Top-Down Interpretability Through Eigenspectra

Jesse Hoogland, Jan Wehner, Rauno Arike, & Simon C. Marshall

Abstract

Random matrix theory (RMT) offers a host of tools to make sense of neural networks. In this paper, we look at the heavy-tailed random matrix theory developed in [4]. From the spectrum of eigenvalues, it's possible to derive generalization metrics that are independent of data, and to make decompose the training process into five unique phases. Additionally, the theory predicts and tests a key form learning bias known as "self-regularization." In this paper, we extend the results from computer vision to language models, finding many similarities and a few potentially meaningful differences. This provides a glimpse of what more "top-down" interpretability approaches might accomplish: from a deeper understanding of the training process and path-dependence to inductive bias and generalization.

1 Introduction

Understanding the internal dynamics of transformers is critical to more advanced interpretability work. Previously, research into this direction has focused on building a "bottom-up" understanding of how the basic building blocks like individual neurons give rise to higher scale structure ("circuits") [8, 7, 9]. This has provided meaningful breakthroughs towards understanding phenomena like polysemanticity [1] and structures like induction heads [9], but without high level insights it seems unlikely that full interpretability will be possible. More generally, interpretability has been argued to be critical as a tool for achieving AGI Safety [3, 6].

In this paper, we approach interpretability from a different, "top-down" angle that draws on techniques developed in other fields like statistical physics and chaos theory. Instead of looking at individual neurons, we study the spectral properties of entire layers. Further work in this area will complement and augment the bottom-up approaches, allowing for a more complete understanding of models decision making.

In this paper, we will extend the use of Marchenko-Pastur (MP) and Heavy-Tail (HT) Random Matrix Theory (RMT) to study large language models. [4] has demonstrated that the distribution of the eigenvalues (the "spectrum") of the weight matrices is related to high-level properties of the model such as generalisation and self-regularisation [4]. [5] applies these ideas broadly, beyond just image models, but looks only briefly at language models as the focus is on generalisation metrics. Here, we will use the tool set developed by [4] to analyse publicly available models. We extend upon the work in [5], which only briefly touches on GPT-2 small [10] by, analysing GPT-2 at multiple sizes, investigate elements of the architecture separately and understand the impacts of fine-tuning.

In section 2 we refresh the reader's understanding of RMT. In section 3, we present our key results: the ways in which the properties of the spectrum of transformer-based language models resemble those of computer vision (CV) models and the ways in which they differ. As in CV models, we find that LLMs's weight matrices are well modeled by heavy-tail random matrices. However, there appear to be additional "quasi-null" eigenvalues that do not show up in previously studied CV models. In the same section, we demonstrate that we are unable to reproduce the "rank collapse" phase conjectured by [4] when fine-tuning a LLM on a task designed to encourage over-fitting with a large regularizer. Finally, we study the spectral properties across models of different sizes.

Preprint. Under review.

2 Random Matrix Theory: Preliminaries

"Bulk universality" is the conjecture that the macroscopic properties of random matrices do not depend on their microscopic details. Formally, in the limit of infinitely sized matrices, the eigenvalue spectrum depends *only* on the global symmetries of the matrix. Similar to the central limit theorem guaranteeing that a sum of many independent random variables is Gaussian, the bulk universality conjecture suggests that a matrix of many independently random variables has an ordered and predictable eigenvalue spectrum.

Although the conjecture has been proved for particular cases[11], the general form remains a conjecture — though a surprisingly robust one at that. Even after relaxing many of the key assumptions, the empirical conclusions often remain valid. As such, random matrices are a powerful tool to model complex systems — from nuclear energy levels to weight matrices of deep neural networks [12, 4].

2.1 Theory TL;DR

The weight matrix of a particular layer, dictates the action of that layer. If it were randomly generated (each element drawn from a gaussian) then the statistics are well known. It isn't randomly generated. We can analyse properties of the model through the eigenspectra[1] of the weights. By starting from a random matrix we can analyse how the model deviates from this, looking at a plot of the eigenspectra allows us to see how it deviates, depending on the curve of that plot we can infer things about properties, such as the generalisation performance.

