Supplementary Material

A Proof for The Two-Hidden-Layer Case

Here we provide the postponed proof.

Proof. Set $\tilde{\nu}_\ell = (\gamma_\ell^{-1}\,\cdot\,)_*\nu_\ell = \alpha_\ell\delta_1 + (1-\alpha_\ell)\delta_0$. Then we have

$$\mu_3 = (q_2 + \sigma_3^2\gamma_2\,\cdot\,)_*\left[\tilde{\nu}_2 \boxtimes (q_1 + \sigma_2^2\gamma_1 q_0\,\cdot\,)_*\tilde{\nu}_1\right]. \tag{A.1}$$

By replacing $\sigma_{\ell+1}^2\gamma_\ell q_{\ell-1}q_\ell^{-1}$ with γ_ℓ $(\ell = 1, 2)$, we may assume that $q_\ell = \sigma_\ell = 1$. Write $\xi = (1+\gamma_1\,\cdot\,)_*\tilde{\nu}_1$. Since $S_{\tilde{\nu}_2\boxtimes\xi}(z) = S_{\tilde{\nu}_2}(z)S_\xi(z)$, we have $h_\xi^{-1}(z) = h_{\tilde{\nu}_2\boxtimes\xi}^{-1}(z)S_{\tilde{\nu}_2}(z)$. Hence $h_{\tilde{\nu}_2\boxtimes\xi}(z)$ is the solution of the following equation on w:

$$w = h_\xi(zS_{\tilde{\nu}_2}(w)). \tag{A.2}$$

Note that

$$S_{\tilde{\nu}_2}(z) = \frac{z+1}{z+\alpha_2},\ h_\xi(z) = z\left[\frac{1-\alpha_1}{z-1} + \frac{\alpha_1}{z-1-\gamma_1}\right] - 1. \tag{A.3}$$

Thus the solution of (A.2) is given by

$$w = \frac{g(z) \pm z\sqrt{-f(z)}}{2(z-1)(1+\gamma_1-z)}, \tag{A.4}$$

where $f(z) = (\lambda_+ - z)(z-\lambda_-)$, $\lambda_\pm = 1 + \gamma_1\left(\sqrt{\alpha_1(1-\alpha_2)} \pm \sqrt{\alpha_2(1-\alpha_1)}\right)^2$, and $g(z) = (z-1)(z - 2(1+\gamma_1)\alpha_2) - \gamma_1(\alpha_1-\alpha_2)z$. By $G(z) = (h(z)+1)/z$ and by the condition $\Im G(z) < 0$ if $\Im z > 0$, we have

$$G_{\tilde{\nu}_2\boxtimes\xi}(z) = \frac{1}{z}\left[1 + \frac{g(z)}{2(z-1)(1+\gamma_1-z)}\right] + \frac{\sqrt{-f(z)}}{2(z-1)(1+\gamma_1-z)}. \tag{A.5}$$

Note that $1 \le \lambda_- \le \lambda_+ \le 1+\gamma_1$. By the Stieltjes inversion, the absolutely continuous part of $\nu\boxtimes\mu$ is

$$-\frac{1}{\pi}\lim_{y\to+0}\Im G_{\tilde{\nu}_2\boxtimes\xi}(x+y\sqrt{-1}) = \frac{\sqrt{f(x)}}{2\pi(x-1)(1+\gamma_1-x)}\mathbf{1}_{\{f\ge0\}}(x)\ (x\in\mathbb{R}). \tag{A.6}$$

Te weights of the atoms are given by

$$\lim_{y\to+0} zG_{\tilde{\nu}_2\boxtimes\xi}(y\sqrt{-1}) = 1-\alpha_2, \tag{A.7}$$

$$\lim_{y\to+0}(z-1)G_{\tilde{\nu}_2\boxtimes\xi}(1+y\sqrt{-1}) = (\alpha_2-\alpha_1)^+, \tag{A.8}$$

$$\lim_{y\to+0}(z-1-\gamma_1)G_{\tilde{\nu}_2\boxtimes\xi}(1+\gamma_1+y\sqrt{-1}) = (\alpha_1+\alpha_2-1)^+, \tag{A.9}$$

where $a^+ = \max(a, 0)$ for $a\in\mathbb{R}$. By Belinschi (2003), the free multiplicative convolution $\tilde{\nu}_2\boxtimes\xi$ has no singular continuous part. Hence $\tilde{\nu}_2\boxtimes\xi$ is the sum of the absolutely continuous part (A.6) and the pure point part (A.7, A.8, A.9) as follows:

$$(\tilde{\nu}_2\boxtimes\xi)(dx) = (1-\alpha_2)\delta_0(dx) + (\alpha_2-\alpha_1)^+\delta_1(dx) + (\alpha_1+\alpha_2-1)^+\delta_{1+\gamma_1}(dx) \tag{A.10}$$

$$+ \frac{\sqrt{(\lambda_+-x)(x-\lambda_-)}}{2\pi(x-1)(1+\gamma_1-x)}\mathbf{1}_{[\lambda_-,\lambda_+]}(x)(dx). \tag{A.11}$$

