Nonlinear random matrix theory for neural networks

Zhichao Wang
Department of Mathematics, UC San Deigo
zhw036@ucsd.edu

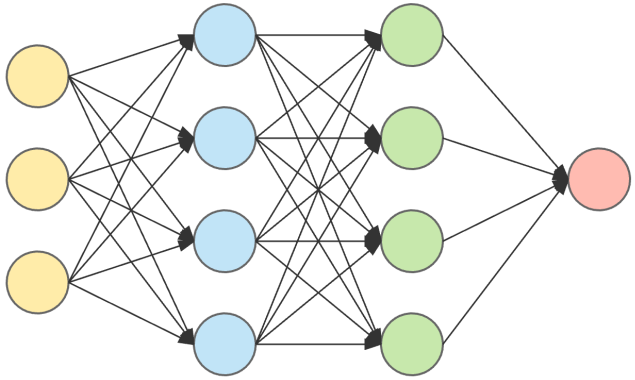
Frontier Probability Days 2021

Motivations from deep learning theory

Fully-connected feedforward neural network

Function $f_\theta : \mathbb{R}^{d_0} \to \mathbb{R}$, $\mathbf{x} \mapsto f_\theta(\mathbf{x})$, defined by

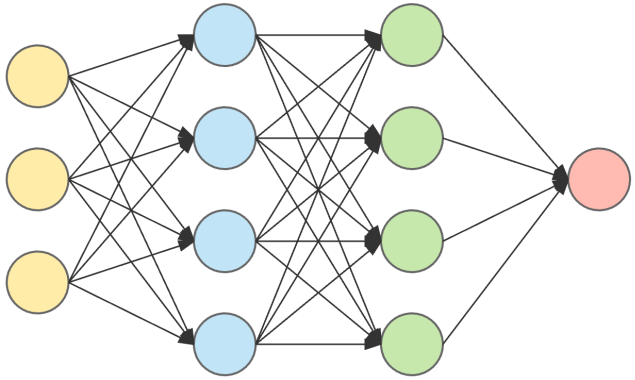
$$f_\theta(\mathbf{x}) = \mathbf{w}^\top \frac{1}{\sqrt{d_L}} \sigma\Big(W_L \frac{1}{\sqrt{d_{L-1}}} \sigma\Big(\dots \frac{1}{\sqrt{d_2}} \sigma\Big(W_2 \frac{1}{\sqrt{d_1}} \sigma(W_1 \mathbf{x}) \Big) \Big) \Big).$$



Fully-connected feedforward neural network

Function $f_\theta : \mathbb{R}^{d_0} \to \mathbb{R}$, $\mathbf{x} \mapsto f_\theta(\mathbf{x})$, defined by

$$f_\theta(\mathbf{x}) = \mathbf{w}^\top \frac{1}{\sqrt{d_L}} \sigma\Big(W_L \frac{1}{\sqrt{d_{L-1}}} \sigma\Big(\dots \frac{1}{\sqrt{d_2}} \sigma\Big(W_2 \frac{1}{\sqrt{d_1}} \sigma(W_1 \mathbf{x}) \Big) \Big) \Big).$$

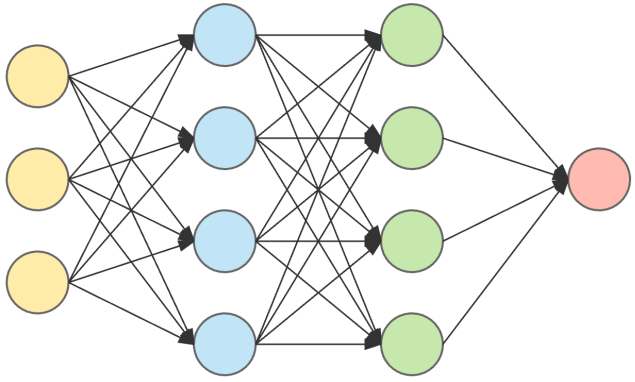


- $W_1 \in \mathbb{R}^{d_1 \times d_0}$, $W_2 \in \mathbb{R}^{d_2 \times d_1}$, …, $W_L \in \mathbb{R}^{d_L \times d_{L-1}}$, and $\mathbf{w} \in \mathbb{R}^{d_L}$. Training parameters: $\theta = (W_1, \dots, W_L, \mathbf{w})$.

Fully-connected feedforward neural network

Function $f_\theta : \mathbb{R}^{d_0} \to \mathbb{R}$, $\mathbf{x} \mapsto f_\theta(\mathbf{x})$, defined by

$$f_\theta(\mathbf{x}) = \mathbf{w}^\top \frac{1}{\sqrt{d_L}} \sigma\Big(W_L \frac{1}{\sqrt{d_{L-1}}} \sigma\Big(\dots \frac{1}{\sqrt{d_2}} \sigma\Big(W_2 \frac{1}{\sqrt{d_1}} \sigma(W_1 \mathbf{x}) \Big) \Big) \Big).$$

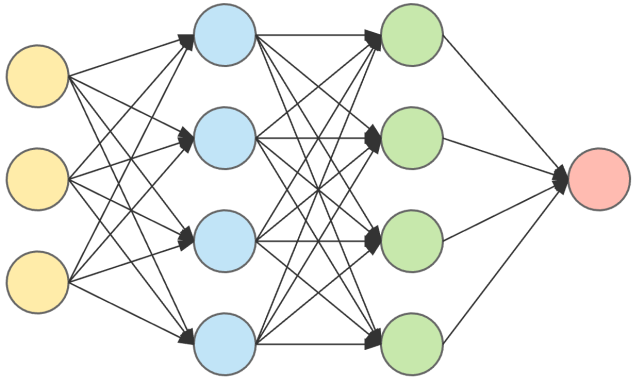


- $W_1 \in \mathbb{R}^{d_1 \times d_0}$, $W_2 \in \mathbb{R}^{d_2 \times d_1}$, …, $W_L \in \mathbb{R}^{d_L \times d_{L-1}}$, and $\mathbf{w} \in \mathbb{R}^{d_L}$. Training parameters: $\theta = (W_1, \dots, W_L, \mathbf{w})$.
- Training samples: $X_0 = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{d_0 \times n}$.