2.2 Marchenko-Pastur (MP) Theory for Rectangular Matrices

The primary objects of focus in random matrix theory are the density of eigenvalues $\rho(\lambda)$ (the "bulk") and the distribution of the maximal eigenvalue $p(\lambda_{max})$ (the "edge").

In the case of random square symmetric matrices, the bulk distribution is given by the Wigner Semicircle Law and the edge distribution is given by the Tracy Widom (TW) Law. However, the weight matrices, $\mathbf{W} \in \mathcal{R}^{N \times M}$, of neural networks are neither square nor symmetric. To derive similar results, we turn to Marchenko-Pastur (MP) theory.

Instead of looking at the eigenvalues, we study the squared singular values (i.e., the eigenvalue of $\mathbf{X} = \mathbf{W}^T\mathbf{W}$). When the elements of $\mathbf{W}$ are sampled from a Gaussian distribution,

$$W_{ij} \sim N(0, \sigma^2_{\text{mp}}) \tag{1}$$

MP theory derives the following limiting form for the bulk density:

$$\rho_N(\lambda) := \frac{1}{N}\sum_{i=1}^{M}\delta(\lambda - \lambda_i) \tag{2}$$

$$\lim_{N\to\infty}\rho_N(\lambda) = \frac{Q}{2\pi\sigma^2_{\text{mp}}}\frac{\sqrt{(\lambda^+ - \lambda)(\lambda - \lambda^-)}}{\lambda}, \text{if } \lambda \in [\lambda^-, \lambda^+] \tag{3}$$

$$0, \text{otherwise.} \tag{4}$$

Above, $Q = N/M \geq 1$ is the aspect ratio of the matrix, and the minimum/maximum eigenvalues, $\lambda^{\pm}$, are given by

$$\lambda^{\pm} = \sigma^2_{\text{mp}}\left(1 \pm \frac{1}{\sqrt{Q}}\right)^2. \tag{5}$$

The important point here is that the density becomes bounded in a finite interval.

[1] The eigenspectra of a matrix provides information about how the matrix acts on given inputs

	Generative Model w/ elements from Universality class	Finite-N Global shape $\rho_N(\lambda)$	Limiting Global shape $\rho(\lambda),\ N \to \infty$	Bulk edge Local stats $\lambda \approx \lambda^+$	(far) Tail Local stats $\lambda \approx \lambda_{max}$
Basic MP	Gaussian	MP, i.e., Eqn. (1)	MP	TW	No tail.
Spiked-Covariance	Gaussian, + low-rank perturbations	MP + Gaussian spikes	MP	TW	Gaussian
Heavy tail, $4 < \mu$	(Weakly) Heavy-Tailed	MP + PL tail	MP	Heavy-Tailed*	Heavy-Tailed*
Heavy tail, $2 < \mu < 4$	(Moderately) Heavy-Tailed (or "fat tailed")	PL** $\sim \lambda^{-(a\mu+b)}$	PL $\sim \lambda^{-(\frac{1}{2}\mu+1)}$	No edge.	Frechet
Heavy tail, $0 < \mu < 2$	(Very) Heavy-Tailed	PL** $\sim \lambda^{-(\frac{1}{2}\mu+1)}$	PL $\sim \lambda^{-(\frac{1}{2}\mu+1)}$	No edge.	Frechet

Table 1. Basic MP theory, and the spiked and Heavy-Tailed extensions we use, including known, empirically-observed, and conjectured relations between them. Boxes marked "*" are best described as following "TW with large finite size corrections" that are likely Heavy-Tailed (Biroli et al., 2007b), leading to bulk edge statistics and far tail statistics that are indistinguishable. Boxes marked "**" are phenomenological fits, describing large ($2 < \mu < 4$) or small ($0 < \mu < 2$) finite-size corrections on $N \to \infty$ behavior. See (Davis et al., 2014; Biroli et al., 2007b;a; Péché; Auffinger et al., 2009; Edelman et al., 2016; Auffinger & Tang, 2016; Burda & Jurkiewicz, 2009; Bouchaud & Potters, 2011; Bouchaud & Mézard, 1997) for additional details.

