Towards Applications of Free Random Variables in Cognitive Science

Maciej A. Nowak

Mark Kac Center for Complex Systems Research,
and Institute of Theoretical Physics
Jagiellonian University

[supported by the TEAMNET POIR.04.04.00- 00-14DE/18-00 grant of the Foundation for Polish Science and by the Priority Research Area DigiWorld under the program Excellence Initiative – Research University at the Jagiellonian University in Kraków.]

Probablistic Operator Algebra Seminar, Berkeley, December 11th, 2023

Outline

- Dataism
- Example 1: Inference of signals from noise
- Example 2: Modelling of the real neuronal networks
- Example 3: Taming Deep Networks (Machine Learning)
- Conclusions

Dataism

Yuval Noah Harari, in his book *Homo Deus: A Brief History of Tomorrow* [2015], calls an emerging ideology or even **a new form of religion**, in which "information flow" is the "supreme value": "Dataism declares that the universe consists of data flows, and the value of any phenomenon or entity is determined by its contribution to data processing"

Dataism in Computational Neuroscience

1. Contemporary real complex systems gather Big Data: dEEG, fMRI, MEG, optogenetics...
2. Data collected at wide spectrum of temporal and/or spatial resolution
3. ...Number of voxels in a single fMRI snapshot - 10^5, time series length for dEEG recordings "arbitrarily" large: 10^3 signals per second
4. Need for redefining "random variable" – XIX century concept versus XXI century calls
5. Random matrix theory - probability theory where the random variable takes values in the space of matrices.
   Surprising simplification when dimension of the matrix tends to infinity - free random variables [Voiculescu]
6. Practical asymptotics: $8 \equiv \infty$

Example 1: Inference: "Free Poisson" - Wishart's distribution and more

- Let us perform sequential measurements of a vector $\tilde{X}_i$ where $(i = 1, ..., N)$ at the series of times $t = 1, ...M$.
- Each measurement is represented by $\tilde{X}_{it}$ (say, a signal from the i-th electrode).
- Standarize measurements (by subtracting the mean and dividing by the variance for each i).
- The correlation matrix $C_{ij} = \frac{1}{M} \sum_{t=1}^{t=M} X_{it} X_{jt}$ is denoted as $C = \frac{1}{M} X X^{\dagger}$.

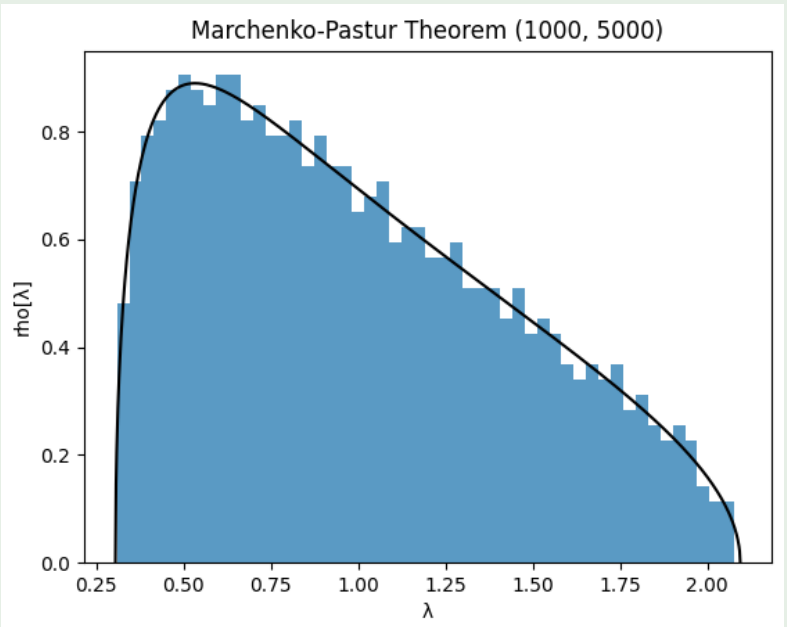
Wishart ensemble

If X_{ij} are i.i.d. Gaussian entries, such an ensemble is called (real or complex) Wishart ensemble, and it represents the benchmark of pure noise (correlation matrix is a unit matrix $\mathbf{1}_N$).

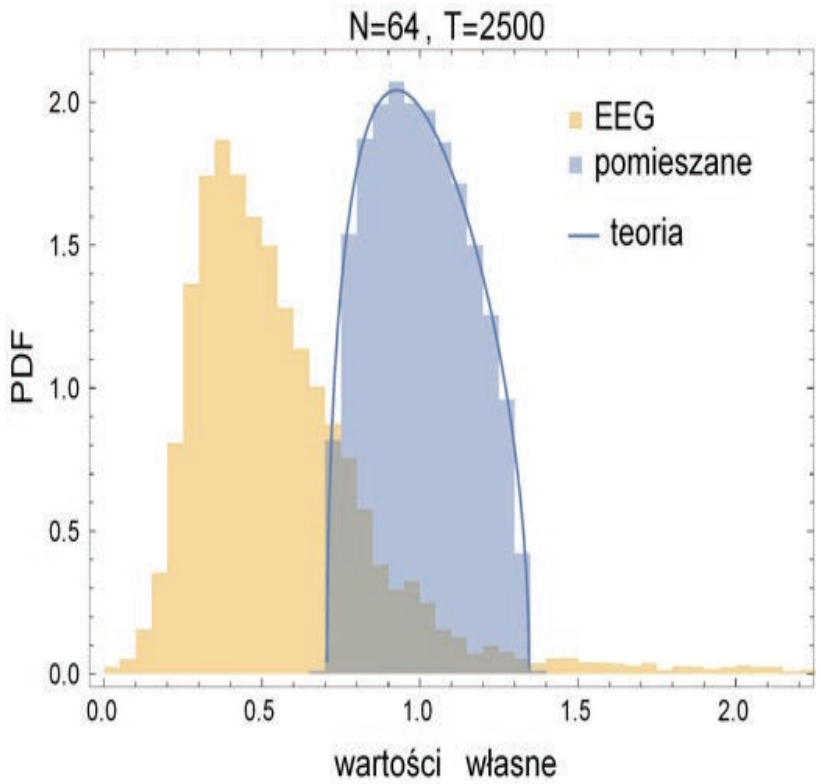
Spectral distribution of Wishart ensemble

- R-transform for Wishart: $R(z) = \frac{1}{1-rz}$, where $r = N/M$ is fixed, whereas N, M tend to infinity.
- Spectral function: $\rho(\lambda) = \frac{1}{\pi r \lambda}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}$ where $\lambda_\pm = (1 \pm \sqrt{r})^2$ (Marcenko-Pastur distribution).
- For $r \to 0$, spectral function tends to Dirac's delta (pure noise).