Fully-connected feedforward neural network

Function $f_\theta : \mathbb{R}^{d_0} \to \mathbb{R}$, $\mathbf{x} \mapsto f_\theta(\mathbf{x})$, defined by

$$f_\theta(\mathbf{x}) = \mathbf{w}^\top \frac{1}{\sqrt{d_L}} \sigma\Big(W_L \frac{1}{\sqrt{d_{L-1}}} \sigma\Big(\dots \frac{1}{\sqrt{d_2}} \sigma\Big(W_2 \frac{1}{\sqrt{d_1}} \sigma(W_1 \mathbf{x}) \Big) \Big) \Big).$$



- $W_1 \in \mathbb{R}^{d_1 \times d_0}$, $W_2 \in \mathbb{R}^{d_2 \times d_1}$, …, $W_L \in \mathbb{R}^{d_L \times d_{L-1}}$, and $\mathbf{w} \in \mathbb{R}^{d_L}$. Training parameters: $\theta = (W_1, \dots, W_L, \mathbf{w})$.
- Training samples: $X_0 = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{d_0 \times n}$.
- Post-activation of each layer: $X_\ell = \frac{1}{\sqrt{d_\ell}} \sigma\left(W_\ell X_{\ell-1}\right) \in \mathbb{R}^{d_\ell \times n}$, for $1 \le \ell \le L$.

Our model and assumptions

- ***Random initialization*** of the weights θ (i.i.d. Gaussian).

Our model and assumptions

- ***Random initialization*** of the weights θ (i.i.d. Gaussian).
- Training samples $X_0 = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ have a conjugate spectrum: $\lim \operatorname{spec} X_0^\top X_0 = \mu_0$, where "lim spec" denotes the limit of the ***eigenvalue distribution***.

Our model and assumptions

- ***Random initialization*** of the weights θ (i.i.d. Gaussian).
- Training samples $X_0 = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ have a conjugate spectrum: $\lim \operatorname{spec} X_0^\top X_0 = \mu_0$, where "lim spec" denotes the limit of the ***eigenvalue distribution***.
- Training samples are ***approximately pairwise orthogonal***.

Our model and assumptions

- ***Random initialization*** of the weights θ (i.i.d. Gaussian).
- Training samples $X_0 = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ have a conjugate spectrum: $\lim \operatorname{spec} X_0^\top X_0 = \mu_0$, where "lim spec" denotes the limit of the ***eigenvalue distribution***.
- Training samples are ***approximately pairwise orthogonal***.
  - Non-white Gaussian inputs $\mathbf{x}_\alpha \sim \mathcal{N}(0, \Sigma)$

Our model and assumptions

- ***Random initialization*** of the weights θ (i.i.d. Gaussian).
- Training samples $X_0 = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ have a conjugate spectrum: $\lim \operatorname{spec} X_0^\top X_0 = \mu_0$, where "lim spec" denotes the limit of the ***eigenvalue distribution***.
- Training samples are ***approximately pairwise orthogonal***.
  - Non-white Gaussian inputs $\mathbf{x}_\alpha \sim \mathcal{N}(0, \Sigma)$
  - $\mathbf{x}_\alpha$ drawn from multi-class Gaussian mixture models

Our model and assumptions

- ***Random initialization*** of the weights θ (i.i.d. Gaussian).
- Training samples $X_0 = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ have a conjugate spectrum: $\lim \operatorname{spec} X_0^\top X_0 = \mu_0$, where "lim spec" denotes the limit of the ***eigenvalue distribution***.
- Training samples are ***approximately pairwise orthogonal***.
  - Non-white Gaussian inputs $\mathbf{x}_\alpha \sim \mathcal{N}(0, \Sigma)$
  - $\mathbf{x}_\alpha$ drawn from multi-class Gaussian mixture models
  - Modified real-world data like CIFAR-10 dataset

Training the neural network

Given $f_\theta(X_0) := (f_\theta(\mathbf{x}_1), \ldots, f_\theta(\mathbf{x}_n))$ with training labels $Y_0 = (y_1, \ldots, y_n)$, and the squared loss $L(\theta) := \frac{1}{2}\|f_\theta(X_0) - Y_0\|_2^2$. Then the continuous time gradient descent shows the evolution of parameters θ_t and $f_{\theta_t}(X_0)$:

$$\frac{d\theta_t}{dt} = -\eta \nabla_{\theta_t} L(\theta_t) = -\eta (f_{\theta_t}(X_0) - Y_0) \nabla_{\theta_t} f_{\theta_t}(X_0)^\top \qquad (1)$$

Training the neural network

Given $f_\theta(X_0) := (f_\theta(\mathbf{x}_1), \ldots, f_\theta(\mathbf{x}_n))$ with training labels $Y_0 = (y_1, \ldots, y_n)$, and the squared loss $L(\theta) := \frac{1}{2}\|f_\theta(X_0) - Y_0\|_2^2$. Then the continuous time gradient descent shows the evolution of parameters θ_t and $f_{\theta_t}(X_0)$:

$$\frac{d\theta_t}{dt} = -\eta \nabla_{\theta_t} L(\theta_t) = -\eta (f_{\theta_t}(X_0) - Y_0) \nabla_{\theta_t} f_{\theta_t}(X_0)^\top \tag{1}$$

Hence, if $u(t) := f_{\theta_t}(X_0) - Y_0 \in \mathbb{R}^{1\times n}$, then

$$\frac{du(t)}{dt} = -\eta u(t) \nabla_{\theta_t} f_{\theta_t}(X_0)^\top \nabla_{\theta_t} f_{\theta_t}(X_0). \tag{2}$$

Training the neural network

Given $f_\theta(X_0) := (f_\theta(\mathbf{x}_1), \ldots, f_\theta(\mathbf{x}_n))$ and training labels $Y_0 = (y_1, \ldots, y_n)$, the squared loss is $L(\theta) := \frac{1}{2}\|f_\theta(X_0) - Y_0\|_2^2$. Then the continuous time gradient descent shows the evolution of

parameters θ_t and $f_{\theta_t}(X_0)$:

$$\frac{d\theta_t}{dt} = -\eta \nabla_{\theta_t} L(\theta_t) = -\eta (f_{\theta_t}(X_0) - Y_0) \nabla_{\theta_t} f_{\theta_t}(X_0)^\top \tag{3}$$

