A Random Matrix Approach to Neural Networks: From Linear to Nonlinear, and from Shallow to Deep

@ Fudan University, Shanghai, China

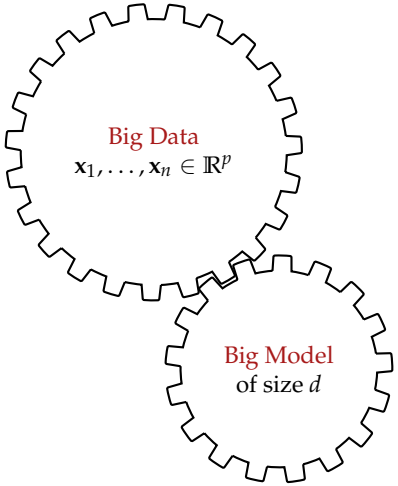
Zhenyu Liao

joint work with H. Tiomoko (UK), R. Couillet (UGA, France), Z. Ling (HUST, China), and Michael W. Mahoney (UC Berkeley, USA)

EIC, Huazhong University of Science and Technology (HUST)

April 22, 2025

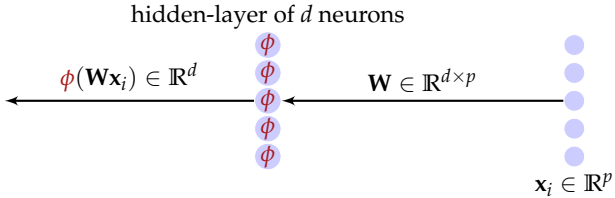
Motivation: understanding large-dimensional machine learning



- **Big Data era**: exploit large n, p, d
- counterintuitive phenomena different from classical asymptotics statistics
- change of understanding of many methods in statistics and machine learning
- **Random Matrix Theory (RMT)** provides the tools!
- In this talk, a review of some recent progress on RMT analysis of **neural networks** models, from linear to nonlinear, and from shallow to deep

Outline

A deep neural network model



- **linear transformation** with first-layer weight matrix $\mathbf{W} \in \mathbb{R}^{d\times p}$
- **nonlinear transformation**: activation function $\phi\colon \mathbb{R} \to \mathbb{R}$ acting entry-wise on $\mathbf{W}\mathbf{x}_i$
- **data representation** at the output of first-layer $\boxed{\mathbf{x}_i \mapsto \phi(\mathbf{W}\mathbf{x}_i)}$
- do the same thing in a layer-by-layer fashion:

$$f(\mathbf{x}_i) = \frac{1}{\sqrt{d_L}}\mathbf{w}^{\mathsf{T}}\phi_L\left(\frac{1}{\sqrt{d_{L-1}}}\mathbf{W}_L\phi_{L-1}\left(\cdots\frac{1}{\sqrt{d_2}}\phi_2\left(\frac{1}{\sqrt{d_1}}\mathbf{W}_2\phi_1(\mathbf{W}_1\mathbf{x}_i)\right)\right)\right), \tag{1}$$

for a large number n of input data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$

Technical challenges and key ideas

Analyze and Optimize Large-scale ML model $\phi(\mathbf{X}, \mathbf{\Theta})$

Objective: Evaluation of $\phi(\mathbf{X}, \mathbf{\Theta})$ via Performance Metric $f(\cdot)$

Technical Challenge 1 High-dimensionality in $\mathbf{X}, \mathbf{\Theta}$	**Key Idea 1** Concentration of $f(\phi(\mathbf{X}, \mathbf{\Theta})) \simeq \mathbb{E}[f(\phi(\mathbf{X}, \mathbf{\Theta}))]$
Technical Challenge 2 Analysis of Eigen-functional	**Key Idea 2** Deterministic Equivalent for Resolvent
Technical Challenge 3 Non-linearity in ML model	**Key Idea 3** High-dimensional linearization of $\phi(\mathbf{X}, \mathbf{\Theta})$

High-dimensional Equivalent

Definition (High-dimensional Equivalent)

Let $\phi(\mathbf{X})$ be a nonlinear model of a random matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$, and let $f(\phi(\mathbf{X}))$ be a scalar functional with entrywise $\phi\colon \mathbb{R}^{p\times n} \to \mathbb{R}^{p\times n}$ and observation map $f\colon \mathbb{R}^{p\times n} \to \mathbb{R}$. We say $\bar{\mathbf{X}}_\phi$ (which can be deterministic or random) is a High-dimensional Equivalent of $\phi(\mathbf{X})$ with respect to $f(\cdot)$ if

$$f(\phi(\mathbf{X})) - f(\bar{\mathbf{X}}_\phi) \to 0, \tag{2}$$

in probability or almost surely as $n, p \to \infty$. We denote $\phi(\mathbf{X}) \overset{f}{\leftrightarrow} \bar{\mathbf{X}}_\phi$ or simply $\phi(\mathbf{X}) \leftrightarrow \bar{\mathbf{X}}_\phi$.

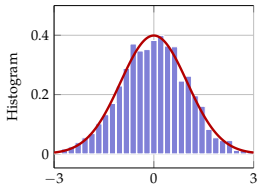
- without (entrywise) nonlinearities, $f(\mathbf{X})$ concentrates around expectation $f(\mathbf{X}) \simeq \mathbb{E}[f(\mathbf{X})]$, and can be assessed through Deterministic Equivalent $f(\bar{\mathbf{X}})$;
- for scalar eigenspectral functionals, Deterministic Equivalent for Resolvent framework provides a unified approach to eigenspectral functionals of random matrices;
- for nonlinear models in two different scaling regimes (LLN versus CLT), $\phi(\mathbf{X})$ can be linearized to yield a Linear Equivalent.

Concentration versus non-concentration behavior

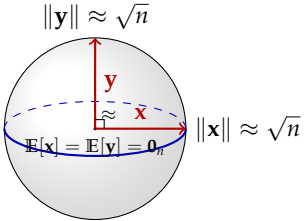
"Concentration" versus "non-concentration" around the mean

Consider two independent random vectors $\mathbf{x} = [x_1, \ldots, x_n]^\top$ and $\mathbf{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$, with i.i.d. entries of zero mean and unit variance. We have the following contrasting observations.

