# Eigenvalue distribution of the Neural Tangent Kernel in the quadratic scaling

Lucas Benigni

Université de Montréal

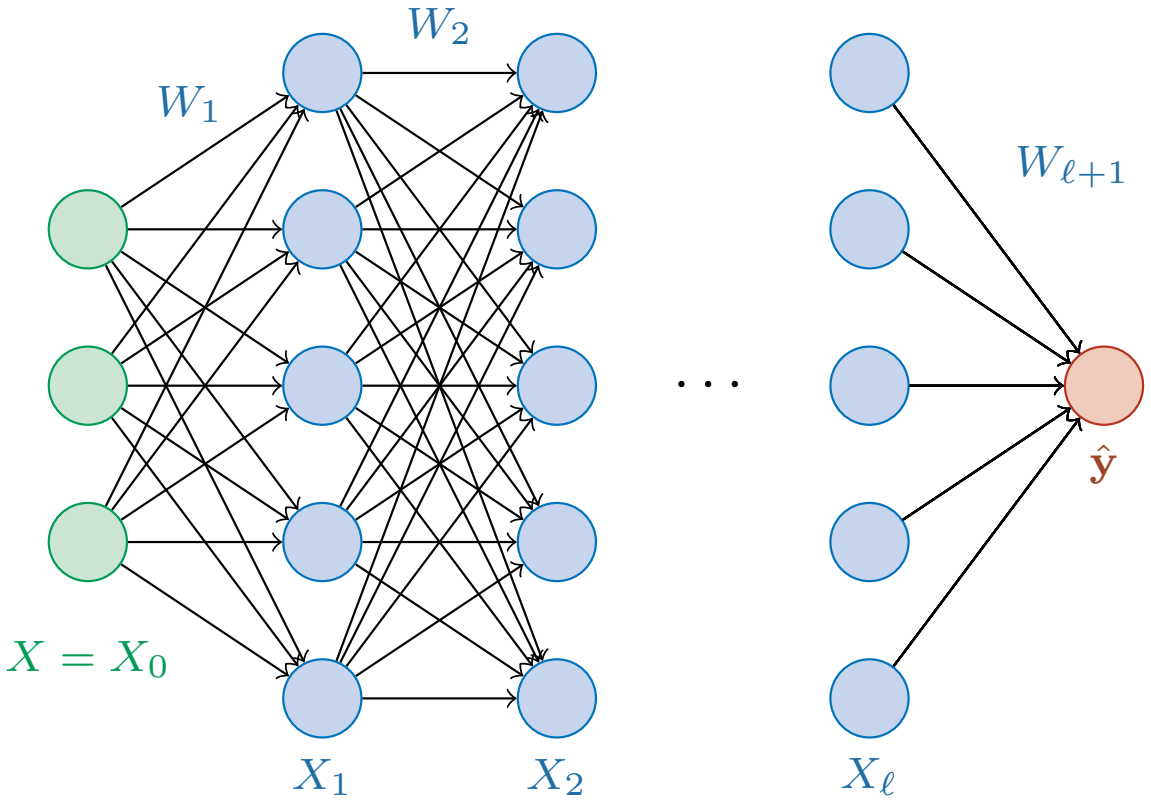
Random Matrices and Applications — ICERM
May 22 2024

Joint work with Elliot Paquette

# Neural networks

- $X \in \mathbb{R}^{n \times d_0}$ : $n$ samples of $d_0$-dimensional **inputs**.
- $W_i \in \mathbb{R}^{d_{i-1} \times d_i}$ : **weight matrices**.
- $\sigma : \mathbb{R} \to \mathbb{R}$ : **activation function** applied entrywise.

**Feed-Forward Neural Network**




$$X_i = \sigma(X_{i-1} W_i)$$

$$\hat{\mathbf{y}} = X_\ell W_{\ell+1}.$$

# Training and Testing

**Supervised learning**: **Input** data $X$ is given with **output target** datasets $\mathbf{y}$

**Training Phase**: Optimise the weights $\Theta = (W_1, \dots, W_{\ell+1})$ :

$$\Theta_{\min} = \operatorname*{argmin}_{\Theta} \mathrm{E}_{\mathrm{Train}} = \operatorname*{argmin}_{\Theta} \mathcal{L}(\hat{\mathbf{y}}(\Theta, X), \mathbf{y})$$

**Loss functions**: $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \dfrac{1}{2n} \displaystyle\sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad -\dfrac{1}{n} \sum_{i=1}^{n} y_i \log(\hat{y}_i), \dots$

**Optimization algorithms**: Gradient flow, gradient descent, SGD, ...

**Testing Phase**: Consider new data $\widetilde{X}$ with output target $\widetilde{\mathbf{y}}$ and see how small is

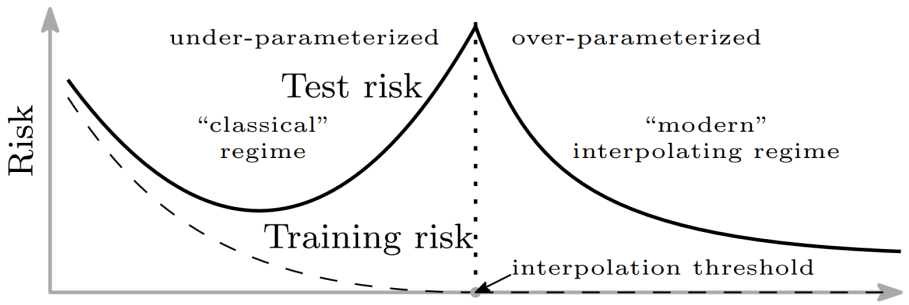
$$\mathrm{E}_{\mathrm{Test}} = \mathcal{L}(\hat{\mathbf{y}}(\Theta_{\min}, \widetilde{X}), \widetilde{\mathbf{y}})$$

**Two-layer neural network**: $X \in \mathbb{R}^{n \times d}$, $W \in \mathbb{R}^{d \times p}$, $\mathbf{a} \in \mathbb{R}^p$,

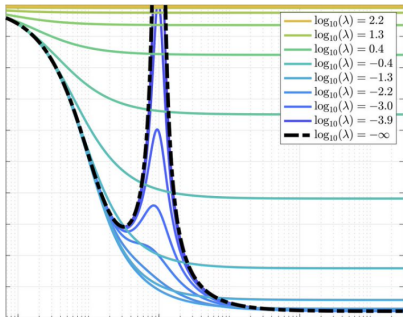
$$\hat{\mathbf{y}} = \sigma(XW)\mathbf{a} \in \mathbb{R}^n, \quad \Theta = (W, \mathbf{a}) \in \mathbb{R}^{d \times p} \times \mathbb{R}^p \cong \mathbb{R}^{p(d+1)}$$

# Overparametrization

Surprising empirical evidence shows that neural networks **generalise well** when the number of parameters of the network is really large. This is illustrated by the **double descent curve** for unregularized losses.




