The Spectrum of Nonlinear Random Matrices for Ultra-Wide Neural Networks

Yizhe Zhu
UC Irvine

Joint work with Zhichao Wang (UC San Diego)

arXiv:2109.09304

Deep Learning Theory Symposium, Simons Institute, Berkeley
December 7, 2021

Fully-connected two-layer neural network

Define two-layer neural network $f_\theta : \mathbb{R}^d \to \mathbb{R}$, $\mathbf{x} \mapsto f_\theta(\mathbf{x})$ by

$$f_\theta(X) = \mathbf{w}^\top \frac{1}{\sqrt{N}} \sigma(WX).$$

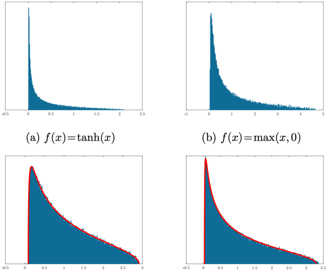
- $X \in \mathbb{R}^{n\times d}$ is the dataset, $W \in \mathbb{R}^{N\times d}$ is the weight matrix. $X_1 = \frac{1}{\sqrt{N}}\sigma(WX)$ is the output of the first hidden layer.
- Training parameters: $\theta = (W, \mathbf{w})$. At initialization, all parameters in θ are drawn from i.i.d. $\mathcal{N}(0, 1)$.
- σ is a Lipschitz function applied entrywise to WX.

Two kernel matrices

1. The **Conjugate Kernel**

$$K^{\mathrm{CK}} = X_1^\top X_1 \in \mathbb{R}^{n\times n}$$

- K^{CK} governs the properties of random feature regression or two-layer network with random first layer weights. [Neal '94], [Williams '97], [Cho, Saul '09], [Rahimi, Recht '09], [Daniely et al '16], [Poole et al '16], [Schoenholz et al '17], [Lee et al '18], [Mei, Montanari '20], ...
- Recently, its limiting spectrum was studied when $N/d \to c_1, d/n \to c_2,\ c_1, c_2 \in (0,\infty)$ in nonlinear random matrices. [Pennington et al '17], [Louart et al '18], [Benigni, Péché '19], [Fan, Wang '20]



(a) $f(x) = \tanh(x)$ (b) $f(x) = \max(x, 0)$

(c) $f(x) = \cos(x)$ (d) $f(x) = x^3 - 3x$

[Benigni, Péché '19]

Two kernel matrices

2. The **Neural Tangent Kernel**

$$
\begin{aligned}
K^{\mathrm{NTK}} :=& (\nabla_\theta f_\theta(X))^\top (\nabla_\theta f_\theta(X)) \in \mathbb{R}^{n\times n} \\
=& X^\top X \odot \left(\frac{1}{N} \sigma'(WX)^\top \operatorname{diag}(\mathbf{w})^2 \sigma'(WX) \right) + X_1^\top X_1.
\end{aligned}
$$

- Training errors evolved during gradient descent is governed by this empirical kernel K^{NTK}. For $N \to \infty$ and fixed n, K^{NTK} converges to its expectation and is fixed over training.

[Jacot, Gabriel, Hongler '18], [Chizat et al '18], [Du et al '19], [Allen-Zhu et al '19], [Lee et al '19], [Arora et al '19], [Adlam et al '20], [Fan, Wang '20], ...

Question:

What are the spectral behaviors of CK and NTK when the width of neural network goes to infinity faster than the training sample size? Namely $N/n \to \infty$ as $n, N \to \infty$ (ultra-wide network).

Semicircle law for sample covariance matrices

Theorem (Bai, Yin '88)

Let $X \in \mathbb{R}^{d\times n}$ be random matrix with i.i.d. entries. If $\mathbb{E}|X_{11}|^4 < \infty$ and $Var(X_{11}) = 1$, then almost surely

$$\lim \operatorname{spec} \sqrt{\frac{d}{n}}\left(\frac{1}{d}X^\top X - \mathrm{Id}\right) = \mu_s,$$

as $d/n \to \infty$ and $n \to \infty$.

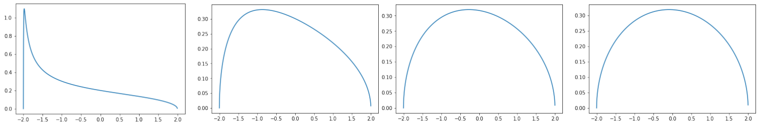


Figure: limiting spectral distributions with increasing $\frac{d}{n}$

Assumptions

$$K^{\mathrm{CK}} = X_1^\top X_1 = \frac{1}{N}\sigma(WX)^\top \sigma(WX).$$

- Approximately pairwise orthogonality of X:

$$\big|\|\mathbf{x}_\alpha\|_2 - 1\big| \le \varepsilon_n, \qquad \big|\mathbf{x}_\alpha^\top \mathbf{x}_\beta\big| \le \varepsilon_n, \qquad n\varepsilon_n^4 \to 0,$$

$$\sum_{\alpha=1}^{n} (\|\mathbf{x}_\alpha\|_2 - 1)^2 \le B^2, \qquad \|X\| \le B.$$

- $\lim \operatorname{spec}(X^\top X) = \mu_0$.
- σ is centered and normalized w.r.t. $\xi \sim N(0,1)$, with bounded σ'' or piecewise linear:

$$\mathbb{E}[\sigma(\xi)] = 0, \qquad \mathbb{E}[\sigma^2(\xi)] = 1, \qquad b_\sigma := \mathbb{E}[\sigma'(\xi)].$$

