subtraction in the last layer with \( T = O(M) \) is a simple way to make the FIM’s largest eigenvalue independent of the width. Similarly, one can expect that the mean subtraction makes the NTK’s largest eigenvalue of \( O(T) \) disappear and the eigenvalue spectrum take a range of \( O(1) \) independent of the width and sample size. Fig. 5 empirically confirms this speculation. We set \( L = 3, C = 2, \alpha_l = 1 \) and used the Gaussian inputs and weights with \( (\sigma_w^2, \sigma_b^2) = (2, 0) \). As shown in Fig. 5 (left), NTK’s eigenvalue spectrum becomes pathologically distorted as the sample size increases. To make an easier comparison of the spectra, the eigenvalues in this figure are normalized by \( 1/T \). As the sample size increases, most of the eigenvalues concentrate close to zero while the largest eigenvalue becomes an outlier. In contrast, Fig. 5 (right) shows that the mean subtraction keeps NTK’s whole spectrum in the range of \( O(1) \) under the condition \( T = O(M) \). The spectrum empirically converged to a fixed distribution in the large \( M \) limit.

5 Metric tensor for input and feature spaces

The above framework for evaluating FIMs is also applicable to metric tensors for input and feature spaces, which are expressed in the matrix form in Fig. 2 (c). Here, we can prove the following theorem:

**Theorem 5.1.** When \( M \) is sufficiently large, the eigenvalue statistics of \( A_k \) are asymptotically evaluated as

\[
m_\lambda \sim \frac{\tilde{\kappa}_1}{M}, \quad s_\lambda \sim \tilde{\alpha} \left( \frac{T - 1}{T} \tilde{\kappa}_2 + \frac{\tilde{\kappa}_1^2}{T} \right),
\]

\[
\lambda_{\text{max}} \sim \tilde{\alpha} \left( \frac{T - 1}{T} \tilde{\kappa}_2 + \frac{\tilde{\kappa}_1}{T} \right),
\]

where \( \tilde{\alpha} := \sum_{l=0}^{L-1} \alpha_l \), and positive constants \( \tilde{\kappa}_1 \) and \( \tilde{\kappa}_2 \) are obtained from the order parameters,

\[
\tilde{\kappa}_1 := \frac{\sigma_w^2}{\tilde{\alpha}} \sum_{l=1}^{L} q_{ll}^l, \quad \tilde{\kappa}_2 := \frac{\sigma_w^2}{\tilde{\alpha}} \sum_{l=1}^{L} q_{st}^l.
\]

The eigenvector of \( A_k \) corresponding to \( \lambda_{\text{max}} \) is \( E[\nabla_h f_k] \).

The theorem is proved in Appendix E. The mean of the eigenvalues asymptotically decreases in the order of \( O(1/M) \), while the variance and largest eigenvalues are of \( O(1) \) for any \( T \). Thus, the spectrum of \( A_k \) is pathologically distorted in the sense that the mean is far from the edge beyond the order difference of \( M \). The local geometry of the whole input \( x \) and feature spaces \( h \) is distorted in the direction of \( E[\nabla_h f_k] \). As the depth increases, \( \lambda_{\text{max}} \) linearly increases while the mean remains unchanged.

In the same way as the FIM, we can also evaluate the eigenvalue statistics of \( A := \sum_k A_k \) (see Appendix E). Furthermore, one can obtain the eigenvalue statistics of the diagonal blocks \( A_k^{ll} \) as follows:

**Theorem 5.2.** When \( M \) is sufficiently large, the eigenvalue statistics of \( A_k^{ll} \) are asymptotically evaluated as

\[
m_\lambda \sim \sigma_w^2 \frac{q_{ll}^{l+1}}{M_l}, \quad s_\lambda \sim \sigma_w^4 \left( \frac{T - 1}{T} (q_{ll}^{l+1})^2 + \frac{(q_{st}^{l+1})^2}{T} \right),
\]

\[
\lambda_{\text{max}} \sim \sigma_w^2 \left( \frac{T - 1}{T} q_{ll}^{l+1} + \frac{q_{st}^{l+1}}{T} \right).
\]

The eigenvector of \( A_k^{ll} \) corresponding to \( \lambda_{\text{max}} \) is \( E[\nabla_h^l f_k] \).
Figure 6: Spectra of metric tensor for input and feature spaces in experiments with deep Tanh networks: (left) the spectra of $A$, (right) the spectra of $A^{00}$. The vertical axis shows the cumulative number of eigenvalues over 100 different networks. The black histograms show the original spectra, while the red dashed ones show the spectra without the $C$ largest eigenvalues. The blue lines represent the theoretical values of the largest eigenvalues.