1. In the one-dimensional case with $n = 1$, random variables "concentrate" around their means, with $\Pr(|x - 0| > t) \leq t^{-2}$ and $\Pr(|y - 0| > t) \leq t^{-2}$ by Markov's inequality.
2. In the multi-dimensional case with $n \gg 1$, random vectors do not "concentrate" around their means, with $\mathbb{E}[\|\mathbf{x} - \mathbf{0}\|^2] = \mathbb{E}[\mathbf{x}^\top \mathbf{x}] = \operatorname{tr}(\mathbb{E}[\mathbf{x}\mathbf{x}^\top]) = n$ and $\mathbb{E}[\|\mathbf{x} - \mathbf{y}\|^2] = \mathbb{E}[\mathbf{x}^\top \mathbf{x} + \mathbf{y}^\top \mathbf{y}] = 2n$.



(a) "Concentration" around the mean



(b) "Non-concentration" around the mean

High-dimensional concentration of scalar observation

- while large random vectors do not "concentrate" round their means, their scalar functionals (often) do
- for a scalar observation map $f\colon \mathbb{R}^n \to \mathbb{R}$ and random vector $\mathbf{x} \in \mathbb{R}^n$, we typically have

$$f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})] \to 0, \tag{3}$$

  with high probability for n large.
- a basic example is the linear function $f(\mathbf{x}) = \mathbf{1}_n^\top \mathbf{x}/n = \frac{1}{n}\sum_{i=1}^n x_i$: By the Large of Large Numbers (LLN) and the Central Limit Theorem (CLT), we have $f(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] + O(n^{-1/2})$ with high probability
- For a random matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$ in the proportional regime with n, p both large, similar holds:

1. just as for vectors, $\mathbf{X}$ does not concentrate, e.g., in a spectral norm sense; for instance, $\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\| \not\to 0$ as $n, p \to \infty$ together.
2. at the same time, scalar (e.g., eigenspectral) functionals $f\colon \mathbb{R}^{p\times n} \to \mathbb{R}$ of the random matrix $\mathbf{X}$ do concentrate; i.e., $f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})] \to 0$ as $n, p \to \infty$.

- This is the key idea of Deterministic Equivalent.

Nonlinear objects in two different scaling regimes

Definition (Two scaling regimes)

Consider a scalar functional $f(\mathbf{x})$ of $\mathbf{x} \in \mathbb{R}^n$, via an observation map $f\colon \mathbb{R}^n \to \mathbb{R}$:

1. **LLN regime**: this holds when $f(\mathbf{x})$ exhibits a LLN-type concentration, strongly concentrating around its mean $\mathbb{E}[f(\mathbf{x})]$, and its distribution function becomes degenerate; that is, it holds when $f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})] \to 0$ in probability or almost surely, as $n \to \infty$.
2. **CLT regime**: this holds when $f(\mathbf{x})$ exhibits a CLT-type concentration, remaining random and maintaining a non-degenerate distribution function; that is, it holds when $\sqrt{n}\,(f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})]) \to \mathcal{N}(0,1)$ in distribution, as $n \to \infty$.

Nonlinear objects in two scaling regimes

Let $\mathbf{x} \in \mathbb{R}^n$ be a random vector such that $\sqrt{n}\mathbf{x}$ has i.i.d. Gaussian entries $\mathcal{N}(0,1)$ (the $\sqrt{n}$ scaling ensures $\mathbb{E}[\|\mathbf{x}\|^2] = 1$). Let $\mathbf{y} \in \mathbb{R}^n$ be a deterministic vector of unit norm $\|\mathbf{y}\| = 1$. Consider two nonlinear objects:

1. **LLN regime**: random variables $f_{\mathrm{LLN}}(\mathbf{x}) = \|\mathbf{x}\|_2^2$ or $f_{\mathrm{LLN}}(\mathbf{x}) = \mathbf{x}^\top \mathbf{y}$ that both exhibit LLN-type concentration (i.e., nearly deterministic for n large), and we are interested in $\phi(f_{\mathrm{LLN}}(\mathbf{x}))$; and
2. **CLT regime**: random variables $f_{\mathrm{CLT}}(\mathbf{x}) = \sqrt{n}(\|\mathbf{x}\|_2^2 - 1)$ or $f_{\mathrm{CLT}}(\mathbf{x}) = \sqrt{n} \cdot \mathbf{x}^\top \mathbf{y}$ that both exhibit CLT-type concentration (they remain inherently random and have non-degenerate distributions for n large), and we are interested in $\phi(f_{\mathrm{CLT}}(\mathbf{x}))$.

Linearization in the two scaling regimes

Theorem (Taylor's theorem)

Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a function that is at least k times continuously differentiable in a neighborhood of some point $\tau \in \mathbb{R}$. Then, there exists $h_k\colon \mathbb{R} \to \mathbb{R}$ such that

$\phi(x) = \phi(\tau) + \phi'(\tau)(x-\tau) + \frac{\phi''(\tau)}{2}(x-\tau)^2 + \ldots + \frac{\phi^{(k)}(\tau)}{k!}(x-\tau)^k + h_k(x)(x-\tau)^k$, *with* $\lim_{x\to\tau} h_k(x) = 0$. *Consequently,* $h_k(x)(x-\tau)^k = o(|x-\tau|^k)$ *as* $x \to \tau$.

Theorem (Hermite polynomial expansion)

The i^{th} normalized Hermite polynomial, $\mathrm{He}_i(t)$, is given by $\mathrm{He}_0(t) = 1, \mathrm{He}_i(t) = \frac{(-1)^i}{\sqrt{i!}} e^{\frac{t^2}{2}} \frac{d^i}{dt^i}\left(e^{-\frac{t^2}{2}}\right), i \geq 1$. *The normalized Hermite polynomials*

1. *are orthogonal with respect to Gaussian measure, i.e.,* $\int \mathrm{He}_m(t)\mathrm{He}_n(t)\mu(dt) = \delta_{mn}$ *for* $\mu(dt) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$; *and*
2. *can be used to formally expand any square-integrable function $\phi \in L^2(\mu)$ as* $\phi(\xi) \sim \sum_{i=0}^{\infty} a_{\phi;i}\mathrm{He}_i(\xi)$, $\quad a_{\phi;i} = \int \phi(t)\mathrm{He}_i(t)\mu(dt) = \mathbb{E}[\phi(\xi)\mathrm{He}_i(\xi)]$, *for $\xi \sim \mathcal{N}(0,1)$. The coefficients $a_{\phi;i}$s are the Hermite coefficients of ϕ:*

$$a_{\phi;0} = \mathbb{E}[\phi(\xi)],\ a_{\phi;1} = \mathbb{E}[\xi\phi(\xi)],\ \sqrt{2}a_{\phi;2} = \mathbb{E}[\xi^2\phi(\xi)] - a_{\phi;0},\ \nu_\phi = \mathbb{E}[\phi^2(\xi)] = \sum_{i=0} a_{\phi;i}^2. \tag{4}$$

