Universal Statistics of Fisher Information in Deep Neural Networks: Mean Field Approach

Ryo Karakida
AIST, Japan

Shotaro Akaho
AIST, Japan

Shun-ichi Amari
RIKEN CBS, Japan

Abstract

The Fisher information matrix (FIM) is a fundamental quantity to represent the characteristics of a stochastic model, including deep neural networks (DNNs). The present study reveals novel statistics of FIM that are universal among a wide class of DNNs. To this end, we use random weights and large width limits, which enables us to utilize mean field theories. We investigate the asymptotic statistics of the FIM's eigenvalues and reveal that most of them are close to zero while the maximum eigenvalue takes a huge value. Because the landscape of the parameter space is defined by the FIM, it is locally flat in most dimensions, but strongly distorted in others. Moreover, we demonstrate the potential usage of the derived statistics in learning strategies. First, small eigenvalues that induce flatness can be connected to a norm-based capacity measure of generalization ability. Second, the maximum eigenvalue that induces the distortion enables us to quantitatively estimate an appropriately sized learning rate for gradient methods to converge.

1 Introduction

Deep learning has succeeded in making hierarchical neural networks perform excellently in various practical applications [1]. To proceed further, it would be beneficial to give more theoretical elucidation as to why and how deep neural networks (DNNs) work well in practice. In particular, it would be useful to not only clarify the individual models and phenomena but also explore various unified theoretical frameworks that could be applied to a wide class of deep networks. One widely used approach for this purpose is to consider deep networks with random connectivity and a large width limit [2–14]. For instance, Poole et al. [3] proposed a useful indicator to explain the expressivity of DNNs. Regarding the trainability of DNNs, Schoenholz et al. [4] extended this theory to backpropagation and found that the vanishing and explosive gradients obey a universal law. These studies are powerful in the sense that they do not depend on particular model architectures, such as the number of layers or activation functions.

Unfortunately, such universal frameworks have not yet been established in many other topics. One is the geometric structure of the parameter space. For instance, the loss landscape without spurious local minima is important for easier optimization and theoretically guaranteed in single-layer models [15], shallow piecewise linear ones [16], and extremely wide deep networks with the number of training samples smaller than the width [17]. Flat global minima have been reported to be related to generalization ability through empirical experiments showing that networks with such minima give better generalization performance [18, 19]. However, theoretical analysis of the flat landscape has been limited in shallow rectified linear unit (ReLU) networks [20, 21]. Thus, a residual subject of interest is to theoretically reveal the geometric structure of the parameter space truly common among various deep networks.

To establish the foundation of the universal perspective of the parameter space, this study analytically investigates the Fisher information matrix (FIM). As is overviewed in Section 2.1, the FIM plays an essential role in the geometry of the parameter space and is a fundamental quantity in both statistics and machine learning.

1.1 Main results

This study analyzes the FIM of deep networks with random weights and biases, which are widely used settings to analyze the phenomena of DNNs [2–14]. First, we

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

analytically obtain novel statistics of the FIM, namely, the mean (Theorem 1), variance (Theorem 3), and maximum of eigenvalues (Theorem 4). These are universal among a wide class of shallow and deep networks with various activation functions. These quantities can be obtained from simple iterative computations of macroscopic variables. To our surprise, the mean of the eigenvalues asymptotically decreases with an order of $O(1/M)$ in the limit of a large network width M, while the variance takes a value of $O(1)$, and the maximum eigenvalue takes a huge value of $O(M)$ by using the $O(\cdot)$ order notation. Since the eigenvalues are non-negative, these results mean that most of the eigenvalues are close to zero, but the edge of the eigenvalue distribution takes a huge value. Because the FIM defines the Riemannian metric of the parameter space, the derived statistics imply that *the space is locally flat in most dimensions, but strongly distorted in others*. In addition, because the FIM also determines the local shape of a loss landscape, the landscape is also expected to be locally flat while strongly distorted.

Furthermore, to confirm the potential usage of the derived statistics, we show some exercises. One is on the Fisher-Rao norm [22] (Theorem 5). This norm was originally proposed to connect the flatness of a parameter space to the capacity measure of generalization ability. We evaluate the Fisher-Rao norm by using an indicator of the small eigenvalues, κ_1 in Theorem 1. Another exercise is related to the more practical issue of determining the size of the learning rate necessary for the steepest descent gradient to converge. We demonstrate that an indicator of the huge eigenvalue, κ_2 in Theorem 4, enables us to *roughly estimate learning rates that make the gradient method converge to global minima* (Theorem 7). We expect that it will help to alleviate the dependence of learning rates on heuristic settings.

1.2 Related works

Despite its importance in statistics and machine learning, study on the FIM for neural networks has been limited so far. This is because layer-by-layer nonlinear maps and huge parameter dimensions make it difficult to take analysis any further. Degeneracy of the eigenvalues of the FIM has been found in certain parameter regions [23]. To understand the loss landscape, Pennington and Bahri [5] has utilized random matrix theory and obtained the spectrum of FIM and Hessian under several assumptions, although the analysis is limited to special types of shallow networks. In contrast, this paper is the first attempt to apply the mean field approach, which overcomes the difficulties above and enables us to identify universal properties of the FIM in various types of DNNs.

