Towards the application of random matrix theory to neural networks

Matteo Favoni, Gert Aarts, Biagio Lucini and Chanju Park

Swansea University Prifysgol Abertawe

Learning as Dyson Brownian motion

The training dynamics of weight matrices in learning algorithms can be understood as a Dyson Brownian motion, hence featuring characteristics of random matrix theory [1].

- $\mathbf{W} \in \mathbb{R}^{M\times N}$ weight matrix in a neural network
- The matrix update rule can be written as

$$\mathbf{W}' = \mathbf{W} - \frac{\alpha}{|\mathcal{B}|}\sum_{b=1}^{|\mathcal{B}|}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{W}}\right)_b + \frac{\alpha}{\sqrt{|\mathcal{B}|}}\sqrt{\text{Var}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{W}}\right)}\boldsymbol{\eta},$$

  where the second term on the rhs is the deterministic part and the third reflects stochasticity, α is the learning rate and $\mathcal{B}$ the batch.
- It is possible to study the symmetric matrix $\mathbf{X} = \mathbf{W}^T\mathbf{W}$. From the update rule for $\mathbf{W}$, it follows this dynamics for the eigenvalues of $\mathbf{X}$:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \underbrace{\alpha K_i}_{\text{Drift term}} + \underbrace{\frac{\alpha^2}{|\mathcal{B}|}\sum_{j\neq i}\frac{g_i^2}{x_i - x_j}}_{\text{Coulomb repulsion}} + \underbrace{\frac{\alpha}{\sqrt{|\mathcal{B}|}}\sqrt{2}g_i\eta_i}_{\text{Diffusion term}}$$

- Stationary distribution: Coulomb gas (derived from Fokker-Planck equation)

$$P_s(x_i) = \frac{1}{\mathcal{Z}}\prod_{i<j}|x_i - x_j|\mathrm{e}^{-\sum_i V_i(x_i)/g_i^2}$$

  with $\mathcal{Z} = \int\prod_i \mathrm{d}x_i P_s(x_i)$ and $K_i = -\frac{\mathrm{d}V_i(x_i)}{\mathrm{d}x_i}$

Wigner's surmise and Wigner's semicircle

We consider the case $N = 2$ and assume that the potential can be written as $V(x_1, x_2) = \frac{x_1^2}{2\sigma_1^2} + \frac{x_2^2}{2\sigma_2^2}$, where x_1 and x_2 are centered around the degenerate eigenvalue κ.

- Partition function $\mathcal{Z} = \frac{1}{N_0}\int \mathrm{d}x_1\mathrm{d}x_2\,|x_1 - x_2|\mathrm{e}^{-\frac{x_1^2}{2\sigma_1^2}-\frac{x_2^2}{2\sigma_2^2}}$
- Transformation: $S = x_1 - x_2, x = \alpha x_1 + \beta x_2$
- α and β are such that the exponent can be written as $AS^2 + Bx^2$
- Probability of separation $P(S) = \frac{1}{\sigma_1^2+\sigma_2^2}S\,\mathrm{e}^{-\frac{S^2}{2(\sigma_1^2+\sigma_2^2)}}$
- We can introduce $s = \frac{S}{\langle S\rangle}$ such that $\langle s\rangle = 1$
- $P(S)\,\mathrm{d}S = P(s)\,\mathrm{d}s \implies P(s) = \frac{\pi}{2}s\mathrm{e}^{-\frac{\pi s^2}{4}}$ Wigner's surmise
- Spectral density $\rho(x) = \left\langle \frac{1}{N}\sum_{i=1}^N \delta(x - x_i)\right\rangle$, x_i eigenvalues

$$\rho(x) = \frac{\mathrm{e}^{-\frac{x^2(\sigma_1^2+\sigma_2^2)}{2\sigma_1^2\sigma_2^2}}}{4\sqrt{2\pi}\sigma_1\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}}\sum_{i=1,2}\left[2\sigma_i^2 + \mathrm{e}^{-\frac{x^2}{2\sigma_i^2}}\sqrt{2\pi}x\sigma_i\text{Erf}\left(\frac{x}{\sqrt{2}\sigma_i}\right)\right]$$

