Stochastic gradient descent and Random Matrix Theory

Gert Aarts

Swansea University
Prifysgol Abertawe

Physics for AI, Oxford, March 2025

Stochastic gradient descent and Random Matrix Theory

Gert Aarts

with **Chanju Park** and Biagio Lucini

PRE 111 (2025) 1, 015303 [2407.16427 [cond-mat.dis-nn]]

and Ouraman Hajizadeh

2411.13512 [cond-mat.dis-nn]

NeurIPS 2024 workshop *ML and the Physical Sciences*

Swansea University
Prifysgol Abertawe

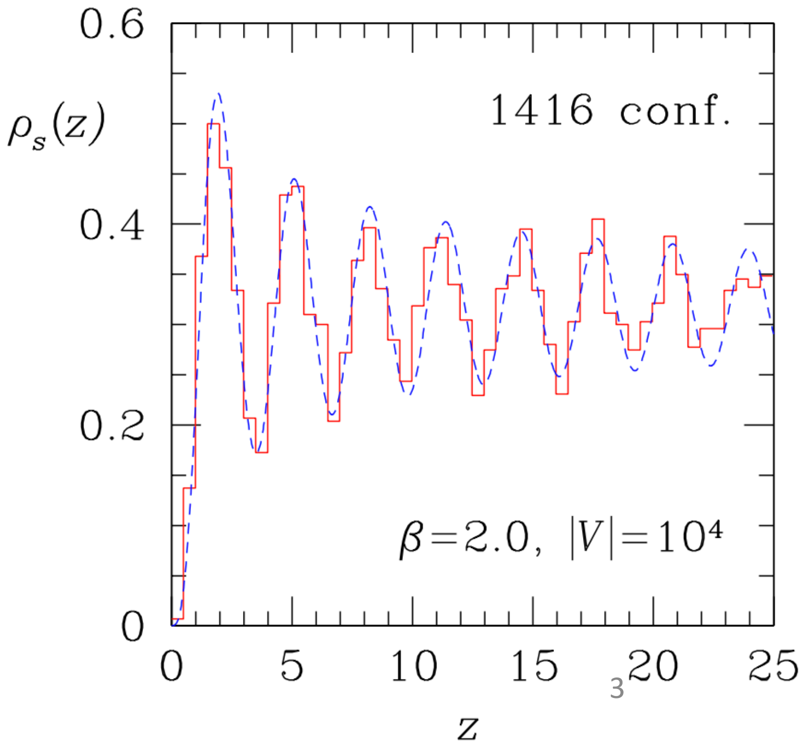
Physics for AI, Oxford, March 2025

Random Matrix Theory (RMT)

- developed by Wigner and Dyson to describe nuclear spectra (1959-1962)
- universal features: level spacing, Coulomb repulsion, Wigner surmise, fluctuations
- non-universal behaviour: spectral density

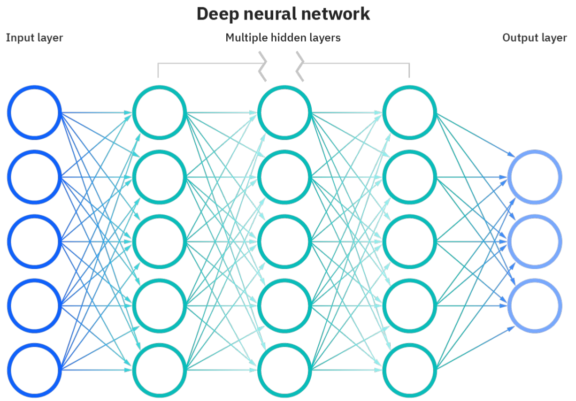
example:

- successfully applied in QCD to describe Dirac operator



JJM Verbaarschot and T Wettig, *Random matrix theory and chiral symmetry in QCD*
Ann. Rev. Nucl. Part. Sci. 50 (2000) 343 [arXiv:hep-ph/0003017 [hep-ph]].

RMT and machine learning

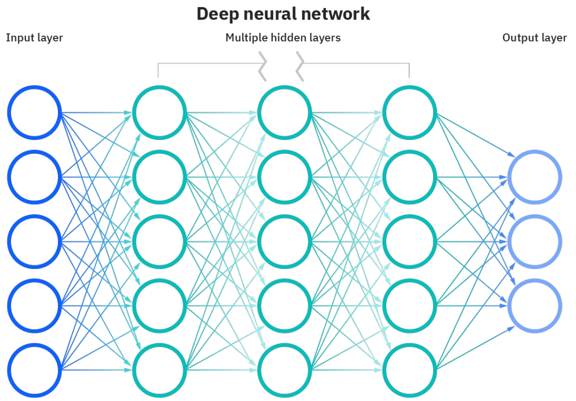


- different context: machine learning and weight matrix dynamics
- neural networks: layers of nodes, connected by weight matrices
- weight matrices are updated using e.g. stochastic gradient descent (SGD)
- stochastic matrix dynamics → Dyson Brownian motion → RMT features

- aim: further understanding of learning by characterising weight matrix dynamics
- identify universal behaviour and limitations of SGD during and after training

Outline

- some general comments on stochastic weight matrix updates
- connection to Dyson Brownian motion and stochastic Coulomb gas
- universal properties of stationary distribution
- application in Restricted Boltzmann Machine (RBM) and Transformer (nano-GPT)
- summary and outlook