$N = 1000, M = 5000 \to r = 0.2$



Marcenko-Pastur distribution - dAEEG experiment



Spectral histograms: orange - experimental data, blue - reshuffled data, line - theory

Beyond pure noise, summary 2

- In general, correlations are more complicated, e.g. $< X_{it} X_{js} >= A_{ij} B_{ts}$ (spatial-temporal correlations).
- Then, true measure is proportional to $e^{-\frac{1}{2} A^{-1} X B^{-1} X^{\dagger}}$.
- Power of FRV: we change variables $\sqrt{A^{-1}} X \sqrt{B^{-1}} \equiv Y$, Then Green's function is calculated with respect to Wishart measure $exp - \frac{1}{2} Y Y^{\dagger}$, but at the cost of generating moments M_k of the type $< \mathrm{tr}[AYBY^{\dagger}]^k >$.
- Using S-transform, we "factorize" the spectrum of A from $YBY^{\dagger}$. Then, using the cyclic property of the trace we factorize B from pure Wishart. Assuming unknown structure of correlators we minimize the error and optimize the predictions for the true correlations A and B from the measured moments.
- Method works for non-Gaussian distributions (e.g. heavy tails) and other estimators than Pearson's.
- For more technical details of such FRV calculus, see e.g. https://arxiv.org/abs/physics/0603024

Example 2: Neural network model

Recurrent neural network (infinite depth limit):

$$\frac{dx_i}{dt} = -x_i + \sum_{j=1}^{N} W_{ij}\phi(x_j)$$

- W_{ij} - a synaptic connectivity matrix
- ϕ a nonlinear activation function
- W a random nonhermitian (Ginibre) matrix

- W elements undergo a diffusion:

$$W_{jk} = w_{jk} + iv_{jk} =$$

$$\frac{1}{\sqrt{2N}} dB^w_{jk} + \frac{i}{\sqrt{2N}} dB^v_{jk}$$

- - transition between the stationary and chaotic dynamics occurs when the largest real part of an eigenvalue exceeds 1.

H. Sompolinsky, A. Chrisanti, H.J. Sommers, *Chaos in Random Neural Networks*, PRL **61** 259 (1988)

Rajan-Abbott Model RAM

- Two types of N neurons: excitatory (E) and inhibitory (I); fractions $f_{I,E}$.
- The synaptic strength of neurons has normal distribution $\mathcal{N}(\mu_{E,I}, \frac{\sigma^2_{E,I}}{N})$ with $\mu_E > 0$ and $\mu_I < 0$
- The synaptic matrix decomposed $W = M + G\Lambda$, where M deterministic, G a Ginibre matrix and Λ diagonal with $\sigma_{I,E}$
- The random part models variability in populations of neurons

RAM model introduces a balance condition: the sum of excitations and inhibitions incoming to neuron is balanced to 0, both on average i.e. $\sum_j M_{ij} = 0$ and at each neuron $\sum_j W_{ij} = 0$.

K. Rajan, L.F. Abbott, *Eigenvalue Spectra of Random Matrices for Neural Networks* PRL **97**, 188104 (2006)

Generic linearization around the fixed point

$$\dot{x_i}(t) = \sum A_{ij} x_j(t) + \xi_i(t)$$

Multivariate Ornstein-Uhlenbeck dynamics with friction A and fluctuations $< \xi_i(t)\xi_j(t^{'}) >= B_{ij}\delta(t - t^{'})$. We introduce $C(\tau, t) =< \delta x(t+\tau)\delta x^T(t) >= e^{A\tau} C(0,t)$ and $C_0 = C(0, t = \infty)$. Then Sylvester (Lyapunov) equation holds (fluctuation-dissipation relation)

$$AC_0 + C_0 A^T = -B$$

For non-normal A

$$C_0 = \int_0^\infty e^{As} B e^{A^T s} ds = -\sum_{k,l} \frac{|R_k >< L_k|B|L_i >< R_i|}{\lambda_k + \bar{\lambda}_l}$$

Violation of FDR and entropy production

$$2\partial_\tau C(\tau) = -\chi(\tau)B + B\chi^T(-\tau) + \Delta(\tau)$$

where $\Delta(\tau) = AC(\tau) - C(\tau)A^T$ and explicitly

$$\Delta(\tau) = -\sum |R_k><L_l|B|L_k><R_l| \frac{\lambda_k - \bar{\lambda}_l}{\lambda_k + \bar{\lambda}_l}(e^{\lambda_k \tau}\theta(\tau) + e^{-\bar{\lambda}_k \tau}\theta(-\tau))$$

$$\bar{\Delta}(\tau) = \frac{1}{N}\mathrm{tr}\Delta(\tau) = -\frac{1}{N}\sum_{k,l} O_{kl} \frac{\lambda_k - \bar{\lambda}_l}{\lambda_l + \bar{\lambda}_k}(e^{\lambda_k \tau}\theta(\tau) + e^{-\bar{\lambda}_k \tau}\theta(-\tau))$$

where $O_{kl} =< L_k|L_l >< R_l|R_k >$ is a Chalker-Mehlig operator.

Entropy production rate per unit time (for $B = 1$)

$$\Phi = -\mathrm{tr} B^{-1} A \Delta(0) = \sum_{k,l} O_{kl} \lambda_k \frac{\lambda_k - \bar{\lambda}_l}{\lambda_l + \bar{\lambda}_k}$$

Different modes are coupled!

Drammatic enhancement, since O_{kl} growths with N. See Fyodorov, Gudowska-Nowak, MAN, Tarnowski, 2310.09018v2 (Nov 2023).

Where is the freeness?

Random matrix theory focused in eigenvalues. Deadly mistake in the case of non- normal matrices.
Conceptual breakthrough - Chalker-Mehlig paper on Ginibre ensemble.

How to address the problem of eigenvectors correctly?

