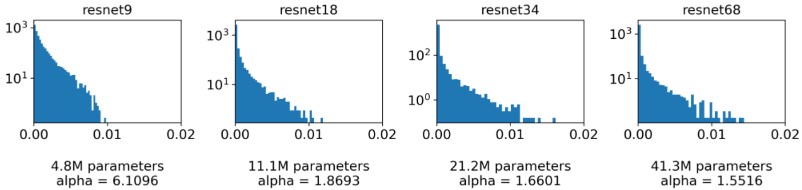
Models of Heavy-Tailed Mechanistic Universality

Liam Hodgkinson (University of Melbourne)
Zhichao Wang (UC Berkeley)
Michael W. Mahoney (UC Berkeley)

Forty-Second International Conference on Machine Learning (ICML 2025)

Heavy-Tailed Phenomena in Machine Learning



- Gradient norms (Simsekli et al., 2019; Hodgkinson et al., 2020) and loss curves (Hestness et al., 2017; Hoffman et al., 2024).

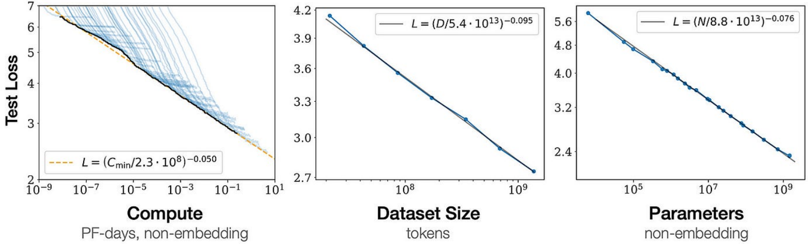
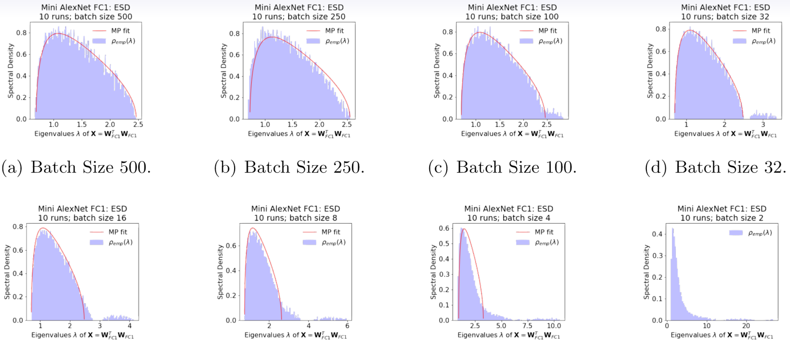


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when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

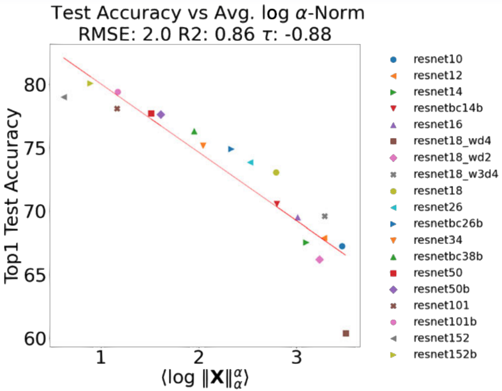
- Power law appears in neural scaling laws (Kaplan et al., 2020; Wei et al., 2022; Defilippis et al., 2024; Paquette et al., 2024; Lin et al., 2024).

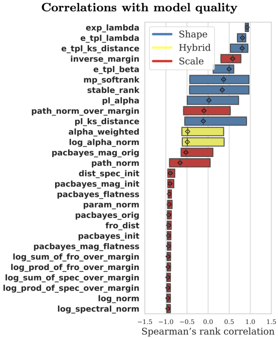


(a) Batch Size 500. (b) Batch Size 250. (c) Batch Size 100. (d) Batch Size 32.