Stochastic weight matrix dynamics

- consider some $M \times N$ weight matrix W
- update (e.g. stochastic gradient descent): $W \to W' = W + \delta W$ with $\delta W = -\alpha \frac{\delta \mathcal{L}}{\delta W}$
- obtained from loss function $\mathcal{L}[W]$, learning rate α
- δW is estimated using a batch $\mathcal{B}$ with batch size $|\mathcal{B}|$: $\delta W_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \delta W_b$
- fluctuations controlled by finite batch size (CLT): $\frac{1}{|\mathcal{B}|} \mathrm{Var}(\delta W)$

Stochastic weight matrix dynamics

- stochastic update $W \to W' = W + \delta W$ becomes

$$\delta W = \delta W_{\mathcal{B}} + \frac{1}{\sqrt{|\mathcal{B}|}} \sqrt{\mathrm{Var}(\delta W)}\eta$$

- or in terms of the gradient of the loss function:

$$W' = W - \alpha \left(\frac{\delta \mathcal{L}}{\delta W} \right)_{\mathcal{B}} + \frac{\alpha}{\sqrt{|\mathcal{B}|}} \sqrt{\mathrm{Var} \left(\frac{\delta \mathcal{L}}{\delta W} \right)} \eta \qquad \eta_{ij} \sim \mathcal{N}(0,1)$$

From rectangular to symmetric matrices

- W is $M \times N$ matrix: singular value decomposition: $W = U\Xi V^T$

  $$UU^T = \mathbb{1} \quad VV^T = \mathbb{1}$$

- singular values: $\xi_i \quad (i = 1 \ldots N)$ [take $N \leq M$ without loss of generality]

- introduce symmetric semi-positive combination: $X = W^T W = VDV^T$

- and focus on the singular/eigenvalues (invariant under left/right rotations on W):

$$D = \Xi^T \Xi = \mathrm{diag}\left(\xi_1^2, \ldots, \xi_N^2\right) = \mathrm{diag}\left(x_1, \ldots, x_N\right)$$

- stochastic dynamics: $X \to X' = X + \delta X_{\mathcal{B}} + \frac{1}{\sqrt{|\mathcal{B}|}} \sqrt{\mathrm{Var}(\delta X)}\eta$

Initialisation: Marchenko-Pastur distribution

- if initial weight matrix $W_{ij} \sim \mathcal{N}(0, \sigma^2)$ then X follows Marchenko-Pastur distribution

$$P_{\rm MP}(x) = \frac{1}{2\pi\sigma^2 Mrx}\sqrt{(x_+ - x)(x - x_-)} \qquad x_- < x < x_+ \qquad r = N/M \leq 1 \quad x_\pm = M\sigma^2\left(1 \pm \sqrt{r}\right)^2$$

- ✓ how to choose σ^2 : distribution should depend on r only, safe to take large N, M limit

- ✓ spectrum is bounded for all r (relevant for RBMs below) : $\sigma^2 = 1/M$ $\quad N \leq M$

$$P_{\rm MP}(x) = \frac{1}{2\pi rx}\sqrt{(x_+ - x)(x - x_-)} \qquad 0 \leq x_- \leq x \leq x_+ \leq 4 \qquad x_\pm = \left(1 \pm \sqrt{r}\right)^2$$

Stochastic matrix dynamics: Dyson Brownian motion and the stochastic Coulomb gas

- framework to consider stochastic matrix dynamics for symmetric matrix X

- Dyson Brownian motion (in continuous time for now, see below):

$$\frac{dX_{ij}}{dt} = K_{ij}(X) + \sqrt{A_{ij}}\eta_{ij}$$

- eigenvalues then evolve according to

$$\frac{dx_i}{dt} = K_i(x_i) + \sum_{j\neq i} \frac{g_i^2}{x_i - x_j} + \sqrt{2}g_i\eta_i$$

$$\equiv K_i^{(\text{eff})}(x_i) + \sqrt{2}g_i\eta_i$$

where $\sqrt{A_{ii}} = \sqrt{2}g_i$

Dyson Brownian motion, stochastic Coulomb gas

- eigenvalues dynamics: $$\frac{dx_i}{dt} = K_i(x_i) + \sum_{j \neq i} \frac{g_i^2}{x_i - x_j} + \sqrt{2} g_i \eta_i$$
- can be derived using 2nd order perturbation theory
- Coulomb term: eigenvalue repulsion [Wigner, Dyson 1959-1962, for nuclear spectra]
- Fokker-Planck equation (FPE) for distribution of eigenvalues:

$$\partial_t P(\{x_i\}, t) = \sum_{i=1}^{N} \partial_{x_i} \left[\left(g_i^2 \partial_{x_i} - K_i^{(\text{eff})}(x_i) \right) \right] P(\{x_i\}, t)$$

Dyson Brownian motion, stochastic Coulomb gas

- FPE: $$\partial_t P(\{x_i\},t) = \sum_{i=1}^{N} \partial_{x_i} \left[\left(g_i^2 \partial_{x_i} - K_i^{(\mathrm{eff})}(x_i) \right) \right] P(\{x_i\},t)$$
- stationary distribution: $$P_s(\{x_i\}) = \frac{1}{Z} \prod_{i<j} |x_i - x_j| \, e^{-\sum_i V_i(x_i)/g_i^2}$$
- with partition function: $$Z = \int dx_1 \ldots dx_N \, P_s(\{x_i\})$$
- and provided drift can be derived from a potential $K_i(x_i) = -\dfrac{dV_i(x_i)}{dx_i}$
- known as Coulomb gas, describes universal features of random matrices

