Heavy-Tailed Universality in Deep Learning

Liam Hodgkinson (University of Melbourne)

RIKEN-AIP Melbourne Workshop



Joint work with

Zhichao Wang and Michael W. Mahoney
(UC Berkeley)



What is Heavy-Tailed Universality?

What are Heavy Tails?

Definition

Heavy-tailed distributions (informally): densities decaying slower than exponential, often exhibiting power-law tails

$$f(x) \sim c\, x^{-\alpha}, \quad x \to \infty,$$

or inverse-Gamma behavior near zero $f(x) \sim c\, x^{\alpha} e^{-\beta/x}, \quad x \to 0^{+}$.

Neural Scaling Laws and Training Curves

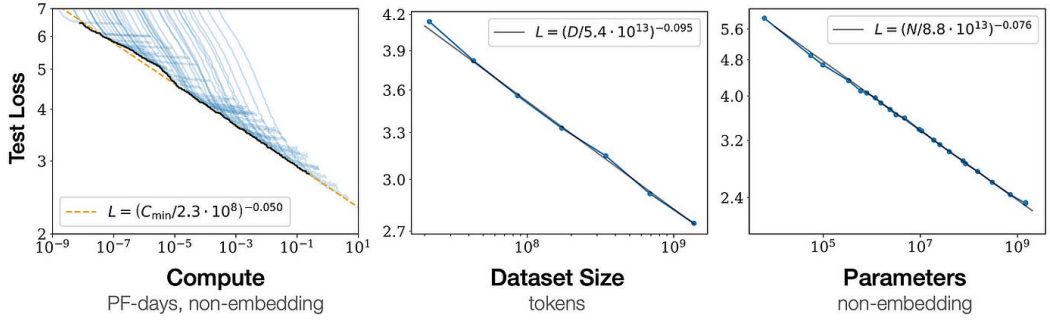
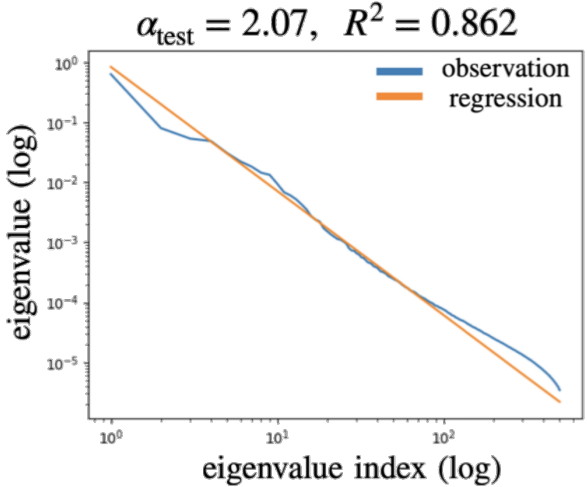


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

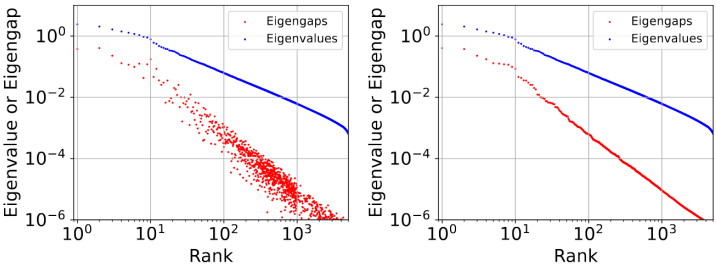
Kaplan et al. (2020); Hoffmann et al. (2022)

Eigenvalues of matrices in neural networks



Activation matrices (Pillaud-Vivien et al., 2018; Agrawal et al., 2022).

Eigenvalues of matrices in neural networks

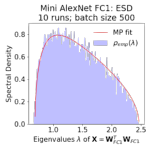


(a) Eigenvalue Rank (b) Eigengap Rank

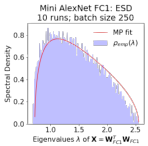
Figure 2: The power-law Hessian eigengaps. Model: LeNet. Datsets: MNIST. Subfigure (a) displayed the eigengaps by original rank indices sorted by eigenvalues. Subfigure (b) displayed the eigengaps by rank indices re-sorted by eigengaps.

Hessian matrices (Xie et al., 2023).