It holds that $\mu_3 = (1+\gamma_2\,\cdot\,)_*(\tilde{\nu}_2\boxtimes\xi)$. We have completed the proof. □

B Out of Initial Spectral Distribution

Fig. 1 shows the detailed results of experiments in Section 5.2. As discussed in Section 5.2, the accuracy followed the boundary (upper figures). However, in the boundary areas with $L/2 < \eta < 0.2$ the loss reduced (lower figures), but the accuracy did not improve. Hence the loss violates the boundary $\eta = L/2$ predicted by Theorem 4.3 and (35), but the accuracy does not. In this section, we discuss the region $L/2 < \eta < 0.2$ of the loss heatmaps.

Recall that we trained the network on benchmark datasets Fashion-MNIST (Xiao et al., 2017) and CIFAR10 (Krizhevsky, 2009). The Fashion-MNIST (resp. CIFAR10) consists of 28×28 (resp. $3 \times 32 \times 32$) dimensional images and 10 class labels. We set $M = 28^2$ for the Fashion-MNIST and $M = 32^2$ for the CIFAR10, by shrinking the first layer in the case of the CIFAR10. We applied the online gradient descent on 500 data, which is uniformly sampled from whole data and fixed, for an epoch (resp. ten epochs) in the Fashion-MNIST (resp. CIFAR10). Recall that we normalized each input so that $\hat{q_0} = 1$ and converted class labels to an orthonormal system in $\mathbb{R}^M$, and use the hard-tanh activation with $s^2 = 0.125$ and $g = 1.0013$ to archive dynamical isometry. After training, we computed the average of MSE loss for each dataset $\mathcal{D}$, which is given by the following:

$$\mathrm{Loss}(x, l) = \frac{1}{2M|\mathcal{D}|} \sum_{(x,\ell)\in\mathcal{D}} ||f_\theta(x) - e_\ell||_2^2, \tag{B.1}$$

where $e_m \in \mathbb{R}^M$ $(m = 1, 2, \ldots, M)$ is the unit vector whose ℓ-th entry is one and the other entries are zero, and the dataset $\mathcal{D}$ is the training dataset, which consists of 500-samples, or the testing dataset of 10000 samples separated from the training dataset. We also computed top-1 accuracy. In Fig. 1, the difference between training and testing accuracy on the CIFAR10 was larger than that on the Fashion-MNIST. The reason for this is because of the overfitting of DNNs to the small dataset consists of only 500 data. However, agreement with the theoretical line was also visible in the testing accuracy on the CIFAR10.