Back to weight matrix dynamics

- stochastic dynamics $X \to X' = X + \delta X_{\mathcal{B}} + \frac{1}{\sqrt{|\mathcal{B}|}}\sqrt{\mathrm{Var}(\delta X)}\eta$
- what can be carried over from Dyson's matrix dynamics? implications? universality?
- eigenvalue equation: $x_i \to x_i' = x_i + \delta x_i + \sum_{j \neq i} \frac{g_i^2}{x_i - x_j} + \sqrt{2} g_i \eta_i$
- make explicit learning rate and batch size dependence

$$\delta x_i = \alpha K_i \qquad g_i = \frac{\alpha}{\sqrt{|\mathcal{B}|}}\tilde{g}_i \qquad \tilde{g}_i \sim \mathrm{Var}(\delta\mathcal{L}/\delta W)\big|_{ii}$$

Back to weight matrix dynamics

- eigenvalue dynamics: $x_i \to x_i' = x_i + \delta x_i + \sum_{j \neq i} \frac{g_i^2}{x_i - x_j} + \sqrt{2} g_i \eta_i$

- insert learning rate and batch size dependence:

$$x_i \to x_i' = x_i + \alpha K_i + \frac{\alpha^2}{|\mathcal{B}|} \sum_{j \neq i} \frac{\tilde{g}_i^2}{x_i - x_j} + \frac{\alpha}{\sqrt{|\mathcal{B}|}} \sqrt{2} \tilde{g}_i \eta_i$$

- no usual scaling of drift and noise with learning rate (Ito calculus: ϵ , $\sqrt{\epsilon}$):

  no obvious continuous time limit (SDE), only in some weak sense

Q Li, C Tai and W E [1511.06251] S Yaida [1810.00004]

- known issue: from SGD to SDE but is in fact blessing (see below)

$$x_i \to x_i' = x_i + \alpha K_i + \frac{\alpha^2}{|\mathcal{B}|}\sum_{j\neq i}\frac{\tilde{g}_i^2}{x_i - x_j} + \frac{\alpha}{\sqrt{|\mathcal{B}|}}\sqrt{2}\tilde{g}_i\eta_i$$

Stationary distribution

- distribution for fixed $\alpha, |\mathcal{B}|$: $P_s(\{x_i\}) = \frac{1}{Z}\prod_{i<j}|x_i - x_j|\, e^{-\sum_i V_i(x_i)/g_i^2}$
- make explicit dependence on learning rate and batch size

$$g_i = \frac{\alpha}{\sqrt{|\mathcal{B}|}}\tilde{g}_i \qquad V_i(x_i) = \alpha\tilde{V}_i(x_i) \qquad \frac{V_i(x_i)}{g_i^2} = \frac{1}{\alpha/|\mathcal{B}|}\frac{\tilde{V}_i(x_i)}{\tilde{g}_i^2}$$

- if drift vanishes at $x_i = x_i^s$, expand potential $\tilde{V}_i(x_i) = \tilde{V}_i(x_i^s) + \frac{1}{2}\Omega_i\,(x_i - x_i^s)^2 + \ldots$
- exponential is Gaussian with variance $\sigma_i^2 = (\alpha/|\mathcal{B}|)\ \ (\tilde{g}_i^2/\Omega_i)$

universal scaling with learning rate and batch size

model-dependent factor

Linear scaling relation

- dependence on $\alpha/|\mathcal{B}|$ in training has been observed before, empirically

✓ P. Goyal, P. Dollár, R.B. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola et al.,
Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour [1706.02677]

✓ S.L. Smith and Q.V. Le,
A Bayesian Perspective on Generalization and Stochastic Gradient Descent [1710.06451]

✓ S.L. Smith, P. Kindermans and Q.V. Le,
Don't Decay the Learning Rate, Increase the Batch Size [1711.00489]

- finds a natural place in the framework of Dyson Brownian motion and Coulomb gas

