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

Supplementary Materials

A Proofs

A.1 Theorem 1

(i) Case of $C = 1$

To avoid complicating the notation, we first consider the case of the single output ($C = 1$). The general case is shown after. The network output is denoted by $f(t)$ here. We denote the Fisher information matrix with full components as

$$F = \sum_{t=1}^{T} \begin{bmatrix} \nabla_W f(t) \nabla_W f(t)^T & \nabla_W f(t) \nabla_b f(t)^T \\ \nabla_b f(t) \nabla_W f(t)^T & \nabla_b f(t) \nabla_b f(t)^T \end{bmatrix} /T, \tag{A.1}$$

where we notice that

$$\nabla_{b_i^l} f(t) = \delta_i^l(t). \tag{A.2}$$

In general, the sum over the eigenvalues is given by the matrix trace, $m_\lambda = \mathrm{Trace}(F)/P$. We also denote the average of the eigenvalues of the diagonal block as $m_\lambda^{(W)}$ for $\nabla_W f \nabla_W f^T$, and $m_\lambda^{(b)}$ for $\nabla_b f \nabla_b f^T$. Accordingly, we find

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

The contribution of $m_\lambda^{(b)}$ is negligible in the large M limit as follows. The first term is

$$\begin{aligned} m_\lambda^{(W)} &= \sum_{t=1}^{T} \mathrm{Trace}(\nabla_W f(t) \nabla_W f(t)^T)/(TP) &\text{(A.4)} \\ &= \sum_{t=1}^{T} \sum_l \sum_{i,j} \delta_i^l(t)^2 h_j^{l-1}(t)^2/(TP). &\text{(A.5)} \end{aligned}$$

We can apply the central limit theorem to summations over the units $\sum_i \delta_i^l(t)^2$ and $\sum_j h_j^{l-1}(t)^2$ independently because they do not share the index of the summation. By taking the limit of $M \gg 1$, we obtain $\sum_i \delta_i^l(t)^2 \sum_j h_j^{l-1}(t)^2/M_{l-1} = \tilde{q}^l \hat{q}^{l-1}$. The variable $\tilde{q}^l$ is computed by the recursive relation (9). Under the Assumption 1, $\hat{q}^{l-1}$ is given by the recursive relation (11). Note that this transformation to the macroscopic variables holds regardless of the sample index t. Therefore, we obtain

$$m_\lambda^{(W)} = \kappa_1/M, \quad \kappa_1 := \sum_{l=1}^{L} \frac{\alpha_{l-1}}{\alpha} \tilde{q}^l \hat{q}^{l-1}, \tag{A.6}$$

where α_l comes from $M_l = \alpha_l M$, and α comes from $P = \alpha M^2$.

In contrast, the contributions of the bias entries are smaller than those of the weight entries in the limit of $M \gg 1$, as is easily confirmed:

$$m_\lambda^{(b)} = \sum_t \mathrm{Trace}(\nabla_b f(t) \nabla_b f(t)^T)/(TP) \tag{A.7}$$
$$= \sum_t \sum_l \sum_i \delta_i^l(t)^2/(TP) \tag{A.8}$$
$$= \sum_l \tilde{q}^l/(\alpha M^2) \quad (\text{when} \ \ M \gg 1). \tag{A.9}$$

$m_\lambda^{(W)}$ is $O(1/M)$ while $m_\lambda^{(b)}$ is $O(1/M^2)$. Hence, the mean $m_\lambda^{(b)}$ is negligible and we obtain $m_\lambda = \kappa_1/M$.