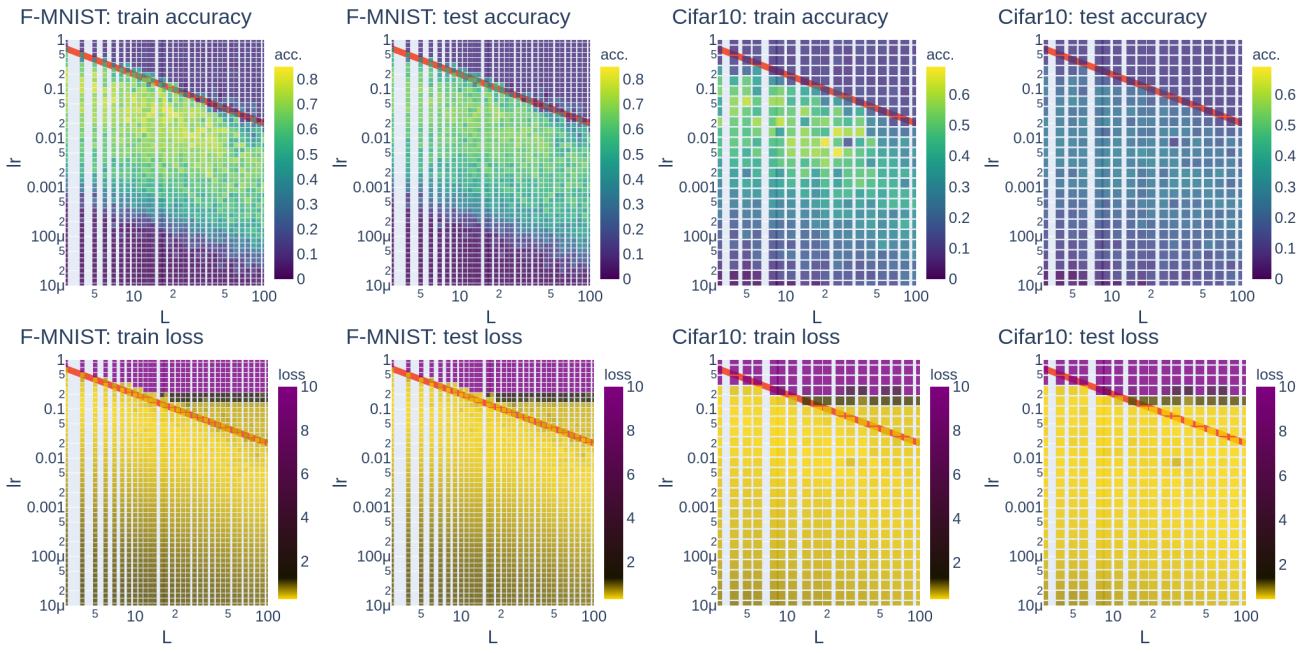

Figure 1: Accuracy heatmaps (upper figures) and loss heatmaps (lower ones) for different value of L (x-axis) and lr $(= \eta)$ (y-axis) after online training. In each figure, the line is $\eta = 2/L$. Each axis is logarithmic. Each network was trained on Fashion-MNIST or CIFAR10. In each experiment, the training dataset consists of 500 data points sampled uniformly from a whole data set, and the testing dataset consists of 10000 samples separated from the training dataset. For the Fashion-MNIST (resp. CIFAR10), the network was trained for a single epoch (resp. ten epoch) training. In each figure of losses, we have rounded off any losses above 10 and show them as 10.

We focus on areas near the boundary $\eta = 2/L$ at large L and we show how different the conditional FIM H_L's spectral distribution during training was from that in the initial state. Fig. 2 shows that eigenvalue distributions

of H_L near the boundary at large L. We observed that most of the eigenvalue distributions shrunk in the areas with $\eta > 2/L$. The shrinkage in the eigenvalue distribution is the reason for the reduction in test loss in the areas.

However, the reason why the appearance of the second boundary around $\eta = 0.2$ appeared has not been revealed yet.

