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

Ryo Karakida [*], **Shotaro Akaho** [†], **and Shun-ichi Amari** [‡]

Abstract

This study analyzes the Fisher information matrix (FIM) by applying mean-field theory to deep neural networks with random weights. We theoretically find novel statistics of the FIM, which are universal among a wide class of deep networks with any number of layers and various activation functions. Although most of the FIM's eigenvalues are close to zero, the maximum eigenvalue takes on a huge value and the eigenvalue distribution has an extremely long tail. These statistics suggest that the shape of a loss landscape is locally flat in most dimensions, but strongly distorted in the other dimensions. Moreover, our theory of the FIM leads to quantitative evaluation of learning in deep networks. First, the maximum eigenvalue enables us to estimate an appropriate size of a learning rate for steepest gradient methods to converge. Second, the flatness induced by the small eigenvalues is connected to generalization ability through a norm-based capacity measure.

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 theoretical elucidation of why and how deep neural networks (DNNs) work well in practice. In particular, it would be useful not only for explaining the individual models and phenomena, but also for exploring some unified theoretical frameworks that could be applied to a wide class of deep networks. We apply the *mean field theory* for this purpose, which originated from statistical physics and has been developed into a tool to analyze neural networks with random weights [2, 3]. For instance, Poole et al. [4] used it on feedforward signals and proposed a useful indicator to explain the expressivity of DNNs. This theory is powerful in the sense that it does not depend on particular model architectures, such as the number of layers or special activation functions. Moreover, regarding the trainability of DNNs, Schoenholz et al. [5] extended mean field theory to backpropagation and found that the vanishing and explosive gradients obey a universal law and that such bad gradients can be prevented in certain parameter regions.

Unfortunately, universal frameworks have not yet been established in many other topics. One such topic is the geometric structure of the parameter space. Some theories have found local minima-free conditions on the loss landscape in special models: single-layer cases [6], shallow piecewise linear cases [7], and extremely wide deep cases [8]. Other theories indicate that flat global minima, which have been empirically shown to give better generalization performance [9, 10], appear in shallow rectified linear unit (ReLU) networks [11, 12]. Although these theories yielded novel and suggestive insights, it is a nontrivial problem to extend each theory to more general deep networks and identify geometric characterization common among various deep networks.

[*] Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan, E-mail: karakida.ryo@aist.go.jp

[†] National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki, Japan, E-mail: s.akaho@aist.go.jp

[‡] RIKEN Center for Brain Science (CBS), Saitama, Japan, E-mail: amari@brain.riken.jp

Preprint. Work in progress.

This paper utilizes the mean-field framework to analytically investigate the Fisher information matrix (FIM) in general deep networks. The FIM plays an essential role in the geometry of the parameter space, and is a fundamental quantity in both statistics and machine learning. It defines a Riemannian metric of information geometry and determines the loss landscape around global minima, as described in Section 2.2. In particular, the maximum eigenvalues of the FIM determine the convergence of steepest gradient methods [13]. The eigenvectors whose eigenvalues are close to zero locally compose flat minima, which lead to better generalization [10, 14]. Empirical studies have reported that many eigenvalues are close to zero while the eigenvalue distribution has a extremely long tail [13, 15]. Despite its importance, theoretical studies on the FIM for neural networks have been limited to the regularity condition [16] and the eigenvalue distribution analytically obtained in special types of shallow networks [17]. This is because layer-by-layer nonlinear maps and huge parameter dimensions make it difficult to take a theoretical analysis any further. Surprisingly, mean field theory, which takes the numbers of units to be sufficiently large, overcomes these difficulties and enables us to identify universal properties of the FIM.

First, we analytically obtain novel statistics of the FIM, namely, the mean (Theorem 1), variance (Theorem 3), and maximum of eigenvalues (Theorem 4). They are universal among a wide class of shallow and deep networks with various activation functions. These quantities can be, interestingly, derived from computations of macroscopic variables (or order parameters in statistical physics). 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 non-zero value of $O(1)$ and the maximum eigenvalue takes a huge value of $O(M)$. Since the eigenvalues are nonnegative, these results mean that most of the eigenvalues are close to zero, but the eigenvalue distribution has an extremely long tail. This conclusion is supported by empirical studies [13, 15].

Second, to confirm the effectiveness of the derived universal statistics further, we consider two interesting exercises on learning theories. One is on the size of the learning rate necessary for the steepest descent gradient to converge. The strongest distortion of the loss landscape induced by the FIM's maximum eigenvalue determines an appropriate learning rate for convergence around the global minimum [13]. We demonstrate that our theory regarding the maximum eigenvalue enables us to *estimate learning rates that prevent the gradient methods from exploding* (Theorem 6). To the best of our knowledge, this is the first theoretical estimation of the learning rate for convergence in a general deep network and we expect that it will help to alleviate the dependence of learning rates on heuristic settings. We also confirm the effectiveness of this estimation in numerical experiments on benchmark datasets. The second exercise is to connect our theory to the problem of generalization ability by using the Fisher-Rao norm, a norm-based capacity measure in statistical learning theory [14]. We show that the Fisher-Rao norm can be transformed into a simple analytical form of macroscopic variables (Theorem 7) and is upper bounded by the small mean of the eigenvalues. This suggests that the flatness induced by the small eigenvalues is connected to better generalization.

2 Preliminaries

2.1 Model architecture

This study investigates a fully-connected feedforward neural network with random weight and bias parameters [4, 5]. 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:

$$u_i^l = \sum_{j=1}^{M_l} W_{ij}^l h_j^{l-1} + b_i^l, \;\; h_i^l = \phi(u_i^l). \tag{1}$$

It includes shallow nets ($L = 2$) and arbitrary deep nets ($L \geq 3$). The activation function $\phi(x)$ is non-decreasing. We also suppose that $\phi(x)$ and its derivative $\phi'(x) := d\phi(x)/dx$ are square-integrable functions on a Gaussian measure. Different layers may have different activation functions. A wide class of activation functions, including the sigmoid-like and (leaky-) ReLU functions, satisfy these conditions. 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. The number of output units is given by a constant $M_L = C$. The parameter set $\{W_{ij}^l, b_i^l\}$ is an ensemble generated by

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

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 complicating the notation, we will set them uniformly as $\sigma^2_{w^l} = \sigma^2_w$ and $\sigma^2_{b^l} = \sigma^2_b$, but they can easily be generalized. In addition, we assume that the input samples, $h^0_i(t) = x_i(t)$ $(t = 1, ..., T)$ 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 generalize the generators of the parameters and inputs to other distributions. This Gaussian assumption is popular in the following theories of neural networks: typical evaluation approach [2, 4], random initialization of training [5, 11], the random matrix formulation [17, 18], and the student-teacher formulation [12, 19].