Hence, if $u(t) := f_{\theta_t}(X_0) - Y_0 \in \mathbb{R}^{1\times n}$, then

$$\frac{du(t)}{dt} = -\eta u(t) \nabla_{\theta_t} f_{\theta_t}(X_0)^\top \nabla_{\theta_t} f_{\theta_t}(X_0). \tag{4}$$

Training the neural network

Given $f_\theta(X_0) := (f_\theta(\mathbf{x}_1), \ldots, f_\theta(\mathbf{x}_n))$ and training labels $Y_0 = (y_1, \ldots, y_n)$, the squared loss is $L(\theta) := \frac{1}{2}\|f_\theta(X_0) - Y_0\|_2^2$. Then the continuous time gradient descent shows the evolution of parameters θ_t and $f_{\theta_t}(X_0)$:

$$\frac{d\theta_t}{dt} = -\eta \nabla_{\theta_t} L(\theta_t) = -\eta (f_{\theta_t}(X_0) - Y_0) \nabla_{\theta_t} f_{\theta_t}(X_0)^\top \tag{5}$$

Hence, if $u(t) := f_{\theta_t}(X_0) - Y_0 \in \mathbb{R}^{1\times n}$, then

$$\frac{du(t)}{dt} = -\eta u(t) \underbrace{\nabla_\theta f_\theta(X_0)^\top \nabla_\theta f_\theta(X_0)}_{\text{Neural Tangent Kernel (NTK)}} . \tag{6}$$

Two random kernel matrices

1. The **Neural Tangent Kernel**

$$K^{\mathrm{NTK}} := \nabla_\theta f_\theta(X_0)^\top \nabla_\theta f_\theta(X_0) = X_L^\top X_L + \sum_{\ell=1}^{L} (S_\ell^\top S_\ell) \odot (X_{\ell-1}^\top X_{\ell-1}),$$

where $X_\ell = \frac{1}{\sqrt{d_\ell}} \sigma\left(W_\ell X_{\ell-1}\right) \in \mathbb{R}^{d_\ell \times n}$, for $1 \leq \ell \leq L$.

[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], ...

Two random kernel matrices

1. The **Neural Tangent Kernel**

$$K^{\mathrm{NTK}} := \nabla_\theta f_\theta(X_0)^\top \nabla_\theta f_\theta(X_0) = X_L^\top X_L + \sum_{\ell=1}^{L} (S_\ell^\top S_\ell) \odot (X_{\ell-1}^\top X_{\ell-1}),$$

where $X_\ell = \frac{1}{\sqrt{d_\ell}} \sigma\left(W_\ell X_{\ell-1}\right) \in \mathbb{R}^{d_\ell \times n}$, for $1 \leq \ell \leq L$.

[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], ...

2. The **Conjugate Kernel** (or equivalent Gaussian process kernel)

$$K_\ell^{\mathrm{CK}} := X_\ell^\top X_\ell \in \mathbb{R}^{n \times n}, \qquad \text{for } 1 \leq \ell \leq L.$$

Two random kernel matrices

1. The **Neural Tangent Kernel**

$$K^{\mathrm{NTK}} := \nabla_\theta f_\theta(X_0)^\top \nabla_\theta f_\theta(X_0) = X_L^\top X_L + \sum_{\ell=1}^{L} (S_\ell^\top S_\ell) \odot (X_{\ell-1}^\top X_{\ell-1}),$$

where $X_\ell = \frac{1}{\sqrt{d_\ell}} \sigma\left(W_\ell X_{\ell-1}\right) \in \mathbb{R}^{d_\ell \times n}$, for $1 \leq \ell \leq L$.

[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], ...

2. The **Conjugate Kernel** (or equivalent Gaussian process kernel)

$$K_\ell^{\mathrm{CK}} := X_\ell^\top X_\ell \in \mathbb{R}^{n \times n}, \qquad \text{for } 1 \leq \ell \leq L.$$

[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], ...

Two random kernel matrices

1. The **Neural Tangent Kernel**

$$K^{\mathrm{NTK}} := \nabla_\theta f_\theta(X_0)^\top \nabla_\theta f_\theta(X_0) = X_L^\top X_L + \sum_{\ell=1}^{L} (S_\ell^\top S_\ell) \odot (X_{\ell-1}^\top X_{\ell-1}),$$

where $X_\ell = \frac{1}{\sqrt{d_\ell}} \sigma\left(W_\ell X_{\ell-1}\right) \in \mathbb{R}^{d_\ell \times n}$, for $1 \leq \ell \leq L$.

[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], ...