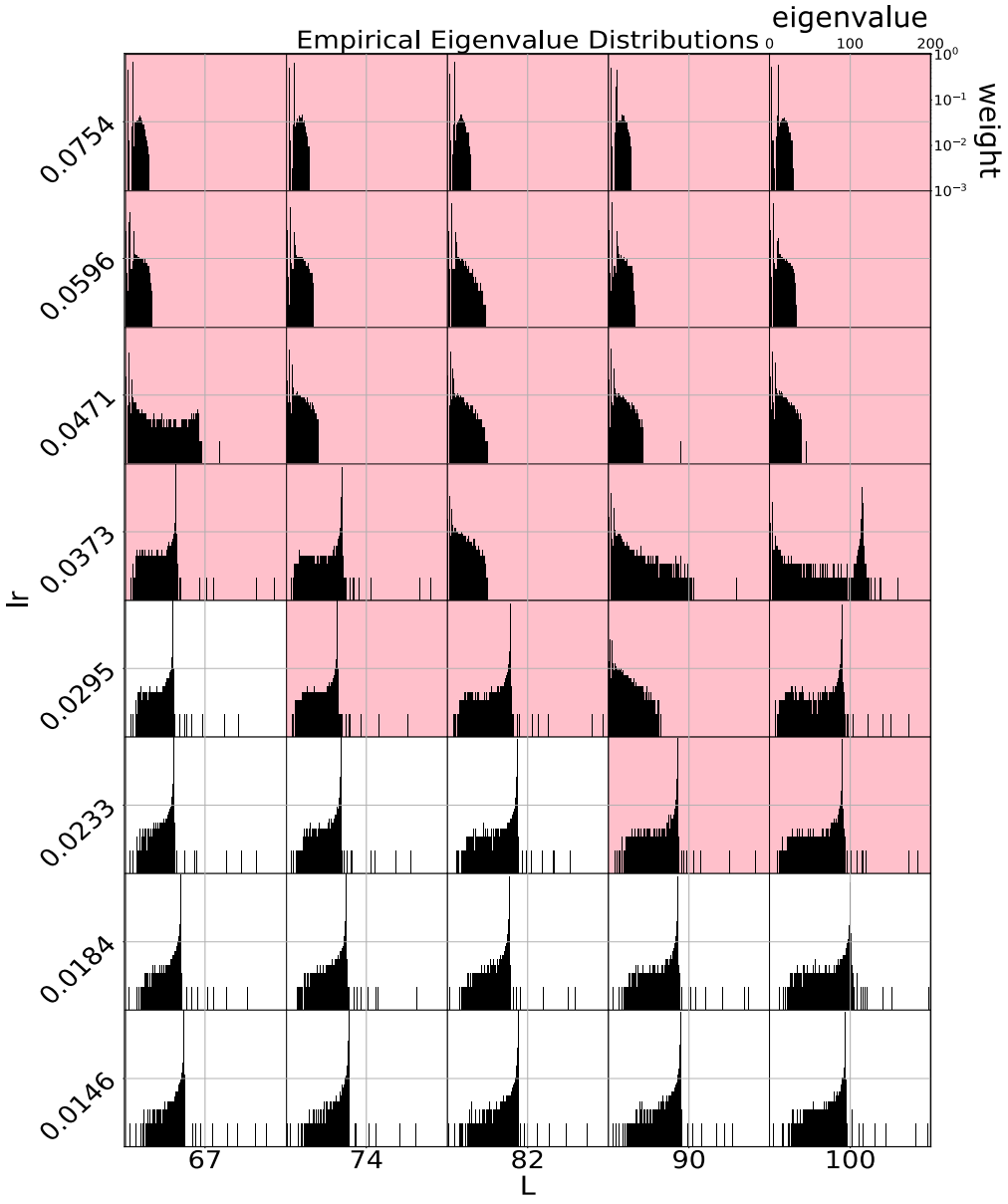

Figure 2: Histograms of eigenvalue distributions of H_L after training 500-steps on Fashion-MNIST. All histograms share the x-axis (eigenvalue) and the y-axis (weight). The y-axis is logarithmic. The outer frame's x-axis represents the depth L and its y-axis represents the learning rate η. Both axes are logarithmic. The histograms in the red region belong to $\eta > 2/L$, and the others belong to $\eta \leq 2/L$.

C A Short Introduction to Asymptotic Freeness for Machine Learning

C.1 Comparison with Classical Probability Theory

Asymptotic freeness is the vital notion in free probability theory. In order to introduce asymptotic freeness to readers in the machine learning community, we explain the freeness by comparing free probability theory to classical probability theory. We refer readers to (Mingo and Speicher, 2017) or (Voiculescu et al., 1992) for the detail.

	r.v.	moments	for multiple r.v.s	independence
classical	X	$\mathbb{E}[X^k]$	joint distribution	decomposition of joint distribution
free	A	$\mathrm{tr}[A^k]$	joint moments	decomposition of joint moments

Table 1: Comparison of free and classical probability theory.

Firstly, consider a matrix $A \in M_M(\mathbb{C})$. We denote by A^* the adjoint matrix, that is, complex-conjugate transpose matrix, of A. Assume that A is self-adjoint, that is, $A^* = A$. Then the spectral distribution, denoted by μ_A, is given by

$$\mu_A = \frac{1}{N} \sum_{\lambda \in \sigma(A)} \delta_\lambda, \tag{C.1}$$

where $\sigma(A) = \{\lambda \in \mathbb{C} \mid A - \lambda I \text{ is invertible } \} \subset \mathbb{R}$, and I is the identity operator. We emphasize that the spectral distribution is determined by its moments. That is, a distribution μ is equal to μ_A if and only if

$$\mathrm{tr}(A^k) = \int x^k \mu(dx) \ (k \in \mathbb{N}), \tag{C.2}$$

where tr is the normalized trace so that $\mathrm{tr}(I) = 1$. In other words, the family of moments $\mathrm{tr}(A^k)$ $(k \in \mathbb{N})$ has the same information as the spectral distribution μ_A. Now consider a counterpart of the spectral distribution at classical probability theory. Let X be a real random variable and μ_X be the distribution of X. If μ_X is compactly supported, the distribution μ_X is determined by the moments:

$$\mathbb{E}[X^k] = \int x^k \mu_X(dx) \ (k \in \mathbb{N}). \tag{C.3}$$

By comparing (C.2) and (C.3), we see that the self-adjoint operator A corresponds to a real random variable and the spectral distribution μ_A corresponds to the distribution of the random variable.