(ii) $C > 1$ **of** $O(1)$

We can apply the above computation of $C = 1$ to each network output ∇f_k $(k = 1, ..., C)$:

$$\mathrm{Trace}(\nabla_\theta f_k \nabla_\theta f_k^T/T)/P = \kappa_1/M. \tag{A.10}$$

Therefore, the mean of the eigenvalues becomes

$$m_\lambda = \sum_k^C \mathrm{Trace}(\nabla_\theta f_k \nabla_\theta f_k^T/T)/P \tag{A.11}$$
$$= C\kappa_1/M. \tag{A.12}$$

■

A.2 Corollary 2

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

$$m_\lambda = \frac{1}{P}\left(\sum_{i;\lambda_i<k} \lambda_i + \sum_{i;\lambda_i \geq k} \lambda_i\right) \tag{A.13}$$
$$\geq \frac{1}{P}\sum_{i;\lambda_i \geq k} \lambda_i \tag{A.14}$$
$$\geq \frac{1}{P} N(\lambda \geq k)k. \tag{A.15}$$

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

$$N(\lambda \geq k) \leq \alpha\kappa_1 CM/k. \tag{A.16}$$

■

A.3 Theorem 3

Let us describe the outline of the proof. One can express the FIM as $F = (BB^T)/T$ by definition. Here, let us consider a dual matrix of F, that is, $F^* := (B^T B)/T$. F and F^* have the same nonzero eigenvalues. Because the sum of squared eigenvalues is equal to $\mathrm{Trace}(F^*(F^*)^T)$, we have $s_\lambda = \sum_{s,t}^T (F_{st}^*)^2/P$. The non-diagonal entry F_{st}^* $(s \neq t)$ corresponds to an inner product of the network activities for different inputs $x(s)$ and $x(t)$, that is, κ_2. The diagonal entry F_{ss}^* is given by κ_1. Taking the summation of $(F_{st}^*)^2$ over all of s and t, we obtain the theorem. In particular, when $T = 1$ and $C = 1$, F^* is equal to the squared norm of the derivative $\nabla_\theta f_\theta$, that is, $F^* = ||\nabla_\theta f_\theta||^2$, and one can easily check $s_\lambda = \alpha\kappa_1^2$.

The detailed proof is given as follows.

(i) Case of $C = 1$

Here, let us express the FIM as $F = \nabla_\theta f \nabla_\theta f^T / T$, where $\nabla_\theta f$ is a $P \times T$ matrix whose columns are the gradients on each input sample, i.e., $\nabla_\theta f(t)$ $(t = 1, ..., T)$. We also introduce a dual matrix of F, that is, F^*:

$$F^* \quad := \quad \nabla_\theta f^T \nabla_\theta f / T. \tag{A.17}$$

Note that F is a $P \times P$ matrix while F^* is a $T \times T$ matrix. We can easily confirm that these F and F^* have the same non-zero eigenvalues.

The squared sum of the eigenvalues is given by $\sum_i \lambda_i^2 = \mathrm{Trace}(F^*(F^*)^T) = \sum_{st} (F^*_{st})^2$. By using the Frobenius norm $||A||_F := \sqrt{\sum_{ij} A_{ij}^2}$, this is $\sum_i \lambda_i^2 = ||F^*||_F^2$. Similar to m_λ, the bias entries in F^* are negligible because the number of the entries is much less than that of weight entries. Therefore, we only need to consider the weight entries. The st-th entry of F^* is given by

$$F^*_{st} = \sum_l \sum_{ij} \nabla_{W^l_{ij}} f(s) \nabla_{W^l_{ij}} f(t) / T \tag{A.18}$$

$$= \sum_l M_{l-1} \tilde{Z}^l(s,t) \hat{Z}^{l-1}(s,t)/T, \tag{A.19}$$

where we defined

$$\hat{Z}^l(s,t) := \frac{1}{M_l} \sum_j h^l_j(s) h^l_j(t), \quad \tilde{Z}^l(s,t) := \sum_i \delta^l_i(s) \delta^l_i(t). \tag{A.20}$$