The proof is given in Appendix F.

Fig. 6 (left) shows typical spectra of $A$ and Fig. 6 (right) those of $A^{00} := \sum_k^C A_{k}^{00}$. We used deep Tanh networks with $M = 500$ and $T = 1000$. The other experimental settings are the same as those in Fig. 3. The pathological spectra appear as the theory predicts. Note that the $\lambda_{max}$’s were distributed due to the finite value of $M$.

Let us remark on the related work in the literature of deep learning. First, Pennington et al. [22] investigated similar metric tensors. Briefly speaking, they used random matrix theory and obtained the eigenvalue spectrum of matrices satisfying $T = 1, M_l = C = M$, and $L \gg 1$. They found that the isometry of the spectrum is helpful to solve the vanishing gradient problem. Second, DNNs are known to be vulnerable to a specific noise perturbation, i.e., the adversarial example [39]. One can speculate that the eigenvector corresponding to $\lambda_{max}$ may be related to adversarial attacks, although such a conclusion will require careful considerations.

6 Discussion

We evaluated the asymptotic eigenvalue statistics of the FIM and its variants in sufficiently wide DNNs. They have pathological spectra in the conventional setting of random initialization and activation functions. This suggests that we need to be careful about the eigenvalue statistics and their influence on the learning when we use large-scale deep networks in naive settings.

It will be straightforward to prove that similar pathological spectra appear in various network architectures because order parameters have already been developed in ResNets [19] and CNNs [20], and we can use them. It is interesting to explore the eigenvalue statistics that the current study cannot capture. Although our study captured some of the basic eigenvalue statistics, it remains to derive the whole spectrum analytically. Random matrix theory enables us to analyze the FIM’s eigenvalue spectrum in a shallow and centered network without bias terms [12]. Extending the theory to more general DNNs seems to be a prerequisite for further progress. In addition, random matrix theory enables us to analyze the spectrum of a special type of Jacobian matrix for backpropagation [22]. The spectrum of the Jacobian matrix with random orthogonal weights differ from one with i.i.d. random Gaussian weights. It would be interesting to investigate the spectra of our metric tensors with other types of random weights, although it seems that applying random matrix theory to general DNNs will be a nontrivial exercise. Furthermore, we assumed a finite number of network output units. In order to deal with multi-label classifications with high dimensionality, it would be helpful to investigate eigenvalue statistics in the wide limit of both hidden and network output layers. Finally, although we focused on the finite depth and regarded order parameters as constants, they can exponentially explode on extremely deep networks in the chaotic regime [18, 41]. The NTK in such a regime has been investigated in [42].
It would also be interesting to explore further connections between the eigenvalue statistics and learning. Recent studies have yielded insights into the connection between the generalization performance of DNNs and the eigenvalues statistics of certain Gram matrices \cite{Nakamura2022, Diakonikolas2020}. The NTK’s eigenvalues affect the convergence of the training and performance of the prediction \cite{Du2018}. We expect that the theoretical foundation of the metric tensors given in this paper will lead to a more sophisticated understanding and development of deep learning in the future.

Appendices

A Derivation of Theorem 3.2

First, we briefly overview the derivation of Theorem 3.1 in \cite{Zaletel2021} and \cite{Nakamura2022}. The essential point is that a Gram matrix has the same non-zero eigenvalues as its dual. One can represent the empirical FIM (22)

\[ F = RR^T, \]

where

\[ R := \frac{1}{\sqrt{T}}[\nabla_{f_1} \nabla_{f_2} \cdots \nabla_{f_C}]. \]  

(A.1)

Its columns are the gradients on each input, i.e., \( \nabla_{f_k}(t) \) for \( t = 1, \ldots, T \). Let us refer to a \( CT \times CT \) matrix \( F^* := R^\dagger R \) as the dual of FIM. Matrices \( F \) and \( F^* \) have the same non-zero eigenvalues by definition. This \( F^* \) can be partitioned into \( T \times T \) block matrices. The \( (k, k') \)-th block is given by

\[ F^*(k, k') = \nabla_{f_k}^\dagger \nabla_{f_{k'}} / T, \] 