Biorthogonality $\langle L_k|R_j\rangle = \delta_{kj}$, completeness $\sum_k |R_k\rangle\langle L_k| = \mathbf{1}$
Invariant under rescaling $|R_k\rangle \to c_k |R_k\rangle$ and $\langle L_k| \to \langle L_k| c_k^{-1}$
The simplest invariant quantity: matrix of overlaps
$O_{ij} = \langle L_i|L_j\rangle\langle R_j|R_i\rangle$ Chalker Mehlig [1998]
Weighted density

$$D(z,w) = \left\langle \frac{1}{N}\sum_{j,k=1}^{N} O_{jk}\delta(z-\lambda_j)\delta(w-\lambda_k)\right\rangle = \tilde{O}_1(z)\delta(z-w)+O_2(z,w)$$

with

$$\tilde{O}_1(z,w) = \left\langle \frac{1}{N}\sum_k O_{kk}\delta^{(2)}(z-\lambda_k)\right\rangle \quad \left(O_1 = \frac{1}{N}\tilde{O}_1\right),$$

$$O_2(z,w) = \left\langle \frac{1}{N}\sum_{j\neq k} O_{jk}\delta^{(2)}(z-\lambda_j)\delta^{(2)}(w-\lambda_k)\right\rangle$$

Sum rules: $\sum_j O_{ij} = 1 \Rightarrow \int d^2w D(z,w) = \rho(z)$

1-point functions [Janik et al., Feinberg and Zee '97]

For the spectral density $\left\langle \frac{1}{N}\sum \delta^{(2)}(z-\lambda_i)\right\rangle$ we need 2D Dirac delta. Identity $\pi\delta^{(2)}(z)=\partial_{\bar z}\frac{1}{z}$. Natural candidate $\mathfrak{g}(z)=\left\langle \frac{1}{N}\mathrm{Tr}(z-X)^{-1}\right\rangle$. Moment expansion valid only outside the spectrum $\longrightarrow$ does not provide the distribution of eigenvalues.
Idea: regularize

$$\mathfrak{g}(z)\to g(z,w)=\left\langle \frac{1}{N}\mathrm{Tr}\frac{\bar z-X^\dagger}{(z-X)(\bar z-X^\dagger)+|w|^2}\right\rangle$$

Problem: how to deal with quadratic denominator? Linearize it

$$G(z)=\begin{pmatrix} G_{11} & G_{12}\\ G_{21} & G_{22}\end{pmatrix}=\left\langle \frac{1}{N}\mathrm{bTr}\begin{pmatrix} z-X & i\bar w\\ iw & \bar z-X^\dagger\end{pmatrix}^{-1}\right\rangle$$ [Janik et al]

$$\left\langle \frac{1}{N}\mathrm{bTr}\begin{pmatrix} \epsilon & z-X\\ \bar z-X^\dagger & \epsilon\end{pmatrix}^{-1}\right\rangle$$ [Feinberg, Zee]

This construction can be written in the resolvent form

$$\mathcal{G} = \left\langle (Q - \mathcal{X})^{-1} \right\rangle, \qquad G(z) = \frac{1}{N} \mathrm{bTr} \mathcal{G}$$

with

$$Q = \begin{pmatrix} z & i\bar{w} \\ iw & \bar{z} \end{pmatrix}, \qquad \mathcal{X} = \begin{pmatrix} X & 0 \\ 0 & X^\dagger \end{pmatrix}$$

Moment expansion

$$\mathcal{G} = Q^{-1} + \left\langle Q^{-1} \mathcal{X} Q^{-1} \right\rangle + \left\langle Q^{-1} \mathcal{X} Q^{-1} \mathcal{X} Q^{-1} \right\rangle + \dots$$

Large N limit: planar diagrams $\longrightarrow$ Schwinger-Dyson equation.
Note that $O_1(z) = -\lim_{|w| \to 0} \frac{1}{\pi} G_{12} G_{21}$, whereas
$\rho(z, \bar{z}) = \lim_{|w| \to 0} \frac{1}{\pi} \partial_{\bar{z}} G_{11}$

2-point functions

Natural candidate

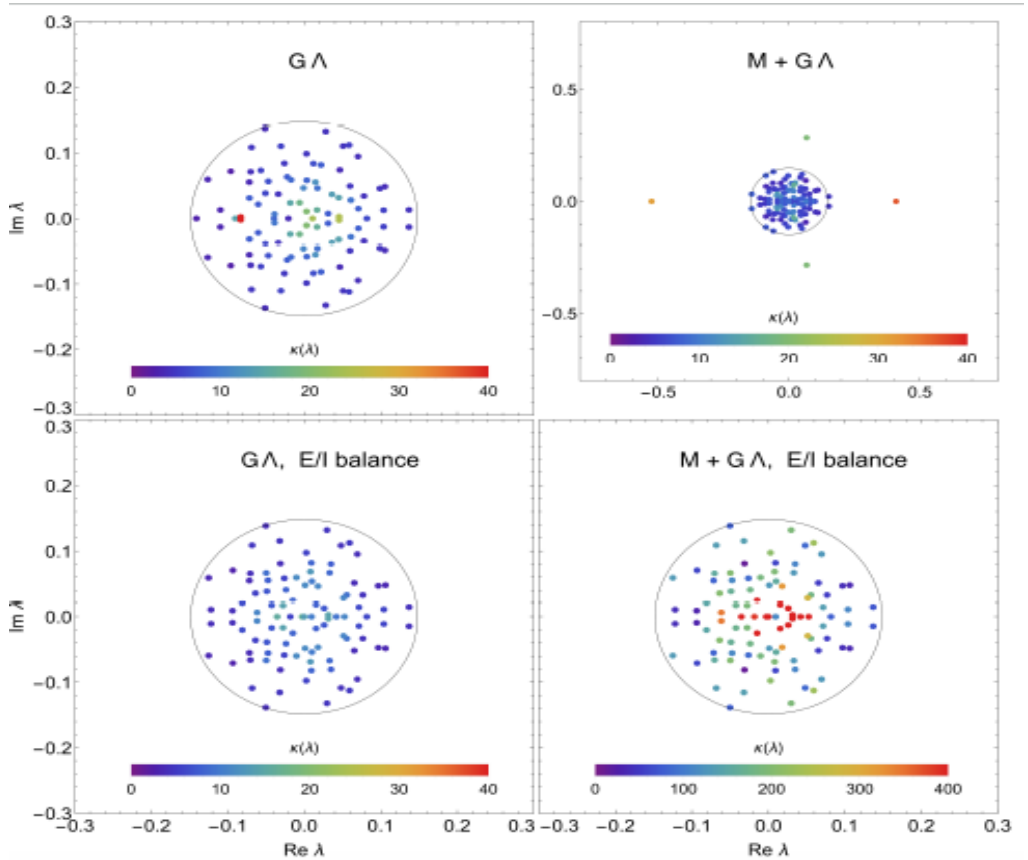
$$\mathfrak{h}(z_1,\bar{z}_2)=\frac{1}{N}\mathrm{Tr}(z_1-X)^{-1}(\bar{z}_2-X^\dagger)^{-1}=\frac{1}{N}\sum_{k,l}O_{kl}\frac{1}{(z_1-\lambda_k)(\bar{z}_2-\bar{\lambda}_l)}$$