We can apply the central limit theorem to $\hat{Z}^{l-1}(s,t)$ and $\tilde{Z}^l(s,t)$ independently because they do not share the index of the summation. For $s \neq t$, we have $\hat{Z}^l = \hat{q}^l_{st} + \mathcal{N}(0, \hat{\gamma}/M)$ and $\tilde{Z}^l = \tilde{q}^l_{st} + \mathcal{N}(0, \tilde{\gamma}/M)$ in the limit of $M \gg 1$, where the macroscopic variables $\hat{q}^l_{st}$ and $\tilde{q}^l_{st}$ satisfy the recurrence relations (10) and (12). Note that the recurrence relation (12) requires the Assumption 1. $\hat{\gamma}$ and $\tilde{\gamma}$ are constants of $O(1)$. Then, for all s and $t (\neq s)$,

$$F^*_{st} = \sum_l M_{l-1} (\tilde{q}^l_{st} + O(1/\sqrt{M}))(\hat{q}^{l-1}_{st} + O(1/\sqrt{M}))/T \tag{A.21}$$

$$= \alpha \kappa_2 M/T + O(\sqrt{M})/T. \tag{A.22}$$

Similarly, for $s = t$, we have $\hat{Z}^l = \hat{q}^l + O(1/\sqrt{M})$, $\tilde{Z}^l = \tilde{q}^l + O(1/\sqrt{M})$ and then $F^*_{ss} = \alpha\kappa_1 M/T + O(\sqrt{M})/T$.

Thus, under the limit of $M \gg 1$, the dual matrix is asymptotically given by

$$F^* = \alpha M K/T + O(\sqrt{M})/T, \quad K := \begin{bmatrix} \kappa_1 & \kappa_2 & \cdots & \kappa_2 \\ \kappa_2 & \kappa_1 & & \vdots \\ \vdots & & \ddots & \kappa_2 \\ \kappa_2 & \cdots & \kappa_2 & \kappa_1 \end{bmatrix}. \tag{A.23}$$

Neglecting the lower order term, we obtain

$$s_\lambda = \sum_{s,t}^T (F^*_{st})^2 / P \tag{A.24}$$

$$= \alpha \left(\frac{T-1}{T} \kappa_2^2 + \frac{1}{T} \kappa_1^2 \right). \tag{A.25}$$

Note that, when $\hat{q}^l_{st} = 0$, κ_2 becomes zero and the lower order term may be non-negligible. In this exceptional case, we have $s_\lambda = \alpha \kappa_1^2/T + O(1/M)$, where the second term comes from the $O(\sqrt{M})/T$ term of Eq. (A.23). Therefore, the lower order evaluation depends on the T/M ratio, although it is outside the scope of this study. Intuitively, the origin of $\hat{q}^l_{st} \neq 0$ is related to the offset of firing activities h^l_i. The condition of $\hat{q}^l_{st} \neq 0$ is satisfied when the bias terms exist or when the activation $\phi(\cdot)$ is not an odd function. In such cases, the firing activities have the offset $\mathrm{E}[h^l_i(t)] \neq 0$. Therefore, for any input samples s and t $(s \neq t)$, we have $\sum_i h^l_i(s) h^l_i(t)/M_l = \hat{q}^l_{st} \neq 0$ and then $\kappa_2 \neq 0$ makes s_λ of $O(1)$.

(ii) $C > 1$ of $O(1)$

Here, we introduce the following dual matrix F^*:

$$F^* := B^T B/T, \tag{A.26}$$
$$B := [\nabla_\theta f_1 \;\; \nabla_\theta f_2 \;\; \cdots \;\; \nabla_\theta f_C], \tag{A.27}$$

where $\nabla_\theta f_k$ is a $P \times T$ matrix whose columns are the gradients on each input sample, i.e., $\nabla_\theta f_k(t)$ $(t = 1, ..., T)$, and B is a $P \times CT$ matrix. The FIM is represented by $F = BB^T/T$. F^* is a $CT \times CT$ matrix and consists of $T \times T$ block matrices,

$$F^*(k, k') := \nabla_\theta f_k^T \nabla_\theta f_{k'}/T, \tag{A.28}$$

for $k, k' = 1, ..., C$.