2. The **Conjugate Kernel** (or equivalent Gaussian process kernel)

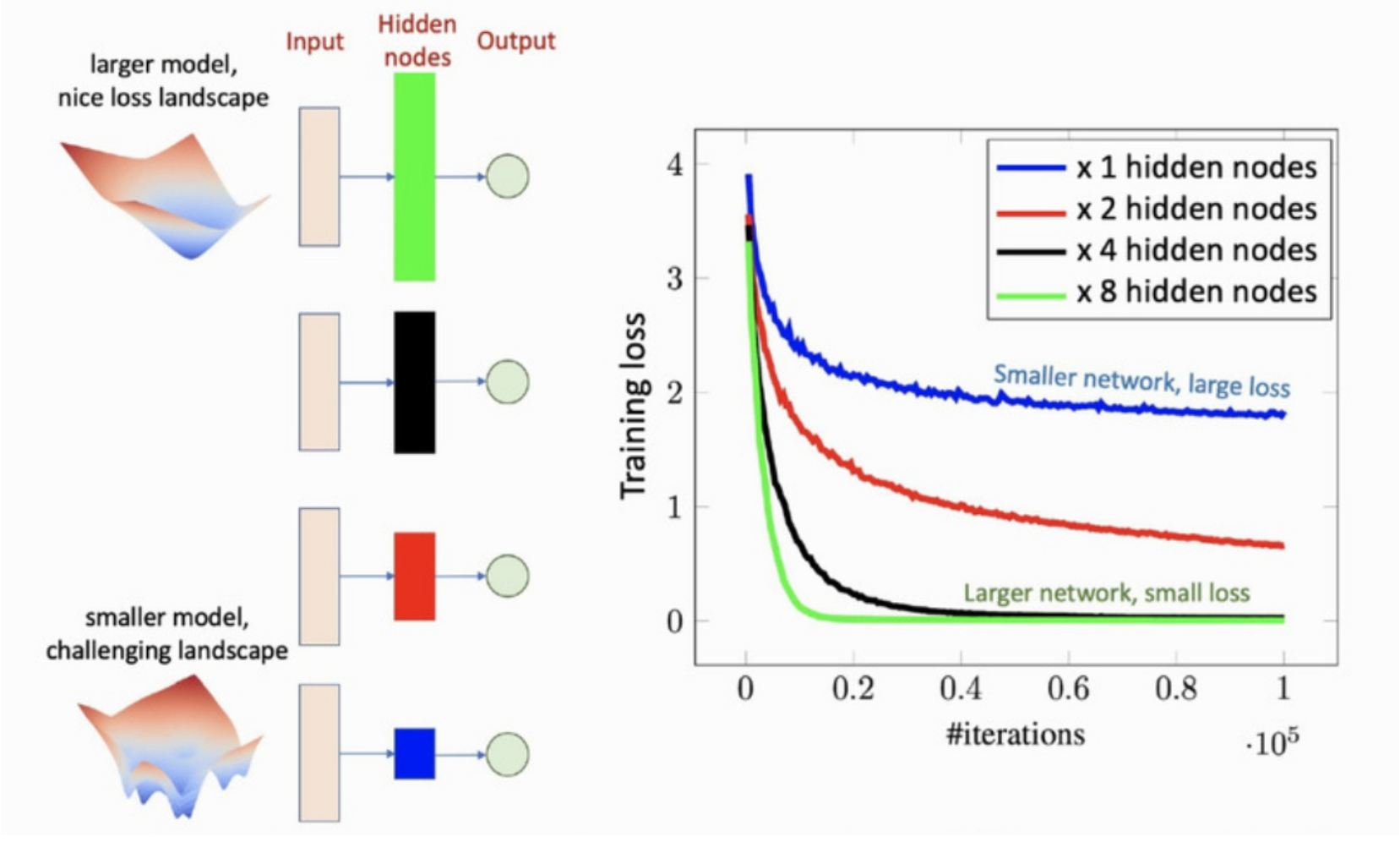
$$K_\ell^{\mathrm{CK}} := X_\ell^\top X_\ell \in \mathbb{R}^{n \times n}, \qquad \text{for } 1 \leq \ell \leq L.$$

[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], ...

[Pennington et al '17], [Louart et al '18], [Benigni, Péché '19], [Piccolo, Schröder '21],...

Overparameterized neural networks

Practical neural networks are typically overparameterized, i.e. widths $d_\ell \to \infty$.



Linear-width regime

Main results

Theorem (Fan, W.)

*For fixed L, almost surely as $n, d_1, \ldots, d_L \to \infty$ under **linear-width** regime where $n/d_\ell \to \gamma_\ell \in (0, \infty)$ for each ℓ,*

$$\lim \operatorname{spec} K^{CK} = \mu_\ell, \qquad \lim \operatorname{spec} K^{NTK} = \mu_{NTK}$$

for probability distributions μ_ℓ and μ_{NTK}, defined by μ_0 and properties of $\sigma(x)$.

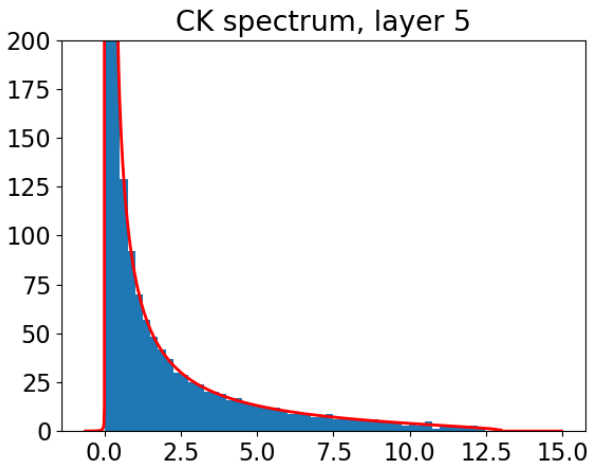
Main results

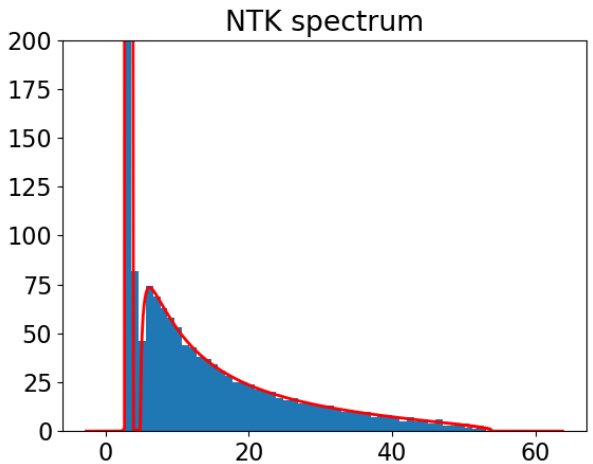
Theorem (Fan, W.)

For fixed L, almost surely as $n, d_1, \ldots, d_L \to \infty$ under ***linear-width*** *regime where $n/d_\ell \to \gamma_\ell \in (0, \infty)$ for each ℓ,*

$$\lim \operatorname{spec} K^{CK} = \mu_\ell, \qquad \lim \operatorname{spec} K^{NTK} = \mu_{NTK}$$

for probability distributions μ_ℓ and μ_{NTK}, defined by μ_0 and properties of $\sigma(x)$.