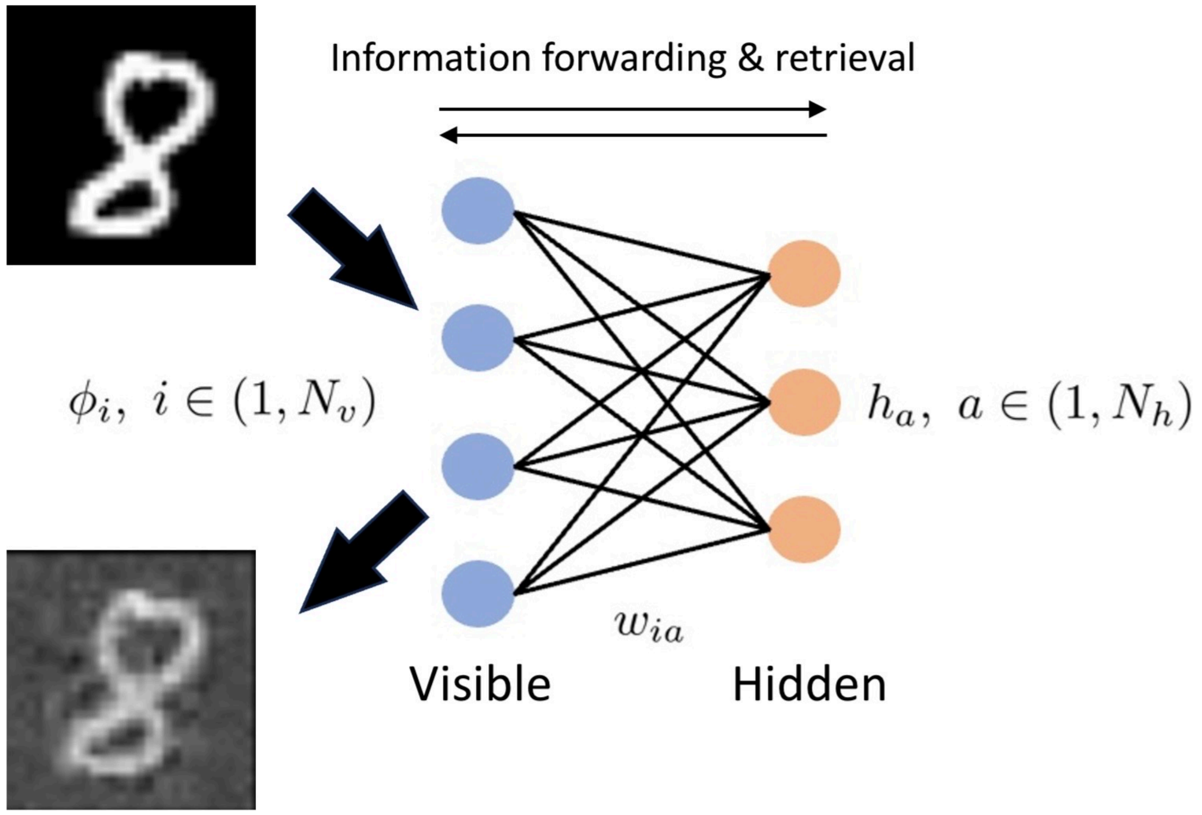
Applications and implications

- so far, the derivation is general: prediction of eigenvalue distribution after learning
- apply to actual ML models to observe universal features and support derivation
  - teacher-student model
  - Gaussian Restricted Boltzmann Machine
  - Transformer

builds on previous analysis of RBM:
GA, B Lucini, C Park, Phys. Rev. D 109 (2024) 034521 [2309.15002 [hep-lat]]
current analysis: PRE 111 (2025) 1, 015303 [2407.16427 [cond-mat.dis-nn]]

GA, O Hajizadeh, B Lucini, C Park
2411.13512 [cond-mat.dis-nn]

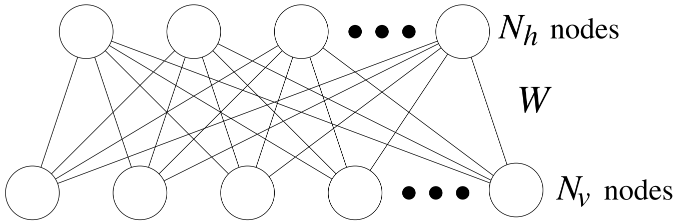
Restricted Boltzmann Machine: generative network



- energy-based method
- probability distribution
- binary or continuous d.o.f.

$$p(\phi, h) = \frac{1}{Z} e^{-S(\phi,h)}$$

$$Z = \int D\phi Dh\, e^{-S(\phi,h)}$$



Scalar field RBM

- distribution: $p(\phi, h) = \frac{1}{Z} e^{-S(\phi,h)}$ $\quad S(\phi, h) = \frac{1}{2}\mu^2\phi^T\phi + \frac{1}{2\sigma_h^2}(h-\eta)^T(h-\eta) - \phi^T W h$
- $M \times N = N_v \times N_h$ weight matrix W
- induced distribution on visible layer $p(\phi) = \int Dh\, p(\phi, h) = \frac{1}{Z}\exp\left(-\frac{1}{2}\phi^T K \phi + J^T \phi\right)$
- kernel $K = \mu^2 \mathbb{1} - \sigma_h^2 W W^T = \mu^2 \mathbb{1} - \sigma_h^2 U \Xi \Xi^T U^T = U\left[\mu^2 \mathbb{1} - \sigma_h^2 \Xi \Xi^T\right] U^T \equiv U D_K U^T$
- eigenvalues $D_K = \mathrm{diag}\big(\underbrace{\mu^2 - \sigma_h^2\xi_1^2, \mu^2 - \sigma_h^2\xi_2^2, \ldots, \mu^2 - \sigma_h^2\xi_N^2}_{N}, \underbrace{\mu^2, \ldots, \mu^2}_{M-N}\big)$

Scalar field RBM as a lattice field theory

- treat RBM as a lattice field theory with bi-linear quadratic action

$$S(\phi, h) = \sum_i \frac{1}{2}\mu_i^2\phi_i^2 + \sum_a \frac{1}{2\sigma^2}(h_a - \eta_a)^2 - \sum_{i,a} \phi_i w_{ia} h_a$$

- induced distribution on visible layer

$$p(\phi) = \int Dh\, p(\phi, h) = \frac{1}{Z}\exp\left(-\frac{1}{2}\sum_{i,j}\phi_i K_{ij}\phi_j + \sum_i J_i\phi_i\right)$$

- all information is stored in quadratic operator, with spectrum

$$D_K = \mathrm{diag}\big(\underbrace{\mu^2 - \sigma_h^2\xi_1^2, \mu^2 - \sigma_h^2\xi_2^2, \ldots, \mu^2 - \sigma_h^2\xi_N^2}_{N}, \underbrace{\mu^2, \ldots, \mu^2}_{M-N}\big)$$

Scalar field RBM as an ultraviolet regulator

- spectrum

$$D_K = \mathrm{diag}\big(\underbrace{\mu^2 - \sigma_h^2\xi_1^2, \mu^2 - \sigma_h^2\xi_2^2, \ldots, \mu^2 - \sigma_h^2\xi_{N_h}^2}_{N_h}, \underbrace{\mu^2, \ldots, \mu^2}_{N_v - N_h}\big)$$