LeCun et al. [24] investigated the Hessian of the loss, which coincides with the FIM at zero training error, and empirically reported that very large eigenvalues exist, i.e., "big killers", which affects the optimization (discussed in Section 4.2). The eigenvalue distribution peaks around zero while its tail is very long; this behavior has been empirically known for decades [25], but its theoretical evidence and evaluation have remained unsolved as far as we know. Therefore, our theory provides novel theoretical evidence that this skewed eigenvalue distribution and its huge maximum appear universally in DNNs.

The theoretical tool we use here is known as the *mean field theory* of deep networks [3, 4, 10–14] as briefly overviewed in Section 2.4. This method has been successful in analyzing neural networks with random weights under a large width limit and in explaining the performance of the models. In particular, it quantitatively coincides with experimental results very well and can predict appropriate initial values of parameters for avoiding the vanishing or explosive gradient problems [4]. This analysis has been extended from fully connected deep networks to residual [11] and convolutional networks [14]. The evaluation of the FIM in this study is also expected to be extended to such cases.

2 Preliminaries

2.1 Fisher information matrix (FIM)

We focus on the Fisher information matrix (FIM) of neural network models, which previous works have developed and is commonly used [26–31]. It is defined by

$$F = \mathrm{E}[\nabla_\theta \log p(x, y; \theta) \nabla_\theta \log p(x, y; \theta)^T], \qquad (1)$$

where the statistical model is given by $p(x, y; \theta) = p(y|x; \theta)p(x)$. The output model is given by $p(y|x; \theta) = \exp(-||y - f_\theta(x)||^2/2)/\sqrt{2\pi}$, where $f_\theta(x)$ is the network output parameterized by θ and $||\cdot||$ is the Euclidean norm. The $q(x)$ is an input distribution. The expectation $\mathrm{E}[\cdot]$ is taken over the input-output pairs (x, y) of the joint distribution $p(x, y; \theta)$. This FIM is transformed into $F = \sum_{k=1}^{C} \mathrm{E}[\nabla_\theta f_{\theta,k}(x) \nabla_\theta f_{\theta,k}(x)^T]$, where $f_{\theta,k}$ is the k-th entry of the output ($k = 1, ..., C$). When T training samples $x(t)$ $(t = 1, ..., T)$ are available, the expectation can be replaced by the empirical mean. This is known as the empirical FIM and often appears in practice [27–31]:

$$F = \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{C} \nabla_\theta f_{\theta,k}(t) \nabla_\theta f_{\theta,k}(t)^T. \qquad (2)$$

This study investigates the above empirical FIM for arbitrary T. It converges to the expected FIM as

$T \to \infty$. Although the form of the FIM changes a bit in other statistical models (i.e., softmax outputs), these differences are basically limited to the multiplication of activations in the output layer [30]. Our framework can be straightforwardly applied to such cases.

The FIM determines the asymptotic accuracy of the estimated parameters, as is known from a fundamental theorem of statistics, namely, the Cramér-Rao bound. Below, we summarize a more intuitive understanding of the FIM from geometric views.

Information geometric view. Let us define an infinitesimal squared distance dr^2, which represents the Kullback-Leibler divergence between the statistical model $p(x, y; \theta)$ and $p(x, y; \theta + d\theta)$ against a perturbation $d\theta$. It is given by

$$dr^2 := \mathrm{KL}(p(x, y; \theta)||p(x, y; \theta + d\theta)) = d\theta^T F d\theta. \quad (3)$$

It means that the parameter space of a statistical model forms a Riemannian manifold and the FIM works as its Riemannian metric, as is known in information geometry [32]. This quadratic form is equivalent to the robustness of a deep network: $\mathrm{E}[||f_{\theta+d\theta}(t) - f_\theta(t)||^2] = d\theta^T F d\theta$. Insights from information geometry have led to the development of natural gradient algorithms [29–31] and, recently, a capacity measure based on the Fisher-Rao norm [22].

Loss landscape view. The empirical FIM (2) determines the local landscape of the loss function around the global minimum. Suppose we have a squared loss function $E(\theta) = (1/2T)\sum_t^T ||y(t) - f_\theta(t)||^2$. The FIM is related to the Hessian of the loss function, $H := \nabla_\theta \nabla_\theta E(\theta)$, in the following way:

$$H = F - \frac{1}{T}\sum_t^T \sum_k^C (y_k(t) - f_{\theta,k}(t)) \nabla_\theta \nabla_\theta f_{\theta,k}(t). \quad (4)$$

The Hessian coincides with the FIM when the parameter converges to the global minimum by learning, that is, the true parameter θ^* from which the teacher signal $y(t)$ is generated by $y(t) = f_{\theta^*}(t)$ or, more generally, with noise (i.e., $y(t) = f_{\theta^*}(t) + \varepsilon_t$, where ε_t denotes zero-mean Gaussian noise) [27]. In the literature on deep learning, its eigenvectors whose eigenvalues are close to zero locally compose flat minima, which leads to better generalization empirically [19, 22]. Modifying the loss function with the FIM has also succeeded in overcoming the catastrophic forgetting [33].

Note that the information geometric view tells us more than the loss landscape. While the Hessian (4) assumes the special teacher signal, the FIM works as the Riemannian metric to arbitrary teacher signals.

