EAI Math Reading Group

Random Matrix Theory 3

Eigenvalue spacing and Applications to Neural Networks

12/11/2023

Outline

1. Chapter 5: Joint distribution and eigenvalue spacing
2. Application 1: “Appearance of RMT in deep learning“
3. Application 2: “Traditional and Heavy-Tailed Self-Regularization in Neural Nets“

Recap: Wishart matrices

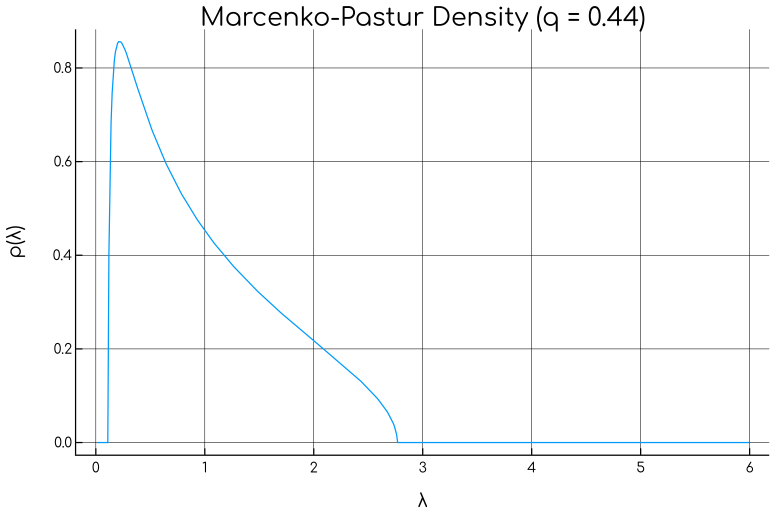
Let $W = \frac{1}{T} H H' \in \mathbb{R}^{N \times N}$, where $H \in \mathbb{R}^{N \times T}$, $H_{ij} \sim N(0,1)$

“Rank” parameter $q = \frac{N}{T} < 1$

Eigenvalue density

$$\rho(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi q \lambda}$$

$$\lambda_\pm = (1 \pm \sqrt{q})^2$$



Joint Eigenvalue distribution

Consider the general class of *rotationally invariant* random matrices

$$P(\boldsymbol{M}) = Z_N^{-1} \exp\left(-\frac{N}{2} \operatorname{tr} V(\boldsymbol{M})\right),$$

where V is called the *potential function*

- Wigner Ensemble: $V(x) = \frac{x^2}{2\sigma^2}$
- Wishart Ensemble: $V(x) = \frac{x+(q-1)\log x}{q}$

Given the eigendecomposition $\boldsymbol{M} = O\Lambda O'$, we compute the joint distribution of the eigenvalues λ_i

Joint Eigenvalue distribution

We need to compute the distribution, we need the jacobian of $M \mapsto (\Lambda, O)$, which introduces a factor $|\det(\Delta)|$, where $\Delta(\boldsymbol{M}) = \left[\frac{\partial \boldsymbol{M}}{\partial \Lambda}, \frac{\partial \boldsymbol{M}}{\partial O}\right]$

tldr:

$$|\det(\Delta)| = \prod_{k<\ell} |\lambda_\ell - \lambda_k|$$

and the joint eigenvalue distribution is given by

$$P(\{\lambda_i\}) \propto \prod_{k<\ell} |\lambda_\ell - \lambda_k| \exp\left(-\frac{N}{2} \sum_{i=1}^{N} V(\lambda_i)\right)$$

Eigenvalue spacing (abridged version)