2.2 Fisher information matrix (FIM)

We focus on a Fisher information matrix (FIM) of neural network models, which previous works have developed [20–24]. The FIM is defined by $F = \mathrm{E}[\nabla_\theta \log p(y|x;\theta) \nabla_\theta \log p(y|x;\theta)^T]$, where the statistical model $p(y|x;\theta)$ with the network output $h^L(x)$ parameterized by θ is given by $p(y|x;\theta) = \exp(-||y - h^L(x)||^2/2)/\sqrt{2\pi}$. Here, we denote the Euclidean norm as $||\cdot||$, the expectation with respect to input-output pairs (x, y) as $\mathrm{E}[\cdot]$, and all network parameters as $\theta = \{W^l_{ij}, b^l_i\}$. When T training samples $(x(t), y(t))$ $(t = 1, ..., T)$ are available and T is sufficiently large, we may replace the expectation by the empirical mean. In this case, we have

$$F = \sum_{k=1}^{C} (\nabla_\theta h^L_k)(\nabla_\theta h^L_k)^T / T, \tag{3}$$

where the derivative $\nabla_\theta h^L_k$ is a $P \times T$ matrix whose columns are the gradients on each input sample, i.e. $\nabla_\theta h^L_k(t)$ $(t = 1, ..., T)$, and P represents the total number of the parameters. Although the form of the FIM changes a bit in other statistical models including softmax outputs, basically, these differences are limited to the multiplication of activations in the output layer [24]. 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, i.e., the Cramér-Rao bound. Below, we summarize a more intuitive understanding of the FIM from geometric views.

Loss landscape view. The empirical FIM (3) determines the local landscape of the loss function at the global minimum. For instance, suppose we have a squared loss function $E(\theta) = (1/2T) \sum_t^T ||y(t) - h^L(t)||^2$. The FIM is related to the Hessian of the loss function in the following way:

$$H := \nabla_\theta \nabla_\theta E(\theta) = F - \frac{1}{T} \sum_t^T \sum_k^C (y_k(t) - h^L_k(t)) \nabla_\theta \nabla_\theta h^L_k(t). \tag{4}$$

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 with noise (i.e., $y(t) = h^L(t) + \varepsilon_t$, where ε_t denotes zero-mean Gaussian noise) [21].

Information geometric view. Let us define an infinitesimal squared distance dr^2, which represents how robust the output of a deep network is against a perturbation $d\theta$ of its parameters: $dr^2 = ||h^L(\theta + d\theta) - h^L(\theta)||^2 = d\theta^T F d\theta$. This quadratic form is equivalent to the Kullback-Liebler divergence between the statistical model $p(y|x;\theta)$ and $p(y|x;\theta + d\theta)$ [20]. It means that the FIM works as a Riemannian metric in the parameter space of a statistical model, studied in information geometry. Insights from information geometry have led to the development of natural gradient algorithms [23–25] and, recently, a capacity measure, called the Fisher-Rao norm [14].

3 Fundamental FIM statistics

Here, we expose mathematical findings that the mean, variance, and maximum of eigenvalues of the FIM (3) are obtained by mean field theory in an asymptotic situation $M \gg 1$ and $T \gg 1$. This situation seems realistic, because modern deep learning requires a large number of units in each layer as well as data samples. Surprisingly, our theorems are universal for networks ranging in size from shallow ($L = 2$) to arbitrarily deep ($L \geq 3$).

The FIM (3) of a deep network is computed by the chain rule in a manner similar to that of the backpropagation algorithm:

$$\frac{\partial h_k^L}{\partial W_{ij}^l} = \delta_i^l \phi(u_j^{l-1}), \tag{5}$$

$$\delta_k^L = \phi'(u_k^L), \;\; \delta_i^l = \phi'(u_i^l) \sum_j \delta_j^{l+1} W_{ji}^{l+1}, \;\; (l = 1, ..., L-1), \tag{6}$$

where $\delta_i^l := \partial h^L / \partial u_i^l$. All the theorems of this paper assume the following:

Assumption 1. *In the limit of $M \gg 1$, (a) the parameter set θ used in the backpropagated signals is independent of those used in the forward signals, (b) the backpropagated signal $\delta_i^{l+1}(t)$ is independent of the forward signal $h_j^l(t)$ ($l = 1, 2, ..., L-1$).*

This assumption hypothesizes that the correlations between backward and feedforward paths are very weak and that we can therefore regard them as totally independent. Although a mathematically rigorous justification is still lacking, previous works incorporated the same assumption into the mean field framework [26, Axiom 3.2] and demonstrated excellent agreements between their theory and experiments on analysis of the backpropagation algorithm [5, 26].

3.1 Mean of eigenvalues

First, we compute the arithmetic mean of the FIM's eigenvalues as $m_\lambda := \sum_{i=1}^P \lambda_i / P$. Let us define the following variables: $\hat{q}^l := \sum_i (h_i^l(t))^2 / M_l$ and $\tilde{q}^l := \sum_i (\delta_i^l(t))^2$. Taking the large M limit and the central limit theorem on u_i^l, the previous studies have proved that these variables obey the following recurrence relations, so they can be easily computed [2, 4, 5]:

$$\hat{q}^{l+1} = \int Du \phi^2 \left(\sqrt{q^{l+1}} u \right), \;\; q^{l+1} = \sigma_w^2 \hat{q}^l + \sigma_b^2, \;\; \hat{q}^0 = 1, \tag{7}$$

$$\tilde{q}^L = \int Du \left[\phi'(\sqrt{q^L} u) \right]^2, \;\; \tilde{q}^l = \sigma_w^2 \tilde{q}^{l+1} \int Du \left[\phi'(\sqrt{q^l} u) \right]^2, \tag{8}$$

for $l = 0, ..., L-1$.[4] The notation $Du = du \exp(-u^2/2)/\sqrt{2\pi}$ implies integration over the standard Gaussian density. The derivation of the recurrence (8) requires Assumption 1(a) [5]. These variables depend only on the variance parameters σ_w^2 and σ_b^2, not on the unit indices. In that sense, $\tilde{q}^l$ and $\hat{q}^l$ are *macroscopic variables* (a.k.a. order parameters in statistical physics).

Then, we find a hidden relation between the macroscopic variables and the statistics of FIM:

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

$$m_\lambda = K_1 / M, \;\; K_1 := C \sum_{l=1}^L \frac{\alpha_{l-1}}{\alpha} \tilde{q}^l \hat{q}^{l-1}, \tag{9}$$

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 m_λ is $O(1/M)$.

The proof is given by Supplementary Material A. The coefficient K_1 is a constant not depending on M, so it is $O(1)$ by using the $O(\cdot)$ order notation. It is easily computed by L iterations of the layer-wise recurrence relations (7)–(8).

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. Let us remind that the FIM determines the local geometry of the parameter space. The theorem suggests that the network output keeps 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 many dimensions. Furthermore, by using Markov's inequality, we can prove that the number of larger eigenvalues is limited as follows:

[4] Rigorously speaking, it is nontrivial exercise to apply the recurrence relations to $\tilde{q}^L$ and $\hat{q}^L$, because we have supposed that the output layer has $O(1)$ units. The integrals in $\tilde{q}^L$ and $\hat{q}^L$, however, give averaged or typical values of the output. Therefore, it would be natural to assume that they obey the recurrence relations.

Corollary 2. *Let us denote the number of eigenvalues satisfying $\lambda \geq k$ by $N(\lambda \geq k)$. For a constant $k > 0$, $N(\lambda \geq k) \leq \alpha K_1 M/k$ holds in the limit of $M \gg 1$.*

The proof is shown in Supplementary Material B. This corollary clarifies that the number of eigenvalues, whose values are $O(1)$, is $O(M)$ at most and thus much smaller than the total number of parameters P.