2.2 Network architecture

This study investigates a fully connected feedforward neural network. The network consists of one input layer with M_0 units, $L-1$ hidden layers ($L \geq 2$) with M_l units per hidden layer ($l = 1, 2, ..., L-1$), and one output layer with M_L units:

$$u_i^l = \sum_{j=1}^{M_{l-1}} W_{ij}^l h_j^{l-1} + b_i^l, \quad h_i^l = \phi(u_i^l). \quad (5)$$

This study focuses on the case of linear outputs, that is, $f_{\theta,k}(x) = h_k^L = u_k^L$. We assume that the activation function $\phi(x)$ and its derivative $\phi'(x) := d\phi(x)/dx$ are square-integrable functions on a Gaussian measure. A wide class of activation functions, including the sigmoid-like and (leaky-) ReLU functions, satisfy these conditions. Different layers may have different activation functions. Regarding the network width, we set $M_l = \alpha_l M \ \ (l \leq L-1)$ and consider the limiting case of large M with constant coefficients α_l. This study mainly focuses on the case where the number of output units is given by a constant $M_L = C$. The higher-dimensional case of $C = O(M)$ is argued in Section 4.3.

The FIM (2) of a deep network is computed by the chain rule in a manner similar to that of the backpropagation algorithm:

$$\frac{\partial f_{\theta,k}}{\partial W_{ij}^l} = \delta_{k,i}^l \phi(u_j^{l-1}), \quad (6)$$

$$\delta_{k,i}^l = \phi'(u_i^l) \sum_j \delta_{k,j}^{l+1} W_{ji}^{l+1}, \quad \delta_{k,k}^L = \phi'(u_k^L), \quad (7)$$

where $\delta_{k,i}^l := \partial f_{\theta,k}/\partial u_i^l$ for $(k = 1, ..., C)$. To avoid the complicated notation, we omit the index of the output unit, i.e., $\delta_i^l = \delta_{k,i}^l$, in the following.

2.3 Random connectivity

The parameter set $\theta = \{W_{ij}^l, b_i^l\}$ is an ensemble generated by

$$W_{ij}^l \sim \mathcal{N}(0, \sigma_{w^l}^2/M_{l-1}), \quad b_i^l \sim \mathcal{N}(0, \sigma_{b^l}^2), \quad (8)$$

and then fixed, where $\mathcal{N}(0, \sigma^2)$ denotes a Gaussian distribution with zero mean and variance σ^2, and we set $\sigma_{w^l} > 0$ and $\sigma_{b^l} > 0$. To avoid complicated notation, we set them uniformly as $\sigma_{w^l}^2 = \sigma_w^2$ and $\sigma_{b^l}^2 = \sigma_b^2$, but they can easily be generalized. It is essential to normalize the variance of the weights by M in order to normalize the output u_i^l to $O(1)$. This setting is similar to how parameters are initialized in practice [34]. We also assume that the input samples $h_i^0(t) = x_i(t) \ \ (t = 1, ..., T)$ are generated in an i.i.d. manner from a standard Gaussian distribution: $x_i(t) \sim \mathcal{N}(0, 1)$. We focus here on

the Gaussian case for simplicity, although we can easily generalize it to other distributions with finite variances.

Let us remark that the above random connectivity is a common setting widely supposed in theories. Analyzing such a network can be regarded as the typical evaluation [2, 3, 5]. It is also equal to analyzing the network randomly initialized [4, 20]. The random connectivity is often assumed in the analysis of optimization as a true parameter of the networks, that is, the global minimum of the parameters [21, 35].

2.4 Mean-field approach

On neural networks with random connectivity, taking a large width limit, we can analyze the asymptotic behaviors of the networks. Recently, this asymptotic analysis is referred to as the mean field theory of deep networks, and we follow the previously reported notations and terminology [3, 4, 11, 12].

First, let us introduce the following variables for feed-forward signal propagations: $\hat{q}^l := \sum_i h_i^l(t)^2/M_l$ and $\hat{q}_{st}^l := \sum_i h_i^l(s)h_i^l(t)/M_l$. In the context of deep learning, these variables have been utilized to explain the depth to which signals can sufficiently propagate. The variable $\hat{q}_{st}^l$ is the correlation between the activations for different input samples $x(s)$ and $x(t)$ in the l-th layer. Under the large M limit, these variables are given by integration over Gaussian distributions because the pre-activation u_l^i is a weighted sum of independent random parameters and the central limit theorem is applicable [2–4]:

$$\hat{q}^{l+1} = \int Du\phi^2\left(\sqrt{q^{l+1}}u\right), \quad q^{l+1} = \sigma_w^2\hat{q}^l + \sigma_b^2, \quad (9)$$

$$\hat{q}_{st}^{l+1} = I_\phi[q^{l+1}, q_{st}^{l+1}], \quad q_{st}^{l+1} = \sigma_w^2\hat{q}_{st}^l + \sigma_b^2, \quad (10)$$

with $\hat{q}^0 = 1$ and $\hat{q}_{st}^0 = 0$ $(l = 0, ..., L-1)$. We can generalize the theory to unnormalized data with $\hat{q}^0 \neq 0$ and $\hat{q}_{st}^0 \neq 0$, just by substituting them into the recurrence relations. The notation $Du = du\exp(-u^2/2)/\sqrt{2\pi}$ means integration over the standard Gaussian density. Here, the notation $I_\cdot[\cdot,\cdot]$ represents the following integral: $I_\phi[a,b] = \int Dz_1Dz_2\phi\left(\sqrt{a}z_1\right)\phi\left(\sqrt{a}(cz_1 + \sqrt{1-c^2}z_2)\right)$ with $c = b/a$. The q_{st}^l is linked to the compositional kernel and utilized as the kernel of the Gaussian process [36].