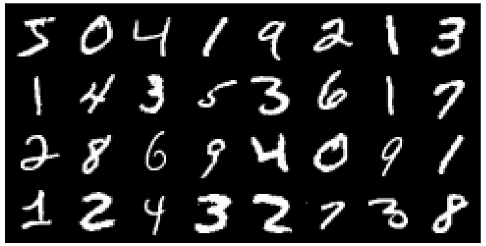
- what if $N_h < N_v$? not all eigenvalues can be reproduced

- role of hyperparameter μ^2 ? if chosen too low, not all eigenvalues can be reproduced

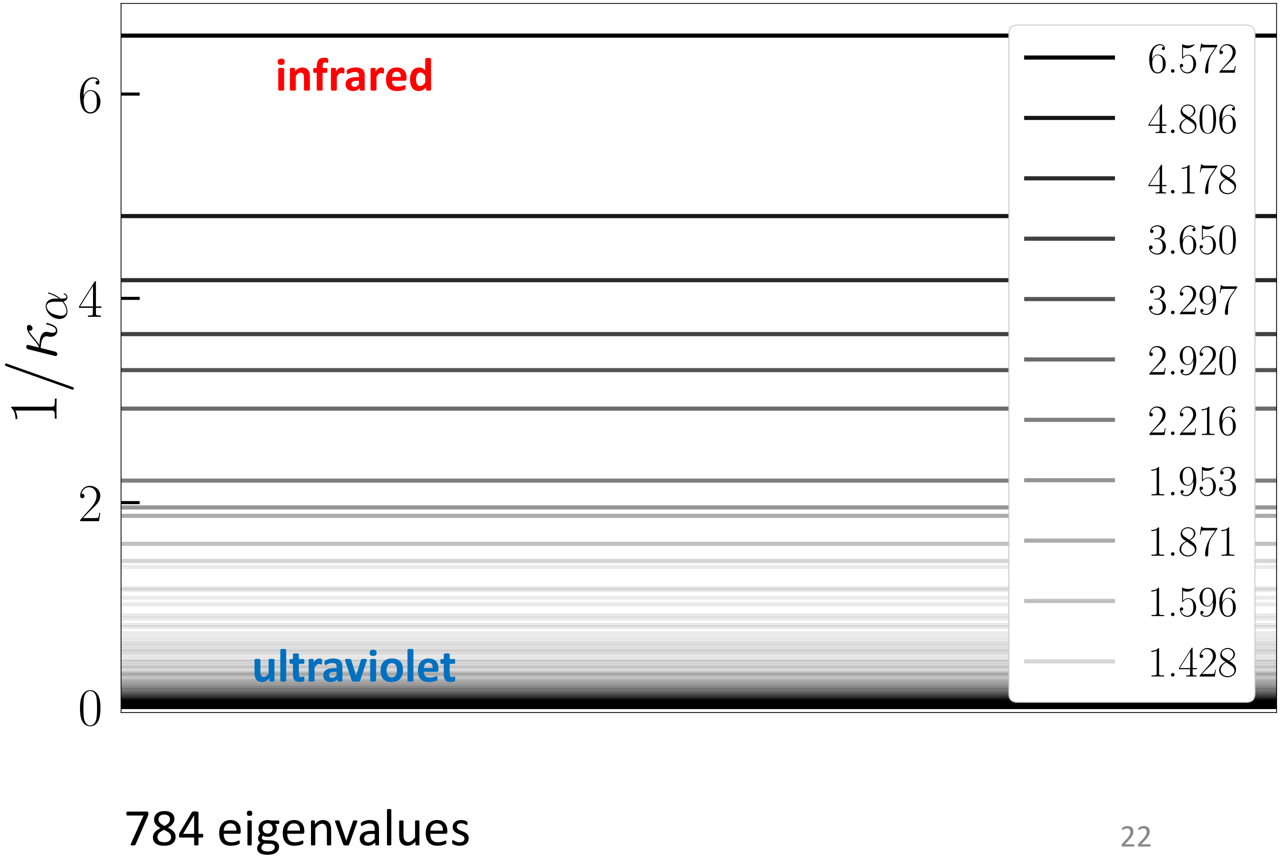
❖ both N_h and μ^2 act as **ultraviolet regulators**

GA, B Lucini, C Park, Phys. Rev. D 109 (2024) 034521 [2309.15002 [hep-lat]]