[Belkin–Hsu–Ma–Mandal '18]



[Mei–Montanari '19]

This phenomenon is now **theoretically understood** for many models in **different regimes** of overparametrization ([Mei–Montanari '19, Hastie–Montanari–Rosset–Tibshirani '19, Deng–Kammoun–Thrampoulidis '19, Montanari–Ruan–Sohn–Yan '19, Liang–Sur '20, Ghorbani–Mei–Misiakiewicz–Montanari '20, Montanari–Zhong '20, Cheng–Montanari '22, Misiakiewicz '22, Mei–Misiakiewicz–Montanari '22, Hu–Lu '22, Xiao–Hu–Misiakiewicz–Pennington '22, Latourelle-Vigeant–Paquette '23, Hu–Lu–Misiakiewicz '24, . . . ] )

See [Misiakiewicz–Montanari '23] for a recent review.

# Regimes of overparametrization

**Two-layer neural network**: $X \in \mathbb{R}^{n\times d}$, $W \in \mathbb{R}^{d\times p}$, $\mathbf{a} \in \mathbb{R}^p$,

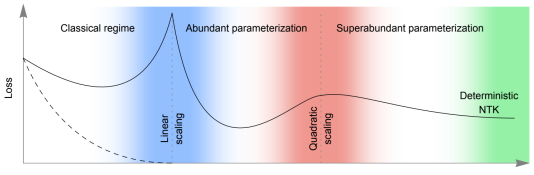
$$\hat{\mathbf{y}} = \sigma(XW)\mathbf{a} \in \mathbb{R}^n, \quad \Theta = (W, \mathbf{a}) \in \mathbb{R}^{d\times p} \times \mathbb{R}^p \cong \mathbb{R}^{p(d+1)}$$

**Interpolation threshold**:

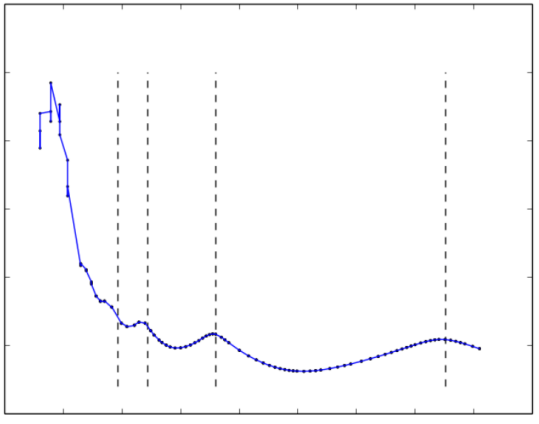
number of samples $n \asymp$ number of parameters $p(d+1)$

**Quadratic scaling**

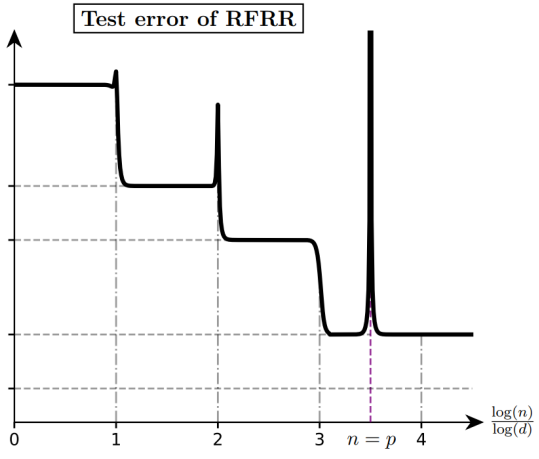
Different **polynomial scalings** can lead to **multiple descent** phenomenon




[Adlam–Pennington '20]



[Liang–Rakhlin–Zhai '20]




[Misiakiewicz '24]

# Neural Tangent Kernel [Jacot–Gabriel–Hongler '18]

Consider the **least squares** loss:

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{2}\|\hat{\mathbf{y}} - \mathbf{y}\|^2.$$

Optimize using **gradient descent**:

$$\Theta_{t+1} = \Theta_t - \eta_t \sum_{i=1}^{n} (\hat{\mathbf{y}}(\Theta_t, \mathbf{x}_i) - y_i) \nabla_\Theta \hat{\mathbf{y}}(\Theta_t, \mathbf{x}_i).$$

There exists a **regime** of training (depending on **width** and **stepsize**) where we can do an affine approximation of $\hat{\mathbf{y}}$ (**lazy training**),

$$\hat{\mathbf{y}}(\Theta_t, X) = \hat{\mathbf{y}}(\Theta_0, X) + \langle \nabla_\Theta \hat{\mathbf{y}}(\Theta_0, X), \Theta_t - \Theta_0 \rangle + \mathcal{O}\left(\|\Theta_t - \Theta_0\|^2\right).$$

Dynamics are thus determined by the **Neural Tangent Kernel** at initialization

$$\boxed{K(\mathbf{x}_i, \mathbf{x}_j) = \nabla_\Theta \hat{\mathbf{y}}(\Theta_0, \mathbf{x}_i)^\top \nabla_\Theta \hat{\mathbf{y}}(\Theta_0, \mathbf{x}_j)}.$$

$K$ only depends on $\Theta_0$, i.e $W$ and $\mathbf{a}$, at **initialization**!