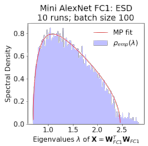
Eigenvalues of matrices in neural networks



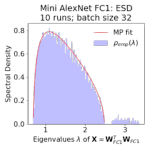
(a) Batch Size 500.



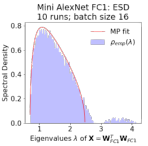
(b) Batch Size 250.



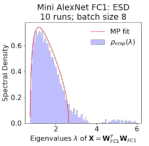
(c) Batch Size 100.



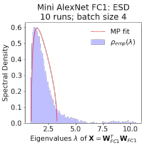
(d) Batch Size 32.



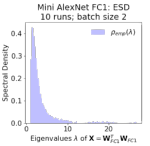
(e) Batch Size 16.



(f) Batch Size 8.

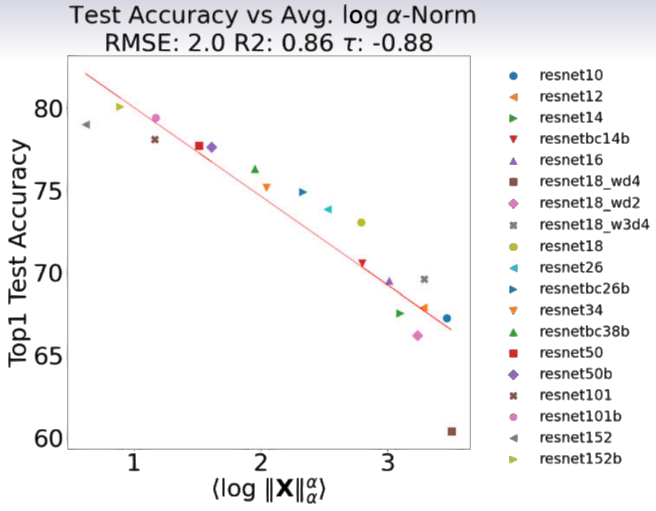


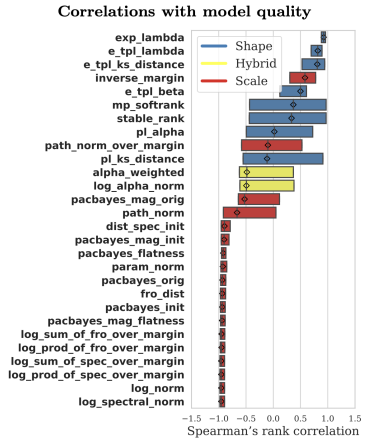
(g) Batch Size 4.



(h) Batch Size 2.

Weight matrices (Martin and Mahoney, 2021).





Strong correlation between heavy-tailed trained weight matrices & model performance (Martin & Mahoney, 2021); useful for layer-wise diagnostics (Zhou et al., 2023; Lu et al, 2024).

Open Questions

- Why do spectral densities of trained feature and weight matrices exhibit heavy-tailed behavior?
- How do data structure, training dynamics, and implicit model bias interplay to produce heavy tails?

Contribution: A drop-in replacement for existing RMT to account for **Heavy-Tailed Universality**.

Theories for Heavy Tails

Possible Approaches for Describing Heavy-Tailed Universality:

- iid Heavy-Tailed Elements: (Arous and Guionnet, 2008) Elements of feature matrices are not independent and heavy-tailed.

Theories for Heavy Tails

Possible Approaches for Describing Heavy-Tailed Universality:

- iid Heavy-Tailed Elements: (Arous and Guionnet, 2008) Elements of feature matrices are not independent and heavy-tailed.
- Kesten Phenomenon: (Hodgkinson and Mahoney, 2021; Vladimirova et al., 2018; Hanin and Nica, 2020) a mechanism discovered by Kesten (1973) for recursive systems:

$$W_{k+1} \approx A_k W_k + B_k$$

Observed for very large step sizes; does not explain common training procedures.