Deformed semicircle law

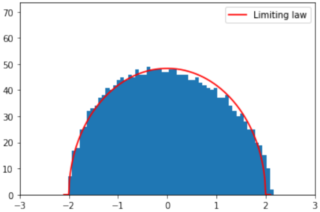
Theorem (Wang, Z. '21)

Under above assumptions, the empirical eigenvalue distribution of

$$\sqrt{\frac{N}{n}}\left(K^{\mathrm{CK}} - \mathbb{E}[K^{\mathrm{CK}}]\right)$$

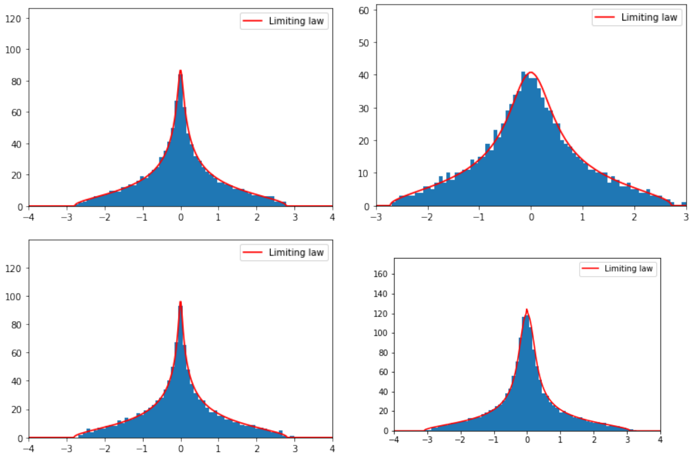
converges weakly to $\mu := \mu_s \boxtimes \left((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_0\right)$ *almost surely as* $N, n, d \to \infty, N/n \to \infty$. *The same result holds for* K^{NTK}.

When $b_\sigma = \mathbb{E}[\sigma'(\xi)] = 0$, $\mu = \mu_s$, independent of μ_0.



$\sigma = \cos(x)$ with normalization, $n = 1.9 \times 10^3$, $d = 2 \times 10^3$ and $N = 2 \times 10^5$

Simulations for Gaussian data



Eigenvalues of $(K^{\mathrm{CK}} - \mathbb{E}K^{\mathrm{CK}})$ and theoretical predictions in red. $\sigma = \frac{e^x}{e^x+1}, x^+, x, \frac{x}{1+e^{-\beta x}}$ with normalization

Ingredients in the proof

- Nonlinear Hanson-Wright inequality: If $\mathbf{y} = \sigma(\mathbf{w}^\top X)^\top, w \sim N(0, I)$, and $\Phi = \mathbb{E}\mathbf{y}\mathbf{y}^\top$ with $\mathbb{E}[\mathbf{y}] = 0$, then, for any $t > 0$,

$$\mathbb{P}\left(|\mathbf{y}^\top A\mathbf{y} - \operatorname{Tr} A\Phi| \geq t\right) \leq 2\exp\left(-\frac{1}{C}\min\left\{\frac{t^2}{4\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{\lambda_\sigma^2\|X\|^2\|A\|}\right\}\right).$$

[Louart et al '18]

- Using Hermite polynomial expansion of σ, we can approximate $\mathbb{E}[X_1^\top X_1] = \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\top X)^\top \sigma(\mathbf{w}^\top X)]$ by

$$\Phi_0 := b_\sigma^2 X^\top X + (1 - b_\sigma^2)\mathrm{Id} + \text{ low-norm terms.}$$

Non-asymptotic bound

Theorem (Wang, Z. '21)

Assume $\sum_{i=1}^n (\|\mathbf{x}_i\|^2 - 1)^2 \le B^2$*, and* $\mathbb{E}\sigma(\xi) = 0$*,* σ *is* λ_σ*-Lipschitz. With probability at least* $1 - 4e^{-2n}$*,*

$$\left\|K^{CK} - \mathbb{E}K^{CK}\right\| \le C\left(\sqrt{\frac{n}{N}} + \frac{n}{N}\right)\lambda_\sigma^2\|X\|^2 + 32B\lambda_\sigma^2\|X\|\sqrt{\frac{n}{N}}.$$

$$\implies \left\|K^{\mathrm{CK}} - \mathbb{E}K^{\mathrm{CK}}\right\| = \Theta\left(\sqrt{n/N}\right) \quad \text{w.h.p.}$$

Similar bounds hold for K^{NTK}.

Random feature regression

- Training labels are given by $\mathbf{y} = X^\top \beta^* + \boldsymbol{\varepsilon}$, $\beta^* \sim \mathcal{N}(0, \sigma_\beta^2 \mathrm{Id})$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{Id})$.
- A test data $\mathbf{x} \in \mathbb{R}^d$ is independent with X such that $\tilde{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}] \in \mathbb{R}^{d \times (n+1)}$ is also (ε_n, B)-orthonormal, and $\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top] = \frac{1}{d}\mathrm{Id}$.
- Test error $\mathcal{L}(\hat{f}) := \mathbb{E}_{\mathbf{x}}[|\hat{f}(\mathbf{x}) - f^*(\mathbf{x})|^2]$.

Theorem (Test error approximation)

For any $\varepsilon \in (0, 1/2)$, the difference of test errors satisfies

$$\left(\frac{N}{n}\right)^{\frac{1}{2}-\varepsilon} \left|\mathcal{L}(\hat{f}_\lambda^{(RF)}(\mathbf{x})) - \mathcal{L}(\hat{f}_\lambda^{(\mathbb{E}K^{CK})}(\mathbf{x}))\right| \to 0,$$

in probability, when $N/n \to \infty$ and $n \to \infty$.

[Mei et al. '21]

Thank You!