Same problems $\Rightarrow$ regularization + linearization

$$\mathcal{K}=\left\langle (Q-\mathcal{X})^{-1}\otimes(P^T-\mathcal{X}^T)^{-1}\right\rangle$$

+ proper contraction of indices (like a block-trace) $\Rightarrow 4\times 4$ matrix. One of its entries is the object of our interest.
Details in Arxiv: [1801.02526]
Luckily, for R-diagonal operators results simplify, e.g.
$\mathfrak{h}(z_1,\bar{z}_2)=\frac{1}{z_1\bar{z}_2-r_{out}^2}$

Rajan-Abbott Model RAM, cont.



The balance condition tames outliers, bringing them back to the disk, but drastically increases the sensitivity of the spectrum, measured by the eigenvalue condition number $\kappa(\lambda_i) = ||L_i|| \times ||R_i||$ *here* $L_i(R_i)$ *is left (right) eigenvector to the eigenvalue* λ_i*.*

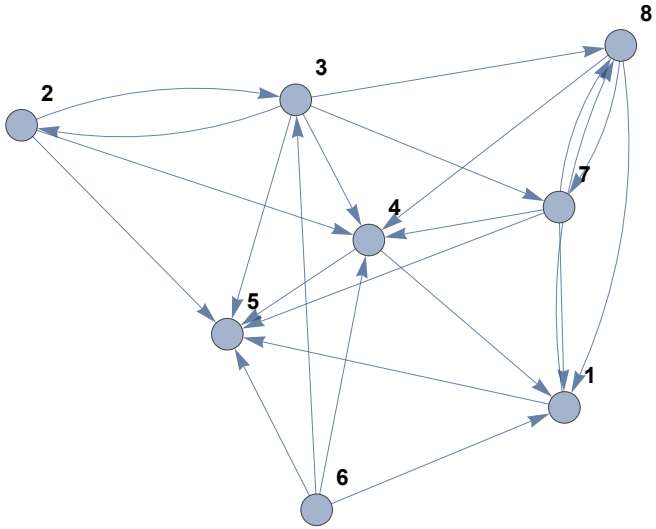
Why non-normal matrices matter?

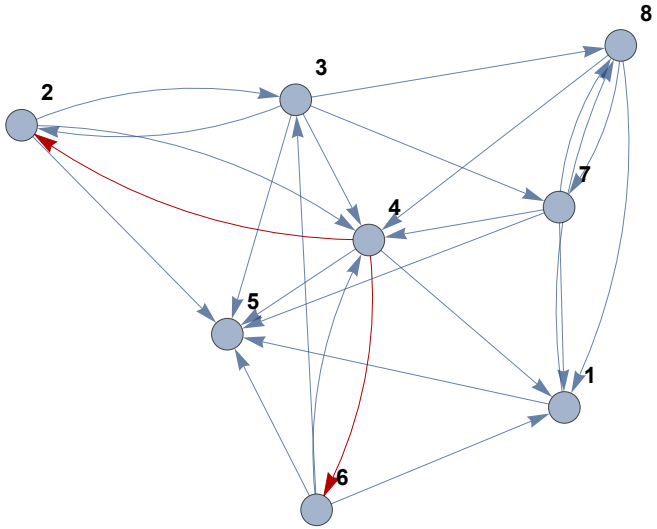
A matrix X is non-normal iff $XX^\dagger \neq X^\dagger X$.
If a non-normal matrix can be diagonalized, it possesses two sets of eigenvectors: right $|R_k>$ (column) and left $<L_k|$ (rows), satisfying eigenequations

$$<L_k|X =<L_k|\lambda_k, \qquad X|R_k>=\lambda_k|R_k>$$

The diagonalization is via similarity transformation $X = S\Lambda S^{-1}$ with S and S^{-1} encoding eigenvectors $X = \sum_k |R_k>\lambda_k<L_k|$.
The eigenvectors are not orthogonal $<R_k|R_l> \neq \delta_{kj}$ but biorthogonal $<L_k|R_j>=\delta_{kj}$ ($\Leftrightarrow S^{-1}S = \mathbf{1}$).
Resolution of identity $\sum_k |R_k><L_k| = \mathbf{1}$ ($\Leftrightarrow SS^{-1} = \mathbf{1}$).

Temporal changes of networks seen as perturbations





Adjacency matrix: $A \rightarrow A' = A + P$ How does the spectrum change? In first order perturbation theory

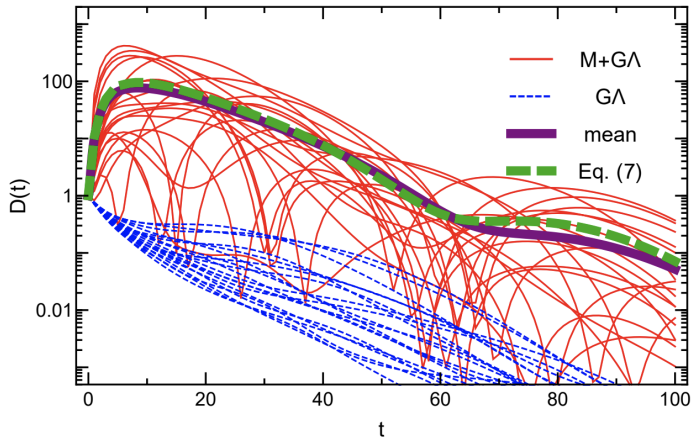
$$\lambda'_k = \lambda_k + < L_k|P|R_k > + \mathcal{O}(||P||^2)$$