3.2 Variance of eigenvalues

Next, let us consider the second moment $s_\lambda := \sum_{i=1}^P \lambda_i^2/P$ and define variables $\hat{q}_{st}^l := \sum_i h_i^l(s)h_i^l(t)/M_l$ and $\tilde{q}_{st}^l := \sum_i \delta_i^l(s)\delta_i^l(t)$. This $\hat{q}_{st}^l$ is the correlation between the activations for different input samples $x(s)$ and $x(t)$ in the l-th layer [2, 4]. The activations due to $x(s)$ and $x(t)$ are not independent, but rather have correlations because they share the same weight vectors even for different inputs. $\tilde{q}_{st}^l$ is the correlation of backpropagated signals [5]. The previous studies proved that $\tilde{q}_{st}^l$ and $\hat{q}_{st}^l$ in the large M limit become macroscopic variables that can be easily computed using the following recurrence relations:

$$\hat{q}_{st}^{l+1} = I_\phi[q^{l+1}, q_{st}^{l+1}], \;\; q_{st}^{l+1} = \sigma_w^2 \hat{q}_{st}^l + \sigma_b^2, \;\; \hat{q}_{st}^0 = 0, \tag{10}$$

$$\tilde{q}_{st}^L = I_{\phi'}[q^L, q_{st}^L], \;\; \tilde{q}_{st}^l = \sigma_w^2 \tilde{q}_{st}^{l+1} I_{\phi'}[q^l, q_{st}^l], \tag{11}$$

for $l = 0, ..., L-1$. Here, the notation $I_\cdot[\cdot, \cdot]$ represents the following integral: $I_\phi[a, b] = \int Dz_1 Dz_2 \phi\left(\sqrt{a} z_1\right) \phi\left(\sqrt{a}(c z_1 + \sqrt{1-c^2} z_2)\right)$ with $c = b/a$. Note that the derivation of $\tilde{q}_{st}^l$ requires Assumption 1(a) [5]. These recurrence relations 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 C). We now show a novel fact that s_λ can be computed from these macroscopic variables:

Theorem 3. *In the limit of $M \gg 1$ and $T \gg 1$, the second moment of the FIM's eigenvalues is bounded by constants, that is, $K_2^{(lower)} \leq s_\lambda \leq K_2^{(upper)}$ where*

$$K_2^{(lower)} := C\alpha^{-1} \left(\sum_{l=1}^L \alpha_{l-1} \tilde{q}_{st}^l \hat{q}_{st}^{l-1} \right)^2, \;\; K_2^{(upper)} := \alpha K_1^2. \tag{12}$$

The macroscopic variables $\hat{q}_{st}^l$ and $\tilde{q}_{st}^l$ can be computed recursively, and s_λ is $O(1)$. [5]

The proof is shown in Supplementary Material D. The proof requires the large T condition in order to neglect the lower order term of $1/T$. From Theorems 1 and 2, we can conclude that the variance of the eigenvalues, $s_\lambda - m_\lambda^2$, is $O(1)$. Because the mean m_λ is $O(1/M)$ and most eigenvalues are close to zero, this result means the eigenvalue distribution has a long tail.

As is shown in Figure 1, we verified our theoretical results by numerical experiments. We investigated three different types of deep networks with parameters generated by Gaussian (2): tanh, ReLU, and linear activations ($L = 3$, $\alpha_l = C = 1$). The input samples were generated using i.i.d. Gaussian samples, and $T = 10^3$. When $P > T$, we calculated s_λ by using the dual matrix F^* shown in Supplementary Material D because F is huge. Theoretical values of m_λ agree very well with the experimental values. We could predict m_λ even for small M. In addition, Theorem 3 gives bounds covering the experimental values of s_λ.

3.3 The maximum eigenvalue

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

Theorem 4. *When $M \gg 1$ and $T \gg 1$, the maximum eigenvalue of the FIM is bounded by*

$$\sqrt{\alpha C^{-1} K_2^{(lower)}} M \leq \lambda_{max} \leq \sqrt{\alpha K_2^{(upper)}} M. \tag{13}$$

[5] Note that we have assumed $\sigma_b > 0$ in the setting (2). This is because deep networks in practice have bias terms and their values are scattered. When $\phi(x)$ is an odd function and $\sigma_b = 0$, $\hat{q}_{st}$ becomes 0 and the lower bound also becomes 0. Therefore, we need to evaluate the lower order terms in such exceptional cases separately.

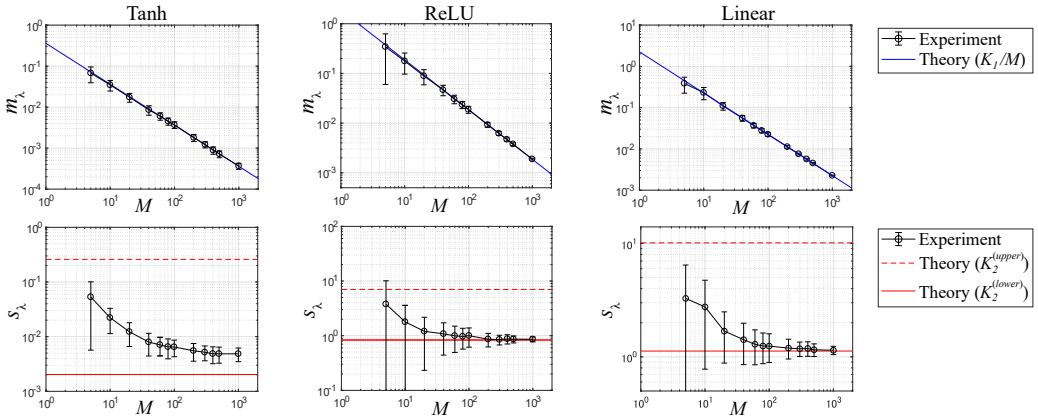

Figure 1: Statistics of the FIM's eigenvalues: means (upper row) and the second moments (lower row). Our theory predicts the results of numerical experiments, indicated by the black points and error bars. The experiments used 100 random ensembles with different seeds. The variances of the parameters were given by $(\sigma_w^2, \sigma_b^2) = (3, 0.64)$ in the tanh case, $(2, 0.5)$ in the ReLU case, and $(1, 0.5)$ in the linear case. Each colored line represents theoretical results.

The proof is shown in Supplemental Material F. From the geometric perspective, this theorem suggests that the local shape of the landscape is strongly distorted in certain direction. Here, let us remark on several previous studies on the Hessian of the loss, which coincides with the FIM at zero training error. LeCun et al. [13] empirically found that very large eigenvalues exist, i.e., "big killers". The eigenvalue distribution peaks around zero while its tail is very long; this behavior has been empirically known for decades, but theoretical explanations of it have been restricted to an analogy with the Marchenko-Pastur law of sample covariance matrices [15]. Therefore, our theory gives new theoretical evidence that this skewed eigenvalue distribution and its huge maximum appear universally in DNNs.

