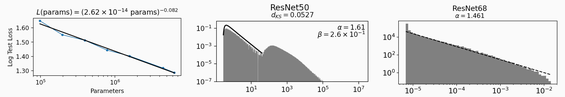
Models for Heavy-Tailed Mechanistic Universality

Liam Hodgkinson[1], Zhichao Wang[2,3], Michael W. Mahoney[2,3,4]
[1]University of Melbourne, [2]UC Berkeley, [3]ICSI, [4]LBNL

Heavy-Tailed Mechanistic Universality



Neural Scaling Laws (GPT-2; OpenWebText) | Spectral Densities (NTK matrix; post-training) | Gradient Norms (during training)

Motivation: Heavy tails appear ***everywhere*** in deep learning.

- Gradients norms, neural scaling laws, loss curves during training.
- Eigenvalues of Gram matrices in neural nets: weights, data covariance, activation, Hessian, Jacobian matrices.
- Strong correlation between heavy-tailed trained weight matrices & model performance: Heavy-Tailed Self-Regularization (HT-SR) theory.

Definition. Heavy-Tails and Inverse Gamma Law

Densities decaying slower than exponential, often ***power-laws*** $f(x) \sim cx^{-\alpha},\ x \to \infty$, e.g., the ***inverse Gamma law***: $f(x) \propto x^{-\alpha}e^{-\beta/x}$.

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?

Theories for Heavy Tails

- **iid Heavy-Tailed Elements:** Elements of feature matrices do not satisfy!
- **Kesten Phenomenon:** Discovered by Kesten (1973) for recursive systems $W_{k+1} \approx A_k W_k + B_k$. Observed for very large step sizes only.
- **Population Covariance:** Power-law in, power-law out (PIPO) principle; power laws come from **data alone**. What about the model?

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 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 & Master Model Ansatz

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

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).

Feature Learning Setup : Stochastic minimization in *two stages*

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

- 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

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

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

Ansatz: For trained feature matrices M (activation matrices, NTK Gram matrices, and Hessian matrices), with parameters $\alpha, \beta > 0$, initial density π:

$$\boldsymbol{q(M) \propto (\det M)^{-\alpha}\exp\left(-\beta\,\mathrm{tr}(\Sigma M^{-1})\right)\pi(M)} \qquad (1)$$

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

Our Main Results

The HTMP Distribution: High temperature beta ensembles imply a drop-in replacement for Marchenko–Pastur (MP).

Theorem. Generalized Marchenko–Pastur

Let M_N follow the High-Temperature Beta Ensemble (2) with parameter $\kappa(N)$. 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 $\kappa(N) \to \kappa \in (0,\infty)$.

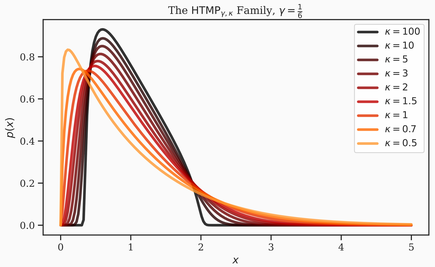


Figure: Plots of the $\mathbf{HTMP}_{\gamma,\kappa}$ densities for $\gamma = 1/6$ and varying κ. $\kappa = 100$ closely resembles the $\kappa \to \infty$ limit (the Marchenko–Pastur density $\mathbf{MP}_\gamma$).

Tail Behavior: The HTMP distribution exhibits the desired spectral behavior.

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 Ansatz,

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

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

Structured Random Matrices & High-Temperature Beta Ensembles

Key Observation: ***Distribution of eigenvectors Q is not uniform*** (non-Haar) due to implicit model biases.

- Consider a variety of matrix structures to understand effect on eigenvalues
- Although $\pi(M)$ could be complicated, much of the behavior is captured by extent of **eigenvalue repulsions**.

High-Temperature Beta-Ensemble: introduce parameter $\kappa \in [0,\infty]$ to capture the joint eigenvalue density for the Master Model Ansatz (1).

$$q_\kappa(\lambda_1,\ldots,\lambda_N) \propto \prod_{i=1}^N \lambda_i^{-\alpha}e^{-\beta\lambda^{-1}} \prod_{i<j}|\lambda_i - \lambda_j|^{\kappa/N} \qquad (2)$$

- As model architecture induces ***more structure***, κ decreases
- Provide ***a numerical algorithm*** to estimate κ.

Structure	$\boldsymbol{\kappa^*}$
(a) Diagonal	0
(b) Commuting block diagonal	$\frac{m}{n} - \frac{1}{2n}$
(c) Symmetric block diagonal	$(m-1)\cdot\frac{mn}{mn-1}$
(d) Kronecker-like matrix	$\frac{n}{m} + \frac{m}{n}$
(e) Symmetric (no structure)	mn

Table: Behavior of κ^* across matrix structures for matrices with $n \times n$ blocks of size $m \times m$. Assuming $n > m$, the first term in provides the dominant behavior.

Applications

Neural Scaling Law: Assuming only HTMP behavior for the features.

- ***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. Minimal-Assumption Neural Scaling Law

Letting $\mu = n^{-\ell}$ with $\ell \in (0,1)$, the **generalization error** $\mathcal{L}$ satisfies with high probability:

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

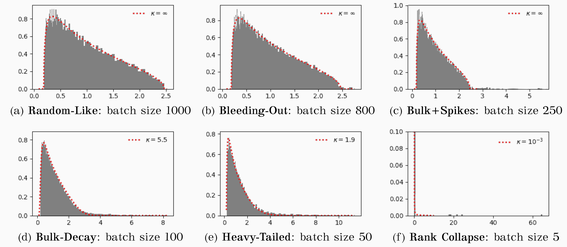
Optimizer Trajectories: *Lower* and *Upper* power-law tails in the distribution of stochastic gradient norms $\|\widehat{\nabla}L_N\|$ during training:

$$\mathbb{P}(\|\widehat{\nabla}L_N\| \le x) \sim C_-\,x^\alpha, \quad x \to 0^+, \qquad \mathbb{P}(\|\widehat{\nabla}L_N\| > x) \sim C_+\,x^{-\beta}, \quad x \to \infty.$$

- ***Model:*** Assume residuals $\bar{Y}$ are Gaussian, NTK matrix M follows inverse-Wishart (or $\mathbf{HTMP}$) independent of $\bar{Y}$.
- ***Application:*** Under these assumptions, $\|\widehat{\nabla}L_N\|$ exhibits both lower and upper power-law tails.

The 5+1 Phases of Learning: Observed by Martin & Mahoney (2017–2020).

- ***Empirical Observation***: Trained weight matrices can exhibit phases:
  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 eigenvalue 0).
- ***Application:*** Consider $A = W^\top W$ with trained weight W, then $\frac{\beta}{\alpha-\kappa/2-1}A$ converges to $\mathbf{HTMP}_{\gamma,\kappa}$.



(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. *Fitted MP/HTMP curves shown in red dashed* with **(a)–(c)** $\kappa = \infty$, **(d)** $\kappa = 5.5$, **(e)** $\kappa = 1.9$, **(f)** $\kappa \to 0^+$