# Neural Tangent Kernel [Jacot–Gabriel–Hongler '18]

**Gradient flow**: Gradient descent as $\eta_t \to 0$,

$$\frac{\mathrm{d}\Theta_t}{\mathrm{d}t} = -\nabla_\Theta \mathcal{L}(\hat{\mathbf{y}}(\Theta_t, X), \mathbf{y}) = -\nabla_\Theta \hat{\mathbf{y}}(\Theta_t, X)(\hat{\mathbf{y}}(\Theta_t, X) - \mathbf{y})$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\mathbf{y}}(\Theta_t, X) = \nabla_\Theta \hat{\mathbf{y}}(\Theta_t, X)^\top \frac{\mathrm{d}\Theta_t}{\mathrm{d}t} = -\nabla_\Theta \hat{\mathbf{y}}(\Theta_t, X)^\top \nabla_\Theta \hat{\mathbf{y}}(\Theta_t, X)(\hat{\mathbf{y}}(\Theta_t, X) - \mathbf{y})$$

With the **linear approximation** $\nabla_\Theta \hat{\mathbf{y}}(\Theta_t, X) \approx \nabla_\Theta \hat{\mathbf{y}}(\Theta_0, X)$,

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\hat{\mathbf{y}}(\Theta_t, X) \approx -K(X)(\hat{\mathbf{y}}(\Theta_t, X) - \mathbf{y})}$$

If we consider the **eigenvalue decomposition** of $K(X) = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$, since

$$\hat{\mathbf{y}}(\Theta_t, X) - \mathbf{y} \approx \mathrm{e}^{-K(X)t}(\hat{\mathbf{y}}(\Theta_0, X) - \mathbf{y}).$$

$\hat{\mathbf{y}}(\Theta_t, X) - \mathbf{y} \to 0$ exponentially fast to 0 with exponential rate $\lambda_i$ in direction $\mathbf{u}_i$.

The spectrum of $K$ thus describes the **speed of the training phase**.

# Computation of the NTK

Since $\Theta = [W, \mathbf{a}]$, the kernel splits in **two parts**

$$K(\mathbf{x}_i, \mathbf{x}_j) = K_1(\mathbf{x}_i, \mathbf{x}_j) + K_2(\mathbf{x}_i, \mathbf{x}_j)$$

$$K_1 = XX^\top \odot \sigma'(XW)\mathrm{Diag}(\mathbf{a})^2\sigma'(XW)^\top, \quad K_2 = \sigma(XW)\sigma(XW)^\top$$

Hadamard product:

$$(A \odot B)_{ij} = A_{ij}B_{ij}.$$

$K_2$ is the **conjugate kernel**: Ridge regression on the **random features model**:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}\in\mathbb{R}^p}{\mathrm{argmin}}\, \mathcal{L}(\sigma(XW)\mathbf{a}, \mathbf{y}) = \underset{\mathbf{a}\in\mathbb{R}^p}{\mathrm{argmin}}\, \frac{1}{2}\|\sigma(XW)\mathbf{a} - \mathbf{y}\|_2^2 + \frac{\gamma}{2}\|\mathbf{a}\|_2^2.$$

Then

$$\hat{\mathbf{a}} = \left(\sigma(XW)^\top\sigma(XW) + \gamma \mathrm{I}_p\right)^{-1}\sigma(XW)^\top\mathbf{y}.$$

Optimizing on the output layer involves the **spectrum** of the conjugate kernel.

# Linear and Quadratic Scaling

Compare the simpler models: $X \in \mathbb{R}^{n\times d}$, $Y \in \mathbb{R}^{n\times p}$ with $i.i.d$ entries $\mathcal{N}(0,1)$, ind.

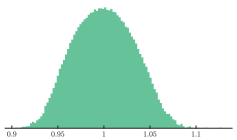
$$M_1 = \frac{1}{d} XX^\top \qquad \text{and} \qquad M_2 = \frac{1}{d} XX^\top \odot \frac{1}{p} YY^\top$$

**Linear Scaling**: $\frac{n}{d} \to \gamma_1, \quad \frac{n}{p} \to \gamma_2.$

$M_1 \to \mathrm{MP}_{\gamma_1}$

[Marchenko–Pastur '67]

$M_2 \to \mathrm{I}_n$



# Linear and Quadratic Scaling

Compare the simpler models: $X \in \mathbb{R}^{n\times d}$, $Y \in \mathbb{R}^{n\times p}$ with $i.i.d$ entries $\mathcal{N}(0,1)$, ind.

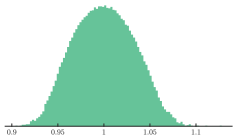
$$M_1 = \frac{1}{d}XX^\top \qquad \text{and} \qquad M_2 = \frac{1}{d}XX^\top \odot \frac{1}{p}YY^\top$$

**Linear Scaling**: $\frac{n}{d} \to \gamma_1, \quad \frac{n}{p} \to \gamma_2.$

$M_1 \to \mathrm{MP}_{\gamma_1}$

[Marchenko–Pastur '67]

$M_2 \to \mathrm{I}_n$



**Quadratic Scaling**: $\frac{n}{dp} \to \gamma_1, \quad \frac{d}{p} \to \gamma_2.$

$M_1 \to$ Low rank



$M_2 \to \mathrm{MP}_{\gamma_1}$

[?]

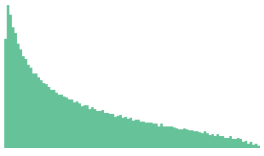
# Linear and Quadratic Scaling

Compare the simpler models: $X \in \mathbb{R}^{n\times d}$, $Y \in \mathbb{R}^{n\times p}$ with $i.i.d$ entries $\mathcal{N}(0,1)$, ind.

$$M_1 = \frac{1}{d}XX^\top \qquad \text{and} \qquad M_2 = \frac{1}{d}XX^\top \odot \frac{1}{p}YY^\top$$

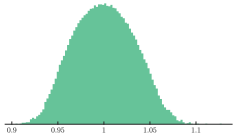
**Linear Scaling**: $\frac{n}{d} \to \gamma_1, \quad \frac{n}{p} \to \gamma_2.$

$M_1 \to \mathrm{MP}_{\gamma_1}$



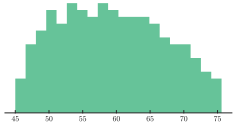
[Marchenko–Pastur '67]

$M_2 \to \mathrm{I}_n$



**Quadratic Scaling**: $\frac{n}{dp} \to \gamma_1, \quad \frac{d}{p} \to \gamma_2.$

$M_1 \to$ Low rank



$M_2 \to \mathrm{MP}_{\gamma_1}$

[?]