The diagonal block $F^*(k, k)$ is evaluated in the same way as the case of $C = 1$. It becomes $\alpha MK/T$ as shown in Eq. (A.23). The non-diagonal block $F^*(k, k')$ has the following st-th entries:

$$F^*(k, k')_{st} = \sum_{ij} \nabla_{W^l_{ij}} f_k^T(s) \nabla_{W^l_{ij}} f_{k'}(t)/T \tag{A.29}$$
$$= M_{l-1}(\sum_i \delta^l_{k,i}(s)\delta^l_{k',i}(t))\hat{Z}^{l-1}(s, t)/T. \tag{A.30}$$

Under the limit of $M \gg 1$, while $\tilde{Z}^l(s, t)$ becomes $\tilde{q}^l_{st}$ of $O(1)$, $(\sum_i \delta^l_{k,i}(s)\delta^l_{k',i}(t))$ becomes zero and its lower order term of $O(1/\sqrt{M})$ appears. This is because the different outputs $(k \neq k')$ do not share the weights W^L_{ij}. We have $\sum_i \delta^L_{k,i}(s)\delta^L_{k',i}(t) = 0$ and then obtain $\sum_i \delta^l_{k,i}(s)\delta^l_{k',i}(t) = 0$ $(l = 1, ..., L-1)$ through the backpropagated chain (7). Thus, the entries of the non-diagonal blocks (A.28) become of $O(\sqrt{M})/T$, and we have

$$F^*(k, k') = \alpha MK/T\delta_{k,k'} + O(\sqrt{M})/T, \tag{A.31}$$

where $\delta_{k,k'}$ is the Kronecker delta.

After all, we have

$$s_\lambda = \sum_{k,k'}^{C} \sum_{s,t}^{T} (F^*(k, k')_{st})^2/P \tag{A.32}$$
$$= C\alpha \left(\frac{T-1}{T}\kappa_2^2 + \frac{1}{T}\kappa_1^2 \right) + CO(1/\sqrt{M}) + C(C-1)O(1/M), \tag{A.33}$$

where the first term comes from the diagonal blocks of $O(M)$ and the second one is their lower order term. The third term comes from the non-diagonal blocks of $O(\sqrt{M})$. As one can see from here, when $C = O(M)$, the thrid term becomes non-negligible. This case is examined in Section 3.4. ■

A.4 Theorem 4

(i) Case of $C = 1$

Because F and F^* have the same non-zero eigenvalues, what we should derive here is the maximum eigenvalue of F^*. As shown in Eq. (A.23), the leading term of F^* asymptotically becomes $\alpha MK/T$ in the limit of $M \gg 1$. The eigenvalues of $\alpha MK/T$ are explicitly obtained as follows: $\lambda_{max} = \alpha \left(\frac{T-1}{T}\kappa_2 + \frac{1}{T}\kappa_1\right) M$ for an eigenvector $e = (1, ..., 1)$, and $\lambda_i = \alpha(\kappa_1 - \kappa_2)M/T$ for eigenvectors $e_1 - e_i$ $(i = 2, ..., T)$ where e_i denotes a unit vector whose entries are 1 for the i-th entry and 0 otherwise. Thus, we obtain $\lambda_{max} = \alpha \left(\frac{T-1}{T}\kappa_2 + \frac{1}{T}\kappa_1\right) M$.

(ii) $C > 1$ of $O(1)$

Let us denote F^* shown in Eq. (A.31) by $F^* := \bar{F}^* + R$. $\bar{F}^*$ is the leading term of F^* and given by a $CT \times CT$ block diagonal matrix whose diagonal blocks are given by $\alpha MK/T$. R denotes the residual term of $O(\sqrt{M})/T$.