Simulated eigenvalues in blue for i.i.d. Gaussian X_0; limit spectrum in red.
$\sigma(x) \propto \tan^{-1}(x)$, $L = 5$, $n = 3000$, $d_0 = 1000$, $d_1 = \ldots = d_5 = 6000$

Marcenko-Pastur map

Let $X \in \mathbb{R}^{d \times n}$ have i.i.d. $\mathcal{N}(0, 1/d)$ entries, let $\Phi \in \mathbb{R}^{n \times n}$ be deterministic and positive semi-definite, and let $n \to \infty$ such that $\lim \operatorname{spec} \Phi = \mu$ and $n/d \to \gamma \in (0, \infty)$. Then the sample covariance matrix $\Phi^{1/2} X^\top X \Phi^{1/2}$ has an almost sure spectral limit,

$$\lim \operatorname{spec} \Phi^{1/2} X^\top X \Phi^{1/2} = \rho_\gamma^{\mathrm{MP}} \boxtimes \mu. \tag{7}$$

(Note that $\mathbb{E}[\Phi^{1/2} X^\top X \Phi^{1/2}] = \Phi$, but this limit distribution is not $\mu = \lim \operatorname{spec} \Phi$.)

Marcenko-Pastur map

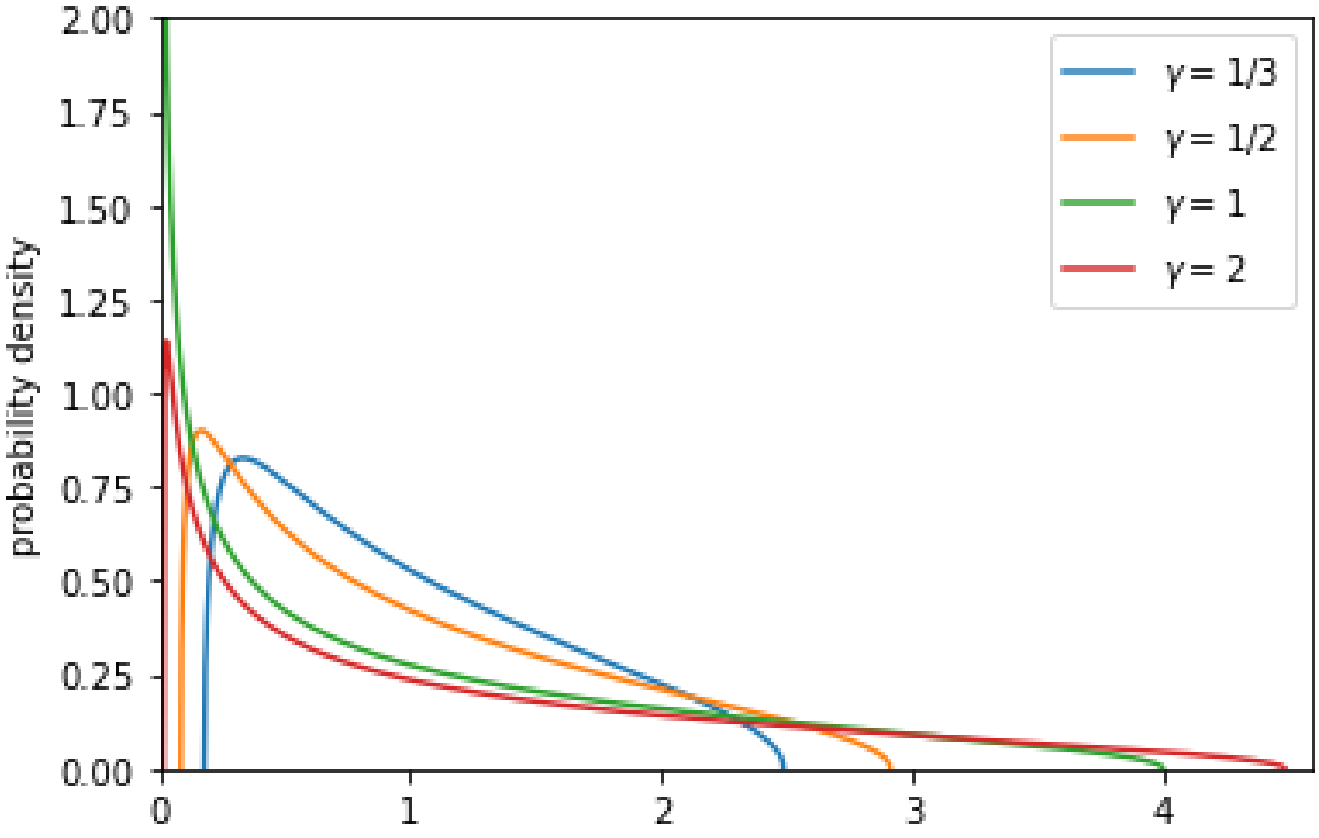
Let $X \in \mathbb{R}^{d \times n}$ have i.i.d. $\mathcal{N}(0, 1/d)$ entries, let $\Phi \in \mathbb{R}^{n \times n}$ be deterministic and positive semi-definite, and let $n \to \infty$ such that $\lim \operatorname{spec} \Phi = \mu$ and $n/d \to \gamma \in (0, \infty)$. Then the sample covariance matrix $\Phi^{1/2} X^\top X \Phi^{1/2}$ has an almost sure spectral limit,

$$\lim \operatorname{spec} \Phi^{1/2} X^\top X \Phi^{1/2} = \rho_\gamma^{\mathrm{MP}} \boxtimes \mu. \tag{7}$$

(Note that $\mathbb{E}[\Phi^{1/2} X^\top X \Phi^{1/2}] = \Phi$, but this limit distribution is not $\mu = \lim \operatorname{spec} \Phi$.)

We will call $\mu \mapsto \rho_\gamma^{\mathrm{MP}} \boxtimes \mu$ the Marcenko-Pastur map of μ with aspect ratio γ.

Marcenko-Pastur distribution



Marcenko-Pastur distribution $\rho_\gamma^{\mathrm{MP}}$ with different ratios γ.

Limit spectral distribution of the CK

Theorem (Fan, W.)

Iteratively for $\ell = 1, \ldots, L$, *define*

$$\mu_\ell = \rho_{\gamma_\ell}^{MP} \boxtimes \left((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\ell-1} \right) \tag{8}$$

where $b_\sigma = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[\sigma'(\xi)]$.[a]

Limit spectral distribution of the CK

Theorem (Fan, W.)