(A.2)

for \( k, k' = 1, \ldots, C \). In the large \( M \) limit, the previous study \cite{Zaletel2021} showed that \( F^* \) asymptotically satisfies

\[ F^*(k, k') = \frac{M}{T} K \delta_{kk'} + \frac{1}{T} o(M), \] 

(A.3)

where \( \delta_{kk'} \) is the Kronecker delta. As is summarized in Lemma A.1 in \cite{Nakamura2022}, the second term of Eq. (A.3) is negligible in the large \( M \) limit. In particular, it is reduced to \( O(\sqrt{M})/T \) under certain condition. The matrix \( K \) has entries given by

\[ K_{st} = \kappa_1 \quad (s = t), \quad \kappa_2 \quad (s \neq t). \] 

(A.4)

Using this \( K \), the previous studies derived the basic eigenvalues statistics \cite{Zaletel2021} and eigenvectors corresponding to \( \lambda_{\text{max}} \) \cite{Nakamura2022}. The matrix \( K \) has the largest eigenvalue \((T - 1)\kappa_2 + \kappa_1\) and its eigenvectors \( \nu_k \in \mathbb{R}^{CT} \) whose entries are given by

\[ (\nu_k)_i := \frac{1}{\sqrt{T}} ((k - 1)T + 1 \leq i \leq kT), \] 

(A.5)

\[ 0 \quad \text{(otherwise)}. \]

The other eigenvalues of \( K \) are given by \( \kappa_1 - \kappa_2 \). We can obtain \( m_\lambda \) from \( C\text{Trace}(F^*(k, k))/P \), \( s_\lambda \) from \( C ||F^*(k, k)||_F^2 / P \) where \( || \cdot ||_F \) is the Frobenius norm, and \( \lambda_{\text{max}} \) from \( \nu_k^\dagger F^*(k, k) \nu_k \). The eigenvector corresponding to \( \lambda_{\text{max}} \) is asymptotically given by \( E[f_k] = RV_k \). If \( T \) is a fixed constant, it is obvious that \( K \)'s eigenvalues determine \( F^* \)'s eigenvalues in the large \( M \) limit. Even if \( T \) increases depending on \( M \), the obtained statistics hold in the large \( M \) and \( T \) limits. That is, we have asymptotically \( m_\lambda \sim \kappa_1 C/M, \quad s_\lambda \sim \alpha \kappa_2^2 C, \) and \( \lambda_{\text{max}} \sim \alpha \kappa_2 M \). As one can see here, the condition of \( \kappa_2 > 0 \) is crucial for our eigenvalue statistics. Non-centered networks guarantee \( \tilde{q}_{1st} > 0 \) and \( \tilde{q}_{kst} > 0 \) which leads to \( \kappa_2 > 0 \). In centered networks, \( \kappa_2 \) can become zero and we need to carefully evaluate the second term of Eq. (A.3).

We can immediately derive Theorem 3.2 in the same way as Theorem 3.1. We can represent the diagonal blocks as \( F^\dagger := R^\dagger R^T \) with

\[ R^\dagger := \frac{1}{\sqrt{T}}[\nabla_{\hat{\omega}} f_1 \nabla_{\hat{\omega}} f_2 \cdots \nabla_{\hat{\omega}} f_C], \] 

(A.6)
and the dual of this Gram matrix as
\[ F^{ll*} := R_l^T R_l, \] (A.7)
where the parameter set \( \theta^l \) means all parameters in the \( l \)-th layer. The \( CT \times CT \) matrix \( F^{ll*} \) can be partitioned into \( T \times T \) block matrices whose \((k,k')\)-th block is given by
\[ F^{ll*}(k,k') = \nabla_{\theta^l} f_k \nabla_{\theta^l} f_{k'}/T, \] (A.8)
for \( k,k' = 1, \ldots, C \). As one can see from the additivity of \( \sum_{l=1}^L F^{ll*}(k,k') = F^*(k,k') \), the following evaluation is part of Eq. (A.3):
\[ F^{ll*}(k,k') = \alpha_{l-1} \frac{M}{T} K^l \delta_{kk'} + \frac{1}{T} o(M), \] (A.9)
where
\[ K^l_{st} := \frac{q_s q_t^{l-1}}{q_s q_t^{l-1}} \quad (s = t), \quad \frac{q_s q_t^{l-1}}{q_s q_t^{l-1}} \quad (s \neq t). \] (A.10)
Thus, we have \( m_\lambda \sim C \text{Trace}(\alpha_{l-1} M^2 K^l)/P_l \) and \( s_\lambda \sim C||\alpha_{l-1} M^2 K^l||^2/P_l \), where the dimension of \( \theta^l \) is given by \( P_l = \alpha L M^2 \). We set \( \alpha L = C/M \) in the last layer. The matrices \( K \) and \( K^l \) have the same eigenvectors corresponding to the largest eigenvalues, i.e., \( \nu_k \). The largest eigenvalue is \( \lambda_{\max} \sim \alpha_{l-1} M^2 \nu_k K^l \nu_k \). The eigenvectors of \( F^{ll*} \) corresponding to \( \lambda_{\max} \) are \( R \nu_k = E[\nabla_{\theta^l} f_k] \).