Secondly, consider multiple matrices. Let $A_1, A_2 \in M_M(\mathbb{C})$ be non-commuting self-adjoint matrices. As an example, consider the distribution of the sum of them. The spectral distribution of the sum $A_1 + A_2$ is determined by its moments

$$\mathrm{tr}[(A_1 + A_2)^k] = \sum_{i_1, i_2, \ldots, i_k \in \{1,2\}} \mathrm{tr}[A_{i_1} A_{i_2} \cdots A_{i_k}] \ (k \in \mathbb{N}). \tag{C.4}$$

Hence the trace of all words on A_1 and A_2 determines the spectral distribution $\mu_{A_1+A_2}$. Now consider its counterpart at classical probability theory. Let X_1 and X_2 be real random variables. Then

$$\mathbb{E}[(X_1 + X_2)^k] = \sum_{m=1}^{k} \binom{k}{m} \mathrm{tr}[X_1^m X_2^{k-m}] \ (k \in \mathbb{N}). \tag{C.5}$$

By comparing (C.4) and (C.5), we need much more information to determine the distribution for non-commuting matrices than for commuting real random variables. Therefore, we extend the definition of the joint distribution in a different way from that in the classical probability theory. For a family of matrices $(A_j)_{j \in J}$, its *joint moments* are trace of all words in the family, which are given by

$$\mathrm{tr}[A_{j_1} A_{j_2} \cdots A_{j_k}] \ (j_1, j_2, \ldots, j_k \in J, \ k \in \mathbb{N}). \tag{C.6}$$

Note that we use the joint moments as a counterpart of the joint distribution, which is a probability distribution in classical probability theory and does not exist for non-commuting matrices.

Lastly, consider the notion of independence. The independence in the classical probability theory means that the joint distribution of multiple random variables decomposed to the product of marginal distributions of each random variable. Since we consider the joint moments for multiple operators, we extend the concept of independence as a decomposition law of joint moments to each operator's moments. The freeness is one of the decomposition laws of joint moments (see Table 1).

C.2 Freeness

Summarizing above, we formulate the free algebra, the tracial state, and introduce the freeness.

Definition C.1. A *unital $*$-algebra over* $\mathbb{C}$ is a nonempty set $\mathcal{A}$ equipped with a triplet $(+, \cdot, *)$ satisfying the following conditions.

1. $(\mathcal{A}, +, *)$ is an associative but possibly noncommutative algebra over $\mathbb{C}$, where $+$ is the addition, $\cdot$ is the multiplication, $0 \in \mathbb{C}$ is equal to the additive identity, and $1 \in \mathbb{C}$ is equal to the multiplicative identity. We omit the symbol $\cdot$ and simply write $x \cdot y$ as xy for $x, y \in \mathcal{A}$.
2. The map $*\colon \mathcal{A} \to \mathcal{A}$ satisfies the following:

$$(xy)^* = y^* x^*, \tag{C.7}$$

$$(\alpha x + \beta y)^* = \overline{\alpha} x^* + \overline{\beta} y^*, \tag{C.8}$$

   for any $x, y \in \mathcal{A}$, $\alpha, \beta \in \mathbb{C}$, where $\overline{\alpha}$ is the complex conjugate of α.

Example C.2. The matrix algebra $M_n(\mathbb{C})$ of $n \times n$ matrices over $\mathbb{C}$ is a unital $*$-algebra for any $n \in \mathbb{N}$, where $A*$ is the adjoint matrix of $A \in M_n(\mathbb{C})$, which is the $n \times n$ matrix obtained from A by taking the transpose and the complex conjugate of the entries. The matrix algebra $M_n(\mathbb{R})$ of $n \times n$ matrices over $\mathbb{C}$ is a subalgebra of $M_n(\mathbb{C})$ as an algebra over $\mathbb{R}$. Now for $A \in M_n(\mathbb{R})$, $A^* = A^\top$.