Next, let us introduce variables for backpropagated signals: $\tilde{q}^l := \sum_i \delta_i^l(t)^2$ and $\tilde{q}_{st}^l := \sum_i \delta_i^l(s)\delta_i^l(t)$. Note that they are defined not by averages but by sums. They remain $O(1)$ because of $C = O(1)$. $\tilde{q}_{st}^l$ is the correlation of backpropagated signals. To compute these quantities, the previous studies assumed the following:

Assumption 1 (Schoenholz et al. [4]). *On the evaluation of the variables $\tilde{q}^l$ and $\tilde{q}_{st}^l$, one can use a different set of parameters, θ for the forward chain (5) and θ' for the backpropagated chain (7), instead of using the same parameter set θ in both chains.*

Mmmmm idk about this assumption

This assumption makes the dependence between $\phi(u_i^l)$ (or $\phi'(u_i^l)$) and δ_j^{l+1}, which share the same parameter set, very weak, and one can regard it as independent. It enables us to apply the central limit theorem to the backpropagated chain (7). Thus, the previous studies [4, 7, 11, 12] derived the following recurrence relations $(l = 0, ..., L-1)$:

$$\tilde{q}^l = \sigma_w^2\tilde{q}^{l+1}\int Du\left[\phi'(\sqrt{q^l}u)\right]^2, \quad (11)$$

$$\tilde{q}_{st}^l = \sigma_w^2\tilde{q}_{st}^{l+1}I_{\phi'}[q^l, q_{st}^l], \quad (12)$$

with $\tilde{q}^L = \tilde{q}_{st}^L = 1$ because of the linear outputs. The previous works confirmed excellent agreements between the above equations and experiments. In this study, we also adopt the above assumption and use the recurrence relations.

The variables $(\hat{q}^l, \tilde{q}^l, \hat{q}_{st}^l, \tilde{q}_{st}^l)$ depend only on the variance parameters σ_w^2 and σ_b^2, not on the unit indices. In that sense, they are referred to as *macroscopic variables* (a.k.a. order parameters in statistical physics). The recurrence relations for the macroscopic variables simply require L iterations of one- and two-dimensional numerical integrals. Moreover, we can obtain their explicit forms for some activation functions (such as the error function, linear, and ReLU; see Supplementary Material B).

3 Fundamental FIM statistics

Here, we report mathematical findings that the mean, variance, and maximum of eigenvalues of the FIM (2) are explicitly expressed by using macroscopic variables. Our theorems are universal for networks ranging in size from shallow $(L = 2)$ to arbitrarily deep $(L \geq 3)$ with various activation functions.

3.1 Mean of eigenvalues

The FIM is a $P \times P$ matrix, where P represents the total number of parameters. First, we compute the arithmetic mean of the FIM's eigenvalues as $m_\lambda := \sum_{i=1}^P \lambda_i/P$. We find a hidden relation between the macroscopic variables and the statistics of FIM:

Theorem 1. *In the limit of $M \gg 1$, the mean of the FIM's eigenvalues is given by*

a is the coefficient for the size of the layer

$$m_\lambda = C\frac{\kappa_1}{M}, \quad \kappa_1 := \sum_{l=1}^{L}\frac{\alpha_{l-1}}{\alpha}\tilde{q}^l\hat{q}^{l-1}, \quad (13)$$

where $\alpha := \sum_{l=1}^{L-1} \alpha_l \alpha_{l-1}$. *The macroscopic variables* $\hat{q}^l$ *and* $\tilde{q}^l$ *can be computed recursively, and notably* m_λ *is* $O(1/M)$.

This is obtained from a relation $m_\lambda = \mathrm{Trace}(F)/P$ (detailed in Supplementary Material A.1). The coefficient κ_1 is a constant not depending on M, so it is $O(1)$. It is easily computed by L iterations of the layer-wise recurrence relations (9) and (11).

Because the FIM is a positive semi-definite matrix and its eigenvalues are non-negative, this theorem means that most of the eigenvalues asymptotically approach zero when M is large. Recall that the FIM determines the local geometry of the parameter space. The theorem suggests that the network output remains almost unchanged against a perturbation of the parameters in many dimensions. It also suggests that the shape of the loss landscape is locally flat in most dimensions.

Furthermore, by using Markov's inequality, we can prove that the number of larger eigenvalues is limited, as follows:

Corollary 2. *Let us denote the number of eigenvalues satisfying* $\lambda \geq k$ *by* $N(\lambda \geq k)$ *and suppose that Assumption 1 holds. For a constant* $k > 0$, $N(\lambda \geq k) \leq \min\{\alpha\kappa_1 CM/k, CT\}$ *holds in the limit of* $M \gg 1$.

The proof is shown in Supplementary Material A.2. When T is sufficiently small, we have a trivial upper bound $N(\lambda \geq k) \leq CT$ and the number of non-zero eigenvalue is limited. The corollary clarifies that even when T becomes large, the number of eigenvalues whose values are $O(1)$ is $O(M)$ at most, and still much smaller than the total number of parameters P.