B Derivation of Theorem 3.3

\( F_{\text{cross}} \) is expressed by
\[ F_{\text{cross}} := RQ R^T, \] (B.1)
where \( Q \) is a \( CT \times CT \) matrix. This \( Q \) can be partitioned into \( T \times T \) block matrices \( Q(k,k') \) whose entries are given by
\[ Q(k,k')_{st} = \{ g_k(t) \delta_{kk'} - g_k(t) g_{k'}(t) \} \delta_{st}, \] (B.2)
for \( k,k' = 1, \ldots, C \). Each block is a diagonal matrix.

The non-zero eigenvalues of \( RQ R^T \) are equivalent to non-zero singular values of \( QR R^T \). Since we have \( F^* = R^T R \), we should investigate the eigenvalues of the following matrix:
\[ F_{\text{cross}}^* := QF^*. \] (B.3)

The mean of the eigenvalues is given by
\[ m_\lambda = \text{Trace}(F_{\text{cross}}^*)/P \]
\[ = \sum_{i,k} \text{Trace}(Q(k,i)F^*(i,k))/P \]
\[ \sim \sum_k \text{Trace}(Q(k,k)F^*(k,k))/P \]
\[ \sim C(1 - \beta_1)\kappa_1/M. \] (B.4)

The third line holds asymptotically, since the order of \( F(k,k) \) in Eq. (A.3) is higher than that of \( F^*(k,k') \) \((k \neq k')\). The fourth line comes from \( \sum_k g_k(t) = 1 \).

The second moment is evaluated as
\[ s_\lambda = \text{Trace}(F_{\text{cross}}^{*2})/P \]
\[ = \sum_k \text{Trace} \left( \sum_{a,b,c} Q(k,a)F^*(a,b)Q(b,c)F^*(c,k) \right)/P \]
\[ \sim \sum_{k,k'} \text{Trace} \left( Q(k,k')F^*(k',k)Q(k',k)F^*(k,k) \right)/P. \] (B.5)

Substituting \( K \) into \( F^*(k,k) \) gives
\[ s_\lambda \sim \frac{\alpha}{T^2} \sum_{k,k'} \sum_s \left( G_{kk'}(s)(\kappa_2^2 \sum_{t(s)} G_{kk'}(t) + \kappa_3^2 G_{kk'}(s)) \right) \]
\[ = \alpha \left( \beta_2 \kappa_2^2 + \beta_3 \frac{\kappa_3^2}{T} \right), \] (B.6)
where $\sum_{t \setminus \{s\}}$ means a summation over $t$ excluding the $s$-th sample and we have defined $G_{kk'}(s) := \delta_{kk'} g_k(s) - g_k(s) g_{k'}(s)$.

Finally, we derive the largest eigenvalue. Let us denote the eigenvectors of $F$ as $v_k := \frac{E[\nabla f_k]}{|E[\nabla f_k]|}$ (B.7).

It is easy to confirm that we have asymptotically $|E[\nabla f_k]|^2 \sim \lambda_{\text{max}}(F)$ (B.7), where the largest eigenvalue of $F$ is denoted as $\lambda_{\text{max}}(F) = \alpha (\frac{T-1}{T} \kappa_2 + \frac{2}{\sqrt{T}}) M$. By definition, $F_{\text{cross}}$’s largest eigenvalue satisfies $\lambda_{\text{max}} \geq x^T F_{\text{cross}} x$ for any unit vector $x$. By taking $x = v_k$, we obtain

$$\lambda_{\text{max}} \geq \lambda_{\text{max}}(F)^{-1} \cdot (Rv_k)^\top F_{\text{cross}} (Rv_k)$$
$$= \lambda_{\text{max}}(F)^{-1} \cdot (F^* v_k)^\top Q (F^* v_k).$$ (B.8)

Because we have asymptotically $F^* v_k = \lambda_{\text{max}}(F) v_k$, the lower bound is given by

$$\lambda_{\text{max}}(F) \geq \lambda_{\text{max}}(F) \cdot (v_k^\top Q v_k)$$
$$= \lambda_{\text{max}}(F) \cdot \frac{1}{T} \sum_i g_i(t) (1 - g_i(t)).$$ (B.9)

Taking the index $k$ that maximizes the right-hand side, we obtain the lower bound of $\lambda_{\text{max}}$.