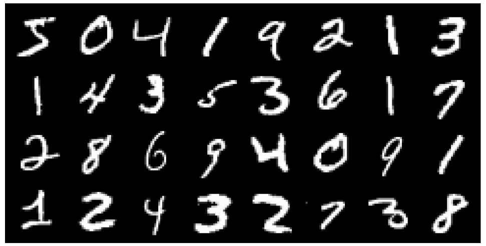
RBM as ultraviolet regulator



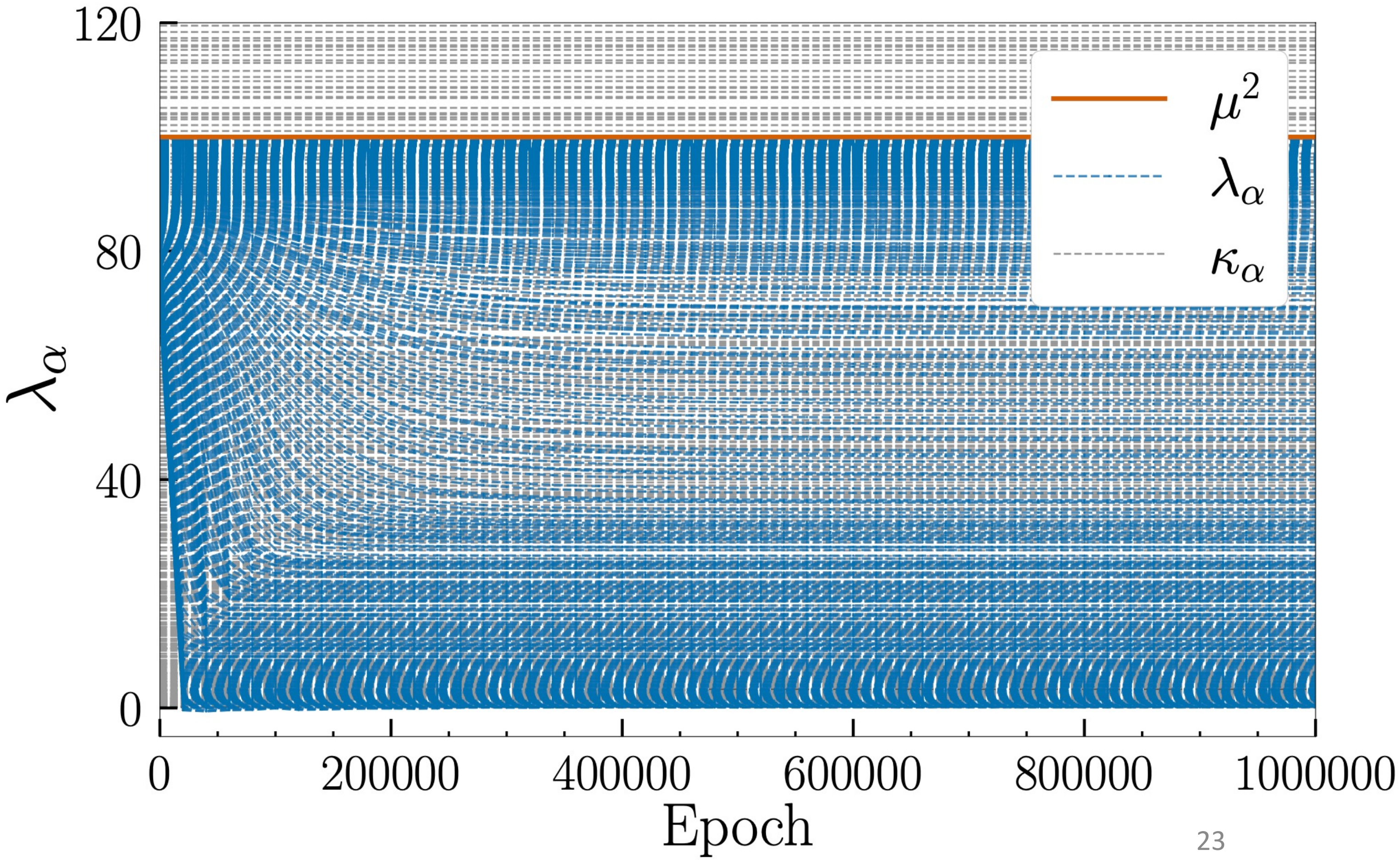
- apply to MNIST data set (28 x 28 images)
- compute spectrum of two-point correlator $K_{ij}^{-1} = \langle \phi_i \phi_j \rangle_{\text{data}}$
- inverse spectrum $1/\kappa$
- infrared safe
- ultraviolet divergent



MNIST with fixed RBM mass



- $N_v = N_h = 784$
- fixed RBM mass $\mu^2 = 100$
- spectrum regulated
- infrared modes learned approximately correctly (see below)



MNIST with $N_h \leq N_v$

what is the effect of including incomplete spectrum?

removal of ultraviolet modes affects generative power

(a) $N_h = 784$

(b) $N_h = 225$

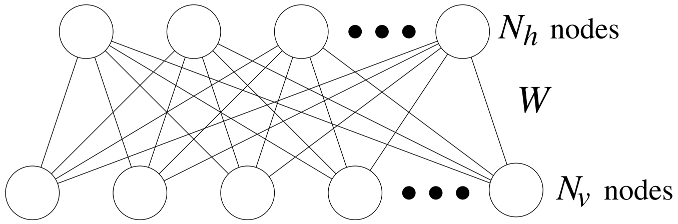
(c) $N_h = 64$

(d) $N_h = 36$

(e) $N_h = 16$

(f) $N_h = 4$

Back to Dyson Brownian motion



- weight matrix is updated using persistent contrastive divergence (PCD)
- maximise likelihood/minimise KL divergence $$\frac{\delta \mathcal{L}}{\delta W_{ia}} = \sigma_h^2 \big(\langle \phi_i \phi_j \rangle_{\text{target}} - \langle \phi_i \phi_j \rangle_{\text{model}} \big) W_{ja}$$
- denote eigenvalues of $X = W^T W$ as x_i
- PCD is stochastic:

$$x_i \to x_i' = x_i + \alpha K_i + \frac{\alpha^2}{|\mathcal{B}|} \sum_{j \neq i} \frac{\tilde{g}_i^2}{x_i - x_j} + \frac{\alpha}{\sqrt{|\mathcal{B}|}} \sqrt{2} \tilde{g}_i \eta_i$$

Back to Dyson Brownian motion

- maximise likelihood/minimise KL divergence $\frac{\delta \mathcal{L}}{\delta W_{ia}} = \sigma_h^2 \big(\langle \phi_i \phi_j \rangle_{\text{target}} - \langle \phi_i \phi_j \rangle_{\text{model}} \big) W_{ja}$