Upper bound
$|\delta\lambda_k| \leqslant ||L_k|| \cdot ||R_k|| \cdot ||P|| = ||P|| \sqrt{< L_k|L_k >< R_k|R_k >}.$
Eigenvalue condition number [Wilkinson 1965]

Transient dynamics

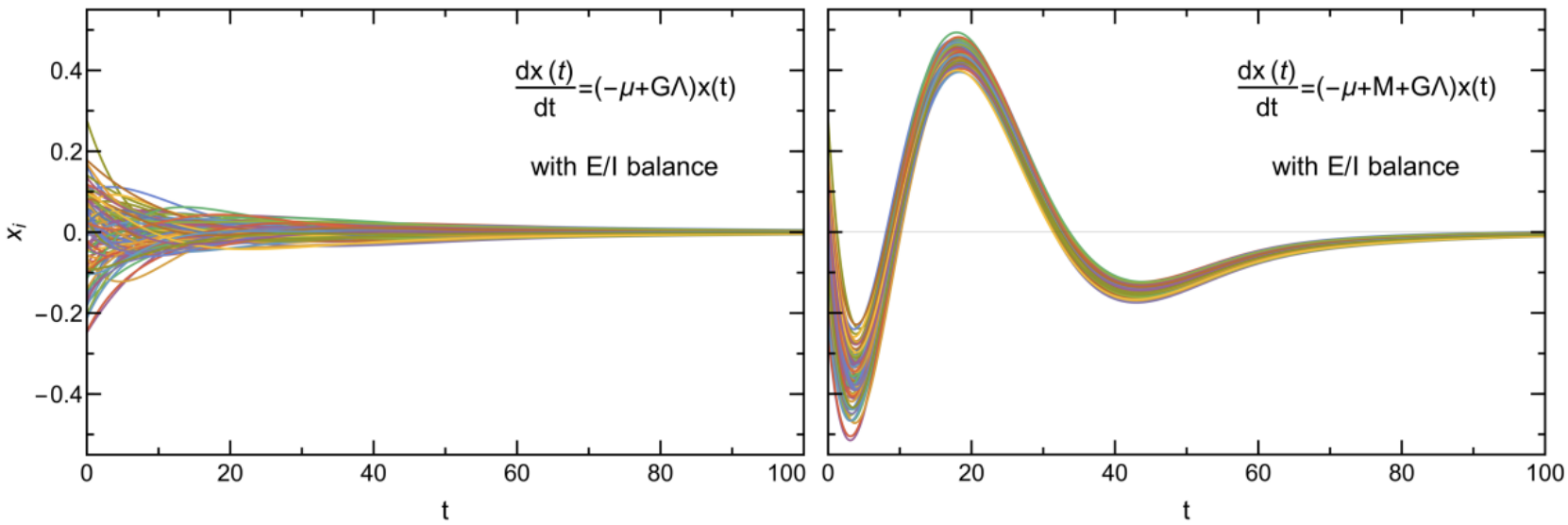
- Consider $\frac{d}{dt}|\psi>=(-\mu+X)|\psi>+\delta(t)|\psi(0)>$. Formal solution reads $|\psi(t)>=e^{(X-\mu)t}|\psi(0)>$.
- We define $D(t)=<\psi(t)|\psi(t)>$, and average this quantity over spikes, so $||\psi(0)||=1$.
- Then, $\overline{D(t)}=e^{-2\mu t}\frac{1}{N}\mathrm{tr}e^{X^{\dagger}t}e^{Xt}=e^{-2\mu t}\frac{1}{N}\sum_{i,j}e^{t(\lambda_i+\bar{\lambda_j})}O_{ij}$, where $O_{ij}=<L_i|L_j><R_j|R_i>$
- Dramatic effect comparing to normal case, since now eigenmodes couple and get amplified by overlaps (!).

Transition to chaos associated with instability at $x = 0$



Relaxation towards the fixed point measured by the squared Euclidean distance $D(t) = \sum_{ij=1}^{N} < x_0|L_i >< R_i|R_j >< L_j|x_0 > e^{-2t+t(\lambda_i+\bar{\lambda_j})}$. Non-orthogonality of eigenmodes mixes them together resulting in oscillatory behavior of $D(t)$ ("interference effects").

Activity of neurons in the linearized dynamics



- Onset of collective dynamics (right) driven by M and the balance condition
- Mean connectivity responsible for synchronization

E.G-N, M.A. Nowak, D.R. Chialvo, J.K. Ochab, W. Tarnowski *From synaptic interactions to collective dynamics in random neuronal networks models* Neural Computations **32** 395 (2020)

Entropy production in Rajan-Abbott model - technicalities

- From our formalism, we get

$$\Phi = \mathrm{tr} \int_0^\infty A e^{As} e^{A^T s} A^T ds - \mathrm{tr} \int_0^\infty A^2 e^{As} e^{A^T s} A^T ds$$

Parametrization $A = -\mu \mathbf{1} + X$ and representation $f(X) = \frac{1}{2\pi i} \oint \frac{f(z)dz}{z-X}$ boils to

$$\Phi = \frac{1}{(2\pi i)^2} \int_0^\infty ds \oint_c dz \oint_c dw (z-\mu)(w-z) e^{s(z+w-2\mu)} R_1(z,w)$$

where we introduce a traced product of resolvents $R_1(z,w) = \mathrm{tr} \frac{1}{z-X} \frac{1}{w-X^T}$. Averaging over randomness yields for above resolvent $1/(zw-1)(1+\nu^2/zw)$,([jhep06(2018)152], Shermann-Morrison formula (Tarnowski, 2011.08215v1)), so explicit calculation is possible in the large N limit, yielding to

$$\Phi = (1+\nu^2)(\mu + \sqrt{\mu^2 - 1})^{-1} \qquad (1)$$

where $A = -\mu \mathbf{1} + X + \nu M$, where M is a ranked one perturbation defined in Rajan-Abbott paper.

Summary 2

- Understanding temporal evolution of non-normal matrix models requires considering the entangled dynamics of both eigenvectors and eigenvalues, contrary to simple evolution of the spectra of normal matrices for which eigenvectors decouple in the presence of the spectral evolution
- General feature of open systems, directed networks (graphs), cross-correlations $XY^\dagger$, timed-lagged correlators etc.
- Transient behaviour crucial in the stability analysis of the network
- Mechanism for synchronization? (memory, learning....)