Theories for Heavy Tails

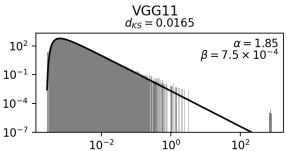
Possible Approaches for Describing Heavy-Tailed Universality:

- iid Heavy-Tailed Elements: (Arous and Guionnet, 2008) Elements of feature matrices are not independent and heavy-tailed.
- Kesten Phenomenon: (Hodgkinson and Mahoney, 2021; Vladimirova et al., 2018; Hanin and Nica, 2020) a mechanism discovered by Kesten (1973) for recursive systems:

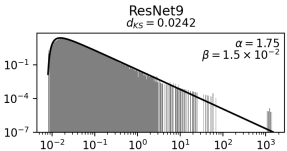
$$W_{k+1} \approx A_k W_k + B_k$$

  Observed for very large step sizes; does not explain common training procedures.
- Population Covariance: power-law in, power-law out (PIPO) principle; power laws come from **data alone**. What about the model?

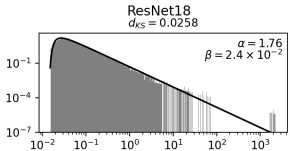
NTK Spectra at Initialization vs. Post-Training



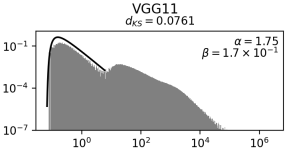
(a) VGG11 Init



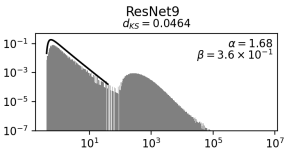
(b) ResNet9 Init



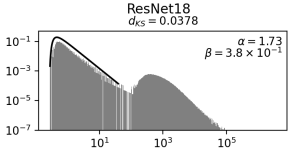
(c) ResNet18 Init



(d) VGG11 Trained



(e) ResNet9 Trained



(f) ResNet18 Trained

Figure: NTK eigenvalue histograms and inverse-Gamma fits near zero. Initialization: mild inverse-Gamma behavior. Post-Training: pronounced heavy-tail

Comparison of Possible Mechanisms

Mechanism	Power Law Elements	Power Law Spectrum	Inverse Gamma
iid Heavy-Tailed Elements	✓	✓	×
Kesten Phenomenon	✓	✓	✓/×
Population Covariance	✓/×	✓	✓/×
Structured Matrices (Ours)	×	✓	✓
Empirical Observations (Features)	×	✓	✓
Empirical Observations (Weights)	×	✓	×

Table: Comparison of various mechanisms: capacity to yield power laws, in feature matrix **elements** and feature matrix **spectral densities**; capacity to yield an inverse Gamma law for the spectral density in a neighborhood of zero.

Modeling Framework

Entropic Regularization Setup

- **Stochastic Minimization Operator**

$$\operatorname*{smin}_{\Theta}^{\pi_\Theta,\tau} f(\Theta) \;:=\; \min_{q\in\mathcal{P}} \left[\mathbb{E}_{q(\Theta)}[f(\Theta)] \;+\; \tau\,\mathrm{KL}(q\,\|\,\pi_\Theta)\right],$$

  where $\mathcal{P}$ is the set of probability densities on the support of π_Θ, and

  - π_Θ is the initial prior (Θ = model coefficients).
  - $\tau > 0$ is the "temperature" (controls early stopping).

- Strong links to Bayesian inference (Germain et al., 2016) and statistical physics of generalization (Mezard and Montanari, 2009).
- Optimizes a **PAC-Bayes bound** on the test error.
- As τ decreases during training, optimizer smoothly interpolates between π_Θ and the final optimal density.

Entropic Regularization Setup

Feature Learning Setup: Stochastic minimization in *two stages*

$$\operatorname*{smin}_{\Theta}^{\pi_\Theta,\tau} L(\Theta,\Phi) \quad \text{and} \quad q(\Phi) \;=\; \operatorname*{argsmin}_{\Phi}^{\pi_\Phi,\eta}\Big[\operatorname*{smin}_{\Theta}^{\pi_\Theta,\tau} L(\Theta,\Phi)\Big].$$

- π_Θ, π_Φ: initial densities of model coefficients Θ and features Φ.
- $\tau, \eta > 0$: "temperatures" control coefficient vs. feature learning rates.

Proposition (Optimal Feature Density)

$$q(\Phi) \;\propto\; \big[\mathcal{Z}_\tau(\Phi)\big]^{\tau/\eta}\,\pi_\Phi(\Phi), \quad \mathcal{Z}_\tau(\Phi) \;=\; \mathbb{E}_{\Theta\sim\pi_\Theta} \exp\big(-L(\Theta,\Phi)/\tau\big).$$

Of particular interest: late stage of training, $\tau, \eta \to 0^+$ with $\tau/\eta \to \rho > 0$.