Iteratively for $\ell = 1, \ldots, L$, *define*

$$\mu_\ell = \rho_{\gamma_\ell}^{MP} \boxtimes \left((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\ell-1} \right) \tag{8}$$

where $b_\sigma = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[\sigma'(\xi)]$.[a] *Then each* $\ell = 1, \ldots, L$,

$$\lim \operatorname{spec} X_\ell^\top X_\ell = \mu_\ell.$$

[a] We normalize σ so that $\mathbb{E}[\sigma(\xi)] = 0$, $\mathbb{E}[\sigma(\xi)^2] = 1$.

Limit spectral distribution of the CK

Theorem (Fan, W.)

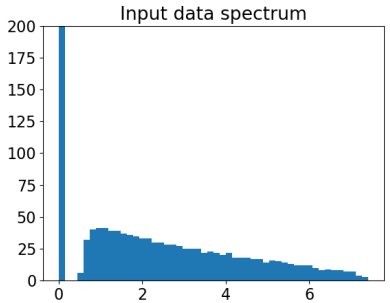
Iteratively for $\ell = 1, \ldots, L$, *define*

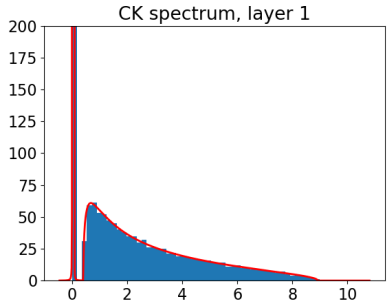
$$\mu_\ell = \rho_{\gamma_\ell}^{MP} \boxtimes \left((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\ell-1} \right) \tag{8}$$

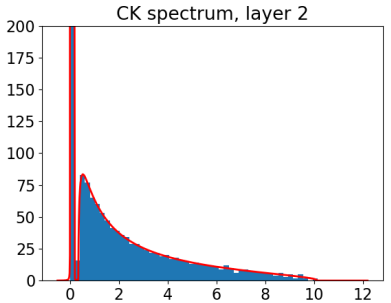
where $b_\sigma = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[\sigma'(\xi)]$.[a] *Then each* $\ell = 1, \ldots, L$,

$$\lim \operatorname{spec} X_\ell^\top X_\ell = \mu_\ell.$$

[a] We normalize σ so that $\mathbb{E}[\sigma(\xi)] = 0$, $\mathbb{E}[\sigma(\xi)^2] = 1$.







Simulation for i.i.d. Gaussian input. $\sigma(x) \propto \tan^{-1}(x)$.

Limit spectral distribution of the CK

Theorem (Fan, W.)

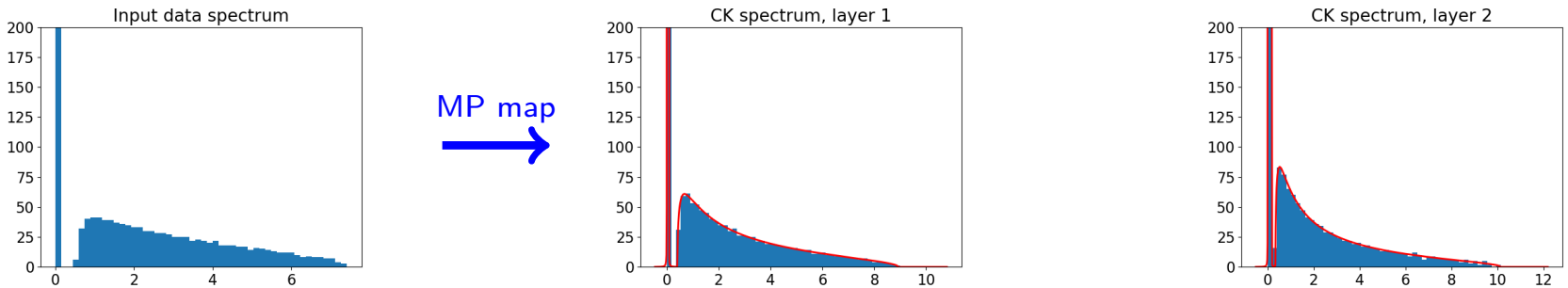
Iteratively for $\ell = 1, \ldots, L$, *define*

$$\mu_\ell = \rho_{\gamma_\ell}^{MP} \boxtimes \left((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\ell-1} \right) \tag{8}$$

where $b_\sigma = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[\sigma'(\xi)]$.[a] *Then each* $\ell = 1, \ldots, L$,

$$\lim \operatorname{spec} X_\ell^\top X_\ell = \mu_\ell.$$

[a] We normalize σ so that $\mathbb{E}[\sigma(\xi)] = 0$, $\mathbb{E}[\sigma(\xi)^2] = 1$.



Simulation for i.i.d. Gaussian input. $\sigma(x) \propto \tan^{-1}(x)$.

Limit spectral distribution of the CK

Theorem (Fan, W.)

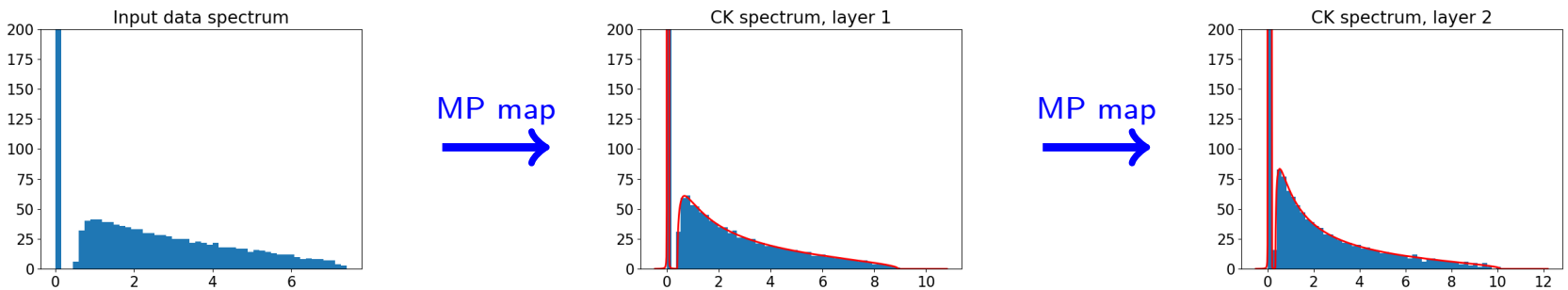
Iteratively for $\ell = 1, \ldots, L$, *define*

$$\mu_\ell = \rho_{\gamma_\ell}^{MP} \boxtimes \left((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\ell-1} \right) \tag{8}$$

where $b_\sigma = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[\sigma'(\xi)]$.[a] *Then each* $\ell = 1, \ldots, L$,