Example 3 - Taming Deep Networks

- Pioneering application of FRV to Machine Learning by Schoenholtz, Ganguli and Pennington (Google AI) in 2017.
- Too small gradients versus too large gradients
- Two universality classes found by S-transform for feed forward networks
- Training successful even for the depth of 200 layers.

- For the residual network of L layers of N neurons, with weight matrix for the l-th layer W^l, and bias vectors b^l, information propagates as

$$x^l = \phi(h^l) + a x^{l-1} \qquad h^l = W x^{l-1} + b_l$$

  where h^l and x^l are pre- and post-activations, ϕ is activation function, a-parameter.

- "Learning" is based by adjusting weights to minimize the error

$$\Delta W^l_{ij} = -\eta \frac{\partial E(x^L, y)}{\partial W^l_{ij}} = -\eta \sum_{k,t} \frac{\partial x^l_t}{\partial W^l_{ij}} \frac{\partial x^L_k}{\partial x^l_t} \frac{\partial E(x^L, y)}{\partial x^L_k}$$

Geometric random walk and input-ouput Jacobian

- $J = \frac{\partial x^L}{\partial x^0} = \left[\prod_{l=0}^{L}(D^l W^l + \mathbf{1}a)\right]$, where $D^l_{ij} = \phi'(h^l)\delta_{ij}$
- 1-dim geometric random walk $x_i = x_{i-1} + w x_{i-1}$, where $< w >= 0$ and $< w^2 >= dt$
- Matricial geometric random walk $W_i = (\mathbf{1} + \sqrt{T/L})W_{i-1}$, $T/L \equiv dt$, $< W >= 0$ and Gaussian (Ginibre) property $< W_{ij} W_{kt} >= dt \frac{1}{N} \delta_{it}\delta_{jk}$.
- Solving the spectrum of $(\prod_{l=1}^{L}(\mathbf{1} + \sqrt{dt} W_l)$ with W_i being GUE or Ginibre Ensemble is a complicated problem (complex spectrum, coupling to eigenvectors), important in math and physics (QFT)
- In the limit of large L and large N, support of eigenvalues solved analytically [Gudowska-Nowak, Janik, Jurkiewicz, Nowak; 2003] using methods inspired by FRV
- Full rigorous solution of the problem [Driver, Hall, Kemp; 2019]

Our results - technicalities

- We study **singular values**, i.e. the spectrum JJ^T.
- Technically, spectrum $\rho(\lambda)$ comes from imaginary part of the resolved $G(z) \sim < Tr(z - JJ^T)^{-1} >$. The resolvent is inferred from **Free Random Variables** techniques, in particular Voiculescu S-transform, which turns out to read for our problem

$$S(z) = \frac{1}{a^{2L}} e^{-c(1+2z)/a^2} \rightarrow a^{2L} G(z) = (zG(z) - 1) e^{(1-2zG(z))/a^2}$$

- Effective cumulant $c = \frac{1}{L} \sum_{l=1}^{L} c_2^l$, where c_2^l is the squared spectral radius of the matrix $D^l W^l$.
- To calculate c for each activation function needed, we apply **dynamical mean field theory** alike Google AI group did.

Sample synthetic data tests

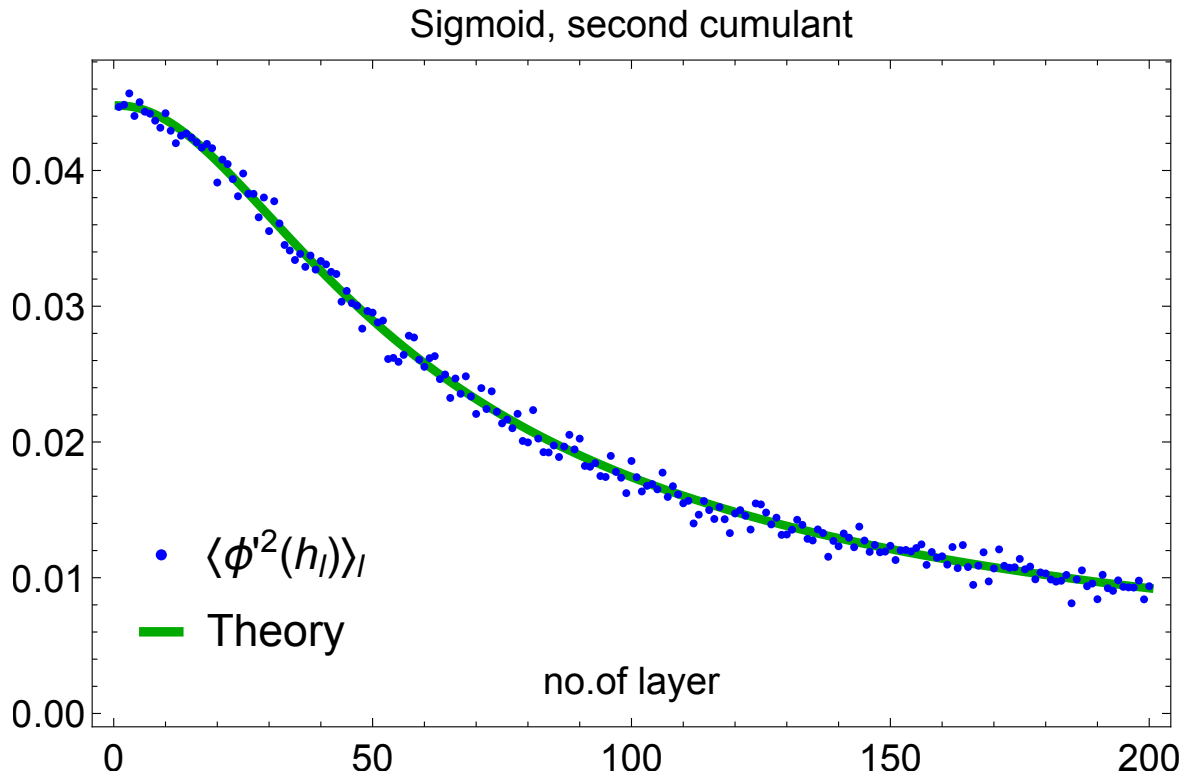


Figure: Verification of the c_2 for sigmoid activation function $\phi(x) = \frac{1}{1+e^{-x}}$.

Sample synthetic data tests - cont.