Example C.3. Let us denote by $\mathbb{C}\langle Z_j \mid j \in J\rangle$ the free $\mathbb{C}$-algebra of indeterminates $(Z_j)_{j\in J}$, which is the algebra of polynomials of noncommutative variables given by the weighted sum of words as follows:

$$\mathbb{C}\langle Z_j \mid j \in J\rangle = \{\alpha_\emptyset 1 + \sum_{k=1}^{\infty} \sum_{j_1,\dots,j_k \in J} \alpha_{j_1,j_2,\dots,j_k} Z_{j_1} Z_{j_2} \dots Z_{j_k}, \tag{C.9}$$

$$\text{where } \alpha_* \in \mathbb{C} \text{ is zero except for finite number of indices } \}. \tag{C.10}$$

We introduce an adjoint operation $*$ on $\mathbb{C}\langle Z_j \mid j \in J\rangle$ using universality of the free product (see (Voiculescu et al., 1992)) by

$$(\alpha_{j_1,j_2,\dots,j_k} Z_{j_1} Z_{j_2} \cdots Z_{j_k})^* = \overline{\alpha_{j_1,j_2,\dots,j_k}} Z_{j_k} \cdots Z_{j_2} Z_{j_1}. \tag{C.11}$$

Next, we introduce the tracial state, which is an abstracted notion of the normalized trace of matrices and the expectation operator to random variables.

Definition C.4. A *tracial state* τ on a unital $*$-algebra $\mathcal{A}$ is a $\mathbb{C}$-valued map satisfying the following conditions.

1. $\tau(1) = 1$.
2. $\tau(\alpha a + \beta b) = \alpha\tau(a) + \beta\tau(b)$ $(a, b \in \mathcal{A}, \alpha, \beta \in \mathbb{C})$.
3. $\tau(a^*) = \overline{\tau(a)}$ $(a \in \mathcal{A})$.
4. $\tau(a^* a) \geq 0$ $(a \in \mathcal{A})$
5. $\tau(ab) = \tau(ba)$ $(a, b \in \mathcal{A})$.

The first condition is the normalization so that the total volume becomes 1. The last one is called the tracial condition.

Definition C.5. A *noncommutative probability space* (NCPS, for short) is a pair of a unital $*$-algebra $\mathcal{A}$ and a tracial state τ on $\mathcal{A}$.

Here we have prepared to introduce the freeness. For readers' convenience, we introduce a simpler version of the freeness that is the minimum necessary to read the primary material.

Definition C.6. (Freeness) Given two families $a = (a_j)_{j\in I}$ and $b = (b_j)_{j\in J}$ of elements in $\mathcal{A}$ are said to be *free* (or *free independent*) with respect to τ if the following decomposition of joint moments follows: For any $k \in \mathbb{N}$, any $p_1, p_2, \ldots, p_k \in \mathbb{C}\langle X_i \mid i \in I\rangle$, and any $q_1, q_2, \ldots, q_k \in \mathbb{C}\langle Y_j \mid j \in J\rangle$, it holds that

$$\tau[p_1(a)q_1(b)p_2(a)q_2(b)\cdots p_k(a)q_k(b)] = 0 \tag{C.12}$$

if $\tau[p_m(a)] = \tau[q_m(b)] = 0$ $(m = 1, 2, \ldots, k)$.

See (Voiculescu et al., 1992) for the full definition of the freeness for more general cases.

Example C.7. Assume that a and b are free elements in a NCPS $(\mathcal{A}, \tau)$, that is, we consider the case $|I| = |J| = 1$ in Definition C.6. Write $x^\circ = x - \tau(x)$ for $x \in \mathcal{A}$. Then

$$\mathrm{Cov}(a, b) := \tau(ab) - \tau(a)\tau(b) = \tau(a^\circ b^\circ) + \tau(a^\circ)\tau(b) + \tau(a)\tau(b^\circ) = 0. \tag{C.13}$$

Here we use the freeness to eliminate the term $\tau(a^\circ b^\circ)$. From this equation, we see that free variables are uncorrelated. The difference between freeness and classical independence appears in the decomposition of higher moments such as $\tau(abab)$:

$$\tau(abab) = \mathbb{V}[a]\mathbb{E}[b]^2 + \mathbb{E}[a]^2\mathbb{V}[b] + \mathbb{E}[a]^2\mathbb{E}[b]^2, \tag{C.14}$$

where $\mathbb{E}[a] = \tau(a)$ and $\mathbb{V}[a] = \tau(a^2) - \tau(a)^2$. The decomposition rule is different from that in the classical probability given by $\mathbb{E}[XYXY] = \mathbb{E}[X^2Y^2] = \mathbb{E}[X^2]\mathbb{E}[Y^2]$ for independent random variables X and Y.

C.3 Infinite Dimensional Approximation of Random Matrices

Here we introduce the relation between freeness and random matrices.