Figure 1: Figure from [5].

2.3 Heavy-Tailed Random Matrix Theory

The MP results assume matrices of *independently* and identically distributed components. Often, the results (for bulk density and edge distribution) continue to apply empirically even in weakly correlated regimes.

As the components become more correlated (as is the case for DNNs), 2 eventually breaks down. Still, we can *model* the spectral properties of DNN weight matrices with an appropriate choice of random matrix universality class. In this case, the relevant universality class consists of matrices whose components are sampled from heavy-tail distributions, which induces heavy tails in the eigenvalue spectrum (see 1).

2.4 Capacity and Generalization Metrics

[4] offer a handful of metrics derived from the empirical spectral density that quantify how much capacity a given layer has.

First, we have the typical linear algebraic rank (the "hard rank"),

$$\mathcal{R}(\mathbf{W}) = \sum_i \delta(\nu_i), \tag{6}$$

is the number of positive singular values $\nu_i > 0$.

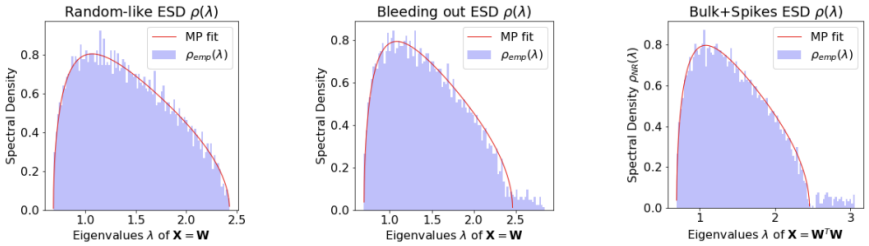
In addition to the hard rank, [4] study the stable rank, defined as:

$$\mathcal{R}_s(\mathbf{W}) = \frac{||\mathbf{W}||_F^2}{||\mathbf{W}||_2^2} = \frac{\sum_i \nu_i^2}{\nu_{\max^2}} = \frac{\sum_i \lambda_i}{\lambda_{\max}}, \tag{7}$$

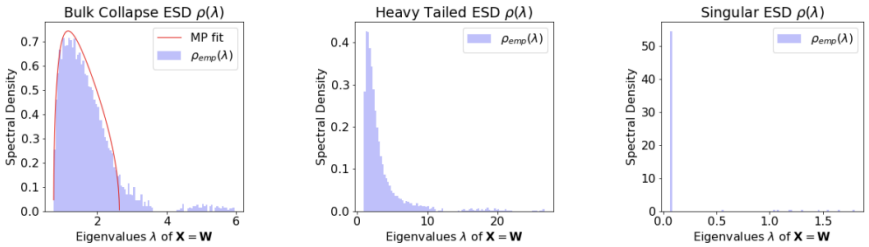
and the matrix entropy (or generalized von-Neumann matrix entropy), defined as:

$$\mathcal{S}(\mathbf{W}) = \frac{-1}{\log(R(\mathbf{W})} \sum_i p_i \log p_i. \tag{8}$$

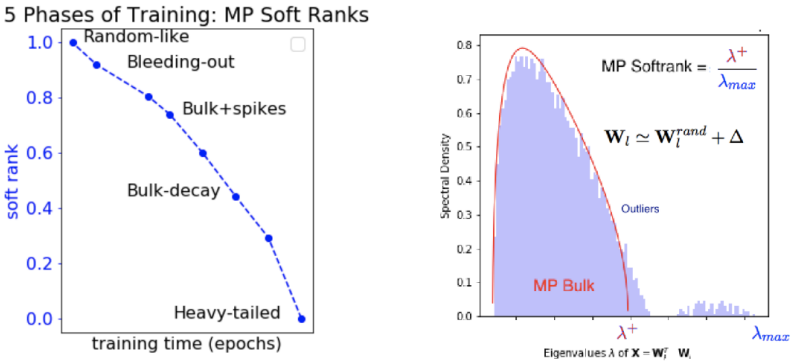
[4] find that for deep computer vision neural networks, the stable rank shrinks with training, even as the hard rank remains full.