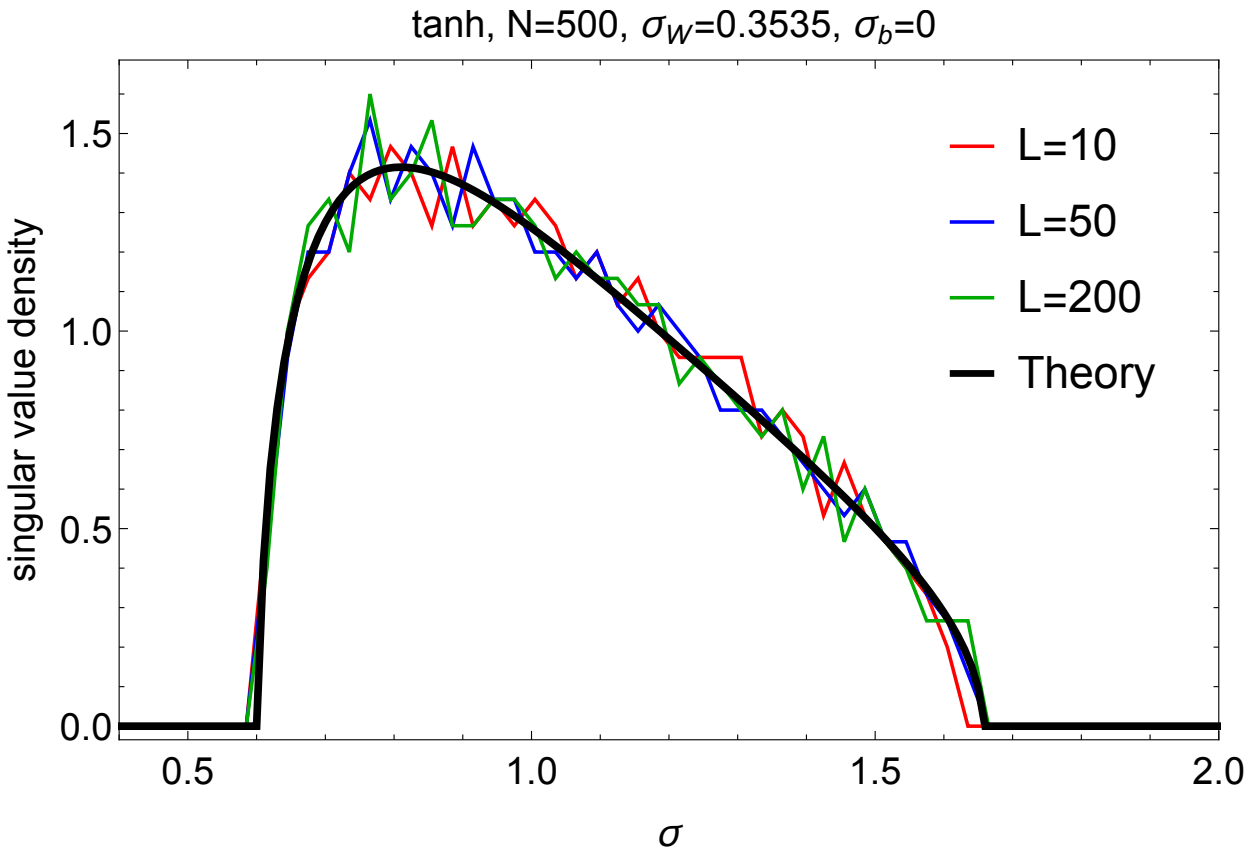


Figure: Singular values of Jacobian for tanh non-linearity. **Note that already $L = 10$ matches well asymptotic result**.

Sample synthetic data tests - cont.