Example of Feature Matrices: Activation Matrix

Activation Matrix (Last Layer).
Consider a neural network with m outputs: $f(x) = W^\top \varphi(x)$, with $\Phi_{ij} = \varphi_j(x_i)$, $W \in \mathbb{R}^{d\times m}$ trained by ridge regression

$$L(W, \Phi) = \|\Phi W - Y\|_F^2 + \mu \|W\|_F^2.$$

where $\mu > 0$, $\Phi_{ij} = \varphi_j(x_i)$ and $Y = (y_i)_{i=1}^n \in \mathbb{R}^{n\times m}$. For

- $\pi_W = \mathcal{N}(0, \sigma^2 I)$ **(initialization)**
- $\Sigma = YY^\top$ **(data covariance)**
- $M = (1 + \frac{2\mu\sigma^2}{\tau})^{-1}\Phi\Phi^\top + \frac{\tau}{2\sigma^2} I$ **(feature matrix)**

Applying the Proposition,

$$q(M) \propto (\det M)^{-\rho m/2} \exp(-\tfrac{1}{2}\rho\sigma^2 \mathrm{tr}(\Sigma M^{-1}))\pi(M)$$

Master Model Ansatz

- **Ansatz:** for trained feature matrices (activations / NTK / Hessian), with parameters $\alpha, \beta > 0$ and initial density π:

$$q(M) \;\propto\; (\det M)^{-\alpha} \exp\!\big(-\beta \operatorname{tr}(\Sigma\, M^{-1})\big)\, \pi(M)$$

  - $\alpha, \beta > 0$ depend on model/optimizer hyperparameters.
  - Σ is label/covariance-related (e.g., $Y\,Y^{\top}$).
  - $\pi(M)$ is the prior "initialization" density of the feature matrix.

Master Model Ansatz

- **Ansatz:** for trained feature matrices (activations / NTK / Hessian), with parameters $\alpha, \beta > 0$ and initial density π:

$$q(M) \;\propto\; (\det M)^{-\alpha} \exp\!\big(-\beta \operatorname{tr}(\Sigma\, M^{-1})\big)\, \pi(M)$$

  - $\alpha, \beta > 0$ depend on model/optimizer hyperparameters.
  - Σ is label/covariance-related (e.g., $Y\,Y^\top$).
  - $\pi(M)$ is the prior "initialization" density of the feature matrix.

- **Key Observation:** The trained feature matrix M generally follows an *inverse-Wishart-type density*.
  1. First consider $\Sigma = I$ to remove the effect of Σ, the density π of feature matrices M at initialization completely determines the density $q(M)$.
  2. Second, we will consider a general Σ to get spectral densities of trained feature/weight matrices.

RMT for Heavy-Tailed Spectral Behavior

Eigenvector Structure and Beta-Ensembles

- **Key Assumption**: *Distribution of eigenvectors Q is not uniform!* (non-Haar) due to implicit model biases.
- Consider variety of matrix structures to understand effect on eigenvalues
- Although $\pi(M)$ could be complicated, much of the behavior of π is captured by extent of **eigenvalue repulsions**.

Use **Beta-Ensemble** (Dumitriu & Edelman, 2002; Forrester, 2010) with parameter $\kappa \in [0, \infty]$ to capture the Master Model Ansatz:

$$q_\kappa(\lambda_1, \ldots, \lambda_N) \propto \prod_{i=1}^{N} \lambda_i^{-\alpha} e^{-\beta\lambda^{-1}} \prod_{i<j} |\lambda_i - \lambda_j|^{\kappa/N}$$

Interpreting κ: Structured Feature Matrices

π uniform over structured matrix classes with block structures ($n \times n$ blocks, each of size $m \times m$, $m \ll n$):

1. **Diagonal**: $\kappa = 0$ (no eigenvector randomness).
2. **Commuting Block-Diagonal**: $\kappa \sim \frac{m}{n}$.
3. **Symmetric Block-Diagonal**: $\kappa \sim m$.
4. **Kronecker-Like** $Q_1 \otimes Q_2$, where $Q_1 \in \mathbb{R}^{m \times m}$ and $Q_2 \in \mathbb{R}^{n \times n}$: $\kappa \sim \frac{n}{m} + \frac{m}{n}$.
5. **Fully Symmetric (no structure)**: $\kappa = mn$.