4 Connections to learning methods

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

4.1 Learning rate for convergence

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

$$\theta_{t+1} \leftarrow \theta_t - \eta \partial E(\theta_t)/\partial\theta + \mu(\theta_t - \theta_{t-1}), \tag{14}$$

where η is a constant learning rate. We have added a momentum term with a coefficient μ, which is often used to train deep networks. Suppose an expansion of the squared loss function $E(\theta)$ of Eq. (4), that is, $E(\theta) \simeq d\theta^T F(\theta^*) d\theta$, where global minimum θ^* achieves the zero training error $E(\theta^*) = 0$. The maximum eigenvalue is dominant over the convergence of learning as follows:

Lemma 5. *A learning rate satisfying $\eta < 2(1+\mu)/\lambda_{max}$ is necessary for the steepest gradient method to converge to the global minimum.*

The proof is shown in Supplementary Material F. This lemma is a generalization of [13]. 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. As is claimed in the previous work [13], $\eta = \eta_c/2$ is the best choice for fastest convergence. Note that, in the online regime, the eigenvalues also determine the bound of the gradient norms and the convergence of learning [27]. Now, the lower bound of λ_{max} in Theorem 4 leads to an upper bound of the critical learning rate. Then, we find,

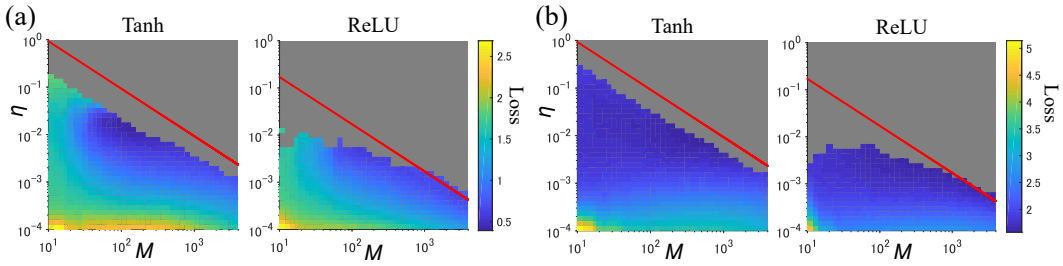

Figure 2: Color map of training losses after one epoch of training: (a) Tanh and ReLU networks on MNIST. (b) Tanh and ReLU networks on CIFAR-10. The losses are averages over five trials. The region where the loss diverges (i.e. is larger than 1000) is in gray. The red line shows the theoretical value of η_c^*. The initial conditions of the parameters were taken from a Gaussian distribution (2) with $(\sigma_w^2, \sigma_b^2) = (3, 0.64)$ in tanh networks and $(2, 0.5)$ in ReLU networks.

Theorem 6. *Suppose that we have a global minimum $\theta^* = \{W^{l*}, b^{l*}\}$, which satisfies Eq. (2) and $E(\theta^*) = 0$. In the limit of $M \gg 1$ and $T \gg 1$, the gradient method never converges to θ^* when*

$$\eta > \eta_c^*, \quad \eta_c^* := 2(1+\mu)/(\sqrt{\alpha C^{-1} K_2^{(lower)}} M). \tag{15}$$

Theorem 6 quantitatively reveals that, the wider the network becomes, the smaller the learning rate we need to set. In addition, $K_2^{(lower)}$ is the squared sum over L constant positive terms, so a deeper network requires a finer setting of the leaning rate and it makes the optimization harder. In contrast, the expressive power of the network grows exponentially as the number of layers increases [4, 28]. We thus expect that there is a trade-off, which decides the adequate number of layers, between trainability and expressive power.

To confirm the validity of Theorem 6 in practice, we investigated learning rates for convergence in training on MNIST and CIFAR-10. As shown in Figure 2, we exhaustively searched training losses after one epoch of training, while changing M and η. We trained deep networks ($L = 4$, $\alpha_l = 1$, $C = 10$) with linear outputs $h_i^L = u_i^L$. We used stochastic gradient descent (SGD) with a mini-batch size of 500, $\mu = 0.9$ and no regularization for simplicity. We used the whole training samples and normalized each sample $x(t)$ to zero mean and variance 1. We found that the losses of the experiments were clearly divided into two areas, one where the SGD exploded (gray area) and the other where it was converging (colored area). The red line is η_c^* theoretically calculated using $K_2^{(lower)}$ on (σ_w^2, σ_b^2) of the initial parameters. In general, global minima θ^* may not satisfy Eq. (2), and the variances (σ_w^2, σ_b^2) may change from the initialization to the global minimum. Nevertheless, we found that training on the regions above η_c^* exploded, just as Theorem 6 predicts. In addition, the explosive region with $\eta < \eta_c^*$ was narrow and, in particular, it got smaller in the ReLU networks in the limit of large M. We also confirmed that our theory can estimate learning rates in experiments on SGD with a smaller mini batch and a network with sigmoid outputs (see Supplementary Material G).

Recently, Schoenholz et al. [5] reported that mean field theory can predict appropriate initial values of parameters which quickly decreases losses on benchmark datasets. Our results suggest that mean field theory is also helpful in estimating an initial learning rate which prevents the gradient update from exploding. 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 for guessing the parameters necessary for learning rates.

4.2 The Fisher-Rao norm and generalization ability

Another natural question is how the statistics of the FIM are related to generalization ability. Recently, Liang et al. [14] proposed a capacity measure related to the FIM, called the Fisher-Rao norm,

$$||\theta||_{FR} = \theta^T F \theta, \tag{16}$$

where θ represents weight parameters. They proved that this norm has desirable properties to explain the high generalization capability of DNNs. In deep linear networks, its generalization error (Rademacher complexity) is upper bounded only by the norm and independent of the network size. In deep ReLU networks, the Fisher-Rao norm serves as a lower bound of the Rademacher complexities induced by other standard norms, such as the path norm [29] and the spectral norm [30]. Although the connection between the flat minima and generalization requires careful consideration of coordinate invariance [31], the Fisher-Rao norm is invariant under a linear coordinate transformation in the ReLU networks and can correctly reflect the flatness.

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 7. *In the limit of $M \gg 1$, the Fisher-Rao norm of a DNN satisfies*

$$\langle ||\theta||_{FR} \rangle_\theta = \sigma_w^2 C \sum_{l=1}^{L} \tilde{q}^l \hat{q}^{l-1}. \tag{17}$$

The proof is shown in Supplementary Material H. As an average evaluation, the norm is reduced to a simpler expression composed of the macroscopic variables. This guarantees that the norm is independent of the network width in the limit of $M \gg 1$, which was empirically conjectured by [14]. In addition, the coefficient K_1 upper-bounds the norm as follows:

$$\langle ||\theta||_{FR} \rangle_\theta \leq \sigma_w^2 \frac{\alpha}{\alpha_{min}} K_1, \tag{18}$$

where $\alpha_{min} = \min_i \alpha_i$. Equality holds in a network with a uniform width $M_l = M$. This inequality suggests that, as the FIM's eigenvalues become small, the norm also becomes small. In other words, the small K_1 realizes both the locally flat landscape and the better generalization ability.

5 Conclusion and discussion

The current work has widened the scope of mean field theory and elucidated the statistics of the Fisher information matrix common to a wide class of deep networks. In addition, we have theoretically evaluated an adequate size of the learning rate by using the derived statistics and it has coincided well with the numerical experiments on the benchmark datasets. Furthermore, we have characterized the Fisher-Rao norm, which is intimately related to the generalization ability of DNNs.

The derived statistics are also of potential importance to 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 [21–24] . Several experiments on the natural gradient in DNNs showed that the choice of damping term ε, introduced in $(F + \varepsilon I)^{-1}\nabla_\theta E(\theta)$, is crucial to its performance [25]. Since we found that the FIM has many eigenvalues close to zero, any naive inversion of it would be very unstable. Therefore, development of more efficient gradient methods will require modification such as damping.

Here, we should also remark on extremely deep networks. Mean field theory shows that macroscopic variables at extreme depths converge to finite values in the vicinity of the special parameter region, i.e. the critical line of the order-to-chaos transition, where deep networks achieve high expressive power and trainability [4, 5]. In contrast, macroscopic variables in residual networks essentially diverge at the extreme depths [26]. Our framework on FIMs is readily applicable to residual networks; what is required will be a careful examination of the order of the network width and the explosion of the macroscopic variables.