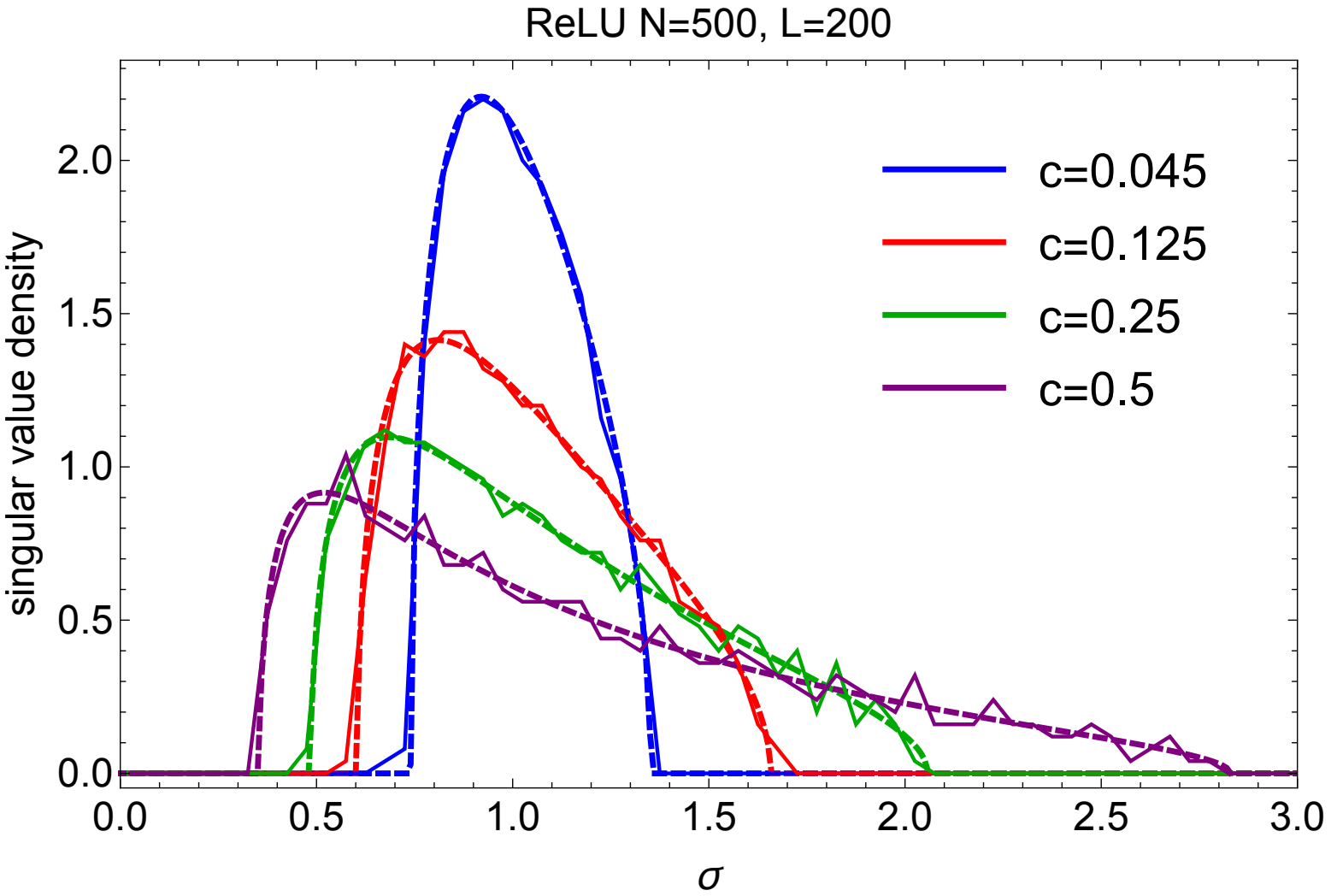


Figure: Singular values of Jacobian for RELU non-linearity for various effective cumulants - theory versus experiment.

Isometry (universality) tested and confirmed on CIFAR-10 benchmark

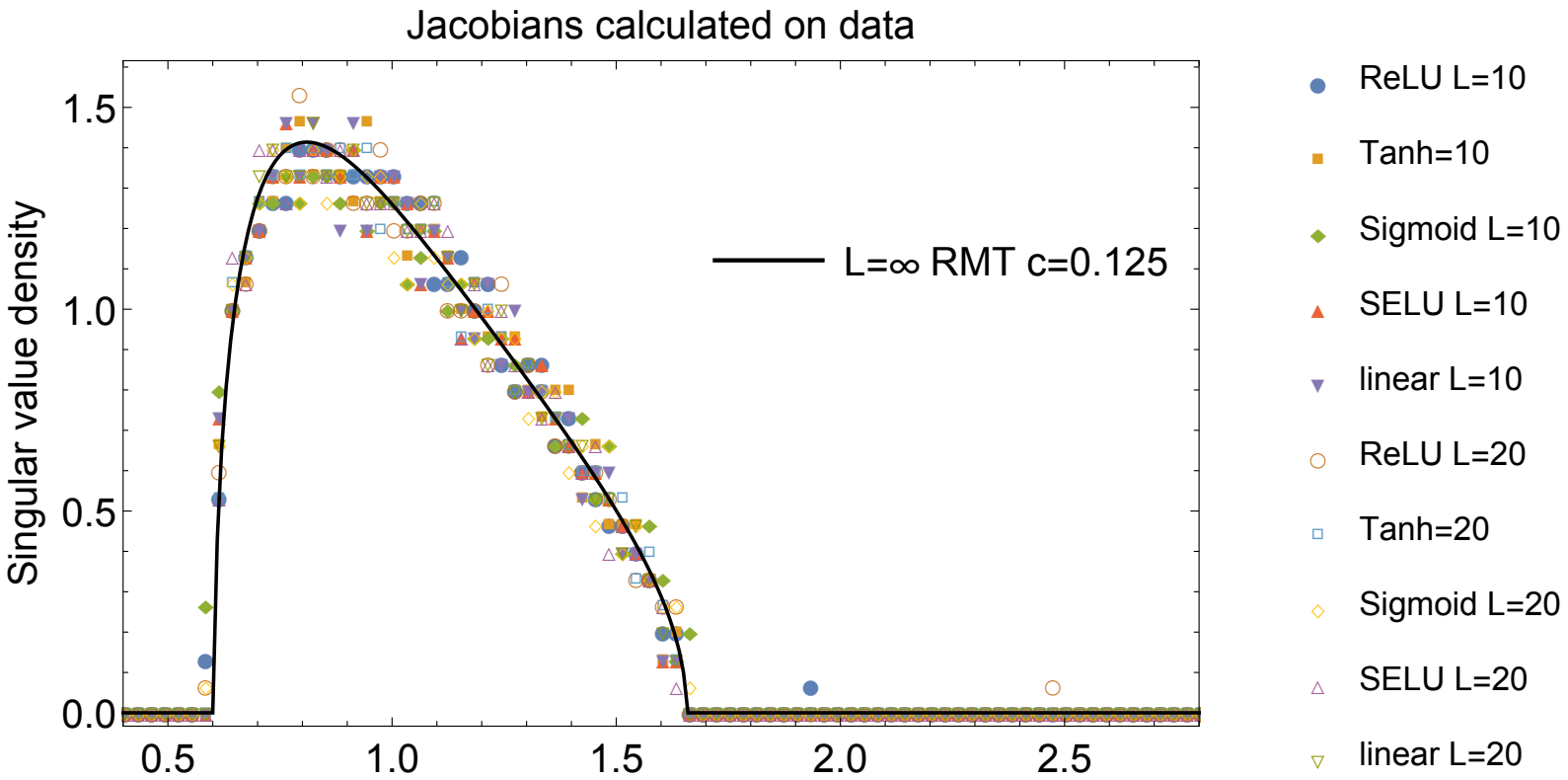


Figure: Jacobians calculated on data for several activation functions and/or numbers of residual blocks.

Summary 3

- Singular spectrum for input-output Jacobian in the limit of large width and depth of the network, is given by **universal formula**.
- Dependence on the type of activation function is encapsulated in a single parameter, therefore ResNet can achieve **dynamical isometry** for many different activation functions.
- Results in **agreement with data**: synthetic data (Random Matrix Theory) and CIFAR-10 classification data.

[More computer science details in Proc. of 22nd International Conference on Artificial Intelligence and Statistics, PMLR 89, 2221-2230, 2019.]

Conclusions

- FRV calculus provides **powerful tool for multivariate statistics** in cognitive science.
- FRV calculus can be used for **analytic modelling of several complex phenomena**.
- FRV concepts are **rather unexploited in cognitive neuroscience**, despite enormous impact on others branches of science and technology.
- FRV can serve as an interlanguage (*lingua franca*) **for different subcommunities in cognitive sciences.**

[Ewa Gudowska-Nowak, MAN, Freeness in cognitive science, https://arxiv.org/abs/2311.04307]