  In case $\sigma_1 = \sigma_2$, the function is called Wigner's semicircle

Teacher-Student model

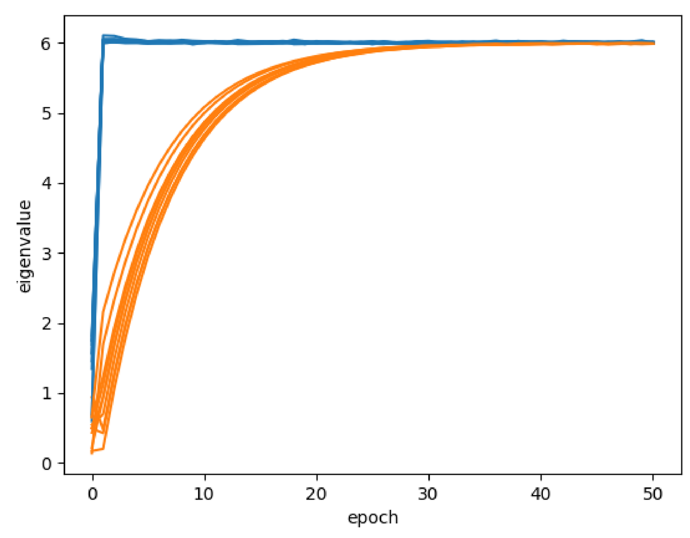
- Dataset $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}$, $i = 1, \dots, N_{\text{samples}}$
- Teacher: $\mathbf{y}_i = \mathbf{Z}\mathbf{W}\mathbf{x}_i$
- Student: $\mathbf{y}_{\text{pred},i} = \mathbf{Z}\mathbf{W}_{\text{pred}}\mathbf{x}_i$
- $\mathbf{x}_i \in \mathbb{R}^N \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, $\mathbf{W}, \mathbf{W}_{\text{pred}} \in \mathbb{R}^{N\times N}$
- $\mathbf{Z} \in \mathbb{R}^{N\times N}$ fixed matrix for both the teacher and the student
- $\mathbf{W}_{\text{pred}}$ optimized by minimizing $\mathcal{L} = \frac{1}{2N_{\text{samples}}}\sum_{i=1}^{N_{\text{samples}}}(\mathbf{y}_i - \mathbf{y}_{\text{pred}})^2$
- We can use the singular value decomposition (SVD) $\mathbf{W}_{\text{pred}} = \mathbf{U}\boldsymbol{\Psi}\mathbf{V}^T$ to write the eigenvalue dynamics

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = -\alpha(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})_{ii}\,x_i + C(t),$$

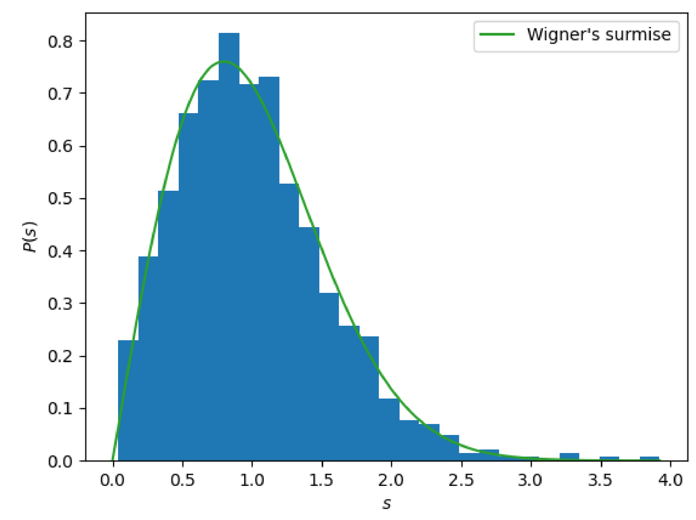
  with $\tilde{\mathbf{Z}} = \mathbf{V}^T\mathbf{Z}$ and $C(t) = \alpha(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}}\mathbf{V}^T\mathbf{X}\mathbf{V})_{ii}$