- denote target distribution has eigenvalues with κ_i

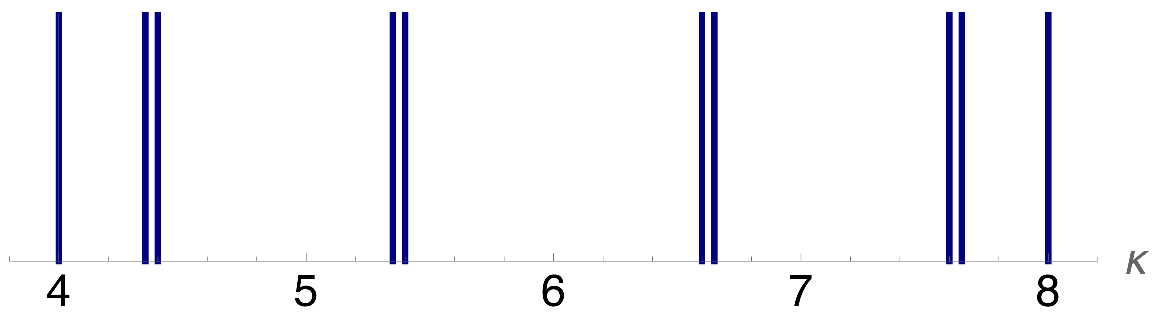
- drift in instantaneous eigen-basis: $K_i(x_i) = \left(\frac{1}{\kappa_i} - \frac{1}{\mu^2 - x_i} \right) x_i$

- fixed point of drift: $x_i^s = \mu^2 - \kappa_i$, spectrum learnt correctly

- where can we observe the effects of RMT?

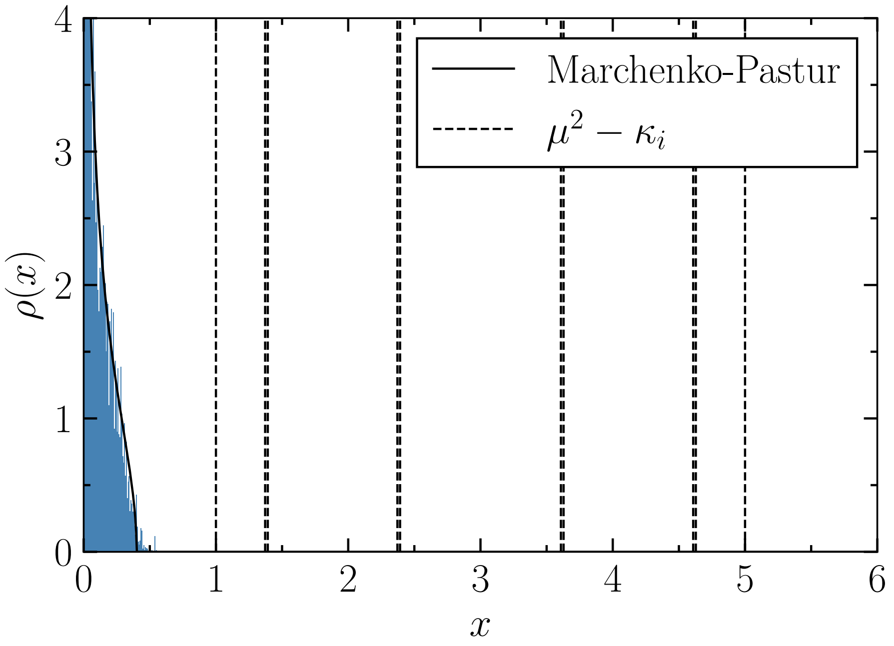
Scalar field RBM

- implement for simple target distribution: scalar field in LFT in 1 dimension
- spectrum is free dispersion relation: $\kappa_k = m^2 + p^2_{\mathrm{lat},k} = m^2 + 2 - 2\cos\left(\frac{2\pi k}{N_v}\right)$
- each mode is doubly degenerate, except lowest and highest one
- example for 10 modes
- degenerate modes split for clarity