The upper bound of $\lambda_{\text{max}}$ immediately comes from a simple inequality for non-negative variables, i.e., $\lambda_{\text{max}} \leq \sqrt{\sum_i \lambda_i^2} = \sqrt{F_{\text{ss}}}$.

**C Derivation of Theorem 3.4**

Define $u_i$ to be the eigenvector of $F_{\text{cross}}$ corresponding to the eigenvalue $\lambda_i (\lambda_1 \geq \cdots \geq \lambda_i \geq \cdots \geq \lambda_P)$. Moreover, let us denote the linear subspace spanned by $\{u_1, \ldots, u_k\}$ as $U_k$ and the orthogonal complement of $U_k$ as $U_k^\perp$. When $k = 0$, we have $U_k^\perp = \mathbb{R}^P$. The dimension of $U_k^\perp$ is $P - k$, and we denote it as $\text{dim}(U_k^\perp) = P - k$. Thus, we have

$$\lambda_r = \max_{\|x\| = 1; x \in U_{r-1}^\perp} x^T F_{\text{cross}} x,$$ (C.1)

for $r = 1, \ldots, P$. Define $V_k$ to be a linear subspace spanned by $k$ eigenvectors of $F$ corresponding to $\lambda_{\text{max}}(F)$, i.e., $\{v_{i_1}, \ldots, v_k\}$. The indices $\{i_1, \ldots, i_k\}$ are chosen from $\{1, \ldots, C\}$ without duplication.

It is trivial to show from the dimensionality of the linear space that the intersection $S_r := \{U_{r-1}^\perp \cap V_C\}$ is a linear subspace satisfying $C - r + 1 \leq \text{dim}(S_r) \leq C$ when $1 \leq r \leq C$. Let us take a unit vector $x$ in $S_r$ as $x = \sum_{i=1}^r a_i v_{i_1}$, where we have defined $r^* := \text{dim}(S_r)$ and the coefficients $a_s$ satisfy $\sum_{s=1}^{r^*} a_s^2 = 1$. In the large $M$ limit, we asymptotically have

$$\lambda_r \geq \max_{\|x\| = 1; x \in S_r} x^T F_{\text{cross}} x$$
$$= \max_{(a_1, \ldots, a_{r^*}) : \sum a_s^2 = 1} \sum_{s,s'} a_s a_{s'} (v_{i_1}^\top Q v_{i_1'}) \cdot \lambda_{\text{max}}(F)$$
$$\geq v_{i_1}^\top Q v_{i_1} \cdot \lambda_{\text{max}}(F)$$
$$= \frac{1}{T} \sum_t g_{i_1}(t) (1 - g_{i_1}(t)) \cdot \lambda_{\text{max}}(F),$$ (C.2)

where $\lambda_{\text{max}}(F)$ is of $O(M)$. This holds for all of $r = 1, \ldots, C$ and we can say that there exist at least $C$ large eigenvalues of $O(M)$. 

17
The eigenvalue statistics are easily derived from the leading term \( \bar{\Theta} \) in the same way as Eq. (A.3), the \((k, k')\)-th block of the NTK is asymptotically given by

\[
\Theta(k, k') = \alpha K' \delta_{kk'} + o(1),
\]

for \( k, k' = 1, ..., C \). In contrast to Eq. (A.3), the NTK parameterization makes \( F^* \) multiplied by \( 1/M \). The negligible term of \( o(1) \) is reduced to \( O(1/\sqrt{M}) \) under the condition summarized in [37]. The entries of \( \bar{K}' \) are given by

\[
K'_{st} := \kappa'_1 (s = t), \quad \kappa'_2 (s \neq t),
\]