Definition C.8. Let $A_i(M), B_j(M) \in M_M(\mathbb{C})$ $(M \in \mathbb{N}, i \in I, j \in J)$. Then the families $(A_i)_{i\in I}$ and $(B_j)_{j\in J}$ are said to be *asymptotically free* as $M \to \infty$ if there exists $\mathcal{A}, \tau$, and $a_i \in \mathcal{A}$ $(i \in I)$ and $b_j \in \mathcal{A}$ $(j \in J)$ so that

$$\lim_{M\to\infty} \mathrm{tr}(A_i(M)^k) = \tau(a_i^k) \ (k \in \mathbb{N}, i \in I), \tag{C.15}$$

$$\lim_{M\to\infty} \mathrm{tr}(B_j(M)^k) = \tau(b_j^k) \ (k \in \mathbb{N}, j \in J), \tag{C.16}$$

$$\text{and, } (a_i)_{i\in I} \text{ and } (b_j)_{j\in J} \text{ are free.} \tag{C.17}$$

Here we introduce a known result in free probability theory.

Proposition C.9 ((Hiai and Petz, 2000, Prop. 3.5))**.** *For each $M \in \mathbb{N}$, consider the following matrices. Let $U(M)$ be random matrix uniformly distributed on $M \times M$ unitary matrices (resp. orthogonal matrices). Let $A(M)$ and $B(M)$ be complex (resp. real) self-adjoint random matrices independent of $U(M)$. Assume that there exist two compactly supported distributions μ and ν such that the following limits hold almost surely.*

$$\lim_{M\to\infty} \mathrm{tr}[A(M)^k] = \int x^k \mu(dx) \ (k \in \mathbb{N}), \tag{C.18}$$

$$\lim_{M\to\infty} \mathrm{tr}[B(M)^k] = \int x^k \nu(dx) \ (k \in \mathbb{N}). \tag{C.19}$$

Under the above conditions, it holds that $B(M)$ and $U(M)^*A(M)U(M)$ are asymptotically free as $M \to \infty$ almost surely. Furthermore, when $A(M)$ and $B(M)$ are positive definite, then the limit distribution of $B(M)^{1/2}U(M)^*A(M)U(M)B(M)^{1/2}$ is the multiplicative free convolution $\mu \boxtimes \nu$.

Note that random matrices $A(M)$ and $B(M)$ do not have to be independent in Proposition C.9.

C.4 Application to the FIM

Recall that the propagation of the conditional FIM is given by the following equation:

$$H_{\ell+1} = \hat{q}_\ell I + W_{\ell+1} D_\ell H_\ell D_\ell W_{\ell+1}^\top. \tag{C.20}$$

Let $A_\ell := W_\ell^* H_\ell W_\ell = \hat{q}_{\ell-1} I + D_{\ell-1} H_{\ell-1} D_{\ell-1}$. Note that $W_\ell^* = W_\ell^\top$. Then, firstly, W_ℓ and A_ℓ are independent. Secondly, recall that D_ℓ and (W_ℓ, W_ℓ^*) are assumed to be asymptotic free (see Assumption 3.1). Thirdly, each W_ℓ is uniformly distributed on orthogonal matrices. By the above conditions, it holds that $W_\ell A_\ell W_\ell^*$ and D_ℓ^2 are asymptotic free as the limit $M \to \infty$ by Proposition C.9. Then the limit spectral distribution of $D_\ell H_\ell D_\ell = D_\ell W_\ell A_\ell W_\ell^* D_\ell$ is equal to $\mu_\ell \boxtimes \nu_\ell$, where μ_ℓ (resp ν_ℓ) is the limit spectral distribution of H_ℓ (resp. D_ℓ^2). Then we get the following desired recursive equation:

$$\mu_{\ell+1} = (q_\ell + \sigma_{\ell+1}^2 \cdot)_* (\nu_\ell \boxtimes \mu_\ell). \tag{C.21}$$

D Analysis

Firstly, let us review on the following proposition known in free probability theory.

Proposition D.1 (Belinschi (2003))**.** *Let μ and ν be compactly supported probability distributions on $\mathbb{R}$. Then $\mu \boxtimes \nu$ has an atom at $c \in \mathbb{R}$ if and only if the following three conditions hold : (i) $a \in \mathbb{R}$ (resp. $b \in \mathbb{R}$) is an atom of μ (resp. ν), (ii) $c = ab$, and (iii) $\mu(\{a\}) + \nu(\{b\}) - 1 > 0$. Furthermore, if c is an atom then $\mu \boxtimes \nu(\{c\}) = \mu(\{a\}) + \nu(\{b\}) - 1$.*

Then we have the following recurrence equation of the maximum eigenvalue.