It would also be interesting to extend our theory to batch normalization, which empirically allows larger learning rates and achieves better generalization performance [32]. The global structure of the parameter space should be also explored. As the small mean of the eigenvalues suggests, the typical shape of the loss landscape would be locally flat in many dimensions. Therefore, we can hypothesize that the parameters are globally connected through the flat dimensions and compose manifolds of flat minima. It would also be fruitful to investigate the eigenvalues of the Hessian (4) with a large error and to theoretically quantify the negative eigenvalues that lead to the existence of saddle points and local minima-free landscapes [33]. We expect that further studies will establish a mathematical foundation of deep learning from the perspective of mean field theory.

Acknowledgments

This work was 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] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural networks. *Physical review letters*, 61(3):259, 1988.

[4] 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.

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

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

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

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

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

[10] 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.

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

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

[13] 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.

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

[15] 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.

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

[17] 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.

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

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

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

[21] 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.

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

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

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

[25] 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.

[26] Ge 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.

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

[28] 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.

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

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

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

[32] 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.

[33] 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 *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 2933–2941, 2014.

Supplementary Materials

In the following proofs, we omit the layer index L of h_i^L as $h_i = h_i^L$.

A Proof of Theorem 1

Case of $C = 1$

To avoid complicating the notation, we first consider the case of $C = 1$. The general case is shown after. The sum over the eigenvalues is given by the matrix trace, $m_\lambda = \mathrm{Tr}(F)/P$. We denote the Fisher information matrix with full components as

$$F = \begin{bmatrix} \nabla_W h \nabla_W h^T & \nabla_W h \nabla_b h^T \\ \nabla_b h \nabla_W h^T & \nabla_b h \nabla_b h^T \end{bmatrix} /T, \tag{A.1}$$

where we notice that

$$\nabla_{b_i^l} h = \delta_i^l. \tag{A.2}$$

In addition, we denote the average of the eigenvalues of $\nabla_W h \nabla_W h^T$ as $m_\lambda^{(W)}/T$, and that of $\nabla_b h \nabla_b h^T$ as $m_\lambda^{(b)}/T$. Accordingly, we find

$$m_\lambda = m_\lambda^{(W)} + m_\lambda^{(b)}. \tag{A.3}$$

The first term is

$$m_\lambda^{(W)} = \mathrm{Tr}(\nabla_W h \nabla_W h^T/T)/P \tag{A.4}$$

$$= \sum_l \langle \sum_{i,j} \delta_i^l(t)^2 h_j^{l-1}(t)^2 \rangle /P, \tag{A.5}$$

where $\langle \cdot \rangle = \sum_t \cdot /T$ represents the empirical average with respect to T samples. Under Assumption 1(b), we can compute$\sum_i \delta_i^l(t)^2$ and $\sum_j h_j^{l-1}(t)^2$ independently in Eq. (A.5). By taking the mean field limit ($M \gg 1$), we obtain

$$m_\lambda^{(W)} = K/M, \;\; K := \sum_{l=1}^{L} \frac{\alpha_{l-1}}{\alpha} \tilde{q}^l \hat{q}^{l-1}. \tag{A.6}$$

In contrast, the contributions of the bias terms are smaller than those of the weight terms in the mean field limit ($M \gg 1$):

$$m_\lambda^{(b)} = \mathrm{Tr}(\langle \nabla_b h \nabla_b h^T \rangle)/P \tag{A.7}$$

$$= \sum_l \sum_i^M \langle (\delta_i^l)^2 \rangle /P \tag{A.8}$$

$$= \sum_l \alpha_l \tilde{q}^l/(\alpha M^2) \quad (\text{when } \; M \gg 1). \tag{A.9}$$

Note that $m_\lambda^{(W)}$ is $O(1/M)$, but $m_\lambda^{(b)}$ is $O(1/M^2)$. Hence, the mean $m_\lambda^{(b)}$ is negligible and we obtain $m_\lambda = K/M$.

General case of C

Next, we consider an empirical Fisher information matrix of the multi-class cases. In the framework of mean field theory, each ∇h_k^L reduces to the same macroscopic variables, independently of the microscopic indices k:

$$\mathrm{Tr}(\nabla_\theta h_k \nabla_\theta h_k^T/T)/P = K/M. \tag{A.10}$$

Therefore, the mean of the eigenvalues becomes

$$m_\lambda = CK/M, \tag{A.11}$$

and hence, $K_1 := CK$. ■

B Proof of Corollary 2

Because the FIM is a positive semi-definite matrix, its eigenvalues are non-negative. For a constant $k > 0$, we obtain

$$m_\lambda = \frac{1}{P}\left(\sum_{i;\lambda_i<k}\lambda_i + \sum_{i;\lambda_i\geq k}\lambda_i\right) \tag{B.1}$$

$$\geq \frac{1}{P}\sum_{i;\lambda_i\geq k}\lambda_i \tag{B.2}$$

$$\geq \frac{1}{P}N(\lambda \geq k)k. \tag{B.3}$$

This is known as Markov's inequality. When $M \gg 1$, combining this with Theorem 1 immediately yields Corollary 2:

$$N(\lambda \geq k) \leq \alpha K_1 M/k. \tag{B.4}$$

■

C Analytical recurrence relations

C.1 Erf networks

Consider the following error function as an activation function $\phi(x)$:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)dt. \tag{C.1}$$

The error function well approximates the tanh function and has a sigmoid-like shape. For a network with $\phi(x) = \mathrm{erf}(x)$, the recurrence relations for macroscopic variables do not require numerical integrations as follows.

(i) $\hat{q}^l$ and $\tilde{q}^l$: Note that we can analytically integrate the error functions over a Gaussian distribution:

$$\int_0^\infty Dx\mathrm{erf}(ax)\mathrm{erf}(bx) = \frac{1}{\pi}\tan^{-1}\frac{\sqrt{2}ab}{\sqrt{a^2+b^2+1/2}}. \tag{C.2}$$

Hence, the recurrence relations for the feedforward signals (7) have the following analytical forms:

$$\hat{q}^{l+1} = \frac{2}{\pi}\tan^{-1}\left(\frac{q^{l+1}}{\sqrt{q^{l+1}+1/4}}\right), \;\; q^{l+1} = \sigma_w^2\hat{q}^l + \sigma_b^2. \tag{C.3}$$

Because the derivative of the error function is Gaussian, we can also easily integrate $\phi'(x)$ over the Gaussian distribution and obtain the following analytical representations of the recurrence relations (8):

$$\tilde{q}^l = \frac{2\tilde{q}^{l+1}\sigma_w^2}{\pi\sqrt{q^l+1/4}}, \;\; \tilde{q}^L = \frac{2}{\pi\sqrt{q^L+1/4}}. \tag{C.4}$$

(ii) $\hat{q}^l_{st}$ and $\tilde{q}^l_{st}$:

To compute the recurrence relations for the feedforward correlations (10), note that we can generally transform $I_\phi[a,b]$ into

$$I_\phi[a,b] = \int Dy\left(\int Dx\phi(\sqrt{a-b}x + \sqrt{b}y)\right)^2. \tag{C.5}$$

For the error function,

$$\int Dx\phi(\sqrt{a-b}x + \sqrt{b}y) = \mathrm{erf}\frac{\sqrt{b}y}{\sqrt{1+2a-2b}}, \tag{C.6}$$

and we obtain

$$I_\phi[a,b] = \frac{2}{\pi}\tan^{-1}\frac{2b}{\sqrt{(1+2a)^2-(2b)^2}}. \tag{C.7}$$

Substituting Eq. (C.7) into Eq. (10) leads to the analytical form of the recurrence relation for $\hat{q}^l_{st}$.