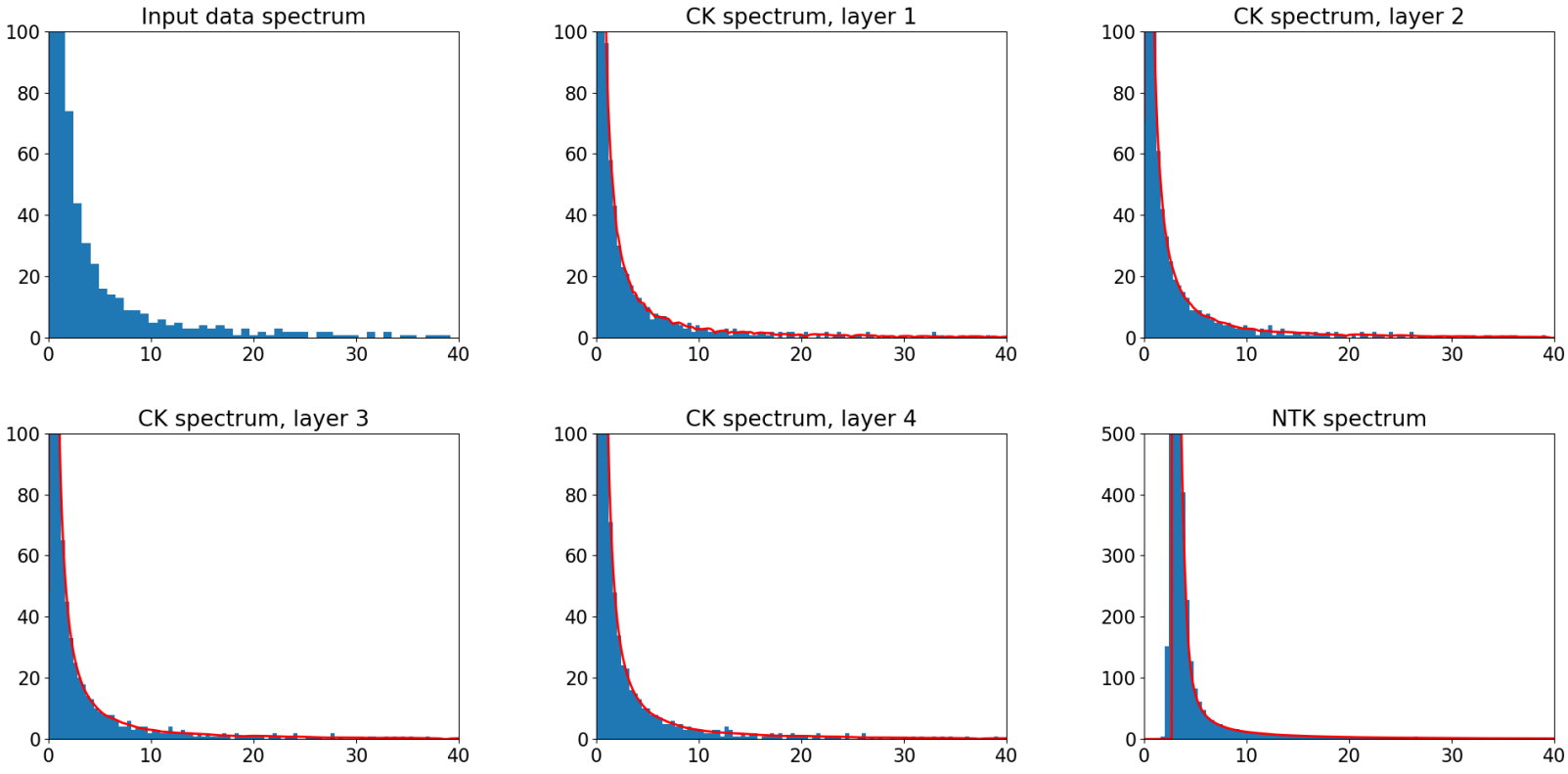
$$\lim \operatorname{spec} X_\ell^\top X_\ell = \mu_\ell.$$

[a] We normalize σ so that $\mathbb{E}[\sigma(\xi)] = 0$, $\mathbb{E}[\sigma(\xi)^2] = 1$.



Simulation for i.i.d. Gaussian input. $\sigma(x) \propto \tan^{-1}(x)$.

Simulations for input images from CIFAR-10



5000 random training images from CIFAR-10, w/ top 10 PCs removed to improve pairwise orthogonality

$\sigma(x) \propto \tan^{-1}(x)$, $L = 5$, $n = 5000$, $d_0 = 3072$, $d_1 = \ldots = d_5 = 10000$

Ultra-wide regime

Question:

What are the behaviors of CK and NTK when the width of neural network goes to infinity faster than the training sample size? Namely $n/d_1 \to 0$ (ultra-wide regime).

Question:

What are the behaviors of CK and NTK when the width of neural network goes to infinity faster than the training sample size? Namely $n/d_1 \to 0$ (ultra-wide regime).

Notice that $X_1^\top X_1 = \frac{1}{d_1}\sigma(W_1 X_0)^\top \sigma(W_1 X_0)$ is a generalized sample covariance matrix with covariance

$$\Phi_1 := \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\top X_0) \otimes \sigma(\mathbf{w}^\top X_0)],$$

Question:

What are the behaviors of CK and NTK when the width of neural network goes to infinity faster than the training sample size? Namely $n/d_1 \to 0$ (ultra-wide regime).