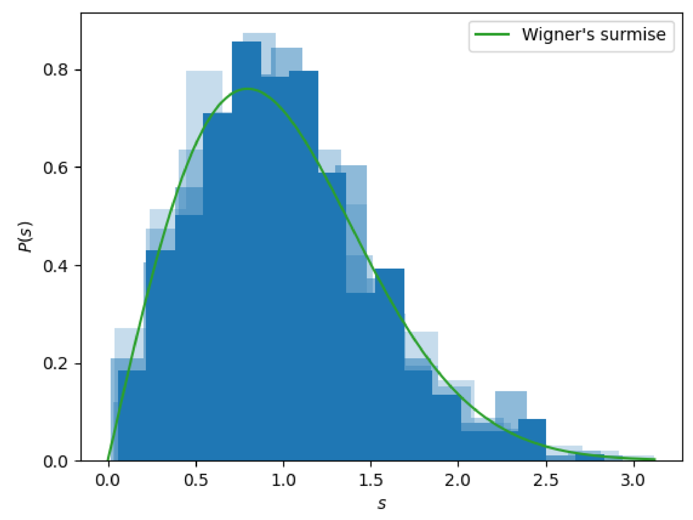
Results

- The role of the hidden layer is to regulate the speed of the eigenvalues

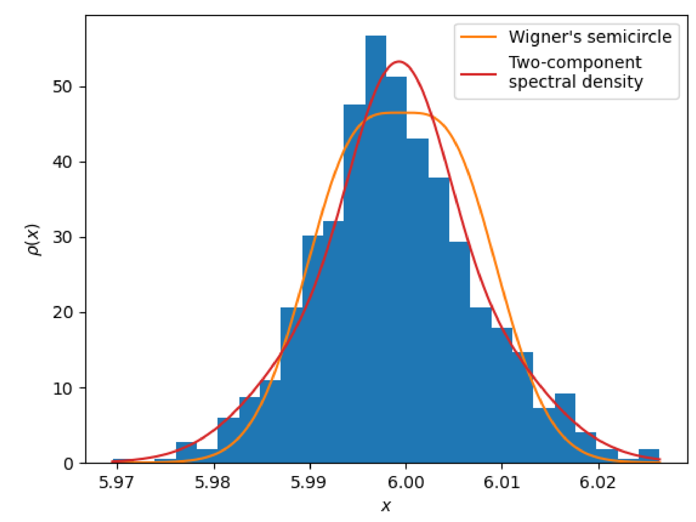


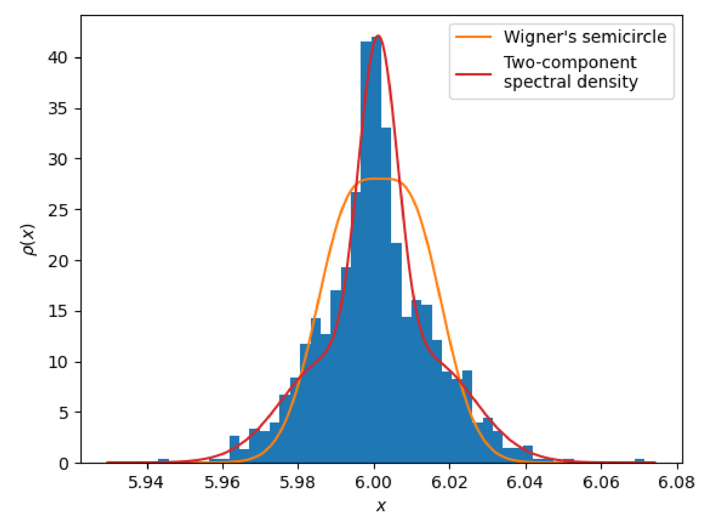
- The Wigner's surmise is found for $N = 2$ and its universality is checked for $N = 10$ with 4 doubly-degenerate eigenvalues



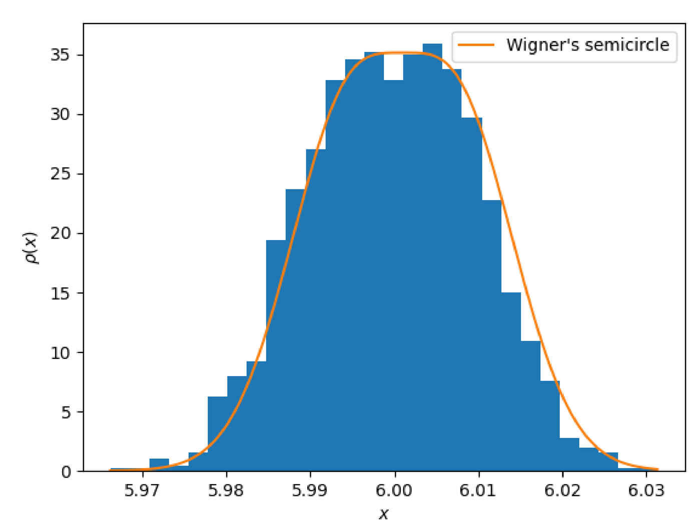


- The two-component spectral density is a better fit than the Wigner's semicircle in presence of the hidden layer





- When $\mathbf{Z} = \mathbb{1}$, i.e. there is no hidden layer, the Wigner's semicircle is found



Conclusions and outlook

Conclusions

- In the TS model that we examined, the hidden layer regulates the speed at which the eigenvalues are moving
- This leads to a generalized form of the Wigner's semicircle for the spectral density, while keeping intact the Wigner's surmise

Next steps

- Collect larger statistics to show the linear scaling rule
- Include the effect of activation functions
- Study the infinite-width limit
- Let the hidden layer be learnable and add multiple layers

References

[1] "Stochastic weight matrix dynamics during learning and Dyson Brownian motion" G. Aarts, B. Lucini, C. Park, [arXiv:2407.16427]