In general, the maximum eigenvalue is denoted by the spectral norm $||\cdot||_2$, that is, $\lambda_{max} = ||F^*||_2$. Using the triangle inequality, we have

$$\lambda_{max} \leq ||\bar{F}^*||_2 + ||R||_2, \tag{A.34}$$

We can obtain $||\bar{F}^*||_2 = \alpha\left(\frac{T-1}{T}\kappa_2 + \frac{1}{T}\kappa_1\right)M$ because the maximum eigenvalues of the diagonal blocks are the same as the case of $C=1$. Its eigenvector is given by a CT-dimensional vector $e=(1,...,1)$. Regarding $||R||_2$, this is bounded by $||R||_2 \leq ||R||_F = \sqrt{C^2\sum_{st}(O(\sqrt{M})/T)^2} = O(C\sqrt{M})$. Therefore, when $C=O(1)$, we can neglect $||R||_2$ of $O(\sqrt{M})$ compared to $||\bar{F}^*||_2$ of $O(M)$.

On the other hand, we can also derive the lower bound of λ_{max} as follows. In general, we have

$$\lambda_{max} = \max_{\mathbf{v};||\mathbf{v}||^2=1} \mathbf{v}^T F^* \mathbf{v}. \tag{A.35}$$

Then, we find

$$\lambda_{max} \geq \mathbf{v}_1^T F^* \mathbf{v}_1, \tag{A.36}$$

where v_1 is a CT-dimensional vector whose first T entries are $1/\sqrt{T}$ and the others are 0, that is, $v_1 = (1,...,1,0,...,0)/\sqrt{T}$. We can compute this lower bound by taking the sum over the entries of $F^*(1,1)$, which is equal to Eq. (A.23):

$$\lambda_{max} \geq \left(\frac{T-1}{T}\kappa_2 + \frac{1}{T}\kappa_1\right)M. \tag{A.37}$$

Finally, we find that the upper bound (A.34) and lower bound (A.37) asymptotically take the same value of $O(M)$, that is, $\lambda_{max} = \left(\frac{T-1}{T}\kappa_2 + \frac{1}{T}\kappa_1\right)M$.

■

A.5 Case of $C = O(M)$

The mean of eigenvalues m'_λ is derived in the same way as shown in Section A.1 (ii), that is, $m'_\lambda = C\kappa_1/M$.

Regarding the second moment s'_λ, the lower order term becomes non-negligible as remarked in Eq. (A.33). We evaluate this s'_λ by using inequalities as follows:

$$s'_\lambda = ||F^*||_F^2/P \tag{A.38}$$

$$= \left(\sum_k^C ||\nabla_\theta f_k^T \nabla_\theta f_k||_F^2 + \sum_{k,k'}^C ||\nabla_\theta f_k^T \nabla_\theta f_{k'}||_F^2\right)/P \tag{A.39}$$

$$\geq \sum_k^C ||\nabla_\theta f_k^T \nabla_\theta f_k||_F^2/P. \tag{A.40}$$

As shown in Section A.3, for any k, we obtain $||\nabla_\theta f_k^T(s)\nabla_\theta f_k(t)||_F^2/P = \alpha\left(\frac{T-1}{T}\kappa_2^2 + \frac{1}{T}\kappa_1^2\right)$ in the limit of $M \gg 1$. Thus, the lower bound becomes the same form as s_λ, That is, $s_\lambda = C\alpha(\frac{T-1}{T}\kappa_2^2 + \frac{1}{T}\kappa_1^2)$. In contrast, the upper bound is given by

$$s'_\lambda = ||F||_F^2/P \tag{A.41}$$

$$= ||\sum_k^C F_k||_F^2/P \tag{A.42}$$

$$\leq (\sum_k^C ||F_k||_F)^2/P, \tag{A.43}$$

where F_k denotes the FIM of the k-th output, i.e., $F_k := \sum_t \nabla_\theta f_k(t)\nabla_\theta f_k(t)^T/T$. Therefore, the upper bound is reduced to the summation over s_λ of $C=1$. In the limit of $M \gg 1$, we obtain $s'_\lambda \leq C^2||F_k||_F^2/P = C^2\alpha\left(\frac{T-1}{T}\kappa_2^2 + \frac{1}{T}\kappa_1^2\right) = Cs_\lambda$.