Notice that $X_1^\top X_1 = \frac{1}{d_1}\sigma(W_1 X_0)^\top \sigma(W_1 X_0)$ is a generalized sample covariance matrix with covariance

$$\Phi_1 := \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\top X_0) \otimes \sigma(\mathbf{w}^\top X_0)],$$

where $\mathbf{w} \in \mathcal{N}(0, \mathrm{Id})$. Hence, as $n/d_1 \to 0$, $X_1^\top X_1$ will concentrate around its expectation

$$\mathbb{E}[X_1^\top X_1] = \Phi_1.$$

Concentration inequality

Theorem (W., Zhu)

With probability at least $1 - 4e^{-2n}$,

$$\left\| X_1^\top X_1 - \Phi_1 \right\| \leq C \|X_0\| \sqrt{\frac{n}{d_1}},$$

where $C > 0$ *is a universal constant. Moreover, with high probability*

$$\lambda_{\min} \left(X_1^\top X_1 \right) \geq 1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2 - o(1),$$

when $n/d_1 \to 0$ *and* $n \to \infty$.

Deformed semicircle law for two-layer neural networks

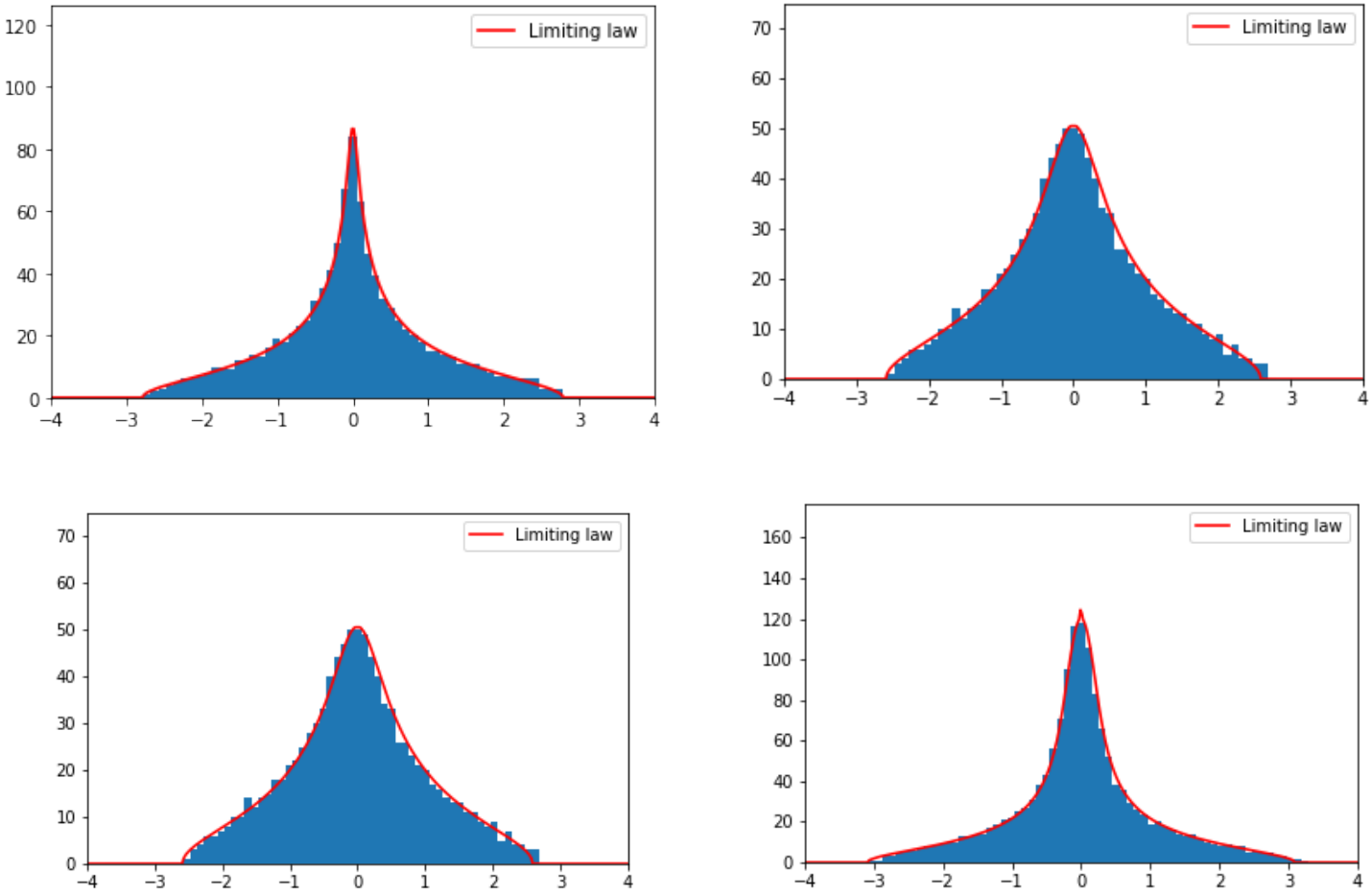
Theorem (W., Zhu)

*As $n, d_0, d_1 \to \infty$ under **ultra-wide** regime where $n/d_1 \to 0$, the empirical eigenvalue distribution of*

$$\sqrt{\frac{d_1}{n}}\left(X_1^\top X_1 - \mathbb{E}[X_1^\top X_1]\right) \text{ and } \sqrt{\frac{d_1}{n}}\left(K^{NTK} - \mathbb{E}[K^{NTK}]\right)$$

both converge weakly to $\mu_s \boxtimes \left((1-b_\sigma^2) + b_\sigma^2 \cdot \mu_0\right)$ almost surely, where μ_s is the standard semicircular law.

Simulations for Gaussian data



Eigenvalues of $X_1^\top X_1$ with different activation functions and theoretical predictions in red.

References

Linear-width regime:

Zhou Fan, Zhichao Wang, "Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks." In Advances in Neural Information Processing Systems, 2020. arXiv:2005.11879.

Ultra-wide regime:

Zhichao Wang, Yizhe Zhu, "Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks." arXiv:2109.09304.

Thanks for your listening!