(a) RANDOM-LIKE. (b) BLEEDING-OUT. (c) BULK+SPIKES.



(d) BULK-DECAY. (e) HEAVY-TAILED. (f) RANK-COLLAPSE.



(a) Depiction of how decreasing MP Soft Rank corresponds to different phases of training.

(b) Depiction of each how each phase is related to the MP Soft Rank.

Figure 2: Figure 14 and 15 from [4] depicting spectral densities of a given latex

2.5 5+1 Phases of Training

Throughout training, [4] find that the computer vision models studied undergo five phases during training that correspond to different values of the capacity metrics introduced in the previous section. These phases are also clearly visible in the empirical spectral density (see figure 2).

First (in the random-like phase), the model's weights are well-modeled by a random matrix (as we would expect from typical weight initialization schemes). During the next two phases (bleeding out and bulk+spikes), a small number of eigenvalues leaks out of the bulk towards higher values. This is well modeled by a perturbative correction to MP theory (the spiked covariance model) [4]. Next, the larger spikes become heavy-tailed, and, finally, the edge disappears, with the entire distribution becoming heavy tailed.

2.6 Self-Regularization

As the authors of [4] point out, these phases have clear interpretations in terms of "self-regularization."

The authors model a weight matrix $\mathbf{W}$ as the sum or random component and signal-bearing component,

$$\mathbf{W} = \mathbf{W}_{\text{rand}} + \mathbf{\Delta}_{\mathbf{sig}} \tag{9}$$

The random component $\mathbf{W}_{\text{rand}}$ is approximated by a MP-random matrix, while the signal-bearing component $\mathbf{\Delta}_{\text{sig}}$ is strongly correlated.

During training, for larger models, $\mathbf{W}_{\text{rand}}$ decreases steadily, while $\mathbf{\Delta}_{\text{sig}}$ increases. Its strong correlations mean it is well-modeled by a heavy tail random matrix.

During the first few phases, most of the eigenvalues of the weight matrix remain uninformative and stuck in the bulk. Only a few, signal-bearing eigenvalues bleed out past the bulk and are able to carry information.

Later, the edge vanishes, and there is no more clean separation between signal and noise. This mixing between noise and signal acts as a kind of implicit regularization ("self-regularization"). [4] go on to make (and verify) predictions of how self-regularization varies with hyperparameters like batch size.

3 Results

Our first result is simply showing the theory developed in [4] largely extends into transformers, although we notice a key difference, so called "quasi-null-eigenvalues", showing that the larger model is failing to make full use of its potential, leaving around 1-6% of its eigenvalues (and hence expressive power) unused. In subsection 3.2 we show that the spectra of attention vs MLP (multi-layer perception) are quantitatively different, achieving different σ parameters despite being trained concurrently, this result has implications for the scaling hypothesis. In subsection 3.3 we unsuccessfully induce rank collapse. Finally, we apply the techniques to larger models, with more time this could be developed into more information about the scaling hypothesis.

3.1 Extension of results into transformer models

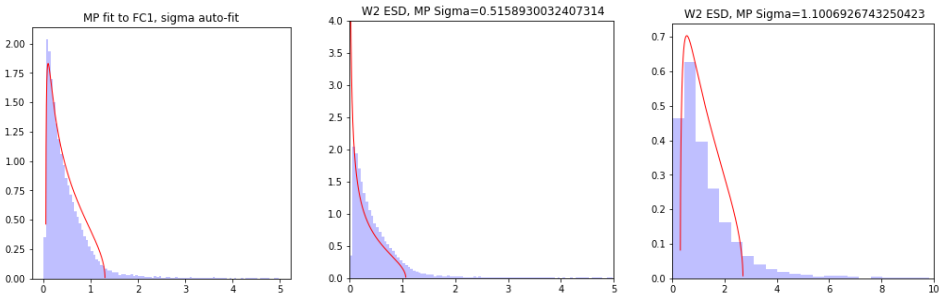
First we verify that the key results of [4] transfer into this new regime by verifying that the spectra is still modeled well. We then go on to find novel "quasi-null-eigenvalues" (QNE) that appear in the transformer spectra.