# Linear and Quadratic Scaling

Compare the simpler models: $X \in \mathbb{R}^{n\times d}$, $Y \in \mathbb{R}^{n\times p}$ with $i.i.d$ entries $\mathcal{N}(0,1)$, ind.

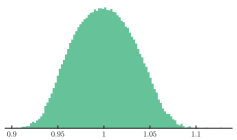
$$M_1 = \frac{1}{d}XX^\top \qquad \text{and} \qquad M_2 = \frac{1}{d}XX^\top \odot \frac{1}{p}YY^\top$$

**Linear Scaling**: $\frac{n}{d} \to \gamma_1, \quad \frac{n}{p} \to \gamma_2.$

$M_1 \to \mathrm{MP}_{\gamma_1}$

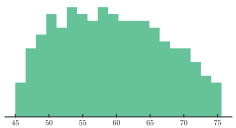
[Marchenko–Pastur '67]

$M_2 \to \mathrm{I}_n$




**Quadratic Scaling**: $\frac{n}{dp} \to \gamma_1, \quad \frac{d}{p} \to \gamma_2.$

$M_1 \to$ Low rank




$M_2 \to \mathrm{MP}_{\gamma_1}$

[?]

**Quadratic scaling** seems natural for the structure of the Neural Tangent Kernel

# Main result

**Quadratic scaling**: $\frac{n}{dp} \to \gamma_1$, $\frac{p}{d} \to \gamma_2$.

$$\boxed{\mathrm{NTK}_n = \frac{1}{d} XX^\top \odot \frac{1}{p} \sigma'\left(\frac{1}{\sqrt{d}} XW\right) \mathrm{diag}(\mathbf{a})^2 \sigma'\left(\frac{1}{\sqrt{d}} XW\right)^\top \in \mathbb{R}^{n \times n}}$$

**Main Assumptions**:

- $X$, $W$, and $\mathbf{a}$ are independent.
- Unstructured data: $X$ *i.i.d* matrix with $\mathbb{E}[x_{11}] = 0$, $\mathbb{E}[x_{12}^2] = 1$, $\mathbb{E}[x_{11}^4] < \infty$.
- Gaussian weights: $W$ *i.i.d* matrix with $w_{ij} \sim \mathcal{N}(0, 1)$.
- $\mathbf{a}$ is an *i.i.d* vector with $\mathbb{E}[a_1] = 0$, $\mathbb{E}[a_1^4] < \infty$, $a_{11}^2 \sim \nu$.
- $\sigma'$ is pseudo–Lipschitz: $|\sigma'(x) - \sigma'(y)| \leqslant K|x - y|(1 + |x|^\alpha + |y|^\alpha)$.
- Denote $\alpha_{\sigma'} = \mathbb{E}_{\mathcal{N}(0,1)}\left[z\sigma'(z)\right]$, $\quad \beta_{\sigma'} = \sqrt{\mathbb{E}_{\mathcal{N}(0,1)}\left[\sigma'(z)^2\right]}$

## Theorem (B.–Paquette '24+)

*The empirical eigenvalue distribution of* $\mathrm{NTK}_n$ *converges weakly in probability to a deterministic measure* $\mu$ *which can be written as*

$$\mu = \mu_{\mathrm{MP}}^{\gamma_1} \boxtimes \left(\alpha_{\sigma'}^2 (\mu_{\mathrm{MP}}^{\gamma_2} \boxtimes \nu) * (\mu_{\mathrm{MP}}^{\gamma_2} \boxtimes \nu) \boxplus (\beta_{\sigma'}^2 - \alpha_{\sigma'}^2)\nu\right).$$

# Limiting distribution

$$\boxed{\mu = \mu_{\mathrm{MP}}^{\gamma_1} \boxtimes \left(\alpha_{\sigma'}^2 (\mu_{\mathrm{MP}}^{\gamma_2} \boxtimes \nu) * (\mu_{\mathrm{MP}}^{\gamma_2} \boxtimes \nu) \boxplus (\beta_{\sigma'}^2 - \alpha_{\sigma'}^2)\nu\right).}$$

**Marchenko–Pastur map** [Voiculescu '86, Bai–Silverstein '95] : For a distribution $\nu$, the Stieltjes transform of $\mu_{\mathrm{MP}}^{\gamma} \boxtimes \nu$ is given by

$$s(z) = \frac{1}{-z + \displaystyle\int \frac{x \mathrm{d}\nu(x)}{1 + \gamma x s(z)}}, \qquad \left(\text{Eigenvalues of } \frac{1}{d} X D X^\top \text{ with } \mathrm{limspec}(D) \to \nu\right).$$

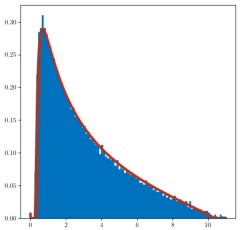
**Free additive convolution** [Voiculescu '86] : For two distributions $\mu$, $\nu$, we have

$$\mu \boxplus \nu = \mathrm{limspec}(A + U B U^\top), \quad \text{if } \mathrm{limspec}(A) = \mu,\ \mathrm{limspec}(B) = \nu,\ U \sim \mathrm{Haar}(\mathrm{O}_n(\mathbb{R})).$$

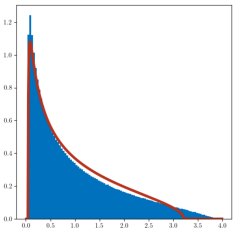
**Convolution** If $X \sim \mu$ and $Y \sim \nu$, $X$ and $Y$ independent then

$$X + Y \sim \mu * \nu.$$

$\sigma'(x) = x$



$\sigma'(x) = \cos(x)$



# Nonlinear random matrix models

**Kernel matrices**: $X \in \mathbb{R}^{n\times d}, \quad M_1 = \sigma(XX^\top)$

- Linear width $\frac{n}{d} \to \gamma$:
  - Eigenvalue distribution [El Karoui '10, Cheng–Singer '12, Do–Vu '12] :

$$\boxed{a_\sigma \mu_{\mathrm{MP}}^{\gamma} \boxplus b_\sigma \mu_{\mathrm{sc}}.}$$