3.2 Variance of eigenvalues

Next, let us consider the second moment $s_\lambda := \sum_{i=1}^{P} \lambda_i^2/P$. We now demonstrate that s_λ can be computed from the macroscopic variables:

Theorem 3. *Suppose that Assumption 1 holds. In the limit of* $M \gg 1$, *the second moment of the FIM's eigenvalues is*

$$s_\lambda = C\alpha \left(\frac{T-1}{T} \kappa_2^2 + \frac{1}{T} \kappa_1^2 \right), \tag{14}$$

$$\kappa_2 := \sum_{l=1}^{L} \frac{\alpha_{l-1}}{\alpha} \tilde{q}_{st}^l \hat{q}_{st}^{l-1}. \tag{15}$$

The macroscopic variables $\hat{q}_{st}^l$ *and* $\tilde{q}_{st}^l$ *can be computed recursively, and* s_λ *is* $O(1)$.[1]

[1]Let us remark that we have assumed $\sigma_b > 0$ in the setting (8). If one considers a case of no bias term ($\sigma_b = 0$), odd activations $\phi(x)$ lead to $\hat{q}_{st}^l = 0$ and $\kappa_2 = 0$. In such exceptional cases, we need to evaluate the lower order terms of s_λ and λ_{max} (outside the scope of this study).

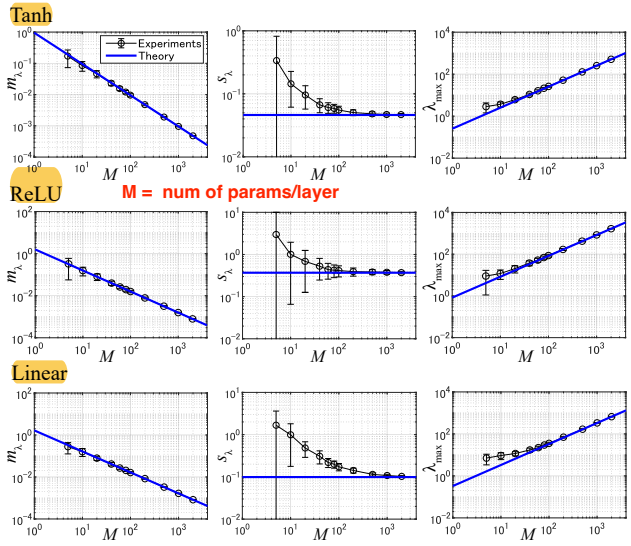


Figure 1: Statistics of FIM eigenvalues: means (left), second moments (center), and maximum (right). Our theory predicts the results of numerical experiments, indicated by the black points and error bars. The experiments used 100 random ensembles with different seeds. The variances of the parameters were given by $(\sigma_w^2, \sigma_b^2) = (3, 0.64)$ in the tanh case, $(2, 0.1)$ in the ReLU case, and $(1, 0.1)$ in the linear case. Each colored line represents theoretical results obtained in the limit of $M \gg 1$.

The proof is shown in Supplementary Material A.3.

From Theorems 1 and 3, we can conclude that the variance of the eigenvalue distribution, $s_\lambda - m_\lambda^2$, is $O(1)$. Because the mean m_λ is $O(1/M)$ and most eigenvalues are close to zero, this result means that the edge of the eigenvalue distribution takes a huge value.

3.3 Maximum eigenvalue

As we have seen so far, the mean of the eigenvalues is $O(1/M)$, and the variance is $O(1)$. Therefore, we can expect that at least one of the eigenvalues must be huge. Actually, we can show that the maximum eigenvalue (that is, the spectral norm of the FIM) increases in the order of $O(M)$ as follows.

Theorem 4. *Suppose that Assumption 1 holds. In the limit of* $M \gg 1$, *the maximum eigenvalue of the FIM is*

$$\lambda_{max} = \alpha \left(\frac{T-1}{T} \kappa_2 + \frac{1}{T} \kappa_1 \right) M. \tag{16}$$

The λ_{max} is derived from the dual matrix F^* (detailed in Supplemental Material A.4). If we take the limit $T \to \infty$, we can characterize the quantity κ_2 by the maximum eigenvalue as $\lambda_{max} = \alpha\kappa_2 M$. Note that λ_{max} is independent of C. When $C = O(M)$, it may depend on C, as shown in Section 3.4.

This theorem suggests that the network output changes dramatically with a perturbation of the parameters in certain dimensions and that the local shape of the loss landscape is strongly distorted in that direction. Here, note that λ_{max} is proportional to α, which is the summation over L terms. This means that, when the network becomes deeper, the parameter space is more strongly distorted.

Does this mean we could take out most of the parameters in the space that doesn't matter and still preserve most of the performance?

We confirmed the agreement between our theory and numerical experiments, as shown in Fig. 1. Three types of deep networks with parameters generated by random connectivity (8) were investigated: tanh, ReLU, and linear activations ($L = 3$, $\alpha_l = C = 1$). The input samples were generated using i.i.d. Gaussian samples, and $T = 10^2$. When $P > T$, we calculated the eigenvalues by using the dual matrix F^* (defined in Supplementary Material A.3) because F^* is much smaller and its eigenvalues are easy to compute. The theoretical values of m_λ, s_λ and λ_{max} agreed very well with the experimental values in the large M limit. We could predict m_λ even for small M. In addition, In Supplementary Material C.1, we also show the results of experiments with fixed M and changing T. The theoretical values coincided with the experimental values very well for any T as the theorems predict.

4 Connections to learning strategies

Here, we show some applications that demonstrate how our universal theory on the FIM can potentially enrich deep learning theories. It enables us to quantitatively measure the behaviors of learning strategies as follows.