By running the techniques described in section 2 we are able to reproduce the plots, now for transformer models. The spectra plots of publicly-available GPT-2 small are placed alongside AlexNet spectra in figure 3. The characteristic heavy-tail bulk-curve is clearly visible with no clean separation between bulk and signal. This serves as strong evidence that the results of [4] continue to apply in this regime.

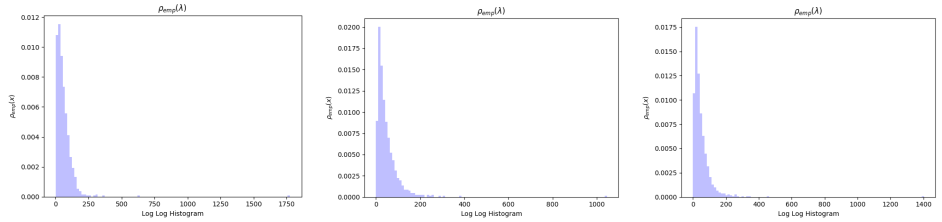
Deviating from the results seen in [4] or [5] we find that between 1 and 6% eigenvalues have values significantly below the bulk. The action of these eigenvalues is then much much smaller, so small that they have no measurable effect on the function being implemented. This is directly counter to the results of [4] which find, for medium-sized models, a handful of eigenvalues far outside the bulk. We name these small eigenvalues "quasi-null-eigenvalues" as they essentially map into the null space. This represent unused expressivity in the model. Future work needs to be done to determine if the number of these QNE's decreases through training or increase.

3.2 Attention vs MLP heads

We demonstrate in figure 4 that the different components of our model (Attention layers vs MLP layers), take on different spectra. We plot a variety of different layers, noteably the MLP layers spectra bulk are always tighter than the attention layers. This implies that MLP layers have a heavier tail than the attention layers. This could suggest that further training of the attention layers but not the MLP layers would result in higher performance. Further development of this work would be able to further pry apart the behaviour of either components scaling performance, a step which is crucial to correctly estimating the closeness to AGI and thus the threat level.



(a) AlexNet FC 1 (b) AlexNet FC 2 (c) AlexNet FC 3



(d) GPT-2 MLP 1 (e) GPT-2 MLP 6 (f) GPT-2 MLP 12

Figure 3: Comparison of Alex Net's fully connected layers and GPT-2 small's MLP layers. Top three plots borrowed from [4]. Our plots are left in log space to more clearly demonstrate the QNE

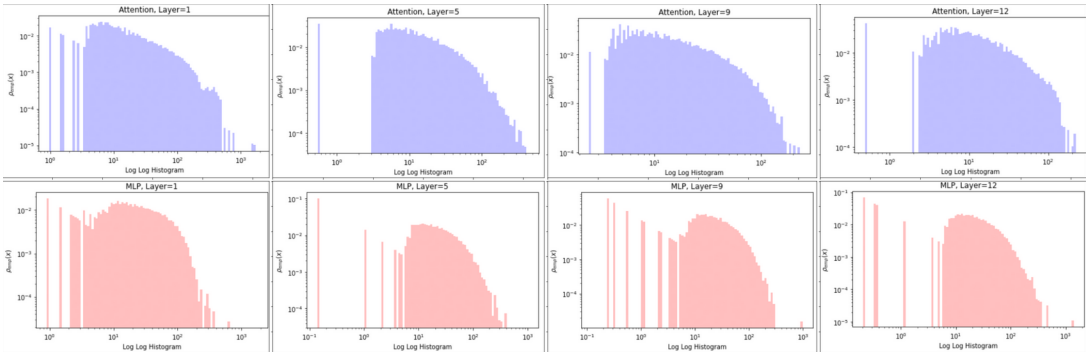


Figure 4: A comparison of the spectral properties between the attention-layer weights and MLP weights.