  - Largest eigenvalue [Fan–Montanari '15] : $\sigma$ odd, $\lambda_1 \to$ edge.
- Polynomial width $\frac{n}{d^\ell} \to \gamma$:
  - Eigenvalue distribution [Lu–Yau '22, Misiakiewicz '22, Dubova–Lu–McKenna–Yau '23] :

$$\ell \in \mathbb{N} : \boxed{a'_\sigma \mu_{\mathrm{MP}}^{\gamma'} \boxplus b'_\sigma \mu_{\mathrm{sc}},} \quad \ell \notin \mathbb{N} : \boxed{b'_\sigma \mu_{\mathrm{sc}}.}$$

**Conjugate Kernel**: $X \in \mathbb{R}^{n\times d}, W \in \mathbb{R}^{d\times p}, \quad M_2 = \sigma(XW)\sigma(XW)^\top$

- Linear width $\frac{n}{d} \to \gamma_1$, $\frac{n}{p} \to \gamma_2$:
  - Eigenvalue distribution [Louart–Liao–Couillet '17, Pennington–Worah '17, B.–Péché '19, Péché '19, Fan–Wang '20, Piccolo–Schröder '21, Chouard '22] :

$$\boxed{\mu_{\mathrm{MP}}^{\gamma_2} \boxtimes (a_\sigma \mu_{\mathrm{MP}}^{\gamma_1} + b_\sigma).}$$

  - Largest eigenvalue [B.–Péché '22, Wang–Wu–Fan '24] : Possible outliers.
- Overparametrized $\frac{n}{d} \to 0$ [Wang–Zhu '21 '22] : $\boxed{\mu_{\mathrm{sc}} \boxtimes (a_\sigma + b_\sigma \mu_{\mathrm{MP}}^{\gamma_1}).}$
- Polynomial width $\frac{n}{d^{\ell_1}} \to \gamma_1$, $\frac{n}{p^{\ell_2}} \to \gamma_2$ [Misiakiewicz '24] .

# Nonlinear random matrix models

**Multi-layer Conjugate Kernel**: $X_\ell = \sigma(X_{\ell-1} W_\ell), \quad M_3 = X_\ell X_\ell^\top$.

- Linear width [B.–Péché '19, Fan–Wang '20] : $\boxed{\mu_{\mathrm{MP}}^{\gamma_\ell} \boxtimes (a_\sigma \mu_{\ell-1} + b_\sigma)}$.

**Neural Tangent Kernel**: $M_4 = XX^\top \odot \sigma'(XW)\sigma'(XW)^\top + XX^\top$

- Linear width, multi-layer [Fan–Wang '20] : Same as $\boxed{a_\sigma \mathrm{I}_n + XX^\top}$
- Overparametrized [Wang–Zhu '21] : $\boxed{\mu_{\mathrm{sc}} \boxtimes (a_\sigma + b_\sigma \mu_{\mathrm{MP}}^{\gamma_1})}$.

This list is not exhaustive, one can consider other models such as the **Hessian** [Liao–Mahoney '21] , **spiked models** ($f(X + A)$ with $A$ low rank and) [Guionnet–Ko–Krakala–Zdeborová '23, Feldman '23] ,... or other questions than the eigenvalue distribution!

To understand the **test error**, one has to understand the whole matrix, including the eigenvectors [Ghorbani–Mei–Misiakiewicz–Montanari '19 '21, Mei–Montanari '19, Montanari – Zhong '20, Wang–Zhu '22, Hu–Lu '22, Schröder–Cui–Dmitriev–Loureiro '23, Latourelle-Vigeant–Paquette '23]

# Matrix Dyson Equation for Marchenko–Pastur

**Linearization trick** [Haagerup–Thorbjørnsen '02, Anderson '11] For the **resolvent**:

$$\left(\frac{1}{d}XX^\top - z\mathrm{I}_n\right)^{-1} = \begin{bmatrix} -z\mathrm{I}_n & \frac{1}{\sqrt{d}}X \\ \frac{1}{\sqrt{d}}X^\top & -\mathrm{I}_d \end{bmatrix}^{-1}_{1,1}.$$

**Matrix Dyson Equation**: Deterministic equation [Alt–Ajanki–Erdős–Schröder–Krüger '19] .

If $\underline{X} = X - \mathbb{E}[X]$

$$\left(\begin{bmatrix} -z\mathrm{I}_n & \frac{1}{\sqrt{d}}\mathbb{E}[X] \\ \frac{1}{\sqrt{d}}\mathbb{E}[X^\top] & -\mathrm{I}_d \end{bmatrix} - \mathcal{S}(M)\right)^{-1} = M, \quad \mathcal{S}(M) = \frac{1}{d}\mathbb{E}\begin{bmatrix} 0 & \underline{X} \\ \underline{X}^\top & 0 \end{bmatrix} M \begin{bmatrix} 0 & \underline{X} \\ \underline{X}^\top & 0 \end{bmatrix}.$$

Suppose $X$ $i.i.d$ standard Gaussian, $\underline{X} = X$,

$$\mathcal{S}(M) = \frac{1}{d}\mathbb{E}\begin{bmatrix} XM_{[2,2]}X^\top & XM_{[2,1]}X \\ X^\top M_{[1,2]}X^\top & X^\top M_{[1,1]}X \end{bmatrix} \approx \frac{1}{d}\begin{bmatrix} \operatorname{Tr} M_{[2,2]}\mathrm{I}_n & 0 \\ 0 & \operatorname{Tr} M_{[1,1]}\mathrm{I}_d \end{bmatrix}.$$

Becomes a **scalar equation**:

$$m(z) := \frac{1}{n}\operatorname{Tr} M_{[1,1]} = \frac{1}{-z - \frac{1}{d}\operatorname{Tr} M_{[2,2]}} = \frac{1}{-z + \frac{1}{1+\frac{n}{d}m(z)}}$$