Finally, because the derivative of the error function is Gaussian, we can also easily obtain

$$I_{\phi'}[a,b] = \frac{4}{\pi\sqrt{(1+2a)^2-(2b)^2}}. \tag{C.8}$$

Substituting Eq. (C.8) into Eqs. (11) leads to the analytical forms of the recurrence relations for $\tilde{q}^l_{st}$.

C.2 ReLU networks

We define a ReLU activation as $\phi(x) = 0 \ \ (x < 0), \ \ x \ \ (0 \leq x)$. For a network with this ReLU activation function, the recurrence relations for the macroscopic variables require no numerical integrations. In particular, we can analytically obtain $\hat{q}^l$, $\tilde{q}^l$, and K_1 as follows.

(i) $\hat{q}^l$ and $\tilde{q}^l$: We can explicitly perform the integrations in the recurrence relations (7) and (8):

$$\hat{q}^{l+1} = \hat{q}^l\sigma_w^2/2 + \sigma_b^2/2, \tag{C.9}$$

$$\tilde{q}^l = \tilde{q}^{l+1}\sigma_w^2/2, \ \ \tilde{q}^L = 1/2. \tag{C.10}$$

(ii) $\hat{q}^l_{st}$ and $\tilde{q}^l_{st}$: We can explicitly perform the integrations for $\hat{q}^l_{st}$ and $\tilde{q}^l_{st}$ and obtain

$$I_\phi[a,b] = \frac{a}{2\pi}(\sqrt{1-c^2} + c\pi/2 + c\sin^{-1}c), \tag{C.11}$$

$$I_{\phi'}[a,b] = \frac{a}{2\pi}(\pi/2 + \sin^{-1}c), \tag{C.12}$$

where $c = b/a$. Substituting them into (10) and (11) leads to analytical forms of the recurrence relations for $\hat{q}^l_{st}$ and $\tilde{q}^l_{st}$.

C.3 Linear networks

We define a linear activation as $\phi(x) = x$. For a network with this linear activation function, the recurrence relations for the macroscopic variables do not require numerical integrations. Moreover, we can analytically obtain all of the macroscopic variables, K_1, $K_2^{(upper)}$, and $K_2^{(lower)}$ as follows.

(i) $\hat{q}^l$ and $\tilde{q}^l$: We can explicitly perform the integrations in the recurrence relations (7) and (8):

$$q^l = q^{l-1}\sigma_w^2 + \sigma_b^2, \tag{C.13}$$

$$\tilde{q}^l = \tilde{q}^{l+1}\sigma_w^2, \ \ \tilde{q}^L = 1. \tag{C.14}$$

(ii) $\hat{q}^l_{st}$ and $\tilde{q}^l_{st}$: We can explicitly perform the integrations in the recurrence relations (10) and (11):

$$\hat{q}^{l+1}_{st} = \hat{q}^l_{st}\sigma_w^2 + \sigma_b^2, \tag{C.15}$$

$$\tilde{q}^l_{st} = \tilde{q}^{l+1}_{st}\sigma_w^2, \ \ \tilde{q}^L_{st} = 1. \tag{C.16}$$

D Proof of Theorem 3

D.1 Derivation of the lower bound $K_2^{(lower)}$

Case of $C = 1$:

By using the Frobenius norm $||A||_F = \sqrt{\sum_{ij} A_{ij}^2}$, we have $s_\lambda = ||F||_F^2/P$, in general. The non-negativity of the norm gives the lower bound

$$s_\lambda = s_\lambda^{(W)} + \frac{1}{P}(2||\langle\nabla_W h \nabla_b h^T\rangle||_F^2 + ||\langle\nabla_b h \nabla_b h^T\rangle||_F^2) \tag{D.1}$$

$$\geq s_\lambda^{(W)}, \tag{D.2}$$

where $s_\lambda^{(W)} = ||\langle \nabla_W h \nabla_W h^T \rangle||_F^2 / P$. Note that this inequality is asymptotically tight, and equality holds when $M \gg 1$ because the number of bias parameters is much smaller than that of the weight parameters.

Instead of F, consider

$$F^* := \frac{1}{T}(\nabla_W h)^T (\nabla_W h). \tag{D.3}$$

Because F^* and F have the same nonzero eigenvalues, we have

$$s_\lambda^{(W)} = ||F^*||_F^2 / P, \tag{D.4}$$

which leads to

$$\begin{aligned} s_\lambda^{(W)} &= ||F^*||_F^2 / P & \text{(D.5)} \\ &= \frac{1}{PT^2} \sum_{s,t} (\sum_l \sum_{ij} \nabla_{W_{ij}^l} h(s) \nabla_{W_{ij}^l} h(t))^2 & \text{(D.6)} \\ &= \frac{1}{PT^2} \sum_{s,t} (\sum_l \sum_i \delta_i^l(s) \delta_i^l(t) \sum_j h_j^{l-1}(s) h_j^{l-1}(t))^2. & \text{(D.7)} \end{aligned}$$

Thus, computing the variance of the FIM's eigenvalues reduces to computing correlations between different input samples $x(s)$ and $x(t)$.

We define

$$\hat{Z}_{st}^l := \frac{1}{M_l} \sum_j h_j^l(s) h_j^l(t), \quad \tilde{Z}_{st}^l := \sum_i \delta_i^l(s) \delta_i^l(t). \tag{D.8}$$

For $s \neq t$, we have $\hat{Z}_{st}^l = \hat{q}_{st}^l$ and $\tilde{Z}_{st}^l = \tilde{q}_{st}^l$ in the limit of $M \gg 1$ where the macroscopic variables $\hat{q}_{st}^l$ and $\tilde{q}_{st}^l$ satisfy the recurrence relations (10) and (11). For $s = t$, we have $\hat{Z}_{tt}^l = \hat{q}^l$ and $\tilde{Z}_{tt}^l = \tilde{q}^l$ satisfying the recurrence relations (7) and (8).

Using $\hat{Z}_{st}^l$ and $\tilde{Z}_{st}^l$, we obtain

$$s_\lambda^{(W)} = \frac{M^2}{P} \left\{ \langle \sum_l \alpha_{l-1} \tilde{Z}_{st}^l \hat{Z}_{st}^{l-1} \rangle^2 + \mathrm{V}_{st}[\sum_l \alpha_{l-1} \tilde{Z}_{st}^l \hat{Z}_{st}^{l-1}] \right\}, \tag{D.9}$$

where we have defined the expectation over the samples s and t by $\langle a(s,t) \rangle := (1/T^2) \sum_{s,t} a(s,t)$, and the variance by $\mathrm{V}_{st}[a(s,t)] := (1/T^2) \sum_{s,t} a(s,t)^2 - \langle a(s,t) \rangle^2$. Regarding the first term of Eq. (D.9), we obtain

$$\begin{aligned} \langle \sum_l \alpha_{l-1} \tilde{Z}_{st}^l \hat{Z}_{st}^{l-1} \rangle &= \sum_l \alpha_{l-1} \langle \tilde{Z}_{st}^l \hat{Z}_{st}^{l-1} \rangle & \text{(D.10)} \\ &= \sum_l \alpha_{l-1} \frac{1}{T^2} \left(\sum_{s \neq t} \tilde{Z}_{st}^l \hat{Z}_{st}^{l-1} + \sum_t \tilde{Z}_{tt}^l \hat{Z}_{tt}^{l-1} \right) & \text{(D.11)} \end{aligned}$$