4.1 The Fisher-Rao norm

Recently, Liang et al. [22] proposed the Fisher-Rao norm for a capacity measure of generalization ability:

$$||\theta||_{FR} = \theta^T F \theta, \tag{17}$$

where θ represents weight parameters. They reported that this norm has several desirable properties to explain the high generalization capability of DNNs. In deep linear networks, its generalization capacity (Rademacher complexity) is upper bounded by the norm. In deep ReLU networks, the Fisher-Rao norm serves as a lower bound of the capacities induced by other norms, such as the path norm [37] and the spectral norm [38]. The Fisher-Rao norm is also motivated by information geometry, and invariant under node-wise linear rescaling in ReLU networks. This is a desirable property to connect capacity measures with flatness induced by the rescaling [39].

Here, to obtain a typical evaluation of the norm, we define the average over possible parameters with fixed variances (σ_w^2, σ_b^2) by $\langle\cdot\rangle_\theta = \int \prod_i D\theta_i(\cdot)$, which leads to the following theorem:

Theorem 5. *Suppose that Assumption 1 holds. In the limit of $M \gg 1$, the Fisher-Rao norm of DNNs satisfies*

$$\langle||\theta||_{FR}\rangle_\theta \leq \sigma_w^2 \frac{\alpha}{\alpha_{min}} C\kappa_1, \tag{18}$$

where $\alpha_{min} = \min_i \alpha_i$. Equality holds in a network with a uniform width $M_l = M$, and then we have $\langle||\theta||_{FR}\rangle_\theta = \sigma_w^2(L-1)C\kappa_1$.

The proof is shown in Supplementary Material A.6. Although what we can evaluate is only the average of the norm, it can be quantified by κ_1. This guarantees that the norm is independent of the network width in the limit of $M \gg 1$, which was empirically conjectured by [22].

Recently, Smith and Le [40] argued that the Bayesian factor composed of the Hessian of the loss function, whose special case is the FIM, is related to the generalization. Similar analysis to the above theorem may enable us to quantitatively understand the relation between the statistics of the FIM and the indicators to measure the generalization ability.

4.2 Learning rate for convergence

Consider the steepest gradient descent method in a batch regime. Its update rule is given by

$$\theta_{t+1} \leftarrow \theta_t - \eta \frac{\partial E(\theta_t)}{\partial \theta} + \mu(\theta_t - \theta_{t-1}), \tag{19}$$

where η is a constant learning rate. We have added a momentum term with a coefficient μ because it is widely used in training deep networks. Assume that the squared loss function $E(\theta)$ of Eq. (4) has a global minimum θ^* achieving the zero training error $E(\theta^*) = 0$. Then, the FIM's maximum eigenvalue is dominant over the convergence of learning as follows:

Lemma 6. *A learning rate satisfying $\eta < 2(1+\mu)/\lambda_{max}$ is necessary for the steepest gradient method to converge to the global minimum θ^*.*

The proof is given by the expansion around the minimum, i.e., $E(\theta^* + d\theta) = d\theta^T F d\theta$ (detailed in Supplementary Material A.7). This lemma is a generalization of LeCun et al. [24], which proved the case of $\mu = 0$. Let us refer to $\eta_c := 2(1+\mu)/\lambda_{max}$ as the critical learning

rate. When $\eta > \eta_c$, the gradient method never converges to the global minimum. The previous work [24] also claimed that $\eta = \eta_c/2$ is the best choice for fastest convergence around the minimum. Although we focus on the batch regime, the eigenvalues also determine the bound of the gradient norms and the convergence of learning in the online regime [41].

Then, combining Lemma 6 with Theorem 4 leads to the following:

Theorem 7. *Suppose that Assumption 1 holds. Let a global minimum θ^* be generated by Eq. (8) and satisfying $E(\theta^*) = 0$. In the limit of $M \gg 1$, the gradient method never converges to θ^* when*

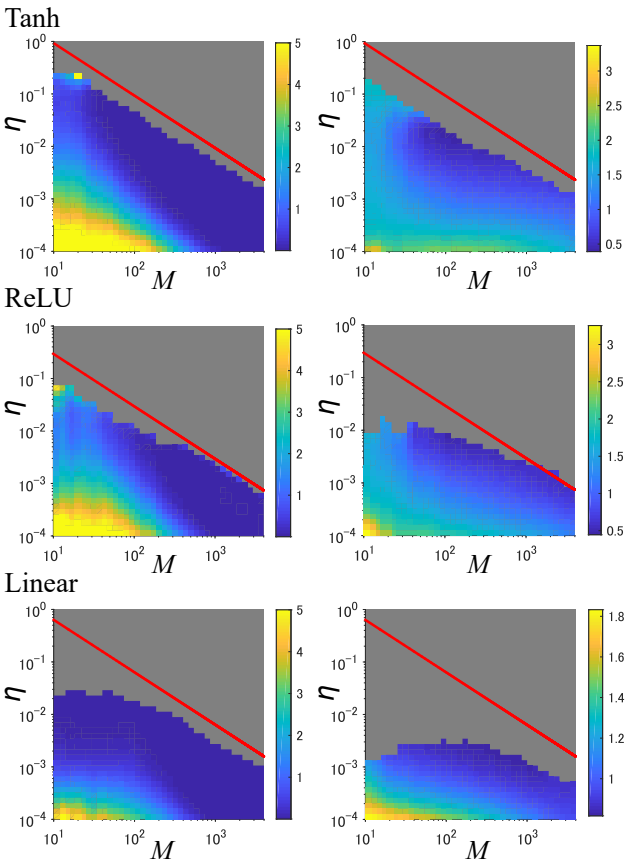
$$\eta > \eta_c, \quad \eta_c := \frac{2(1+\mu)}{\alpha\left(\frac{T-1}{T}\kappa_2 + \frac{1}{T}\kappa_1\right) M}. \tag{20}$$