# NTK is a Gram matrix

Consider $\sigma'(x) = x$ and $a_i = \pm 1$ (for simplicity). Then the model becomes

$$\mathrm{NTK} = \frac{1}{d} X X^\top \odot \frac{1}{p} X W W^\top X^\top.$$

**Kernel Trick** [Grothendieck '53] : For $\varphi : \mathbb{R} \to \mathbb{R}$ and $\mathcal{H}_0$, there exists $\Phi : \mathcal{H}_0 \to \mathcal{H}$,

$$\varphi\left(\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{H}_0}\right) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle_{\mathcal{H}}$$

$$\mathrm{NTK}_{ij} = \frac{1}{pd} \langle \mathbf{x}_i, \mathbf{x}_j \rangle \langle W^\top \mathbf{x}_i, W^\top \mathbf{x}_j \rangle$$

Set

$$\omega^i = \mathbf{x}_i \otimes W^\top \mathbf{x}_i \in \mathbb{R}^d \otimes \mathbb{R}^p =: \mathbb{R}^{d,p}.$$

Then

$$\boxed{\mathrm{NTK}_{ij} = \frac{1}{pd} \langle \omega^i, \omega^j \rangle_{\mathbb{R}^d \otimes \mathbb{R}^p}} =: \frac{1}{pd} \sum_{a=1}^{d} \sum_{b=1}^{p} \omega^i_{ab} \omega^j_{ab} \qquad \textbf{Gram matrix}$$

We want to understand the covariance structure, as a **4-tensor**,

$$\boxed{\mathrm{Cov}(\omega^i, \omega^j) \in \mathbb{R}^{d,p} \otimes \mathbb{R}^{d,p}}$$

# Covariance tensor

**Conditionally on W**, we compute

$$\mathbb{E}_X\left[\omega^i \otimes \omega^j\right] - \mathbb{E}_X\left[\omega^i\right] \otimes \mathbb{E}_X\left[\omega^j\right].$$

For the **expectation** of $\omega^i$, we have

$$\mathbb{E}_X\left[\omega^i\right]_{ab} = \mathbb{E}_X\left[x_{ia}\sum_{k=1}^{d} W_{kb}x_{ik}\right] = W_{ab}.$$

And the **four moment** tensor

$$E_X[\omega^i \otimes \omega^j] = W \otimes W + \delta_{ij}\tau_{23}(\mathrm{I}_d \otimes W^\top W) + \delta_{ij}\tau_{24}(W \otimes W) + \ldots$$

where ... corresponds to the 4-th order term (not important). Finally

$$\boxed{\mathrm{Cov}(\omega^i, \omega^j) = \delta_{ij}\left(\tau_{23}(\mathrm{I}_d \otimes W^\top W) + \tau_{24}(W \otimes W)\right) =: \delta_{ij}\mathcal{Q}}$$

**Braid operator**:

$$\tau_{23}(\mathcal{A})_{abcd} = \mathcal{A}_{acbd}, \qquad \tau_{24}(\mathcal{A})_{abcd} = \mathcal{A}_{adcb}.$$

# Tensor Dyson Equation

**Linearization Trick**: Define the 3-tensor $\mathcal{A} = \frac{1}{\sqrt{pd}} \sum_{i=1}^{n} \omega^i \otimes \mathrm{e}_i$. then

$$\mathcal{A}^\top \mathcal{A} = \frac{1}{pd} \sum_{i,j=1}^{n} \langle \omega^i, \omega^j \rangle \mathrm{e}_i \otimes \mathrm{e}_j = \mathrm{NTK}.$$

$$(\mathcal{A}^\top \mathcal{A} - z\mathrm{I}_n)^{-1} = \begin{bmatrix} -z\mathrm{I}_n & \mathcal{A}^\top \\ \mathcal{A} & -\mathcal{I}_{d,p} \end{bmatrix}^{-1}_{1,1} \quad \text{with} \quad \mathcal{I}_{d,p} = \tau_{23}(\mathrm{I}_d \otimes \mathrm{I}_p) \text{ or } (\mathcal{I}_{d,p})_{pqrs} = \delta_{pr}\delta_{qs}.$$

**Tensor Dyson Equation**:

$$\left( \begin{bmatrix} -z\mathrm{I}_n & \mathbb{E}\left[\mathcal{A}^\top\right] \\ \mathbb{E}\left[\mathcal{A}\right] & -\mathcal{I}_{n,d} \end{bmatrix} - \mathfrak{S}(\mathcal{M}) \right)^{-1} = \mathcal{M}$$

**Superoperator**: For $\underline{\mathcal{A}} = \mathcal{A} - \mathbb{E}\left[\mathcal{A}\right] \in \mathbb{R}^{n,d} \otimes \mathbb{R}^n$,

$$\mathfrak{S}(\mathcal{M}) = \mathbb{E} \begin{bmatrix} \underline{\mathcal{A}}^\top \mathcal{M}_{[2,2]} \underline{\mathcal{A}} & \underline{\mathcal{A}}^\top \mathcal{M}_{[2,1]} \underline{\mathcal{A}}^\top \\ \underline{\mathcal{A}} \mathcal{M}_{[1,2]} \underline{\mathcal{A}} & \underline{\mathcal{A}} \mathcal{M}_{[1,1]} \underline{\mathcal{A}}^\top \end{bmatrix}.$$

For simplicity, think of $\mathfrak{S}$ as **block-diagonal**.

# Superoperator computation

If we recall the **covariance tensor**

$$\boxed{\mathcal{Q} = \tau_{23}(\mathrm{I}_d \otimes W^\top W) + \tau_{24}(W \otimes W)}$$

Then we can compute

$$(\underline{\mathcal{A}}^\top \mathcal{M}_{[2,2]} \underline{\mathcal{A}})_{ij} \approx \frac{\delta_{ij}}{pd} \langle \mathcal{Q}, \mathcal{M}_{[2,2]} \rangle, \quad (\underline{\mathcal{A}} \mathcal{M}_{[1,1]} \underline{\mathcal{A}}^\top) \approx \frac{1}{pd} \operatorname{Tr}(\mathcal{M}_{[1,1]}) \mathcal{Q}$$

**Tensor Dyson Equation** becomes

$$\mathcal{M}_{[1,1]} = \frac{1}{-z - \frac{1}{pd}\langle \mathcal{Q}, \mathcal{M}_{[2,2]} \rangle} \mathrm{I}_n, \quad \mathcal{M}_{[2,2]} = -\left( \mathcal{I}_{n,d} + \frac{1}{pd} \operatorname{Tr}(\mathcal{M}_{[1,1]}) \mathcal{Q} \right)^{-1}.$$

Denote $m(z) = \frac{1}{n} \operatorname{Tr}(\mathcal{M}_{[1,1]})$, then we have the **self-consistent equation**

$$\boxed{m(z) = \frac{1}{-z + \frac{1}{pd} \langle \mathcal{Q}, (\mathcal{I}_{n,d} + \gamma_1 m(z) \mathcal{Q})^{-1} \rangle}}$$

Idea: Consider the **spectrum** of $\mathcal{Q}$ !