3.3 Null result: Induced Rank Collapse

We compared the weight spectra of a normal pre-trained 12-layer GPT-2 model to the weight spectra of the same model after a 4 epochs of fine tuning with intense regularisation. Disproving our initial hypothesis that fine-tuning may bring the model's weights into the "rank-collapse" phase conjectured in [4], we found that overregularisation made an almost negligible difference to the properties of the spectra. Results from a selection of layers for the attention layers (Figure 5) and MLP layers (Figure 6) are presented, there is no meaningful difference.

The fine-tuning in this section specifically tunes the model to always suggest "KILL" as the next token. This result then also suggests that qualitative changes in the outputs of the model can result from barely noticeable changes in the weights. It also suggests that important changes (importantly e.g. a treacherous turn) in the properties of the models resulting from fine-tuning may not be easily detectable through our techniques.

3.4 Differences in Spectra of Multiple Sized Models

We demonstrate the differences in the weight spectra between the attention and MLP layers of GPT-2 small, 12-layer transformer model and GPT-2 large, a 36-layer transformer model (Figure 7) and

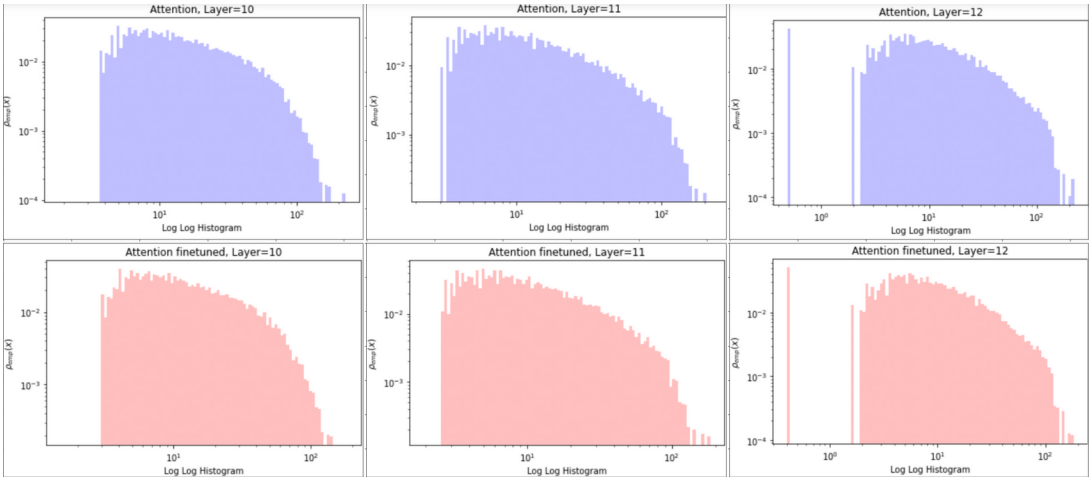


Figure 5: A comparison of the spectral properties between the plain (top) and fine-tuned (bottom) attention-layer weights.