Theorem 7 quantitatively reveals that, the wider the network becomes, the smaller the learning rate we need to set. In addition, α is the sum over L constant positive terms, so a deeper network requires a finer setting of the learning rate and it will make the optimization more difficult. In contrast, the expressive power of the network grows exponentially as the number of layers increases [3, 42]. We thus expect there to be a trade-off between trainability and expressive power.

To confirm the effectiveness of Theorem 7, we performed several experiments. As shown in Fig. 2, we exhaustively searched training losses while changing M and η, and found that the theoretical estimation coincides well with the experimental results. We trained deep networks ($L = 4$, $\alpha_l = 1$, $C = 10$) and the loss function was given by the squared error.

The left column of Fig. 2 shows the results of training on artificial data. We generated training samples $x(t)$ in the Gaussian manner ($T = 100$) and teacher signals $y(t)$ by the teacher network with a true parameter set θ^* satisfying Eq. (8). We used the gradient method (19) with $\mu = 0.9$ and trained the DNNs for 100 steps. The variances (σ_w^2, σ_b^2) of the initialization of the parameters were set to the same as the global minimum. We found that the losses of the experiments were clearly divided into two areas: one where the gradient exploded (gray area) and the other where it was converging (colored area). The red line is η_c theoretically calculated using κ_1 and κ_2 on (σ_w^2, σ_b^2) of the initial parameters. Training on the regions above η_c exploded, just as Theorem 7 predicts. The explosive region with $\eta < \eta_c$ got smaller in the limit of large M.

We performed similar experiments on benchmark datasets and found that the theory can estimate the appropriate learning rates. The results on MNIST are shown in the right column of Fig. 2. As shown in Supplementary Material C.2, the results of training



So loss is lowest right before it diverges??

Figure 2: Color map of training losses: Batch training on artificial data (left column) and SGD training on MNIST (right column). The losses are averages over five trials. The color bar shows the value of the training loss after the training. The region where the loss diverges (i.e., is larger than 1000) is in gray. The red line shows the theoretical value of η_c. The initial conditions of the parameters were taken from a Gaussian distribution (8) with $(\sigma_w^2, \sigma_b^2) = (3, 0.64)$ in tanh networks, $(2, 0.1)$ in ReLU networks, and $(1, 0.1)$ in linear networks.

on CIFAR-10 were almost the same as those of MINIST. We used stochastic gradient descent (SGD) with a mini-batch size of 500 and $\mu = 0.9$, and trained the DNNs for 1 epoch. Each training sample was $x(t)$ normalized to zero mean and variance 1 ($T = 50000$). The initial values of (σ_w^2, σ_b^2) were set to the vicinity of the special parameter region, i.e., the critical line of the order-to-chaos transition, which the previous works [3, 4] recommended to use for achieving high expressive power and trainability. Note that the variances (σ_w^2, σ_b^2) may change from the initialization to the global minimum, and the conditions of the global minimum in Theorem 7 do not hold in general. Nevertheless, the

learning rates estimated by Theorem 7 explained the experiments well. Therefore, the ideal conditions supposed in Theorem 7 seem to hold effectively. This may be explained by the conjecture that the change from the initialization to the global minima is small in the large limit [43].

Theoretical estimations of learning rates in deep networks have so far been limited; such gradients as AdaGrad and Adam also require heuristically determined hyper-parameters for learning rates. Extending our framework would be beneficial in guessing learning rates to prevent the gradient update from exploding.

4.3 Multi-label classification with high dimensionality

This study mainly focuses on the multi-dimensional output of $C = O(1)$. This is because the number of labels is much smaller than the number of hidden units in most practice cases. However, since classification problems with far more labels are sometimes examined in the context of machine learning [44], it would be helpful to remark on the case of $C = O(M)$ here. Denote the mean of the FIM's eigenvalues in the case of $C = O(M)$ as m'_λ and so on. Straightforwardly, we can derive

$$m'_\lambda = m_\lambda, \quad s_\lambda \leq s'_\lambda \leq C s_\lambda, \tag{21}$$

$$\lambda_{max} \leq \lambda'_{max} \leq \sqrt{\alpha C s_\lambda} M. \tag{22}$$

The derivation is shown in Supplementary Material A.5. The mean of eigenvalues has the same form as Eq. (13) obtained in the case of $C = O(1)$. The second moment and maximum eigenvalues can be evaluated by the form of inequalities. We found that the mean is of $O(1)$ while the maximum eigenvalue is of $O(M)$ at least and of $O(M^2)$ at most. Therefore, the eigenvalue distribution is more widely distributed than the case of $C = O(1)$.