Next, we show inequalities for λ_{max}. We have already derived the lower bound (A.37) and this bound holds in the case of $C = O(M)$ as well. In contrast, the upper bound (A.34) may become loose when C is larger than $O(1)$ because of the residual term $||R||_2$. Although it is hard to explicitly obtain the value of $||R||_2$, the following upper bound holds and is easy to compute by using s_λ of Eq. (14). Because the FIM is a positive semi-definite matrix, $\lambda_i \geq 0$ holds by definition. Then, we have $\lambda_{max} \leq \sqrt{\sum_i \lambda_i^2}$. Combining this with $s'_\lambda = \sum_i \lambda_i^2/P$, we have $\lambda_{max} \leq \sqrt{\alpha s'_\lambda} M \leq \sqrt{\alpha C s_\lambda} M$.

■

A.6 Theorem 5

The Fisher-Rao norm is written as

$$||\theta||_{FR} = \sum_{l,ij} \sum_{l',ab} F_{(l,ij),(l',ab)} W_{ij}^l W_{ab}^{l'}, \tag{A.44}$$

where $F_{(l,ij),(l',ab)}$ represents an entry of the FIM, that is, $\sum_k^C \sum_t \nabla_{W_{ij}^l} f_k(t) \nabla_{W_{ab}^{l'}} f_k(t)/T$. Because $F_{(l,ij),(l',ab)}$ includes the random variables W_{ij}^l and $W_{ab}^{l'}$, we consider the following expansion. Note that W_{ij}^l and $W_{ab}^{l'}$ are infinitesimals generated by Eq. (8). Performing a Taylor expansion around $W_{ij}^l = W_{ab}^{l'} = 0$, we obtain

$$F_{(l,ij),(l',ab)}(\theta) = F_{(l,ij),(l',ab)}(\theta^*) + \frac{\partial F_{(l,ij),(l',ab)}}{\partial W_{ij}^l}(\theta^*) W_{ij}^l + \frac{\partial F_{(l,ij),(l',ab)}}{\partial W_{ab}^{l'}}(\theta^*) W_{ab}^{l'} + \text{higher-order terms}, \tag{A.45}$$

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

$$\langle F_{(l,ij),(l',ab)} W_{ij}^l W_{ab}^{l'} \rangle_\theta = \langle F_{(l,ij),(l',ab)}(\theta^*) W_{ij}^l W_{ab}^{l'} \rangle_\theta \tag{A.46}$$

$$= \langle F_{(l,ij),(l',ab)}(\theta^*) \rangle_{\theta^*} \langle W_{ij}^l W_{ab}^{l'} \rangle_{\{W_{ij}^l, W_{ab}^{l'}\}} \tag{A.47}$$

For, $(l,ij) \neq (l',ab)$, the last line becomes zero because of $\langle W_{ij}^l W_{ab}^{l'} \rangle_{\{W_{ij}^l, W_{ab}^{l'}\}} = \langle W_{ij}^l \rangle_{W_{ij}^l} \langle W_{ab}^{l'} \rangle_{W_{ab}^{l'}} = 0$. For $(l,ij) = (l',ab)$, we have $\langle (W_{ij}^l)^2 \rangle_{\{W_{ij}^l\}} = \sigma_w^2 / M_{l-1}$. After all, in the limit of $M \gg 1$, we obtain

$$\langle ||\theta||_{FR} \rangle_\theta = \sum_k^C \frac{\sum_t}{T} \sum_l \langle \sum_i \delta_{k,i}^l(t)^2 \sum_j h_j^{l-1}(t)^2 \rangle_{\theta^*} \frac{\sigma_w^2}{M_{l-1}} \tag{A.48}$$

$$= \sum_k^C \frac{\sum_t}{T} \sigma_w^2 \sum_l \langle \tilde{q}^l \rangle_\theta \langle \hat{q}^{l-1} \rangle_\theta \tag{A.49}$$