Linearization in the two scaling regimes: an example

Example (Distinct linearizations of tanh in two scaling regimes)

Consider $\phi(t) = \tanh(t)$. By Taylor and Hermite polynomial expansion, this nonlinear function is "close" to different quadratic functions, depending on the scaling regime.
Consider $\mathbf{x} \in \mathbb{R}^n$ be a random vector such that $\sqrt{n}\mathbf{x}$ has i.i.d. standard Gaussian entries, and let $\mathbf{y} \in \mathbb{R}^n$ be a deterministic vector of unit norm ($\|\mathbf{y}\| = 1$). Then:

1. **In the LLN regime**, we have for $f_{\text{LLN}}(\mathbf{x}) = \mathbf{x}^\top \mathbf{y}$ that

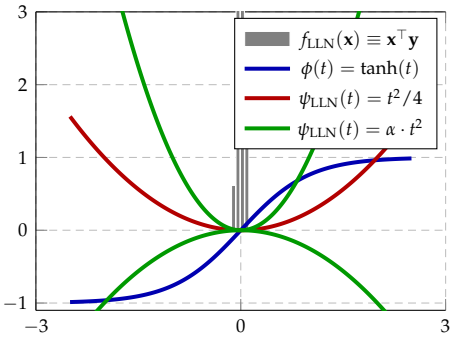
$$\tanh(f_{\text{LLN}}(\mathbf{x})) - \psi_{\text{LLN}}(f_{\text{LLN}}(\mathbf{x})) \to 0, \tag{5}$$

   as $n \to \infty$, with $\psi_{\text{LLN}}(t) = t^2/4$. This is as a consequence of $\tanh(t = 0) = \psi_{\text{LLN}}(t = 0) = 0$. In particular, we also have $\mathbb{E}[\tanh(f_{\text{LLN}}(\mathbf{x}))] \simeq \mathbb{E}[\psi(f_{\text{LLN}}(\mathbf{x}))]$ as a result.

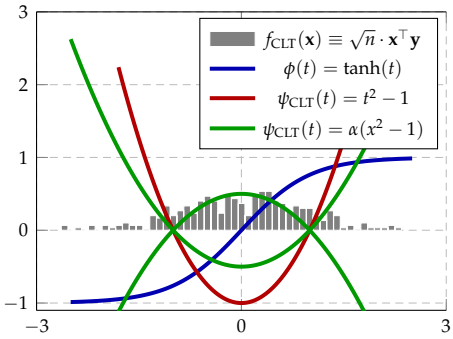
2. **In the CLT regime**, we have for $f_{\text{CLT}}(\mathbf{x}) = \sqrt{n} \cdot \mathbf{x}^\top \mathbf{y}$ that

$$\mathbb{E}[\tanh(f_{\text{CLT}}(\mathbf{x}))] = \mathbb{E}[\psi_{\text{CLT}}(f_{\text{CLT}}(\mathbf{x}))], \tag{6}$$

   in expectation, where the corresponding quadratic function is $\psi_{\text{CLT}}(t) = t^2 - 1$. This follows from the fact that both functions have the same zeroth-order Hermite coefficient, $a_{\tanh;0} = a_{\psi;0} = 0$.



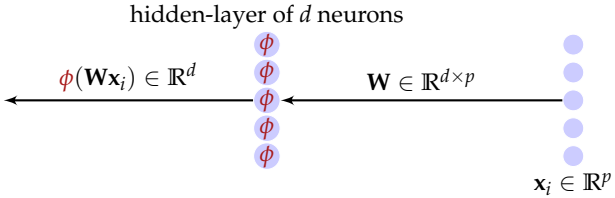
(a) LLN regime



(b) CLT regime

Figure: Different behavior of nonlinear $\phi(f_{\mathrm{LLN}}(\mathbf{x}))$ and $\phi(f_{\mathrm{CLT}}(\mathbf{x}))$ for $\phi(t) = \tanh(t)$ (in **blue**) in the LLN and CLT regime, with $n = 500$. We have $\phi(f_{\mathrm{LLN}}(\mathbf{x})) \simeq \psi_{\mathrm{LLN}}(f_{\mathrm{LLN}}(\mathbf{x}))$ in the LLN regime (as a consequence of $\phi(0) = \psi_{\mathrm{LLN}}(0) = 0$) and $\mathbb{E}[\phi(f_{\mathrm{CLT}}(\mathbf{x}))] = \mathbb{E}[\psi_{\mathrm{CLT}}(f_{\mathrm{CLT}}(\mathbf{x}))]$ in the CLT regime (as a consequence of $a_{\phi;0} = a_{\psi_{\mathrm{CLT}};0} = 0$), with different quadratic functions $\psi_{\mathrm{LLN}}(t) = t^2/4$ and $\psi_{\mathrm{CLT}}(t) = t^2 - 1 = \sqrt{2}\mathrm{He}_2(t)$ in **red**. Note that the these linearizations (in the two different regimes respectively) are not unique and all functions in dashed **green** are also valid linearizations.

Two-layer network with random first layer



Definition (Single-hidden-layer NN model)

Consider a single-hidden-layer NN model with first-layer weights $\mathbf{W} \in \mathbb{R}^{d\times p}$ and second-layer weights $\boldsymbol{\beta} \in \mathbb{R}^d$. For an input vector $\mathbf{x} \in \mathbb{R}^p$, the network output is given by $\hat{y}(\mathbf{x}) = \boldsymbol{\beta}^\top \phi(\mathbf{W}\mathbf{x})$, where $\phi(\cdot)$ is an entrywise activation function. We are interested in the NN performance measured by

1. its **training MSE** $E_{\text{train}} = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}(\mathbf{x}_i))^2 = \frac{1}{n}\|\mathbf{y} - \boldsymbol{\Phi}^\top \boldsymbol{\beta}\|^2$ with $\boldsymbol{\Phi} \equiv \phi(\mathbf{W}\mathbf{X})$ for a training set $(\mathbf{X}, \mathbf{y})$ of size n, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$, $\mathbf{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$; and
2. its **test MSE** $E_{\text{test}} = \frac{1}{n}\sum_{i=1}^{n'} (y'_i - \hat{y}(\mathbf{x}'_i))^2 = \frac{1}{n'}\|\mathbf{y}' - \phi(\mathbf{W}\mathbf{X}')^\top \boldsymbol{\beta}\|^2$ on a test set $(\mathbf{X}', \mathbf{y}')$ of size n', with $\mathbf{X}' = [\mathbf{x}'_1, \ldots, \mathbf{x}'_{n'}] \in \mathbb{R}^{p\times n'}$ and $\mathbf{y}' = [y'_1, \ldots, y'_{n'}]^\top \in \mathbb{R}^{n'}$.