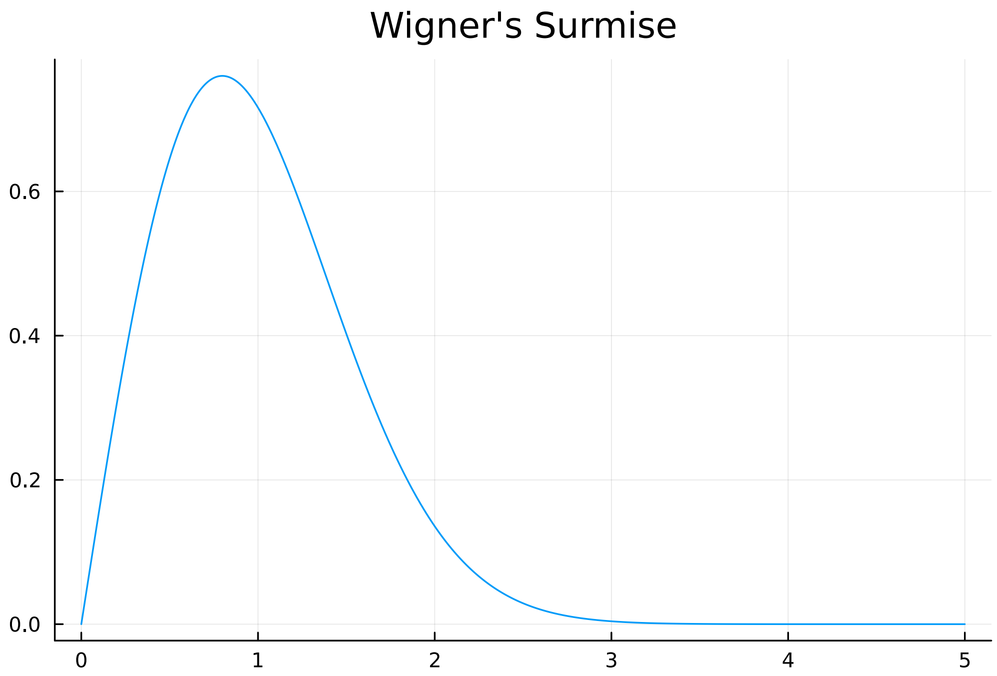
From a Statistical Mechanics point of view, eigenvalues can be interpreted as particles trying to minimize the potential, while under repulsive interations

Likelihood: $P(\{\lambda_i\}) \propto e^{\frac{1}{2}\beta N \mathcal{L}(\{\lambda_i\})}$,

$$\mathcal{L}(\{\lambda_i\}) = -\sum_{i=1}^{N} V(\lambda_i) + \frac{1}{N}\sum_{i\neq j} \log|\ \lambda_i - \lambda_j\ |$$

Wigner's surmise

$$P(|\lambda_i - \lambda_{i-1}| = s) = \frac{\pi}{2} s \exp\left(-\frac{\pi}{4} s^2\right)$$



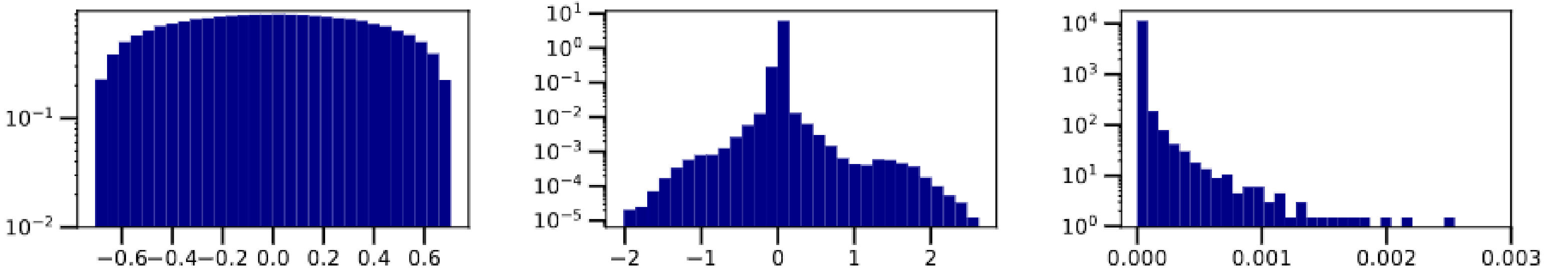
Application 1:

Appearances of Random Matrix theory in Deep Learning

https://arxiv.org/abs/2102.06740

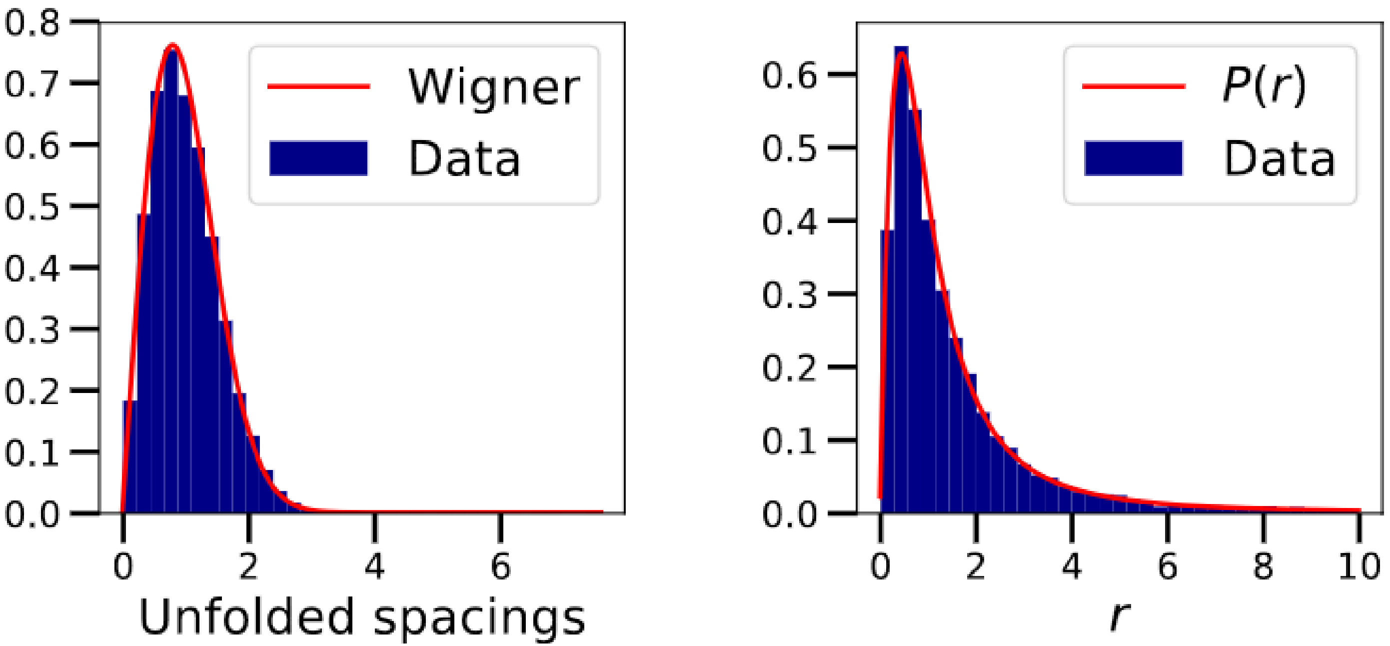
Summary

- Looked the spectral statistics of the *Loss Hessian*
- Spectral Densities do *not* match classical ensembles
- *Distances* between nearest neighbors do[1]
- But so do a lot of matrix ensembles...



(a) Wigner semicircle (b) MLP (c) Logistic Regression

[1]For small networks



(a) Unfolded spacings. Batch-size 64.

(b) Spacing ratios. Entire dataset.

Figure 3: Spacing distributions for the Hessian of a logistic regression trained Resnet-34 embeddings of CIFAR10. Hessians computed over the test set.

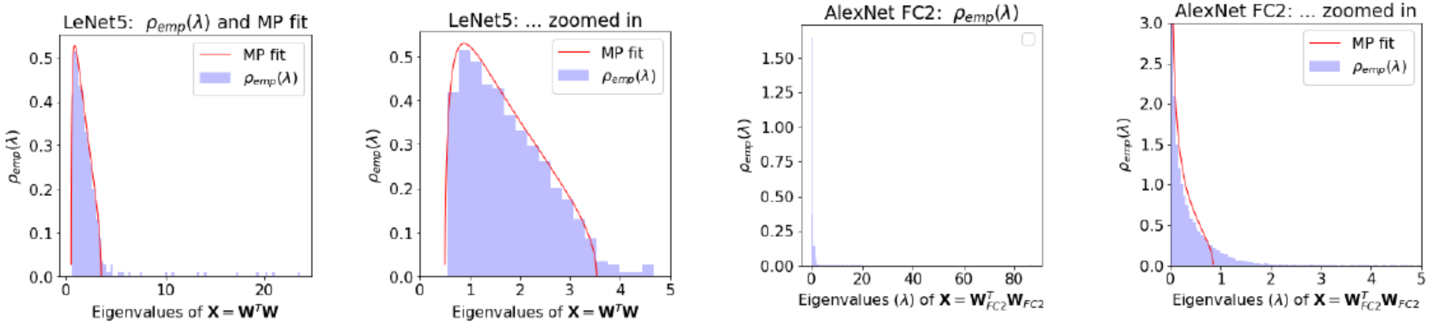
Application 2:

Traditional and Heavy-Tailed Self Regularization in Neural Network Models

https://arxiv.org/abs/1901.08276
https://arxiv.org/abs/1810.01075

Summary

- Looked at singular values of Neural Net weight matrices
- Identify 5 phases of eigenvalue densities
- Older models behave as Marcenko-Pastur
- Newer models have Power law distributions



(a) LeNet5, full (b) LeNet5, zoomed-in (c) AlexNet, full (d) AlexNet, zoomed-in

Strongly correlated matrices

Weight matrix sampled from a Power-Law

$$W_{ij} \sim \frac{1}{x^{1+\mu}},\ \mu > 0$$

- $(4 < \mu)$ $\rho_{N(\lambda)}$ asymptotically MP
- $(2 < \mu < 4)$ $\rho_{N(\lambda)} \sim \lambda^{-(a\mu+b)} \to \lambda^{-1-\frac{\mu}{2}}$
- $(0 < \mu < 2)$ $\rho_{N(\lambda)} \to \lambda^{-1-\frac{\mu}{2}}$, with smaller finite size effects

Strongly correlated matrices

Weight matrix sampled from a Power-Law

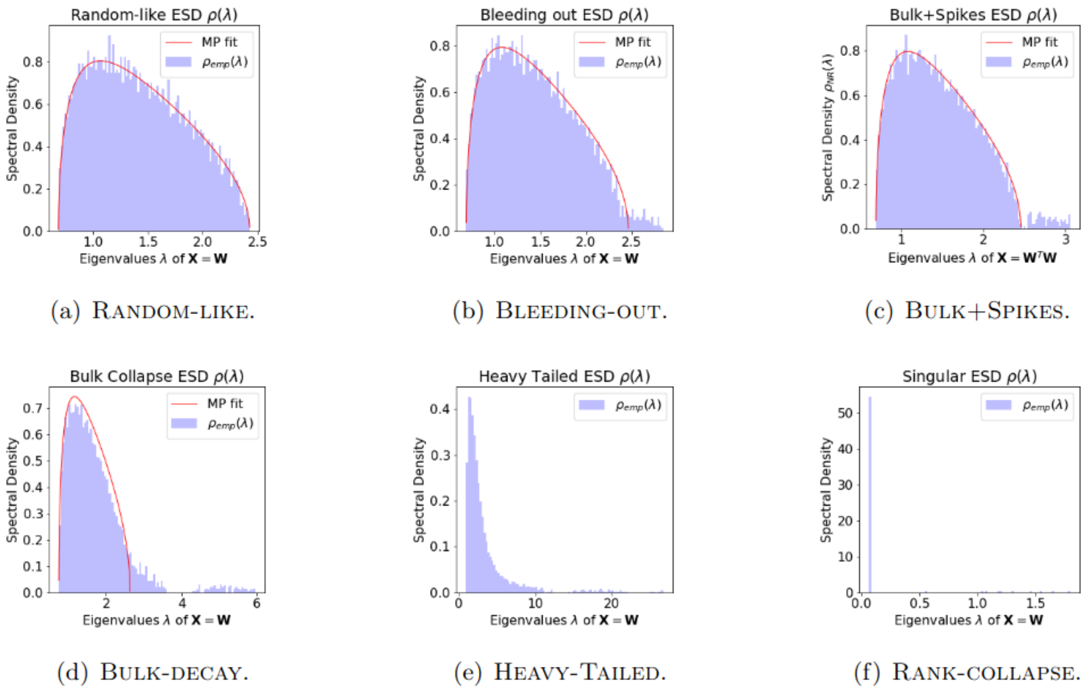
$$W_{ij} \sim \frac{1}{x^{1+\mu}},\ \mu > 0$$