where \( \kappa'_1 \) is composed of two parts. The first part is \( ||\nabla \omega f(t)||^2 = \sigma_w^2 \sum_{i,l,j} \delta_i^l(t)^2 h_j^{l-1}(t)^2 / M_{l-1} \sim \sigma_w^2 \sum_{i,l} \tilde{q}_i^l \tilde{q}_i^{l-1} \). The second term is \( ||\nabla \beta f(t)||^2 = \sigma_b^2 \sum_{l,t} \delta_i^l(t)^2 \sim \sigma_b^2 \sum_{l=1}^{L} \tilde{q}_i^l \). Despite that the number of weights is much larger than the number of biases, the NTK parameterization makes a contribution of \( \nabla \omega f_k \) comparable to that of \( \nabla \beta f_k \). This is in contrast to the evaluation of \( F^* \) in Theorem 3.1, where the contribution of \( \nabla \omega f_k \) is negligible [31]. We can evaluate \( \kappa'_2 \) in the same way.

In the same way as with the FIM, the trace of \( \bar{K}' \) leads to \( m_\lambda \), the Frobenius norm of \( K'_{st} \) leads to \( s_\lambda \), and \( K' \) has the largest eigenvalue \((T - 1)\kappa'_2 + \kappa'_1 \) for arbitrary \( T \). The eigenspace of \( \Theta \) corresponding to \( \lambda_{\text{max}} \) is also the same as \( F^* \). It is spanned by eigenvectors \( \hat{v}_k \) \((k = 1, ..., C)\).

### D Derivation of Theorem 4.1

The NTK is defined as \( \Theta = TF^* \) under the NTK parameterization\(^4\) in the same way as Eq. (A.3), the \((k, k')\)-th block of the NTK is asymptotically given by

\[
\Theta(k, k') = \alpha K' \delta_{kk'} + o(1),
\]

for \( k, k' = 1, ..., C \). In contrast to Eq. (A.3), the NTK parameterization makes \( F^* \) multiplied by \( 1/M \). The negligible term of \( o(1) \) is reduced to \( O(1/\sqrt{M}) \) under the condition summarized in [37]. The entries of \( \bar{K}' \) are given by

\[
K'_{st} := \kappa'_1 (s = t), \quad \kappa'_2 (s \neq t),
\]

where \( \kappa'_1 \) is composed of two parts. The first part is \( ||\nabla \omega f(t)||^2 = \sigma_w^2 \sum_{i,l,j} \delta_i^l(t)^2 h_j^{l-1}(t)^2 / M_{l-1} \sim \sigma_w^2 \sum_{i,l} \tilde{q}_i^l \tilde{q}_i^{l-1} \). The second term is \( ||\nabla \beta f(t)||^2 = \sigma_b^2 \sum_{l,t} \delta_i^l(t)^2 \sim \sigma_b^2 \sum_{l=1}^{L} \tilde{q}_i^l \). Despite that the number of weights is much larger than the number of biases, the NTK parameterization makes a contribution of \( \nabla \omega f_k \) comparable to that of \( \nabla \beta f_k \). This is in contrast to the evaluation of \( F^* \) in Theorem 3.1, where the contribution of \( \nabla \omega f_k \) is negligible [31]. We can evaluate \( \kappa'_2 \) in the same way.

In the same way as with the FIM, the trace of \( \bar{K}' \) leads to \( m_\lambda \), the Frobenius norm of \( K'_{st} \) leads to \( s_\lambda \), and \( K' \) has the largest eigenvalue \((T - 1)\kappa'_2 + \kappa'_1 \) for arbitrary \( T \). The eigenspace of \( \Theta \) corresponding to \( \lambda_{\text{max}} \) is also the same as \( F^* \). It is spanned by eigenvectors \( \hat{v}_k \) \((k = 1, ..., C)\).

### E Derivation of Theorem 5.1

The metric tensor \( A_k \) can be represented as \( A_k = \nabla_h f_k \nabla_h f_k^\top / T \), where \( \nabla_h f_k \) is an \( N_k \times T \) matrix and its columns are the gradients on each input, i.e., \( \nabla_h f_k(t) \) \((t = 1, ..., T)\). Let us introduce the \( T \times T \) dual matrix of \( A_k \), i.e., \( A_k^* := \nabla_h f_k^\top \nabla_h f_k / T \). It has the same non-zero eigenvalues as \( A_k \) by definition. Its \( st\)-th entry is given by

\[
(A_k^*)_st = \nabla_h f_k(s)^\top \nabla_h f_k(t) / T
\]

\[
= \sum_{i=0}^{L-1} \sum_{i',j'} W_{ij}^{i+1} W_{ij'}^{i+1} \delta_j^i(s) \delta_j^{i+1}(t) / T
\]

\[
\sim \sigma_w^2 \sum_{i=1}^{L} \sum_{j} \delta_j^i(s) \delta_j^{i+1}(t) / T,
\]

in the large \( M \) limit. Accordingly, we have

\[
A_k^* = \tilde{\kappa}_1 \bar{A}^* + \frac{1}{T} o(1),
\]

\[
\bar{A}^* := \kappa'_1 (s = t), \quad \kappa'_2 (s \neq t).
\]

\( \tilde{\kappa}_1 \) is positive by definition, and \( \kappa'_2 \) is positive under the condition of the activation functions (ii).