Single-hidden-layer NN model and a Deterministic Equivalent for nonlinear resolvent

- Given first-layer $\mathbf{W}$ and training data $\mathbf{X} \in \mathbb{R}^{p\times n}$, consider the random feature matrix $\mathbf{\Phi} \equiv \phi(\mathbf{WX}) \in \mathbb{R}^{d\times n}$ and regress against the target $\mathbf{y}$ by minimizing the following ridge-regularized MSE

$$L(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}(\mathbf{x}_i))^2 + \frac{\gamma}{2}\|\boldsymbol{\beta}\|_2^2 = \frac{1}{2n}\|\mathbf{y} - \mathbf{\Phi}^\top\boldsymbol{\beta}\|_2^2 + \frac{\gamma}{2}\|\boldsymbol{\beta}\|_2^2, \quad \gamma \geq 0, \tag{7}$$

- solution is uniquely given by $\boldsymbol{\beta}_\gamma = \frac{1}{n}\mathbf{\Phi}\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi} + \gamma\mathbf{I}_n\right)^{-1}\mathbf{y} = \left(\frac{1}{n}\mathbf{\Phi}\mathbf{\Phi}^\top + \gamma\mathbf{I}_d\right)^{-1}\frac{1}{n}\mathbf{\Phi}\mathbf{y}$, for $\gamma > 0$.
- Training MSE is $E_{\text{train}} = \frac{1}{n}\|\mathbf{y} - \mathbf{\Phi}^\top\boldsymbol{\beta}_\gamma\|_2^2 = \frac{\gamma^2}{n}\frac{\partial \mathbf{y}^\top\mathbf{Q}^2(-\gamma)\mathbf{y}}{\partial\gamma}$, with **resolvent** of nonlinear Gram $\mathbf{\Phi}^\top\mathbf{\Phi}$.

$$\mathbf{Q}(-\gamma) \equiv \left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi} + \gamma\mathbf{I}_n\right)^{-1}, \quad \mathbf{\Phi}^\top\mathbf{\Phi} = \phi(\mathbf{X}^\top\mathbf{W}^\top)\phi(\mathbf{WX}). \tag{8}$$

Theorem (Deterministic Equivalent for nonlinear resolvent, [LLC18, Theorem 1])

Let $\mathbf{W} \in \mathbb{R}^{d\times p}$ be a random matrix with i.i.d. sub-gaussian entries of zero mean and unit variance, and let $\mathbf{X} \in \mathbb{R}^{p\times n}$ be independent of $\mathbf{W}$ with $\|\mathbf{X}\|_2 \leq 1$. Then, as $n, p, d \to \infty$ together and for Lipschitz $\phi : \mathbb{R} \to \mathbb{R}$,

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \bar{\mathbf{Q}}(z) = \left(\frac{d}{n}\frac{\mathbf{K}}{1+\delta(z)} - z\mathbf{I}_n\right)^{-1}, \quad \delta(z) = \frac{1}{n}\operatorname{tr}\mathbf{K}\bar{\mathbf{Q}}(z), \quad \mathbf{K} \equiv \mathbb{E}_{\mathbf{w}}[\phi(\mathbf{X}^\top\mathbf{w})\phi(\mathbf{w}^\top\mathbf{X})], \tag{9}$$

where $\delta(z)$ is the unique Stieltjes transform solution, and $\mathbf{K}$ the kernel matrix.

Implications of the Deterministic Equivalent

Scaling law of training MSE

Consider the ridgeless setting with $\gamma = 0$ and the under-parameterized regime with n, p, d all large but $d < n$

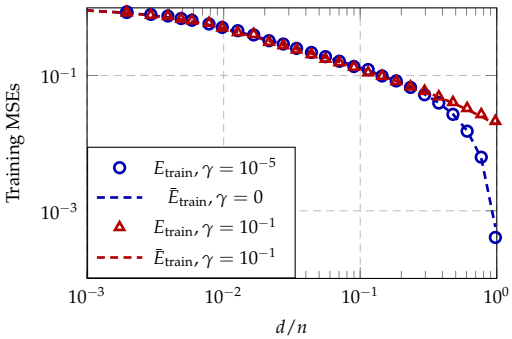
- δ diverges as $\gamma \to 0$, however, $\gamma\delta = \frac{1}{n} \operatorname{tr} \mathbf{K} \left(\frac{d}{n} \frac{\mathbf{K}}{\gamma + \gamma\delta} + \mathbf{I}_n \right)^{-1} \xrightarrow{\gamma \to 0} \theta = \frac{1}{n} \operatorname{tr} \mathbf{K} \left(\frac{d}{n} \frac{\mathbf{K}}{\theta} + \mathbf{I}_n \right)^{-1}$
- explicit scaling laws for the training MSEs that depend on the **eigenspectrum** of $\mathbf{K}$

1. **exponential eigendecay** (e.g., RBF kernel related to cosine activation [RW05]) yields an error decay rate of $\log(n)/n$ (which is slightly slower than the n^{-1} rate of linear models);
2. **polynomial decay** (e.g., Matérn kernel associated with to ReLU activation [Gei+20]) yields an error decay rate of $n^{-1-\beta}$ (with $\beta > 0$), which is faster than the linear case.

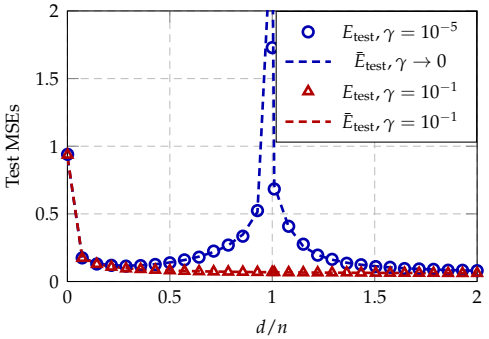
Double descent behavior for test MSE

- it can be checked that both θ and δ diverge as $\gamma \to 0$ at $n/d = 1$.
- thus, the test risk likewise exhibits a singularity at $d/n = 1$.
- mirrors the **double descent** phenomenon for linear models, but applies here to nonlinear NN model, regardless of the activation function or the training/test data.