$$= \sigma_w^2 C \sum_l \tilde{q}^l \hat{q}^{l-1}, \tag{A.50}$$

where the derivation of the macroscopic variables is similar to that of m_λ, as shown in Section A.1. Since we have $\kappa_1 = \sum_l \frac{\alpha_{l-1}}{\alpha} \tilde{q}^l \hat{q}^{l-1}$, it is easy to confirm $\langle ||\theta||_{FR} \rangle_\theta \leq C \sigma_w^2 \alpha / \alpha_{min} C \kappa_1$. When all α_l take the same value, we have $\alpha / \alpha_{min} = L - 1$ and the equality holds. ■

A.7 Lemma 6

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

$$\Delta_{t+1} \leftarrow (I - \eta F) \Delta_t + \mu (\Delta_t - \Delta_{t-1}), \tag{A.51}$$

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

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

$$\bar{\Delta}_{t+1,i} \leftarrow (1-\eta\lambda_i+\mu)\bar{\Delta}_{t,i} - \mu\Delta_{t-1,i}. \tag{A.52}$$

Solving its characteristic equation, we obtain the general solution,

$$\bar{\Delta}_{t,i} = A\lambda_+^t + B\lambda_-^t, \quad \lambda_\pm = (1-\eta\lambda_i+\mu \pm \sqrt{(1-\eta\lambda_i+\mu)^2-4\mu})/2, \tag{A.53}$$

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^*. ■

B Analytical recurrence relations

B.1 Erf networks

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

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

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.

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

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

Hence, the recurrence relations for the feedforward signals (9) have the following analytical forms:

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

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 (11):

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

(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

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

For the error function,

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

and we obtain

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

This is 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

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

This is the analytical forms of the recurrence relations for $\tilde{q}^l_{st}$.

B.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.

(i) $\hat{q}^l$ and $\tilde{q}^l$: We can explicitly perform the integrations in the recurrence relations (9) and (11):

$$\hat{q}^{l+1} = \hat{q}^l \sigma_w^2/2 + \sigma_b^2/2, \tag{B.9}$$
$$\tilde{q}^l = \tilde{q}^{l+1}\sigma_w^2/2, \quad \tilde{q}^L = 1/2. \tag{B.10}$$

(ii) $\hat{q}_{st}^l$ and $\tilde{q}_{st}^l$: We can explicitly perform the integrations in the recurrence relations (10) and (12):

$$I_\phi[a,b] = \frac{a}{2\pi}(\sqrt{1-c^2} + c\pi/2 + c\sin^{-1} c), \tag{B.11}$$
$$I_{\phi'}[a,b] = \frac{a}{2\pi}(\pi/2 + \sin^{-1} c), \tag{B.12}$$

where $c = b/a$.

B.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.

(i) $\hat{q}^l$ and $\tilde{q}^l$: We can explicitly perform the integrations in the recurrence relations (9) and (11):

$$q^l = q^{l-1}\sigma_w^2 + \sigma_b^2, \tag{B.13}$$
$$\tilde{q}^l = \tilde{q}^{l+1}\sigma_w^2, \quad \tilde{q}^L = 1. \tag{B.14}$$

(ii) $\hat{q}_{st}^l$ and $\tilde{q}_{st}^l$: We can explicitly perform the integrations in the recurrence relations (10) and (12):

$$\hat{q}_{st}^{l+1} = \hat{q}_{st}^l \sigma_w^2 + \sigma_b^2, \tag{B.15}$$
$$\tilde{q}_{st}^l = \tilde{q}_{st}^{l+1}\sigma_w^2, \quad \tilde{q}_{st}^L = 1. \tag{B.16}$$