- As model architecture induces *more structure* (fewer free eigenvector degrees of freedom), κ *decreases* $\Rightarrow$ heavier tail in spectrum.
- We provide a numerical algorithm to efficiently estimate κ.

HTMP Distribution

Theorem (Generalized Marchenko–Pastur)

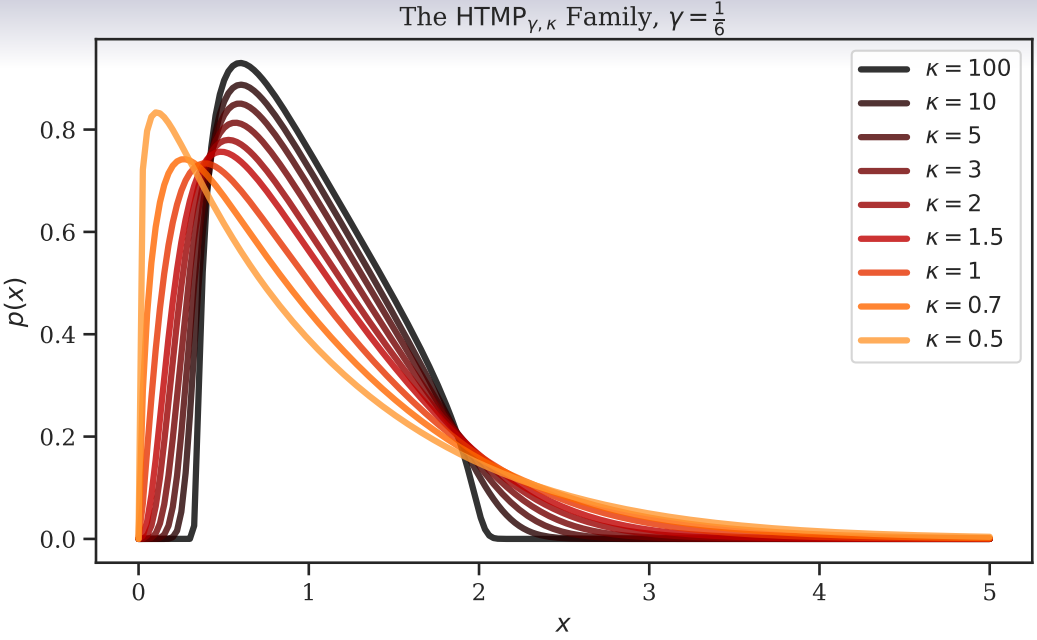
Let M_N *follow* $q_\kappa(\lambda_1, \ldots, \lambda_N) \propto \prod_{i=1}^{N} \lambda_i^{-\alpha} e^{-\beta \lambda_i^{-1}} \prod_{i<j} |\lambda_i - \lambda_j|^{\frac{\kappa(N)}{N}}$ *with parameter* $\kappa(N)$*. Define*

$$\gamma(N) = \frac{\kappa(N)/2}{\alpha - \kappa(N)/2 - 1} \to \gamma \in (0, 1) \quad \text{as } N \to \infty.$$

Then the empirical spectral distribution of $\frac{2\gamma(N)\beta}{\kappa(N)} M_N^{-1}$ *converges to:*

1. $\mathbf{MP}_\gamma$ *(Marchenko-Pastur distribution) if* $\kappa(N) \to \infty$*;*
2. $\mathbf{HTMP}_{\gamma,\kappa}$ *(High-Temperature MP) if* $\kappa(N) \to \kappa \in (0, \infty)$*.*

This beta-ensemble result is derived from Dumitriu and Edelman (2006); Dung and Duy (2021).