(e) Batch Size 16. (f) Batch Size 8. (g) Batch Size 4. (h) Batch Size 2.

- Eigenvalues of Gram matrices in neural nets: data covariance (Sorscher et al., 2022; Zhang et al., 2023), activation (conjugate kernel) (Pillaud-Vivien et al., 2018; Agrawal et al., 2022; Wang et al., 2023), Hessian (Xie et al., 2023), Jacobian (Wang et al., 2023).





- 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).

Heavy-Tailed Mechanistic Universality

Definition

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

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

Existing Approaches for Describing HT-MU:

- iid Heavy-Tailed Elements: (Arous & Guionnet, 2008). Elements of feature matrices are not independent and heavy-tailed in practice.
- Kesten Phenomenon: (Hodgkinson & Mahoney, 2021; Gurbuzbalaban et al, 2021; Vladimirova et al, 2018; Hanin & Nica, 2020) a mechanism discovered by Kesten (1973) for recursive systems.
- Population Covariance: power-law in, power-law out (PIPO) principle.

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?

Need new RMT for **Heavy-Tailed Mechanistic Universality (HT-MU)**.

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.

- **Feature Learning Setup:** Stochastic minimization in *two stages*

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

  - 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)

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

Master Model Ansatz

- **Ansatz:** for trained feature matrices, with parameters $\alpha, \beta > 0$ and initial density π:

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

  - $\alpha, \beta > 0$ 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.

- To get spectral density, change of variables $M \mapsto Q \Lambda Q^\top$ for orthogonal Q and diagonal Λ; so we only need to study the spectral distribution Λ.
- Let $\Sigma = I$ to remove the effect of Σ for now.

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

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

- 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.

The HTMP Distribution

Theorem (Generalized Marchenko–Pastur)

Let M_N follow the high-temperature beta-ensemble. The empirical spectral distribution of M_N^{-1} (appropriately scaled) converges to:

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

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. Then

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

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

- *Inverse-Gamma Law 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 $x \to \infty$.*

5+1 Phases for Trained Weight: HTMP Fits

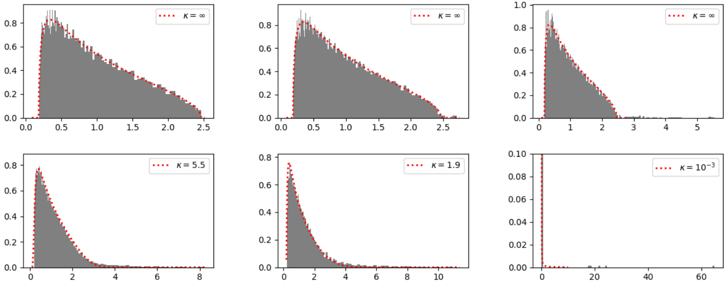


Figure: Weight spectral densities for MiniAlexNet trained on CIFAR-10 with batch sizes 1000, 800, 250, 100, 50, 5 (top to bottom). *Fitted MP/HTMP curves shown in red dashed* with different κ.

As batch size decreases, κ decreases $\Rightarrow$ heavier tail.
(a)–(c): $\kappa = \infty$ for MP or MP+spike behavior.
(d)–(f): Finite κ for heavy tail plus eventual rank collapse.

Neural Scaling Law

- **Setup:** Ridge regression for a fixed set of features Φ.
- **Spectral Assumption:** Feature matrix follows Master Model ($\textbf{HTMP}_{\gamma,\kappa}$).
- *Data-Free Scaling Law:* Predicts test loss decay solely from spectral tail; no access to held-out data required. Previous scaling law work focus on power law in the dataset (e.g., Wei et al (2022); Defilippis et al (2024); Paquette et al (2024); Lin et al (2024)),

Proposition

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

$$\mathcal{L} \asymp n^{-\ell\left(2+\frac{\kappa}{2\gamma}-\frac{\kappa}{2}\right)}, \quad n \to \infty.$$

Thank You!

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