C Additional Experiments

C.1 Dependence on T

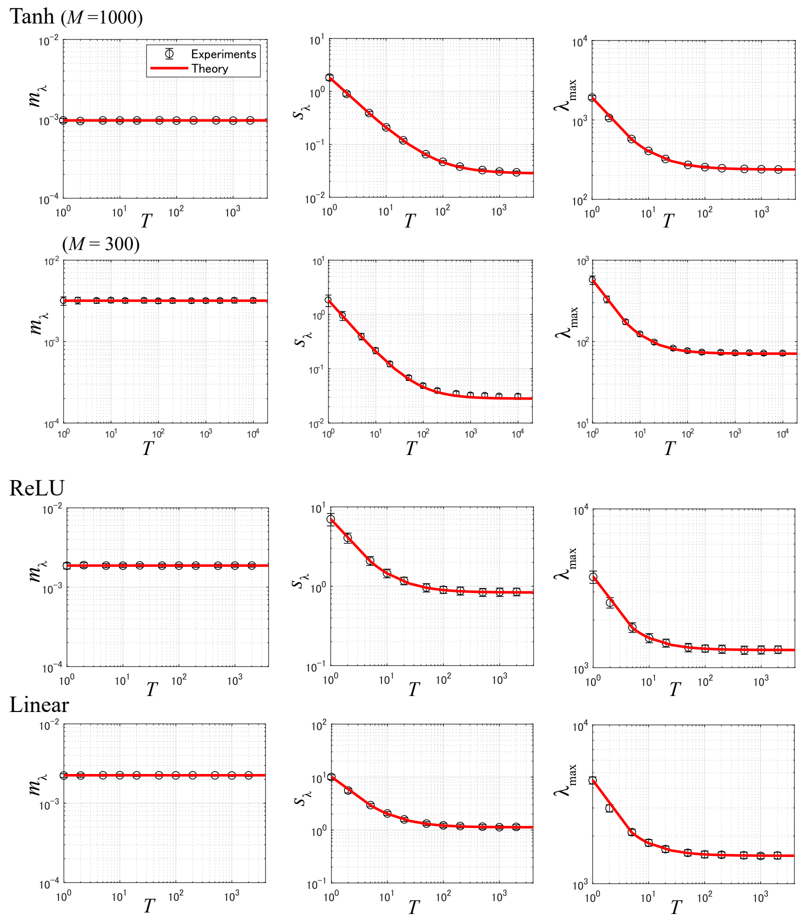

Figure C.1: Statistics of FIM eigenvalues with fixed M and changing T ($L = 3, \alpha_l = C = 1$). The red line represents theoretical results obtained in the limit of $M \gg 1$. The first row shows results of Tanh networks with $M = 1000$. The second row shows those with a relatively small width ($M = 300$) and higher T. We set $M = 1000$ in ReLU and linear networks. The other settings are the same as in Fig. 1.

C.2 Training on CIFAR-10

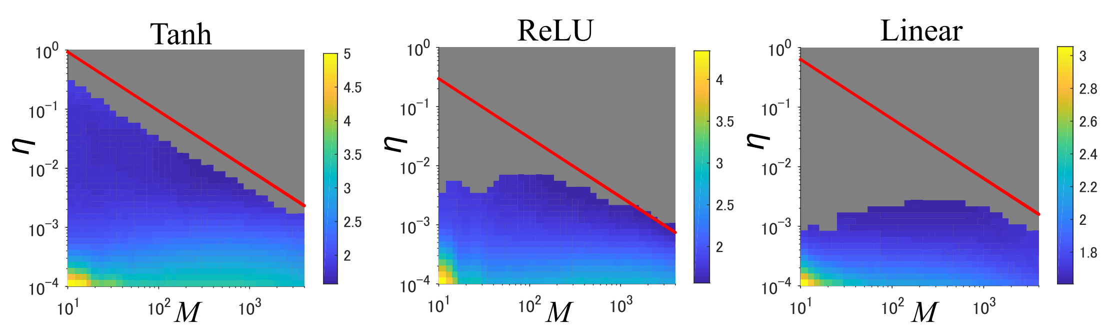

Figure C.2: Color map of training losses after one epoch of SGD training: Tanh, ReLU, and linear networks trained on CIFAR-10.