The eigenvalue statistics are easily derived from the leading term \( \bar{A}^* \). We can derive the mean of the eigenvalues as \( m_\lambda \sim \text{Trace}(\frac{\bar{A}^*}{T} A_k^*) / N_k \) and the second moment as \( s_\lambda \sim ||\frac{\bar{A}^*}{T} A_k^*||_F^2 / N_k \), where \( N_k = \tilde{\alpha} M \). We can determine the largest eigenvalue because we explicitly obtain the eigenvalues of \( \bar{A}^* \); \( \lambda_1 = (T - 1)\tilde{\kappa}_2 + \tilde{\kappa}_1 \) with an eigenvector \( \tilde{v} := (1, \ldots, 1) \) and \( \lambda_i = \tilde{\kappa}_1 - \tilde{\kappa}_2 \) with eigenvectors \( e_i - e_i \) \((i = 2, ..., T)\). The vector \( e_i \) denotes a unit vector whose entries are 1 for the \( i \)-th entry and 0 otherwise. The largest eigenvalue is given by \( \lambda_1 \).

The eigenvector of \( A_k \) corresponding to \( \lambda_{\text{max}} \) is constructed from \( \tilde{v} \). Let us denote by \( v \) an eigenvector of \( A_k \) satisfying \( A_k v = \lambda_{\text{max}} v \). Multiplying both sides by \( \nabla_h f_k^\top \), we get

\[
A_k^*(\nabla_h f_k^\top v) = \lambda_{\text{max}} \cdot (\nabla_h f_k^\top v).
\]

\(^4\)Precisely speaking, the original definition of the NTK [28] is equivalent to the above \( \Theta \) under a simultaneous permutation of rows and columns. However, such permutations hold the same eigenvalues, and one can set \( \Theta \) in the above form without loss of generality.
This means that \( \nabla_h f_k^T v \) is the eigenvector of \( A_k^* \) and equals \( \tilde{v} \). Multiplying both sides of \( \tilde{v} = \nabla_h f_k^T v \) by \( \frac{1}{T} \nabla_h f_k \), we get

\[
E[\nabla_h f_k] = A_k v = \lambda_{\text{max}} v. \tag{E.4}
\]

As a result, we obtain \( v = E[\nabla_h f_k] \) up to a scale factor.

We can also evaluate eigenvalue statistics of \( A = \sum C_k A_k \) in analogy with \( F \). We can represent \( A_k \) as

\[
A_k = \tilde{R} \tilde{R}^T \text{ with } \tilde{R} := \frac{1}{\sqrt{T}} \left[ \nabla_h f_1 \nabla_h f_2 \cdots \nabla_h f_C \right]. \tag{E.5}
\]

Its columns are the gradients on each input, i.e., \( \nabla_h f_k(t) \) \( (t = 1, ..., T) \). Its dual matrix is given by \( A^* := \tilde{R}^T \tilde{R} \). We can perform the same calculation as on the FIM and obtain the following eigenvalue statistics of \( A \):

\[
m_\lambda \sim \bar{\kappa}_1 C, \quad s_\lambda \sim \bar{\alpha} \left( \frac{T - 1}{T} \bar{\kappa}_2 + \frac{\bar{\kappa}_1}{T} \right) C, \quad \lambda_{\text{max}} \sim \bar{\alpha} \left( \frac{T - 1}{T} \bar{\kappa}_2 + \frac{\bar{\kappa}_1}{T} \right). \tag{E.6}
\]

**F Derivation of Theorem 5.2**

The metric tensor \( A_k^{ll} \) can be represented by \( A_k^{ll} = \nabla_h f_k^T \nabla_h f_k^T / T \). Consider its dual, i.e., \( A_k^{ll*} = \nabla_h^T f_k^T \nabla_h f_k / T \) whose \( st \)-th entry is given by

\[
(A_k^{ll*})_{st} = \sum_i \nabla_{h_i} f_k(s) \nabla_{h_i} f_k(t) / T \sim \sigma_w^2 \sum_j \delta_j^{l+1}(s) \delta_j^{l+1}(t) / T. \tag{F.1}
\]