Under Assumption 1(b), we can neglect the dependence between the neighboring layers. Therefore, we can compute $\tilde{Z}_{st}^l$ and $\hat{Z}_{st}^{l-1}$ separately:

$$\begin{aligned} \langle \sum_l \alpha_{l-1} \tilde{Z}_{st}^l \hat{Z}_{st}^{l-1} \rangle &= \sum_l \alpha_{l-1} \frac{1}{T^2} \left(\sum_{s \neq t} \tilde{q}_{st}^l \hat{q}_{st}^{l-1} + \sum_t \tilde{q}^l \hat{q}^{l-1} \right) & \text{(D.12)} \\ &= \sum_l \alpha_{l-1} \frac{1}{T^2} \left(T(T-1) \tilde{q}_{st}^l \hat{q}_{st}^{l-1} + T \tilde{q}^l \hat{q}^{l-1} \right) & \text{(D.13)} \\ &= \sum_l \alpha_{l-1} \tilde{q}_{st}^l \hat{q}_{st}^{l-1} \quad (\text{when } T \gg 1). & \text{(D.14)} \end{aligned}$$

Note that, even under Assumption 1, the second term of Eq. (D.9) is hard to evaluate analytically because it is a sum of all covariances between different layers;

$$\mathrm{V}_{st}[\sum_l \alpha_{l-1}\tilde{Z}^l_{st}\hat{Z}^{l-1}_{st}] = \sum_{l,k}\alpha_{l-1}\alpha_{k-1}\mathrm{Cov}_{st}[\tilde{Z}^l_{st}\hat{Z}^{l-1}_{st}, \tilde{Z}^k_{st}\hat{Z}^{k-1}_{st}]. \tag{D.15}$$

where we have defined the covariance over the samples s and t by by $\mathrm{Cov}_{st}[a(s,t), b(s,t)] := (1/T^2)\sum_{s,t} a(s,t)b(s,t) - \langle a(s,t)\rangle\langle b(s,t)\rangle$. Nevertheless, noting that a variance is always non-negative, we obtain the following lower bound by substituting Eq. (D.14) into Eq. (D.9):

$$s_\lambda \geq \alpha^{-1}\left(\sum_l \alpha_{l-1}\tilde{q}^l_{st}\hat{q}^{l-1}_{st}\right)^2 \tag{D.16}$$

$$:= K_2 \tag{D.17}$$

where the coefficient α^{-1} appears from $M^2/P = \alpha^{-1}$ in the limit of $M \gg 1$.

General case of C:

The lower bound of the general case is reduced to that of $C = 1$ as follows:

$$s_\lambda = ||\sum_i^C \langle \nabla_\theta h_i(t)\nabla_\theta h_i(t)^T\rangle||^2_F/P \tag{D.18}$$

$$= \sum_{i,j}^C \frac{1}{T^2P}\sum_{s,t}(\nabla_\theta h_i(t)^T\nabla_\theta h_j(s))^2 \tag{D.19}$$

$$\geq \sum_i^C \frac{1}{T^2P}\sum_{s,t}(\nabla_\theta h_i(t)^T\nabla_\theta h_i(s))^2 \tag{D.20}$$

$$= CK_2. \tag{D.21}$$

On the last line, we have used the result of $C = 1$ under $M \gg 1$, $T \gg 1$ and Assumption 1. Hence, we obtain the lower bound $K_2^{(lower)} := CK_2$.

D.2 Derivation of the upper bound $K_2^{(upper)}$

Because F is a positive semi-definite matrix by definition, we obtain

$$||F||_F \leq \mathrm{Trace}(F). \tag{D.22}$$

Moreover, using $m_\lambda = \mathrm{Trace}(F)/P$, $s_\lambda = ||F||^2_F/P$, we have

$$s_\lambda \leq \mathrm{Trace}(F)^2/P \tag{D.23}$$

$$= m_\lambda^2 P \tag{D.24}$$

$$= \alpha K_1^2 \tag{D.25}$$

$$:= K_2^{(upper)}, \tag{D.26}$$

where we have used Theorem 1 on the third line. ■

E Proof of Theorem 4

E.1 Derivation of the Lower bound

Let us denote the maximum eigenvalue by λ_{max}. Here, we introduce the following matrix F^*:

$$F^* := B^TB/T, \tag{E.1}$$

$$B := [\nabla_\theta h_1 \ \nabla_\theta h_2 \ \cdots \ \nabla_\theta h_C], \tag{E.2}$$

where $\nabla_\theta h_k$ is a $P \times T$ matrix whose columns are the gradients on each input sample, i.e., $\nabla_\theta h_k^L(t)$ $(t = 1, ..., T)$ and B is a $P \times CT$ matrix. The FIM is represented by $F = BB^T/T$. We can easily confirm that these F and F^* have the same maximum eigenvalue. In general, we can obtain it as follow:

$$\lambda_{max} = \max_{\mathbf{v};||\mathbf{v}||^2=1} \mathbf{v}^T F^* \mathbf{v}. \tag{E.3}$$

Then, we find

$$\lambda_{max} \geq \mathbf{v}_1^T F^* \mathbf{v}_1 \;\; s.t. \;\; \mathbf{v}_1 = (1, 1, ..., 1)^T/\sqrt{CT} \tag{E.4}$$

$$= \frac{1}{CT^2} \sum_k^C \sum_{s,t} \sum_l \sum_j h_j^l(s) h_j^l(t) \sum_i \delta_i^l(s) \delta_i^l(t) \tag{E.5}$$

$$= \frac{M}{T^2} \sum_l \alpha_{l-1} \left(\sum_{s \neq t} \tilde{Z}_{st}^l \hat{Z}_{st}^{l-1} + \sum_t \tilde{Z}_{tt}^l \hat{Z}_{tt}^{l-1} \right). \tag{E.6}$$

Considering $M \gg 1, T \gg 1$ and Assumption 1, we can transform Eq. (E.6) into the following in the same way as Eqs. (D.10)–(D.14):

$$\lambda_{max} \geq M \sum_l \alpha_{l-1} \tilde{q}_{st}^l \hat{q}_{st}^{l-1}. \tag{E.7}$$

By substituting the definition of $K_2^{(lower)}$, the lower bound becomes

$$\lambda_{max} \geq \sqrt{\alpha C^{-1} K_2^{(lower)}} M. \tag{E.8}$$

E.2 Deviation of the Upper bound

Because the FIM is a positive semi-definite matrix, $\lambda_i \geq 0$ holds by definition. Then, we have $\lambda_{max} \leq \sqrt{\sum_i \lambda_i^2}$. Theorem 3 gives $\sqrt{\sum_i \lambda_i^2/P} \leq \sqrt{K_2^{(upper)}}$ and we have $\lambda_{max} \leq \sqrt{\alpha K_2^{(upper)}} M$. ■

F Proof of Lemma 5

Suppose a perturbation around the global minimum: $\theta_t = \theta^* + \Delta_t$. Then, the gradient update becomes

$$\Delta_{t+1} \leftarrow (I - \eta F)\Delta_t + \mu(\Delta_t - \Delta_{t-1}), \tag{F.1}$$

where we have used $E(\theta^*) = 0, \partial E(\theta^*)/\partial\theta = 0$.