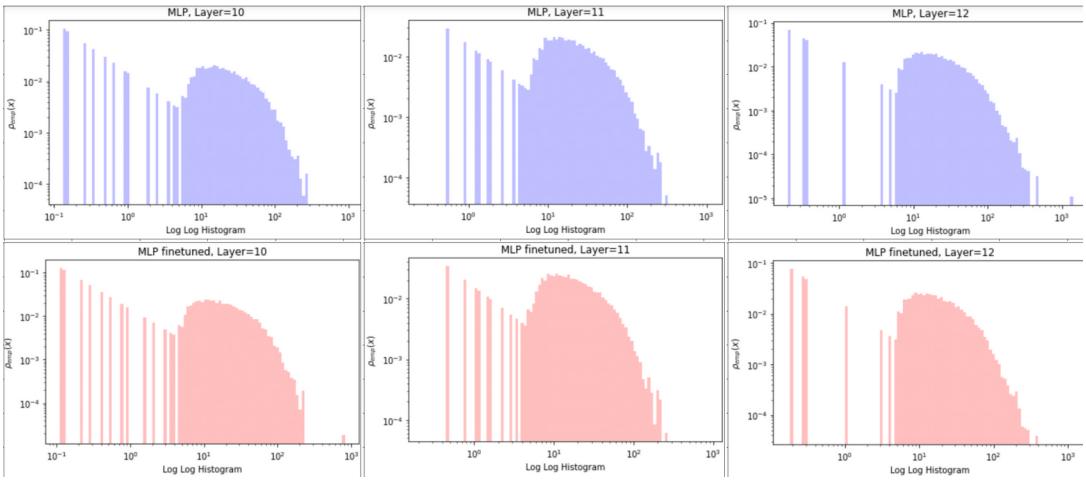


Figure 6: A comparison of the spectral properties between the plain (top) and fine-tuned (bottom) MLP weights.

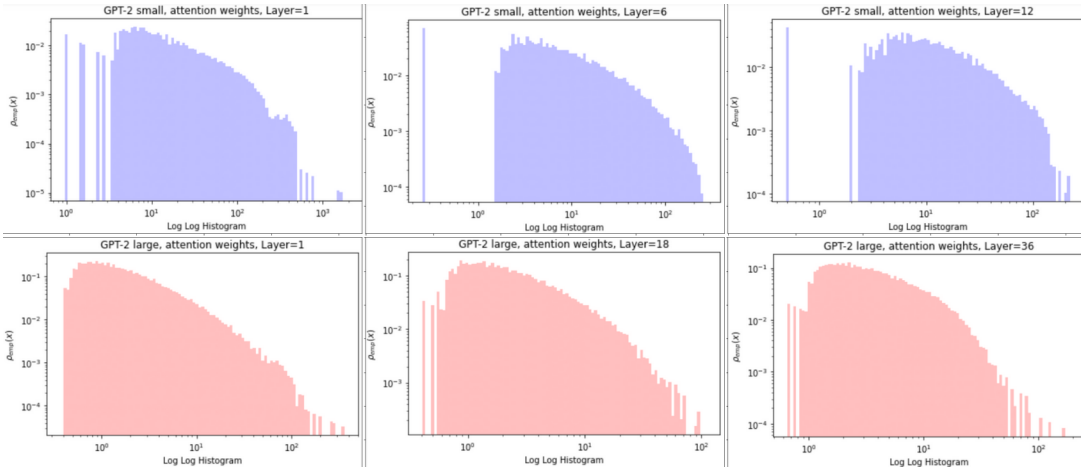


Figure 7: A comparison of the spectral properties between the attention layers of a 12-layer GPT-2 model (top) and a 36-layer GPT-2 model (bottom).