5 Conclusion and discussion

The present work elucidated the asymptotic statistics of the Fisher information matrix (FIM) common among deep networks with any number of layers and various activation functions. The statistics of FIM are characterized by the small mean of eigenvalues and the huge maximum eigenvalue, which are computed by the recurrence relations. This suggests that the parameter space determined by the FIM is locally flat in many directions while highly distorted in certain others. As examples of how one can connect the derived statistics to learning strategies, we suggest the Fisher-Rao norm and learning rates of steepest gradient descents.

We demonstrated that the experiments with the Gaussian prior on the parameters coincided well with the theory. Basically, the mean field theory is based on the central limit theorem with the parameters generated in an i.i.d. manner with finite variances. Therefore, one can expect that the good agreement with the theory is not limited to the experiments with the Gaussian prior. Further experiments will be helpful to clarify the applicable scope of the mean field approach.

The derived statistics are also of potential importance to other learning strategies, for instance, natural gradient methods. When the loss landscape is non-uniformly distorted, naive gradient methods are likely to diverge or become trapped in plateau regions, but the natural gradient, $F^{-1}\nabla_\theta E(\theta)$, converges more efficiently [27–30]. Because it normalizes the distortion of the loss landscape, the naive extension of Section 4.2 to the natural gradient leads to $\eta_c = 2(1+\mu)$ and it seems to be much easier to choose the appropriately sized learning rate. However, we found that the FIM has many eigenvalues close to zero, and the inversion of it would make the gradient very unstable. In practice, several experiments showed that the choice of damping term ε, introduced in $(F + \varepsilon I)^{-1}\nabla_\theta E(\theta)$, is crucial to its performance in DNNs [31]. The development of practical natural gradient methods will require modification such as damping.

It would also be interesting for our framework to quantitatively reveal the effects of normalization methods on the FIM. In particular, batch normalization may alleviate the larger eigenvalues because it empirically allows larger learning rates for convergence [45]. It would also be fruitful to investigate the eigenvalues of the Hessian with a large error (4) and to theoretically quantify the negative eigenvalues that lead to the existence of saddle points and the loss landscapes without spurious local minima [46]. The global structure of the parameter space should be also explored. We can hypothesize that the parameters are globally connected through the locally flat dimensions and compose manifolds of flat minima.

Our framework on FIMs is readily applicable to other architectures such as convolutional networks and residual networks by using the corresponding mean field theories [11, 12]. To this end, it may be helpful to remark that macroscopic variables in residual networks essentially diverge at the extreme depths [11]. If one considers extremely deep residual networks, the statistics will require a careful examination of the order of the network width and the explosion of the macroscopic variables. We expect that further studies will establish a mathematical foundation of deep learning from the perspective of the large limit.

Acknowledgments

This work was partially supported by a Grant-in-Aid for Research Activity Start-up (17H07390) from the Japan Society for the Promotion of Science (JSPS).

References

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

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

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

[4] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *ICLR'2017 arXiv preprint arXiv:1611.01232*, 2016.

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

[6] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2634–2643, 2017.

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

[8] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *International Conference on Machine Learning (ICML)*, pages 2847–2854, 2017.

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

[10] Bo Li and David Saad. Exploring the function space of deep-learning machines. *Physical Review Letters*, 120(24):248301, 2018.

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

[12] 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 *International Conference on Machine Learning (ICML)*, pages 5393–5402, 2018.

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

[14] Minmin Chen, Jeffrey Pennington, and Samuel S Schoenholz. Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks. In *International Conference on Machine Learning (ICML)*, pages 873–882, 2018.

[15] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*, 2016.

[16] Daniel Soudry and Elad Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*, 2017.

[17] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *International Conference on Machine Learning (ICML)*, pages 2603–2612, 2017.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

[19] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *ICLR'2017 arXiv:1609.04836*, 2016.

[20] Itay Safran and Ohad Shamir. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning (ICML)*, pages 774–782, 2016.

[21] Yuandong Tian. An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. In *International Conference on Machine Learning (ICML)*, pages 3404–3413, 2017.

[22] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-Rao metric, geometry, and complexity of neural networks. *arXiv preprint arXiv:1711.01530*, 2017.

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

[24] 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.

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

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

[27] Shun-Ichi Amari, Hyeyoung Park, and Kenji Fukumizu. Adaptive method of realizing natural gradient learning for multilayer perceptrons. *Neural Computation*, 12(6):1399–1409, 2000.

[28] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. *ICLR'2014 arXiv preprint arXiv:1301.3584*, 2013.

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

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

[31] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning (ICML)*, pages 2408–2417, 2015.

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

[33] 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.

[34] 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.

[35] David Saad and Sara A Solla. Exact solution for on-line learning in multilayer neural networks. *Physical Review Letters*, 74(21):4337, 1995.

[36] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *ICLR'2018 arXiv preprint arXiv:1711.00165*, 2017.

[37] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory (COLT)*, pages 1376–1401, 2015.

[38] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6241–6250, 2017.

[39] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning (ICML)*, pages 1019–1028, 2017.

[40] Samuel L Smith and Quoc V Le. Understanding generalization and stochastic gradient descent. *ICLR'2018 arXiv preprint arXiv:1710.06451*, 2017.

[41] Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):9–42, 1998.

[42] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2924–2932, 2014.

[43] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

[44] Jia Deng, Alexander C Berg, Kai Li, and Li Fei-Fei. What does classifying more than 10,000 image categories tell us? In *European Conference on Computer Vision (ECCV)*, pages 71–84. Springer, 2010.

[45] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[46] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2933–2941, 2014.