Lemma D.2. *Fix $L \in \mathbb{N}$. Let $\beta_1 = 1$ and $\beta_\ell = 1 - \sum_{k=1}^{\ell-1}(1-\alpha_k)$ for $\ell \geq 2$. Assume that $\beta_L > 0$. Then for any $\ell \leq L$, the value $||\mu_\ell||_\infty$ is an atom of μ_ℓ with weight β_ℓ. Furthermore, we have $||\mu_\ell||_\infty = q_{\ell-1} + \sigma_\ell^2 \gamma_{\ell-1} ||\mu_{\ell-1}||_\infty$ for $\ell \neq 1$.*

Proof. Let us define $\lambda_\ell \in \mathbb{R}$ recursively by $\lambda_\ell = q_{\ell-1} + \sigma_\ell^2 \gamma_{\ell-1} \lambda_{\ell-1} (\ell \geq 2)$ and $\lambda_1 = q_0$. Firstly we prove that λ_ℓ is an atom with weight β_ℓ of μ_ℓ for $\ell \leq L$. In the case $\ell = 1$, we have $\mu_1 = \delta_1 = \delta_{\lambda_1}$. Fix $\ell > 1$ and assume that $\lambda_{\ell-1}$ is an atom of $\mu_{\ell-1}$ with weight $\beta_{\ell-1}$. Now $\beta_{\ell-1} + \alpha_{\ell-1} - 1 = \beta_\ell \geq \beta_L > 0$. Hence by Proposition D.1, $\nu_{\ell-1} \boxtimes \mu_{\ell-1}$ has an atom $\gamma_{\ell-1}\lambda_{\ell-1}$ with weight β_ℓ. Therefore μ_ℓ has the atom λ_ℓ with weight β_ℓ. The claim follows from the induction on ℓ. To complete the proof, we only need to show that $\lambda_\ell = ||\mu_\ell||_\infty$. Clearly $\lambda_\ell \leq ||\mu_\ell||_\infty$. Note that $||\mu \boxtimes \nu||_\infty \leq ||\mu||_\infty ||\nu||_\infty$. Then $||\mu_\ell||_\infty \leq q_{\ell-1} + \sigma_\ell^2 \gamma_{\ell-1} ||\mu_{\ell-1}||_\infty$. Thus it holds that $||\mu_\ell||_\infty \leq \lambda_\ell$ since $||\mu_1||_\infty = q_0 = \lambda_1$. Hence the claim follows. □

Now we have prepared to prove the desired theorem.

Theorem D.3. *Consider Assumption 4.1 and 4.2. Then for sufficiently larger L, it holds that $||\mu_L||_\infty$ is an atom of μ_L with weight $1 - (L-1)(1-\alpha_{L-1})$, and*

$$\lim_{L\to\infty} L^{-1} ||\mu_L||_\infty = q\varepsilon_2^{-1} \left[1 - \exp\left(-\varepsilon_2\right)\right]. \tag{D.1}$$

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

Proof. Since $\varepsilon_1 < 1$, we have $1 - \alpha_{L-1} < (L-1)^{-1}$ for sufficiently large L. Then $\beta_L = 1 - (L-1)(1-\alpha_{L-1}) > 0$. Hence by Lemma D.2, for any $\ell \leq L$, it holds that $||\mu_\ell||_\infty$ is an atom of μ_ℓ and $||\mu_L||_\infty = q_L \sum_{\ell=0}^{L-1} (\sigma_L^2 \gamma_{L-1})^\ell$. Then by the same discussion as Proposition 4.4, the assertion follows. □

Theorem D.3 shows that the maximum eigenvalue of the conditional FIM H_L is $O(L)$ as $L \to \infty$. Furthermore, we emphasize that the weight $1 - (L-1)(1-\alpha_{L-1}) \sim 1 - \varepsilon_1$ of the maximal eigenvalue $||\mu_L||$ is close to 1. Therefore, eigenvalues of the dual conditional FIM H_L concentrates around $qL(1 + \varepsilon_2/2)$, and the dual FIM approximates the scaled identity operator. Clearly the same property holds for non-zero eigenvalues of the conditional FIM $\mathcal{I}(\theta|x)$.

References

Serban Teodor Belinschi. The atoms of the free multiplicative convolution of two probability distributions. *Integral Equations and Operator Theory*, 46(4):377–386, 2003.

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

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

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

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

Fumio Hiai and Dénes Petz. Asymptotic freeness almost everywhere for random matrices. *Acta scientiarum mathematicarum*, 66(3-4):809–834, 2000. ISSN 0001-6969.