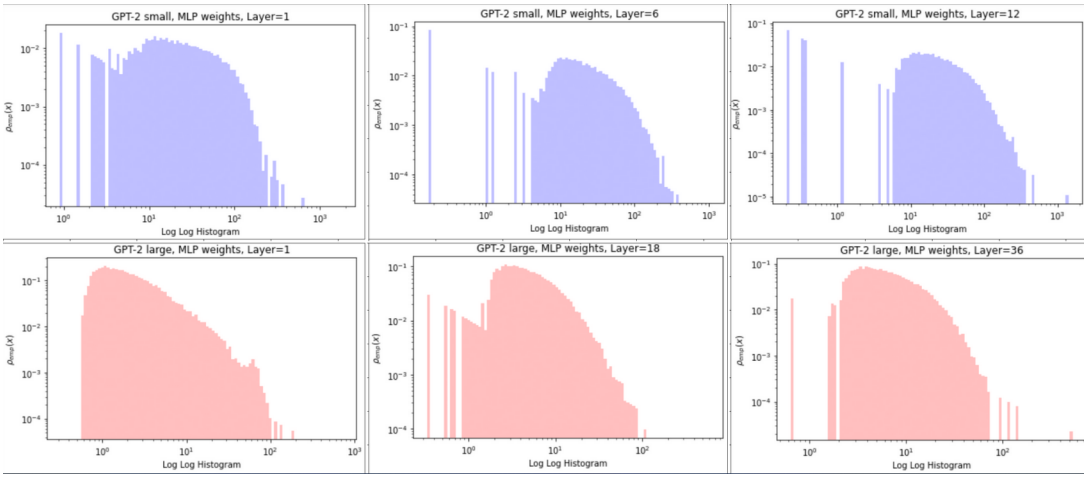


Figure 8: A comparison of the spectral properties between the MLP layers of a 12-layer GPT-2 model (top) and a 36-layer GPT-2 model (bottom).

MLP layers (Figure 8). We found the larger model to have more stable weight spectra across layers and somewhat smaller eigenvalues.

An interesting difference between the spectral properties of these large language models and previously studied computer vision models [4] is the presence of spikes in the low parameter regime (see Figure 8). These are eigenvalues very close to zero. They span an effective kernel, which suggests the model is over-parameterized (we can safely eliminate these nodes without harming performance).

4 Conclusion

The theory of heavy-tail self-regularization has striking implications for understanding deep neural networks and their training. To recap: deeper models are able to represent stronger correlations between weights in a given layer, this smooths out the bulk of the eigenspectrum which, in mixing signal and noise, acts as a kind of regularization. Just as trained neural networks have "inductive biases", self-regularization represents a kind of "learning bias" for neural networks during training.

Understanding these learning biases is key to answering highly alignment relevant questions such as how path-dependent training is [2] and the relation between model size and performance (esp. the scaling hypothesis). Our aim is not simply to understand the workings of trained models but to understand how training and scaling change models.

To this end, we provided evidence towards the universality of heavy-tail matrix theory across different architectures. Though heavy-tailed spectra are omnipresent, we find stark, consistent differences between different kinds of transformer layers. Moreover, we find the existence of quasi-null eigenvalues in transformer weights without analogues in CV (computer vision) models, making this another novel result. We posit that these are related to the model not using the full expressivity available.

More broadly, methods like these suggest the potential of a more "top-down" approach to interpreting neural networks. Integrating "top-down" approaches with existing (e.g. circuits-style) "bottom-up" approaches moves us closer towards a goal of full model interpretability.

Reproducible and Open Source Code

All code has been provided on the open GitHub available at: `https://github.com/jqhoogland/implicit-self-regularization`

An interactive notebook to play around with the spectra generated by each layer of the publicly available model is available here: `https://colab.research.google.com/github/jqhoogland/implicit-self-regularization/blob/master/rmt.ipynb`

The weights of fine-tuned and other models are provided at the following Google Drive link: https://drive.google.com/drive/folders/13TgN31kX82UqX21j3TC9PwUbJSNLIk7U?usp=share_link

Declaration of Time-Spent

The authors declare that this work has been completed within the time requirements of the hackathon.

References

[1] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[2] Vivek Hebbar and Evan Hubinger. Path dependence in ML inductive biases - AI Alignment Forum, 2022.

[3] Evan Hubinger. Chris Olah's views on AGI safety - AI Alignment Forum, 2019.

[4] Charles H. Martin and Michael W. Mahoney. Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning. *arXiv:1810.01075 [cs, stat]*, October 2018. arXiv: 1810.01075.

[5] Charles H. Martin, Tongsu (Serena) Peng, and Michael W. Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1):4122, July 2021. Number: 1 Publisher: Nature Publishing Group.

[6] Evan Murphy. Interpretability's alignment-solving potential: Analysis of 7 scenarios, 2022.

[7] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020.

[8] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017.

[9] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

[10] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and others. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[11] Terence Tao and Van Vu. Random Matrices: The circular Law. *arXiv:0708.2895 [math]*, February 2008. arXiv: 0708.2895.

[12] Eugene P. Wigner. Characteristic Vectors of Bordered Matrices With Infinite Dimensions. *Annals of Mathematics*, 62(3):548–564, 1955. Publisher: Annals of Mathematics.