Numerical results



(a) Training MSE



(b) Test MSE

Figure: Empirical and theoretical training and test MSEs of single-hidden-layer NN model, as a function of d/n, for $\gamma = 10^{-1}$ and $\gamma = 10^{-5}$, with Gaussian $\mathbf{W}$ and ReLU activation $\phi(t) = \max(t, 0)$, $n = 1\,024$ training samples and $n' = 1\,024$ test samples from the MNIST dataset (number 1 and 2).**Figure 3a**: log-log plot of training MSEs averaged over 30 runs. **Figure 3b**: test MSEs averaged over 30 runs on independent test sets of size $\hat{n} = 2\,048$.

Theorem (High-dimensional linearization of kernel matrix)

Let $\mathbf{w} \sim \mathbb{R}^p$ *be standard Gaussian* $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ *and let* $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ *be independently drawn from the unit sphere* $\mathbb{S}^{p-1} \subset \mathbb{R}^p$*. Then, as* $n, p \to \infty$ *with* $p/n \in (0, \infty)$*, the kernel matrix* $\mathbf{K} = \mathbb{E}_{\mathbf{w}}[\phi(\mathbf{X}^\top \mathbf{w})\phi(\mathbf{w}^\top \mathbf{X})]$ *admits the following Linear Equivalent*

$$\mathbf{K} \leftrightarrow \tilde{\mathbf{K}}_\phi, \quad \tilde{\mathbf{K}}_\phi = a_{\phi;0}^2 \mathbf{1}_n \mathbf{1}_n^\top + a_{\phi;1}^2 \mathbf{X}^\top \mathbf{X} + a_{\phi;2}^2 \cdot \frac{1}{p} \mathbf{1}_n \mathbf{1}_n^\top + \left(\nu_\phi - a_{\phi;0}^2 - a_{\phi;1}^2 \right) \mathbf{I}_n, \tag{10}$$

with high probability, up to a spectral norm error $\|\mathbf{K} - \tilde{\mathbf{K}}\|_2 = O(n^{-1/2})$*, where* $a_{\phi;0}, a_{\phi;1}, a_{\phi;2}, \nu_\phi$ *are the Hermite coefficients of* ϕ*.*

- a striking (and perhaps counterintuitive) consequence is that, in the proportional regime with n, p both large and comparable, the eigenvalue distribution of $\mathbf{K}$ becomes independent of the activation function ϕ, up to a scaling and shift
- the eigenspectrum of $\mathbf{K}$ coincides with that of $\mathbf{X}^\top \mathbf{X}$ (which approximates the Marčenko-Pastur law), and depends only on the dimension ratio p/n—provided the data are unstructured and uniformly distributed on the unit sphere.

Linearization of Conjugate Kernel matrix for structured data

Data: K-class Gaussian mixture model (GMM)

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be independently drawn (non-necessarily uniformly) from one of the K classes:

$$\mathcal{C}_a : \sqrt{p}\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a), \quad a \in \{1, \ldots, K\} \tag{11}$$

Large dimensional asymptotics, and non-trivial classification

As $n, p \to \infty$ with $p/n \to c \in (0, \infty)$ and some additional growth-rate assumptions on the difference $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|$ and $\|\mathbf{C}_a - \mathbf{C}_b\|$, $a, b \in \{1, \ldots, K\}$ and $\tau \equiv \sqrt{\operatorname{tr} \mathbf{C}^\circ / p}$ with $\mathbf{C}^\circ \equiv \sum_{a=1}^K \frac{n_a}{n} \mathbf{C}_a$, as $n, p \to \infty$.

Theorem (Asymptotic approximation for conjugate kernels, [AZC22])

For CK matrix $\mathbf{K}_{\mathrm{CK}} = \{\mathbb{E}[\phi(\mathbf{x}_i^\mathsf{T}\mathbf{w})\phi(\mathbf{w}^\mathsf{T}\mathbf{x}_j)]\}_{i,j=1}^n$ *defined above, one has, as* $n, p \to \infty$ *that* $\|\mathbf{K}_{\mathrm{CK}} - \tilde{\mathbf{K}}_{\mathrm{CK}}\| \to 0$, *for some random matrix* $\tilde{\mathbf{K}}_{\mathrm{CK}}$ *dependent of data* $\mathbf{X}$, *of activation* ϕ *but only via the following scalars*

$$\alpha_0 = \mathbb{E}[\phi^2(\tau\xi)] - \mathbb{E}[\phi(\tau\xi)]^2 - \tau\mathbb{E}[\phi'(\tau\xi)]^2, \quad \alpha_1 = \mathbb{E}[\phi'(\tau\xi)]^2, \quad \alpha_2 = \frac{1}{4}\mathbb{E}[\phi''(\tau\xi)]^2$$

and independent of the distribution of $\mathbf{W}$, *as long as normalized to have zero mean and unit variance.*

Main result and the proof

Theorem (Asymptotic approximation for conjugate kernels, [AZC22])

For CK matrix $\mathbf{K}_{\rm CK} = \{\mathbb{E}[\phi(\mathbf{x}_i^\mathsf{T}\mathbf{w})\phi(\mathbf{w}^\mathsf{T}\mathbf{x}_j)]\}_{i,j=1}^n$ *defined above, one has, as* $n, p \to \infty$ *that* $\|\mathbf{K}_{\rm CK} - \tilde{\mathbf{K}}_{\rm CK}\| \to 0$, *for some random matrix* $\tilde{\mathbf{K}}_{\rm CK}$ *dependent of data* $\mathbf{X}$, *of activation* ϕ *but only via the following scalars*

$$\alpha_0 = \mathbb{E}[\phi^2(\tau\xi)] - \mathbb{E}[\phi(\tau\xi)]^2 - \tau\mathbb{E}[\phi'(\tau\xi)]^2, \quad \alpha_1 = \mathbb{E}[\phi'(\tau\xi)]^2, \quad \alpha_2 = \frac{1}{4}\mathbb{E}[\phi''(\tau\xi)]^2$$

and independent of the distribution of $\mathbf{W}$, *as long as normalized to have zero mean and unit variance.*

Proof sketch:

- We are interested in the kernel matrix $\mathbf{K}$, the (i,j) entry of which $\mathbf{K}_{ij} = \mathbb{E}_{\mathbf{w}}[\phi(\mathbf{x}_i^\mathsf{T}\mathbf{w})\phi(\mathbf{w}^\mathsf{T}\mathbf{x}_j)]$.
- Conditioned on $\mathbf{x}_i, \mathbf{x}_j$, $\mathbf{w}^\mathsf{T}\mathbf{x}_i \equiv \|\mathbf{x}_i\| \cdot \xi_i$ and $\mathbf{w}^\mathsf{T}\mathbf{x}_j$ are asymptotically Gaussian, but **correlated**!
- Gram-Schmidt to de-correlate $\mathbf{w}^\mathsf{T}\mathbf{x}_j = \frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\|\mathbf{x}_i\|}\xi_i + \sqrt{\|\mathbf{x}_j\|^2 - \frac{(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)^2}{\|\mathbf{x}_i\|^2}}\tilde{\xi}_j$, for Gaussian $\tilde{\xi}_j$ now **independent** of ξ_j
- Use the fact $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j = O(p^{-1/2})$ and $\|\mathbf{x}_i\|^2 \approx \tau/2 = O(1)$, Taylor-expand to "linearize" $\phi(\cdot)$ to order $o(n^{-1})$
- Since $\|\mathbf{A}\|_2 \le n\|\mathbf{A}\|_{\max}$, with $\|\mathbf{A}\|_{\max} = \max_{ij}|\mathbf{A}_{ij}|$, obtain **spectral** approximation $\tilde{\mathbf{K}}$.

[1] Hafiz Tiomoko Ali, **Zhenyu Liao**, and Romain Couillet. "Random matrices in service of ML footprint: ternary random features with no performance loss". In: *International Conference on Learning Representations (ICLR 2022)*. 2022

Practical consequence of the theory

According to theorem, allowed to choose arbitrary weights $\mathbf{W}$ and activation ϕ, without affecting $\mathbf{K}$ asymptotically, under the following conditions:

- weights $\mathbf{W}$ have independent entries with zero mean and unit variance
- activation ϕ has the same few parameters as the original net

$$\alpha_0 = \mathbb{E}[\phi^2(\tau\xi)] - \mathbb{E}[\phi(\tau\xi)]^2 - \tau\mathbb{E}[\phi'(\tau\xi)]^2, \quad \alpha_1 = \mathbb{E}[\phi'(\tau\xi)]^2, \quad \alpha_2 = \frac{1}{4}\mathbb{E}[\phi''(\tau\xi)]^2, \tag{12}$$

In particular,

- **sparse and binarized** (e.g., Bernoulli distributed) weights $\mathbf{W}$ instead of dense Gaussian weights

$$[\mathbf{W}]_{ij} = 0 \text{ with proba } \varepsilon \in [0,1), \quad [\mathbf{W}]_{ij} = \pm(1-\varepsilon)^{-1/2} \text{ each with proba } 1/2 - \varepsilon/2, \tag{13}$$

- **sparse quantized** (e.g., binarized) activation ϕ shares the same α_0, α_1, and α_2

Numerical results

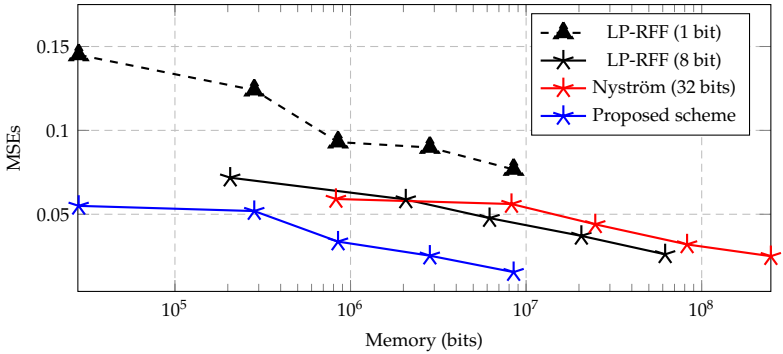


Figure: Test mean square errors of ridge regression on quantized single-hidden-layer random nets for different numbers of features $N \in \{5.10^2, 10^3, 5.10^3, 10^4, 5.10^4\}$, using LP-RFF, Nyström approximation, versus the proposed approach, on the Census dataset, with $n = 16\,000$ training samples, $n_{\text{test}} = 2\,000$ test samples, and data dimension $p = 119$.

CK of fully-connected random deep neural networks

- everyone cares **more** about deep neural networks
- with some additional efforts, extension to fully-connected **deep** neural networks of depth L,

$$f(\mathbf{x}) = \frac{1}{\sqrt{d_L}}\mathbf{w}^\top \phi_L\left(\frac{1}{\sqrt{d_{L-1}}}\mathbf{W}_L\phi_{L-1}\left(\cdots\frac{1}{\sqrt{d_2}}\phi_2\left(\frac{1}{\sqrt{d_1}}\mathbf{W}_2\phi_1(\mathbf{W}_1\mathbf{x})\right)\right)\right), \tag{14}$$

again for random $\mathbf{W}_1, \ldots, \mathbf{W}_L$ and activations $\phi_1(\cdot), \ldots, \phi_L(\cdot)$.

Theorem (Asymptotic approximation for conjugate kernels, informal)

Under the same condition, define output features of layer $\ell \in \{1, \ldots, L\}$, as

$$\mathbf{\Sigma}_\ell = \frac{1}{\sqrt{d_\ell}}\phi_\ell\left(\frac{1}{\sqrt{d_{\ell-1}}}\mathbf{W}_\ell\phi_{\ell-1}\left(\cdots\frac{1}{\sqrt{d_2}}\phi_2\left(\frac{1}{\sqrt{d_1}}\mathbf{W}_2\phi_1(\mathbf{W}_1\mathbf{X})\right)\right)\right). \tag{15}$$

we have for the Conjugate Kernel $\mathbf{K}_{\mathrm{CK},\ell}$ at layer ℓ defined as

$$\mathbf{K}_{\mathrm{CK},\ell} = \mathbb{E}[\mathbf{\Sigma}_\ell^\top \mathbf{\Sigma}_\ell] \in \mathbb{R}^{n\times n}, \tag{16}$$

that $\|\mathbf{K}_{\mathrm{CK},\ell} - \tilde{\mathbf{K}}_{\mathrm{CK},\ell}\| \to 0$, some random matrix $\tilde{\mathbf{K}}_{\mathrm{CK},\ell}$ dependent of data, of activation ϕ_ℓ but only via a few parameters, and independent of the distribution of $\mathbf{W}$, as long as of normalized to have zero mean and unit variance.