# Spectrum of $\mathcal{Q}$

Consider the **singular value decomposition** of $W$

$$W = U\Sigma V^\top$$

Then

$$\begin{aligned}(\tau_{23}(\mathrm{I}_d \otimes W^\top W)\mathbf{u}^{\boldsymbol{i}} \otimes \mathbf{v}^{\boldsymbol{j}})_{k\ell} &= \sum_{q=1}^{d}\sum_{r=1}^{p} \delta_{kq}(W^\top W)_{\ell r} u_k^i v_r^j \\ &= \sigma_j^2 (\mathbf{u}^{\boldsymbol{i}} \otimes \mathbf{v}^{\boldsymbol{j}})_{k\ell}\end{aligned}$$

Similarly, ($\sigma_i = 0$ for $i > p \wedge d$)

$$\tau_{24}(W \otimes W)\mathbf{u}^{\boldsymbol{i}} \otimes \mathbf{v}^{\boldsymbol{j}} = \sigma_i\sigma_j \mathbf{u}^{\boldsymbol{j}} \otimes \mathbf{v}^{\boldsymbol{i}}$$

On Span($\mathbf{u}^{\boldsymbol{i}} \otimes \mathbf{v}^{\boldsymbol{j}}, \mathbf{u}^{\boldsymbol{j}} \otimes \mathbf{v}^{\boldsymbol{i}}$), $\mathcal{Q}$ acts as

$$\begin{bmatrix} \sigma_j^2 & \sigma_i\sigma_j \\ \sigma_i\sigma_j & \sigma_i^2 \end{bmatrix} \longrightarrow \text{Eigenvalues: } \sigma_i^2 + \sigma_j^2, 0 \longrightarrow \mu_{\mathrm{MP}}^{\gamma_2} * \mu_{\mathrm{MP}}^{\gamma_2}$$

$$\text{MP map:} \quad m(z) = \frac{1}{-z + \dfrac{1}{2pd}\displaystyle\sum_{i,j} \dfrac{\sigma_i^2 + \sigma_j^2}{1 + \gamma_1 m(z)(\sigma_i^2 + \sigma_j^2)}}$$

# Spectrum of $\mathcal{Q}$

Consider the **singular value decomposition** of $W$

$$W = U\Sigma V^\top$$

Then

$$\begin{aligned}(\tau_{23}(\mathrm{I}_d \otimes W^\top W)\mathbf{u}^{\boldsymbol{i}} \otimes \mathbf{v}^{\boldsymbol{j}})_{k\ell} &= \sum_{q=1}^{d}\sum_{r=1}^{p} \delta_{kq}(W^\top W)_{\ell r} u_k^i v_r^j \\ &= \sigma_j^2 (\mathbf{u}^{\boldsymbol{i}} \otimes \mathbf{v}^{\boldsymbol{j}})_{k\ell}\end{aligned}$$

Similarly, ($\sigma_i = 0$ for $i > p \wedge d$)

$$\tau_{24}(W \otimes W)\mathbf{u}^{\boldsymbol{i}} \otimes \mathbf{v}^{\boldsymbol{j}} = \sigma_i\sigma_j \mathbf{u}^{\boldsymbol{j}} \otimes \mathbf{v}^{\boldsymbol{i}}$$

On Span($\mathbf{u}^{\boldsymbol{i}} \otimes \mathbf{v}^{\boldsymbol{j}}, \mathbf{u}^{\boldsymbol{j}} \otimes \mathbf{v}^{\boldsymbol{i}}$), $\mathcal{Q}$ acts as

$$\begin{bmatrix} \sigma_j^2 & \sigma_i\sigma_j \\ \sigma_i\sigma_j & \sigma_i^2 \end{bmatrix} \longrightarrow \text{Eigenvalues: } \sigma_i^2 + \sigma_j^2, 0 \longrightarrow \mu_{\mathrm{MP}}^{\gamma_2} * \mu_{\mathrm{MP}}^{\gamma_2}$$

$$\text{MP map:} \quad m(z) = \frac{1}{-z + \displaystyle\int \frac{x \mathrm{d}\mu_{\mathrm{MP}}^{\gamma_2} * \mu_{\mathrm{MP}}^{\gamma_2}(x)}{1 + \gamma_1 m(z) x}}$$

# Adding the nonlinearity

**General nonlinearity** : Write $\sigma'$ as

$$\sigma'(x) = \alpha_{\sigma'} x + \varphi(x), \quad \mathbb{E}_{\mathcal{N}(0,1)}\left[z\varphi(z)\right] = 0.$$

Model becomes ($a_i = \pm 1$ for simplicity)

$$M = \frac{1}{d} XX^\top \odot \left(\frac{\alpha_\varphi}{\sqrt{d}} XW + \varphi\left(\frac{1}{\sqrt{d}} XW\right)\right)\left(\frac{\alpha_\varphi}{\sqrt{d}} XW + \varphi\left(\frac{1}{\sqrt{d}} XW\right)\right)^\top$$

Using a **leave-one-out** strategy, same spectrum as, for $\widetilde{X}$ $i.i.d$ Gaussian,

$$\boxed{\widetilde{M} = \frac{1}{d} XX^\top \odot \left(\frac{\alpha_\varphi}{\sqrt{d}} XW + \beta_{\sigma'} \varphi\left(\widetilde{X}\right)\right)\left(\frac{\alpha_\varphi}{\sqrt{d}} XW + \beta_{\sigma'} \varphi\left(\widetilde{X}\right)\right)^\top}$$

Gives an **additional** term in the covariance tensor $\mathcal{Q}$,

$$\mathcal{Q} = \frac{\alpha_\varphi^2}{d}\left(\tau_{23}\left(\mathrm{I}_d \otimes W^\top W\right) + \tau_{24}\left(W \otimes W\right)\right) + (\beta_{\sigma'}^2 - \alpha_{\sigma'}^2)\boldsymbol{\tau_{23}\left(\mathrm{I}_d \otimes \mathrm{I}_p\right)}$$