The HTMP$_{\gamma,\kappa}$ Family, $\gamma = \frac{1}{6}$
$\kappa = 100$
$\kappa = 10$
$\kappa = 5$
$\kappa = 3$
$\kappa = 2$
$\kappa = 1.5$
$\kappa = 1$
$\kappa = 0.7$
$\kappa = 0.5$
$p(x)$
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
x

Main Theorem: Tail Behavior for Trained Features

Theorem (Spectral Density of Trained Feature Matrix)

Let ρ_N be the ESD of a trained feature matrix M_N, and μ_Σ the spectral measure of label covariance Σ. Under the Master Model,

$$\rho_N(\lambda) \xrightarrow[N\to\infty]{} (\mu_\Sigma \boxtimes \rho)(\lambda),$$

where $\boxtimes$ is multiplicative free convolution, ρ is either $\lambda^{-2}\,\rho_{\mathrm{MP}}(\lambda^{-1})$ (if $\kappa = \infty$) or $\lambda^{-2}\,\rho_{\mathrm{HTMP}}(\lambda^{-1})$ (if $\kappa < \infty$). Additionally,

- *Inverse-Gamma near zero: If $\kappa < \infty$, density $\rho(x) \sim x^{-\frac{\kappa}{2\gamma}-1-\frac{\kappa}{2}} \exp\left(-\frac{\beta_-}{x}\right)$ as $x \to 0^+$.*
- *Power-law Tail: $\rho(x) \sim x^{-\frac{\kappa}{2\gamma}-1+\frac{\kappa}{2}}$ for $x \to \infty$.*

Applications

Application 1: Neural Scaling Laws

- **Setup**: Ridge regression on activation matrix $\Phi \in \mathbb{R}^{n\times d}$, $m = 1$:

$$\hat{w} = \operatorname*{argmin}_w \; L(w) \;=\; \frac{1}{n}\|\Phi w - Y\|^2 + \frac{\mu}{n}\|w\|^2.$$

  Assume $y_i = w_*^\top \varphi(x_i)$, and $\mathbb{E}_x[\varphi(x)\varphi(x)^\top] = I$.

- **Spectral Assumption**: $\Phi\Phi^\top$ follows $\mathbf{HTMP}_{\gamma,\kappa}$ (Master Model).

Proposition

*Let $\mu = n^{-\ell}$ with $\ell \in (0,1)$. Then, the **Generalization Error** satisfies*

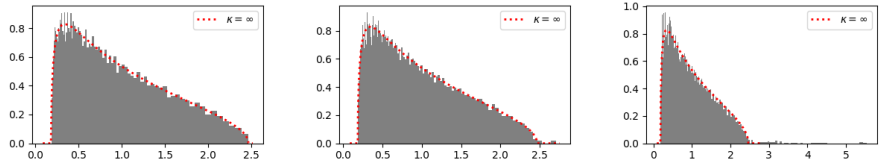
$$\mathcal{L} := \mathbb{E}_{x,w_*}[(\varphi(x)^\top \hat{w} - y)^2] \asymp n^{-\ell\left(2+\frac{\kappa}{2\gamma}-\frac{\kappa}{2}\right)}, \quad n \to \infty$$

with high probability.

Application 2: 5+1 Phases of Trained Weight Matrices

- **Empirical Observation** (Martin and Mahoney, 2019, 2020, 2021; Yang et al., 2023; Zhou et al., 2023): Trained weight matrices can exhibit 5+1 Phases of Training:
  1. Random-Like (MP bulk, no outliers).
  2. Bleeding-Out (MP bulk with emerging spikes).
  3. Bulk+Spikes (distinct spikes outside bulk).
  4. Bulk-Decay (bulk extends, no finite support).
  5. Heavy-Tailed (power-law tail).
  6. Rank-Collapse (mass at zero eigenvalue).
- **Application**: Weight spectral densities converge to $\mathbf{HTMP}_{\gamma,\kappa}$.

5+1 Phases for Trained Weight: HTMP Fits



(a) **Random-Like**: batch size 1000 (b) **Bleeding-Out**: batch size 800 (c) **Bulk+Spikes**: batch size 250

(d) **Bulk-Decay**: batch size 100 (e) **Heavy-Tailed**: batch size 50 (f) **Rank Collapse**: batch size 5

Figure: Weight spectral densities for MiniAlexNet, CIFAR-10; batch sizes 1000, 800, 250, 100, 50, 5 (top to bottom). *Fitted MP/HTMP curves shown in red dashed* with **(a)–(c)** $\kappa = \infty$, **(d)** $\kappa = 5.5$, **(e)** $\kappa = 1.9$, **(f)** $\kappa \to 0^+$