Theorem (Asymptotic approximation for CK matrices, formal, [Gu+22])

Let $\tau_0, \tau_1, \ldots, \tau_L \geq 0$ be a sequence of non-negative numbers satisfying the following recursion:

$$\tau_\ell = \sqrt{\mathbb{E}[\phi_\ell^2(\tau_{\ell-1}\xi)]}, \quad \xi \sim \mathcal{N}(0,1), \quad \ell \in \{1, \ldots, L\}. \tag{17}$$

Further assume that the activation functions $\phi_\ell(\cdot)$s are "centered," such that $\mathbb{E}[\phi_\ell(\tau_{\ell-1}\xi)] = 0$. Then, for the CK matrix $\mathbf{K}_{\mathrm{CK},\ell}$ of layer $\ell \in \{1, \ldots, L\}$ defined in (16), as $n, p \to \infty$, one has that:

$$\boxed{\|\mathbf{K}_{\mathrm{CK},\ell} - \tilde{\mathbf{K}}_{\mathrm{CK},\ell}\| \to 0, \quad \tilde{\mathbf{K}}_{\mathrm{CK},\ell} \equiv \alpha_{\ell,1}\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{V}\mathbf{A}_\ell\mathbf{V}^\mathsf{T} + (\tau_\ell^2 - \tau_0^2\alpha_{\ell,1})\mathbf{I}_n,} \tag{18}$$

almost surely, with $\mathbf{V} = [\mathbf{J}/\sqrt{p},\ \boldsymbol{\psi}] \in \mathbb{R}^{n\times(K+1)}$, $\mathbf{A}_\ell = \begin{bmatrix} \alpha_{\ell,2}\mathbf{t}\mathbf{t}^\mathsf{T} + \alpha_{\ell,3}\mathbf{T} & \alpha_{\ell,2}\mathbf{t} \\ \alpha_{\ell,2}\mathbf{t}^\mathsf{T} & \alpha_{\ell,2} \end{bmatrix} \in \mathbb{R}^{(K+1)\times(K+1)}$, for class label vectors $\mathbf{J} = [\mathbf{j}_1, \ldots, \mathbf{j}_K] \in \mathbb{R}^{n\times K}$, "second-order" data fluctuation vector $\boldsymbol{\psi} \in \mathbb{R}^n$, second-order data statistics $\mathbf{t} = \{\operatorname{tr}\mathbf{C}_a^\circ/\sqrt{p}\}_{a=1}^K \in \mathbb{R}^K$ and $\mathbf{T} = \{\operatorname{tr}\mathbf{C}_a\mathbf{C}_b/p\}_{a,b=1}^K \in \mathbb{R}^{K\times K}$, as well as non-negative $\alpha_{\ell,1}, \alpha_{\ell,2}, \alpha_{\ell,3}$ satisfying

$$\alpha_{\ell,1} = \mathbb{E}[\phi_\ell'(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,1}, \quad \alpha_{\ell,2} = \mathbb{E}[\phi_\ell'(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,2} + \frac{1}{4}\mathbb{E}[\phi_\ell''(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,4}^2, \tag{19}$$

$$\alpha_{\ell,3} = \mathbb{E}[\phi_\ell'(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,3} + \frac{1}{2}\mathbb{E}[\phi_\ell''(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,1}^2. \tag{20}$$

with $\alpha_{\ell,4} = \mathbb{E}\left[(\phi_\ell'(\tau_{\ell-1}\xi))^2 + \phi_\ell(\tau_{\ell-1}\xi)\phi_\ell''(\tau_{\ell-1}\xi)\right]\alpha_{\ell-1,4}$ for $\xi \sim \mathcal{N}(0,1)$.

Implications

$$\|\mathbf{K}_{\mathrm{CK},\ell} - \tilde{\mathbf{K}}_{\mathrm{CK},\ell}\| \to 0,\ \tilde{\mathbf{K}}_{\mathrm{CK},\ell} \equiv \alpha_{\ell,1}\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{V}\mathbf{A}_\ell\mathbf{V}^\mathsf{T} + (\tau_\ell^2 - \tau_0^2\alpha_{\ell,1})\mathbf{I}_n,\ \mathbf{A}_\ell = \begin{bmatrix} \alpha_{\ell,2}\mathbf{t}\mathbf{t}^\mathsf{T} + \alpha_{\ell,3}\mathbf{T} & \alpha_{\ell,2}\mathbf{t} \\ \alpha_{\ell,2}\mathbf{t}^\mathsf{T} & \alpha_{\ell,2} \end{bmatrix}. \quad (21)$$

Can (already) say something on the features obtained from **random** DNNs:

- $\alpha_{\ell,1}$ weighting first-order statistics of input data $\mathbf{X}$, i.e., $\mathbb{E}[\mathbf{X}] = \mathbf{M}$
- $\alpha_{\ell,2}, \alpha_{\ell,3}$ weighting second-order statistics, i.e., $\mathbf{t} = \{\operatorname{tr}\mathbf{C}_a^\circ/\sqrt{p}\}_{a=1}^K$ and $\mathbf{T} = \{\operatorname{tr}\mathbf{C}_a\mathbf{C}_b/p\}_{a,b=1}^K$

A few **qualitative** remarks on the depth:

- intrinsic limitation of shallow NN: for $\ell = 1$, one has $\alpha_{1,3} = 2\alpha_{1,2}$ independently of the choice of the first layer activation; for DNN of depth $L \geq 2$, **no** much constraint
- deeper NNs are stronger nonlinear feature extractors: one has that $\frac{\alpha_{\ell,2}}{\alpha_{\ell,1}} \geq \frac{\alpha_{\ell-1,2}}{\alpha_{\ell-1,1}}$, $\frac{\alpha_{\ell,3}}{\alpha_{\ell,1}} \geq \frac{\alpha_{\ell-1,3}}{\alpha_{\ell-1,1}}$, $\ell \geq 1$
- limitation of even or odd activation: $\alpha_{\ell,1} = 1$ for all even activation, and $\frac{\alpha_{\ell,2}}{\alpha_{\ell,1}} = \frac{\alpha_{\ell-1,2}}{\alpha_{\ell-1,1}}$, $\frac{\alpha_{\ell,3}}{\alpha_{\ell,1}} = \frac{\alpha_{\ell-1,3}}{\alpha_{\ell-1,1}}$ for all odd function so that the CK at any layer ℓ is (asymptotically) the linear kernel $\mathbf{X}^\mathsf{T}\mathbf{X}$
  $\Rightarrow$ ReLU-type activations that are **neither even nor odd** are good!