**Shifts** the distribution in the MP map,

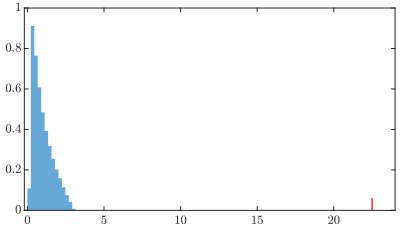
$$\boxed{\mu_{\mathrm{MP}}^{\gamma_1} \boxtimes \left(\alpha_\varphi^2 \mu_{\mathrm{MP}}^{\gamma_2} * \mu_{\mathrm{MP}}^{\gamma_2} + (\beta_{\sigma'} - \alpha_{\sigma'}^2)\right)}$$

**Adding a:** Consider the SVD of $WD$, additional term becomes $\tau_{23}(\mathrm{I}_d \otimes D^2)$

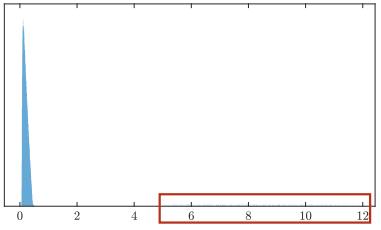
# Conclusion

Questions beyond asymptotic e.e.d:

- **Extreme eigenvalues**: $\lambda_{\min}, \lambda_{\max}$ are related to the convergence rate of gradient descent.




- **Outliers or "minibulk"**: $\mathcal{O}(d)$ eigenvalues outside of the main bulk.



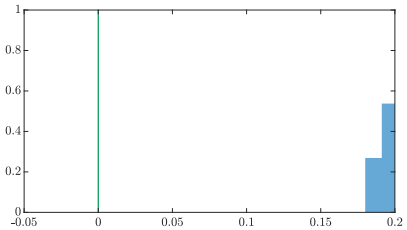

- **Other scalings** in the quadratic regime: $\frac{n}{pd} \to \gamma_1$, $\frac{d}{p^\ell} \to \gamma_2$.

- **Test error** at the interpolation threshold: Beyond eigenvalue distribution!
  - Overparametrized regime: $\frac{n}{dp} \to 0$ [Montanari–Zhong '20]
  - Random Feature Model [Misiakiewicz '24]

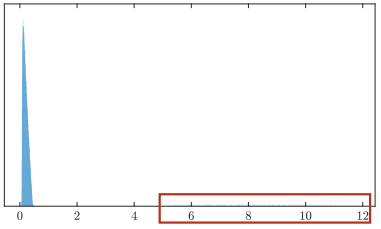
# Conclusion

Questions beyond asymptotic e.e.d:

- **Extreme eigenvalues**: $\lambda_{\min}, \lambda_{\max}$ are related to the convergence rate of gradient descent.




- **Outliers or "minibulk"**: $\mathcal{O}(d)$ eigenvalues outside of the main bulk.



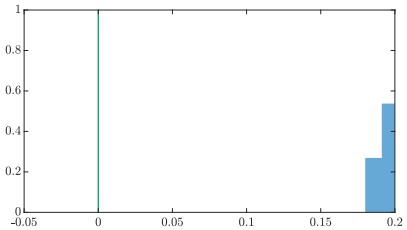

- **Other scalings** in the quadratic regime: $\frac{n}{pd} \to \gamma_1$, $\frac{d}{p^\ell} \to \gamma_2$.

- **Test error** at the interpolation threshold: Beyond eigenvalue distribution!
  - Overparametrized regime: $\frac{n}{dp} \to 0$ [Montanari–Zhong '20]
  - Random Feature Model [Misiakiewicz '24]

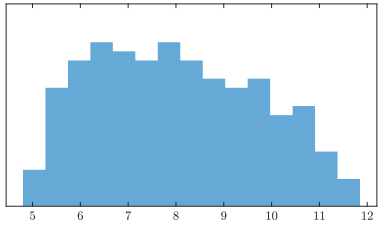
# Conclusion

Questions beyond asymptotic e.e.d:

- **Extreme eigenvalues**: $\lambda_{\min}, \lambda_{\max}$ are related to the convergence rate of gradient descent.




- **Outliers or "minibulk"**: $\mathcal{O}(d)$ eigenvalues outside of the main bulk.



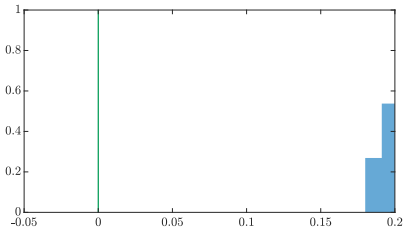

- **Other scalings** in the quadratic regime: $\frac{n}{pd} \to \gamma_1$, $\frac{d}{p^\ell} \to \gamma_2$.

- **Test error** at the interpolation threshold: Beyond eigenvalue distribution!
  - Overparametrized regime: $\frac{n}{dp} \to 0$ [Montanari–Zhong '20]
  - Random Feature Model [Misiakiewicz '24]

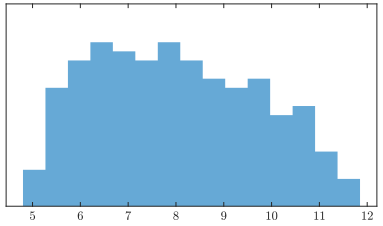
# Conclusion

Questions beyond asymptotic e.e.d:

- **Extreme eigenvalues**: $\lambda_{\min}, \lambda_{\max}$ are related to the convergence rate of gradient descent.




- **Outliers or "minibulk"**: $\mathcal{O}(d)$ eigenvalues outside of the main bulk.




- **Other scalings** in the quadratic regime: $\frac{n}{pd} \to \gamma_1$, $\frac{d}{p^\ell} \to \gamma_2$.

- **Test error** at the interpolation threshold: Beyond eigenvalue distribution!
  - Overparametrized regime: $\frac{n}{dp} \to 0$ [Montanari–Zhong '20]
  - Random Feature Model [Misiakiewicz '24]

**Thank you for your attention!**