Conclusions

- **Master Model**: A unified RMT framework (Master Model Ansatz) that captures heavy-tailed spectral behavior of trained feature matrices from a Bayesian perspective.
- **HTMP Ensemble**: High-temperature MP ($\mathbf{HTMP}_{\gamma,\kappa}$) arises for structured feature matrices; interpolates between MP ($\kappa \to \infty$) and heavy-tailed regimes ($\kappa \to 0^+$).
- **Key Insights**
  1. *Data Contribution*: Heavy-tailed population covariance $\Sigma \implies$ heavy-tailed trained spectra (PIPO).
  2. *Eigenvector Structure*: Architectural bias (smaller κ) $\implies$ heavier tails.
  3. *Training Dynamics*: As $\tau, \eta \to 0$, HTMP hyperparameters α, β, κ evolve, explaining transitions (5+1 phases).
- **Applications**
  - Neural scaling laws (ridge regression) predicted by HTMP exponents.
  - 5+1 training phases fit by tuning κ for HTMP.

Thank You!



Liam Hodgkinson, Zhichao Wang, Michael W. Mahoney.
"Models of Heavy-Tailed Mechanistic Universality"
https://arxiv.org/abs/2506.03470.

MP vs. HTMP

- The Marchenko-Pastur distribution $\mathbf{MP}_\gamma$ with parameter $\gamma \in (0,1)$ is absolutely continuous on $(0,\infty)$ with finite support only on the interval $I_\gamma = [\gamma_-, \gamma_+]$ where $\gamma_\pm = (1 \pm \sqrt{\gamma})^{1/2}$. The corresponding probability density function is given by

$$\rho_\gamma(x) = \frac{1}{2\pi} \frac{\sqrt{(\gamma_+ - x)(x - \gamma_-)}}{\gamma x}, \qquad x \in I_\gamma.$$

- The high-temperature Marchenko-Pastur distribution $\mathbf{HTMP}_{\gamma,\kappa}$ is a probability distribution on $(0,\infty)$ with a probability density function

$$\rho_{\gamma,\kappa}(x) = \frac{\kappa}{2\gamma} \frac{1}{\Gamma(\kappa/2+1)\Gamma(\kappa/2\gamma)} \frac{\left(\frac{\kappa x}{2\gamma}\right)^{\frac{\kappa}{2\gamma}-1-\frac{\kappa}{2}} e^{-\frac{\kappa x}{2\gamma}}}{|U(\kappa/2, -\frac{\kappa}{2\gamma}+1+\frac{\kappa}{2}; -\kappa x/2\gamma)|^2}$$

where $U(a,b;z)$ denote the Tricomi confluent hypergeometric function.

References I

G. B. Arous and A. Guionnet. The spectrum of heavy tailed random matrices. *Communications in Mathematical Physics*, 278(3):715–751, 2008.

I. Dumitriu and A. Edelman. Global spectrum fluctuations for the β-hermite and β-laguerre ensembles via matrix models. *Journal of Mathematical Physics*, 47(6), 2006.

T. H. Dung and T. K. Duy. Beta laguerre ensembles in global regime. *Osaka Journal of Mathematics*, 58(2):435–450, 2021.

P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. *Advances in Neural Information Processing Systems*, 29, 2016.

B. Hanin and M. Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020.

L. Hodgkinson and M. W. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning*, pages 4262–4274, 2021.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

H. Kesten. Random difference equations and renewal theory for products of random matrices. *Acta Mathematica*, 131(1):207–248, 1973.

C. H. Martin and M. W. Mahoney. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pages 4284–4293, 2019.

C. H. Martin and M. W. Mahoney. Heavy-tailed universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 505–513. SIAM, 2020.

C. H. Martin and M. W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *The Journal of Machine Learning Research*, 22(1):7479–7551, 2021.

M. Mezard and A. Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

M. Vladimirova, J. Arbel, and P. Mesejo. Bayesian neural networks become heavier-tailed with depth. In *NeurIPS 2018-Thirty-second Conference on Neural Information Processing Systems*, pages 1–7, 2018.

References II

Y. Yang, R. Theisen, L. Hodgkinson, J. E. Gonzalez, K. Ramchandran, C. H. Martin, and M. W. Mahoney. Test accuracy vs. generalization gap: Model selection in NLP without accessing training or testing data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3011–3021, 2023.

Y. Zhou, T. Pang, K. Liu, C. H. Martin, M. W. Mahoney, and Y. Yang. Temperature balancing, layer-wise weight analysis, and neural network training. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.