Fully-connected deep nets: CK, NTK, and beyond

- happy with the study of (limiting) CK for random DNN models
- extension to NTK via intrinsic **connection** between CK and NTK [JGH18]

$$\mathbf{K}_{\mathrm{NTK},\ell}(\mathbf{X}) = \mathbf{K}_{\mathrm{CK},\ell}(\mathbf{X}) + \mathbf{K}_{\mathrm{NTK},\ell-1}(\mathbf{X}) \circ \mathbf{K}'_{\mathrm{CK},\ell}(\mathbf{X}), \quad \mathbf{K}_{\mathrm{NTK},0}(\mathbf{X}) = \mathbf{K}_{\mathrm{CK},0}(\mathbf{X}) = \mathbf{X}^{\mathsf{T}}\mathbf{X}, \tag{22}$$

  and some additional efforts
- **convergence** and **generalization** theory via NTK [JGH18]: for
  - (i) sufficiently wide nets
  - (ii) trained with gradient descent of sufficiently small step size
- NTK is **determined** at random initialization and remains unchanged during training, and applies to **explicitly** characterize DNN convergence and generalization properties
- we now have a theory for **trained nonrandom DNNs** and can be used for DNN compression!

[2] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 31. NIPS'18. Curran Associates, Inc., 2018, pp. 8571–8580

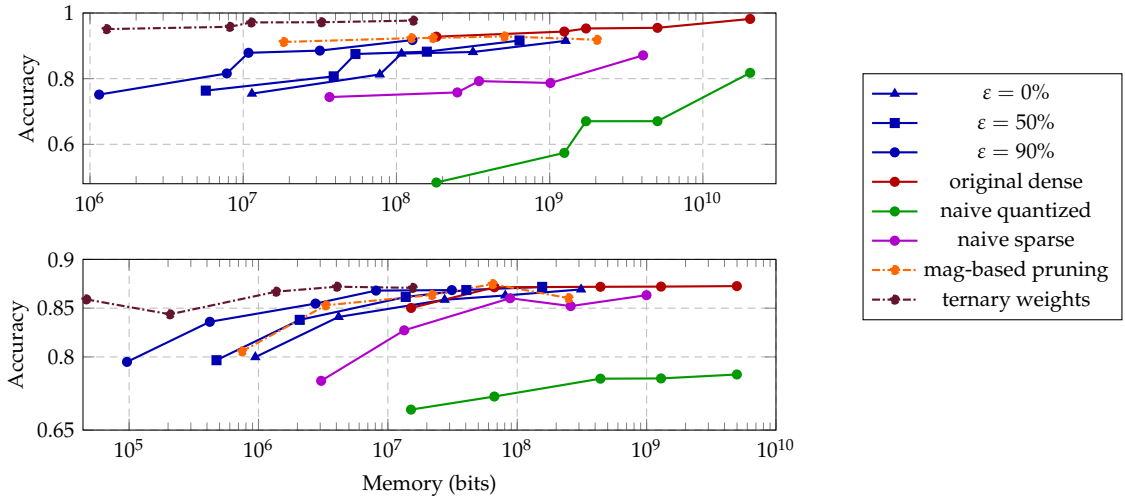


Figure: Test accuracy of classification on MNIST (**top**) and CIFAR10 (**bottom**) datasets. **Blue**: proposed NTK-LC approach with different levels of sparsity $\varepsilon \in \{0\%, 50\%, 90\%\}$, **purple**: heuristic sparsification approach by uniformly zeroing out 80% of the weights, **green**: heuristic quantization approach with binary activation $\phi(t) = 1_{t<-1} + 1_{t>1}$, **red**: original network, **orange**: NTK-LC without activation quantization, and **brown**: magnitude-based pruning with same sparsity level as **orange**. Memory varies due to the **change of layer width** of the network.

Take-away

Take-away messages:

- a unified RMT analysis to ML via High-dimensional (Deterministic and Linear) Equivalent
- for GMM input data, RMT allows for precise characterization of (the CKs of) **random** shallow and deep neural networks
- extends to **NTKs**, providing access to trained DNNs, but **only** in the "lazy" NTK regime

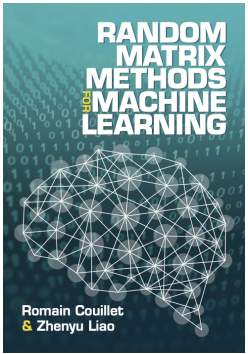
References:

- Hafiz Tiomoko Ali, **Zhenyu Liao**, and Romain Couillet. "Random matrices in service of ML footprint: ternary random features with no performance loss". In: *International Conference on Learning Representations (ICLR 2022)*. 2022
- Lingyu Gu, Yongqi Du, Yuan Zhang, Di Xie, Shiliang Pu, Robert Qiu, and **Zhenyu Liao**. ""Lossless" Compression of Deep Neural Networks: A High-dimensional Neural Tangent Kernel Approach". In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., 2022, pp. 3774–3787 (**Please** refer to the ArXiv version on `https://arxiv.org/abs/2403.00258` that fixed typos in Theorems 1 and 2 from the NeurIPS 2022 proceeding version.)
- Zenan Ling, Longbo Li, Zhanbo Feng, Yixuan Zhang, Feng Zhou, Robert C. Qiu, and **Zhenyu Liao**. "Deep Equilibrium Models Are Almost Equivalent to Not-so-deep Explicit Models for High-dimensional Gaussian Mixtures". In: *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*. Vol. 235. PMLR, 21–27 Jul 2024, pp. 30585–30609
- **Zhenyu Liao**, and Michael W. Mahoney, "Random Matrix Theory for Deep Learning: Beyond Eigenvalues of Linear Models". 2025.

RMT for machine learning: from theory to practice!

Random matrix theory (RMT) for machine learning:

- **change of intuition** from small to large dimensional learning paradigm!
- **better understanding** of existing methods: why they work if they do, and what the issue is if they do not
- **improved novel methods** with performance guarantee!



- book "*Random Matrix Methods for Machine Learning*"
- by Romain Couillet and **Zhenyu Liao**
- Cambridge University Press, 2022
- a pre-production version of the book and exercise solutions at `https://zhenyu-liao.github.io/book/`
- MATLAB and Python codes to reproduce all figures at `https://github.com/Zhenyu-LIAO/RMT4ML`

Thank you! Q & A?