- $(4 < \mu)$ $\rho_{N(\lambda)}$ asymptotically MP
- $(2 < \mu < 4)$ $\rho_{N(\lambda)} \sim \lambda^{-(a\mu+b)} \rightarrow \lambda^{-1-\frac{\mu}{2}}$
- $(0 < \mu < 2)$ $\rho_{N(\lambda)} \rightarrow \lambda^{-1-\frac{\mu}{2}}$, with smaller finite size effects

NB. In physics, power-laws indicate emergence of non-random/fractal structure, long range correlations

Phases of Self-Regularization

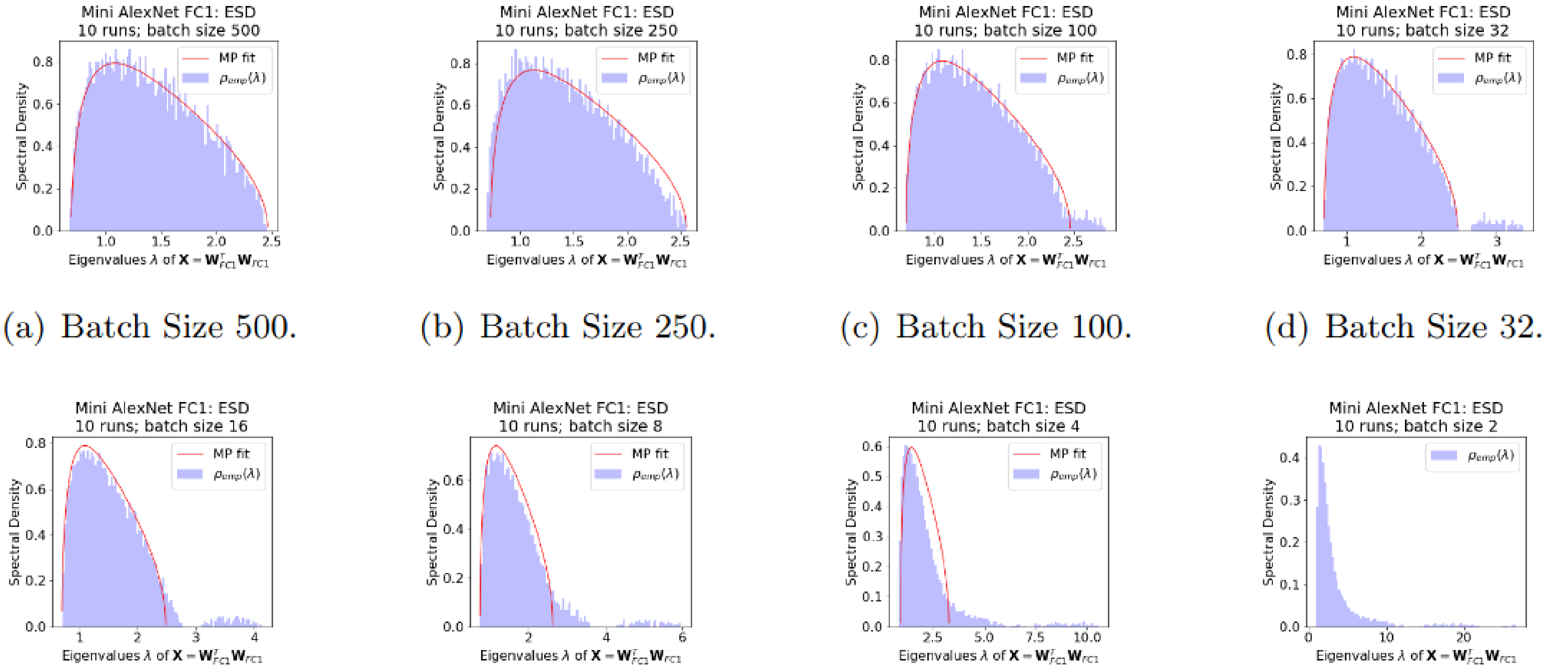
Assuming $W = W^{\text{rand}} + \Delta^{\text{sig}}$ (noise + signal)



(a) Random-like. (b) Bleeding-out. (c) Bulk+Spikes.

(d) Bulk-decay. (e) Heavy-Tailed. (f) Rank-collapse.

Explaining the generalization gap



(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.