In the large \( M \) limit, we have asymptotically

\[
A_k^{ll*} = \frac{1}{T} A_k^{ll*} + \frac{1}{T} o(1), \quad \bar{A}^{ll*} := \sigma_w^2 \delta_{l+1}^{l+1} (s = t), \quad \sigma_w^2 \delta_{l+1}^{l+1} (s \neq t). \tag{F.2}
\]

The rest of calculations are the same as in Theorem 5.1. We can also derive the eigenvalue statistics of the summation \( A^{ll} := \sum C_k A_k^{ll} \); the mean and second moment are multiplied by \( C \), similar to the case of \( A \) shown in Eqs. (E.6).

**References**

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

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

[3] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998.

[4] Hyeongeun Park, Shun-ichi Amari, and Kenji Fukumizu. Adaptive natural gradient learning algorithms for various stochastic models. *Neural Networks*, 13(7):755–764, 2000.

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

[6] Yann Ollivier. Riemannian metrics for neural networks I: feedforward networks. *Information and Inference: A Journal of the IMA*, 4(2):108–153, 2015.

[7] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 2408–2417, 2015.
[8] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[9] Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer, 1998.

[10] Levent Sagun, Utku Evci, V Ugur Gunev, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv:1706.04454*, 2017.

[11] Vardan Papyan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians. In *Proceedings of International Conference on Machine Learning (ICML)*, 2019.

[12] 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.

[13] Kenji Fukumizu. A regularity condition of the information matrix of a multilayer perceptron network. *Neural Networks*, 9(5):871–879, 1996.

[14] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 888–896, 2019.

[15] Shun-ichi Amari. A method of statistical neurodynamics. *Kybernetik*, 14(4):201–215, 1974.

[16] Jonathan Kadmon and Haim Sompolinsky. Optimal architectures in a solvable model of deep networks. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 4781–4789, 2016.

[17] Ben Poole, Subhaneil Lahiri, Maithreyi Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Proceedings of Advances In Neural Information Processing Systems (NIPS)*, pages 3360–3368, 2016.

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

[19] Greg Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 2865–2873. 2017.

[20] 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.

[21] 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.

[22] 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.

[23] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 2798–2806, 2017.

[24] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 2634–2643, 2017.
[25] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In Proceedings of Advances In Neural Information Processing Systems (NIPS), pages 2253–2261, 2016.

[26] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. ICLR 2018 arXiv:1711.00165, 2018.

[27] Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. ICLR’2018, arXiv:1804.11271, 2018.

[28] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In Proceedings of Advances in Neural Information Processing Systems (NeurIPS), pages 8580–8589, 2018.

[29] 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.

[30] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. arXiv:1904.11955, 2019.

[31] 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, 2019.

[32] Yann LeCun, Ido Kanter, and Sara A Solla. Eigenvalues of covariance matrices: Application to neural-network learning. Physical Review Letters, 66(18):2396, 1991.

[33] Ke Sun and Frank Nielsen. Lightlike neuromanifolds, occam’s razor and deep learning. arXiv preprint arXiv:1905.11027, 2019.

[34] Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi. Fisher information and natural gradient learning of random deep networks. In Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS), pages 694–702, 2019.

[35] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In international conference on artificial intelligence and statistics (AISTATS), pages 249–256, 2010.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In The IEEE International Conference on Computer Vision (ICCV), December 2015.

[37] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. The normalization method for alleviating pathological sharpness in wide neural networks. arXiv preprint arXiv:1906.02926, 2019.

[38] Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. ICLR’2018, arXiv:1802.08760, 2018.

[39] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[40] Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation. arXiv preprint arXiv:1905.12558, 2019.

[41] Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, and Samuel S Schoenholz. A mean field theory of batch normalization. ICLR’2019 arXiv:1902.08129, 2019.

[42] Arthur Jacot, Franck Gabriel, and Clément Hongler. Freeze and chaos for dnns: an NTK view of batch normalization, checkerboard and boundary effects. arXiv:1907.05715, 2019.

[43] Taiji Suzuki. Fast generalization error bound of deep learning from a kernel perspective. In International Conference on Artificial Intelligence and Statistics (AISTATS), pages 1397–1406, 2018.