Consider a coordinate transformation from Δ_t to $\bar{\Delta}_t$ which diagonalizes F. It does not change the stability of the gradients. Accordingly, we can update the i-th component as follows:

$$\bar{\Delta}_{t+1,i} \leftarrow (1 - \eta\lambda_i + \mu)\bar{\Delta}_{t,i} - \mu\Delta_{t-1,i}. \tag{F.2}$$

Solving its characteristic equation, we obtain the general solution,

$$\bar{\Delta}_{t,i} = A\lambda_+^t + B\lambda_-^t, \;\; \lambda_\pm = (1 - \eta\lambda_i + \mu \pm \sqrt{(1 - \eta\lambda_i + \mu)^2 - 4\mu})/2, \tag{F.3}$$

where A and B are constants. This recurrence relation converges if and only if $\eta\lambda_i < 2(1 + \mu)$ for all i. Therefore, $\eta < 2(1 + \mu)/\lambda_{max}$ is necessary for the steepest gradient to converge to θ^*. ■

G Additional Experiments

G.1 Sigmoid outputs

Consider the case where the output units are given by $h_i^L = \mathrm{Sigmoid}(u_i^L)$. This experimental setting is similar to Figure 2, except we set $\mu = 0$ here. Note that the loss, i.e., the squared loss with the sigmoid output, is always bounded and the regions above η_c^* do not explode by definition. As shown in Figure G.1, the theoretical values η_c^* well estimate the boundary between the lower loss and higher loss regions.

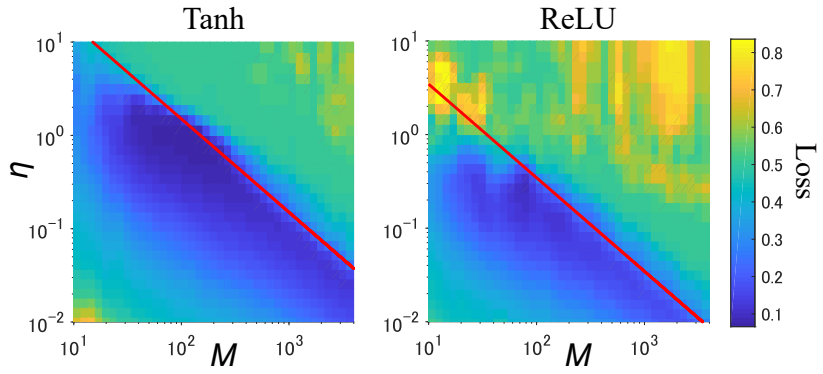

Figure G.1: Color map of training losses after one epoch of training: (a) Tanh and ReLU networks with sigmoid outputs trained on MNIST.

G.2 SGD with a smaller mini batch

Theorem 6 is based on steepest gradient descent on the batch regime. Nevertheless, it can explain the experimental results even in SGD training with a smaller mini-batch size of 20 (see Figure G.2).

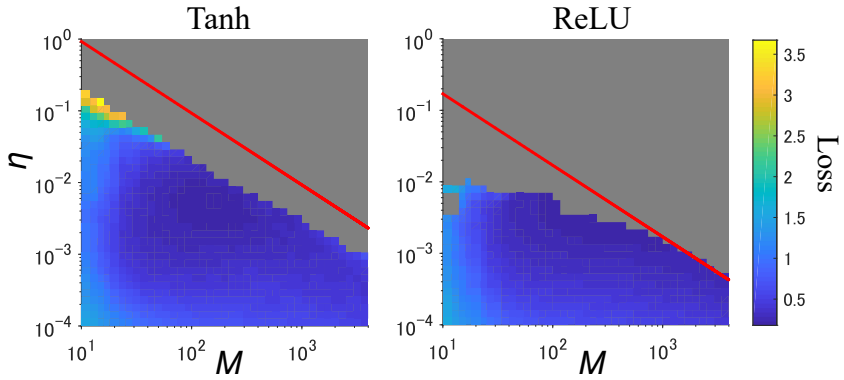

Figure G.2: Color map of training losses after one epoch of training with a smaller mini batch: Tanh and ReLU networks trained on MNIST.

H Proof of Theorem 7

The Fisher-Rao norm is written as

$$||\theta||_{FR} = \sum_{l,ij} \sum_{k,ab} F_{(l,ij),(k,ab)} W^l_{ij} W^k_{ab}, \tag{H.1}$$

where $F_{(l,ij),(k,ab)}$ represents an entry of the FIM, that is, $\sum_s^C \langle \nabla_{W^l_{ij}} h_s \nabla_{W^k_{ab}} h_s \rangle$.

(i) Case of $(l,ij) \neq (k,ab)$: Note that W^l_{ij} and W^k_{ab} are infinitesimals generated by Eq. (2). Performing a Taylor expansion around W^l_{ij} and W^k_{ab}, we obtain

$$F_{(l,ij),(k,ab)}(\theta) = F_{(l,ij),(k,ab)}(\theta^*) + \frac{\partial F_{(l,ij),(k,ab)}}{\partial W^l_{ij}}(\theta^*) W^l_{ij} + \frac{\partial F_{(l,ij),(k,ab)}}{\partial W^k_{ab}}(\theta^*) W^k_{ab} + \text{higher-chorder terms}, \tag{H.2}$$

where θ^* is the parameter set $\{W_{ij}^l, b_i^l\}$ with $W_{ij}^l = W_{ab}^k = 0$. By substituting the above expansion into the Fisher-Rao norm and taking the average $\langle\cdot\rangle_\theta$, we obtain

$$\langle F_{(l,ij),(k,ab)} W_{ij}^l W_{ab}^k \rangle_\theta = \langle F_{(l,ij),(k,ab)}(\theta^*) W_{ij}^l W_{ab}^k \rangle_\theta \tag{H.3}$$

$$= \langle F_{(l,ij),(k,ab)}(\theta^*) \rangle_{\theta^*} \langle W_{ij}^l W_{ab}^k \rangle_{\theta=\{W_{ij}^l, W_{ab}^k\}} \tag{H.4}$$

$$= 0. \tag{H.5}$$

In the last line, we have used $\langle W_{ij}^l W_{ab}^k \rangle_{\theta=\{W_{ij}^l, W_{ab}^k\}} = \langle W_{ij}^l \rangle_{W_{ij}^k} \langle W_{ab}^k \rangle_{W_{ab}^k} = 0$.

(ii) Case of $(l, ij) = (k, ab)$ **:** Here, considering the parameter set $\{W_{ij}^l, b_i^l\}$ with $W_{ij}^l = 0$ as θ^* and performing a Taylor expansion around W_{ij}^l, we obtain

$$\langle F_{(l,ij),(l,ij)} (W_{ij}^l)^2 \rangle_\theta = \frac{\sigma_w^2}{M_{l-1}} \langle F_{(l,ij),(l,ij)}(\theta^*) \rangle_{\theta^*}. \tag{H.6}$$

In addition, the limit of $M \gg 1$ makes $\langle F_{(l,ij),(l,ij)}(\theta) \rangle_\theta$ asymptotically equal to $\langle F_{(l,ij),(l,ij)}(\theta^*) \rangle_{\theta^*}$. Then, the norm becomes

$$||\theta||_{FR} = \sigma_w^2 C \sum_l \langle \sum_i (\delta_i^l)^2 \frac{1}{M_{l-1}} \sum_j (\phi_j^{l-1})^2 \rangle_\theta. \tag{H.7}$$

Under Assimption 1(b), the dependence between the neighboring layers becomes negligible and we obtain

$$||\theta||_{FR} = \sigma_w^2 C \sum_l \langle \tilde{q}^l \rangle_\theta \langle \hat{q}^{l-1} \rangle_\theta \tag{H.8}$$

$$= \sigma_w^2 C \sum_l \tilde{q}^l \hat{q}^{l-1}. \tag{H.9}$$

■