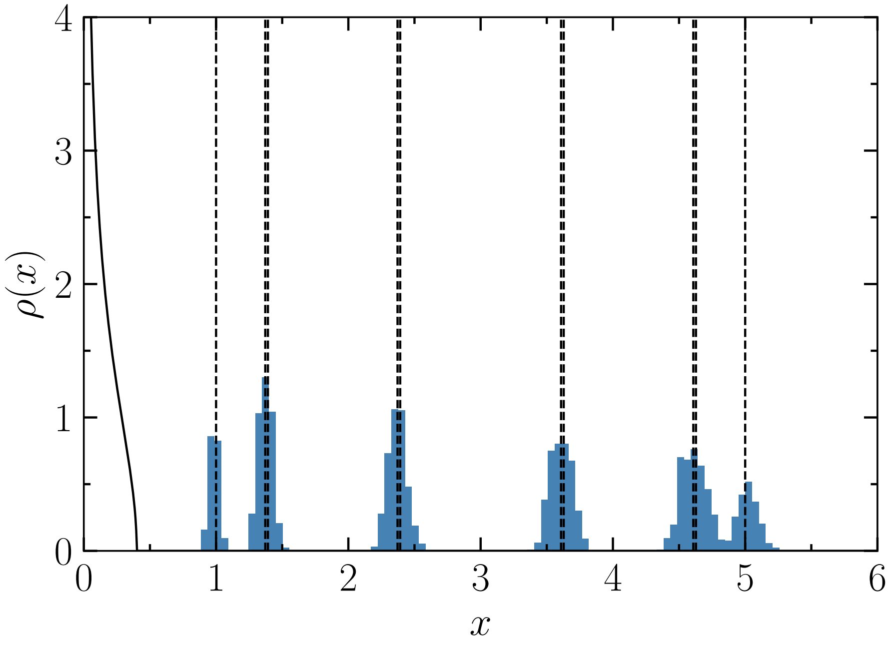
RBM evolution

weight matrix updates using persistent contrastive divergence with mini-batches



initial Marchenko-Pastur distribution

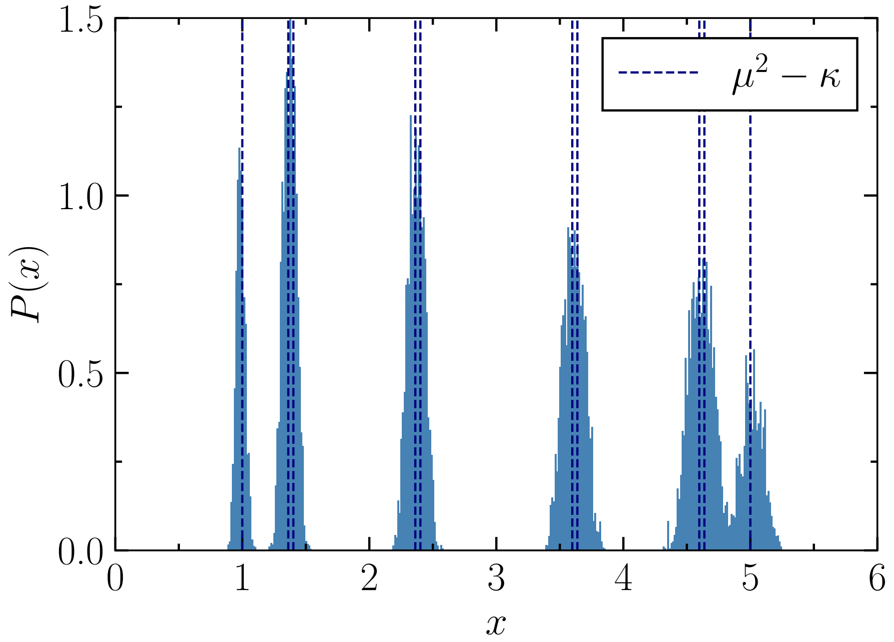
➔



towards target spectrum

RBM evolution and RMT universality

- weight matrix updates using persistent contrastive divergence with mini-batches
- no sharp lines, distributions around target spectrum
- test predictions from RMT:
  - induced Coulomb term and eigenvalue repulsion (universal)
  - potential from drift (non-universal)



Universal RMT predictions

- consider two degenerate modes only: Coulomb gas description

$$Z = \frac{1}{N_0} \int dx_1 dx_2 \, |x_1 - x_2| \, e^{-V(x_1,x_2)} \qquad V(x_1, x_2) = \frac{1}{2\sigma^2} \left[(x_1-\kappa)^2 + (x_2-\kappa)^2 \right]$$

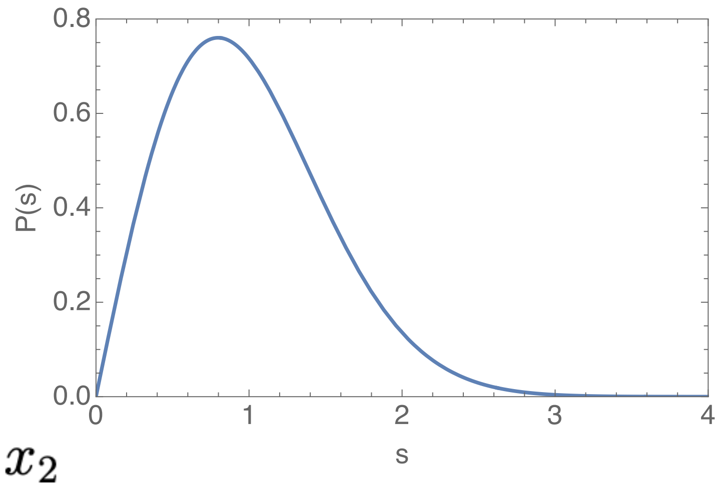
- eigenvalues x_1, x_2 cannot both be equal to κ due to Coulomb repulsion

- two ways to detect this: Wigner surmise and Wigner semi-circle

- Wigner surmise: distribution for level spacing $S = x_1 - x_2$ $\qquad P(S) = \frac{S}{2\sigma^2} e^{-S^2/(4\sigma^2)}$

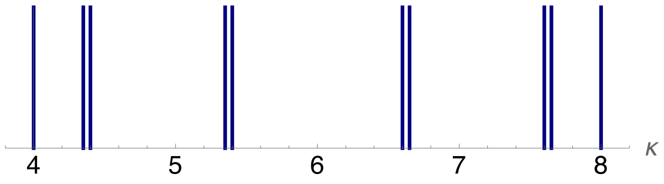
- mean level spacing $\langle S \rangle = \int_0^\infty dS \, S P(S) = \sqrt{\pi}\sigma$ $\qquad s = S/\langle S \rangle$

Wigner surmise



- distribution $P(S) = \frac{S}{2\sigma^2} e^{-S^2/(4\sigma^2)}$ for level spacing $S = x_1 - x_2$
- mean level spacing $\langle S \rangle = \int_0^\infty dS\, S P(S) = \sqrt{\pi}\sigma$
- Wigner surmise for $s = S/\langle S \rangle$: $P(s) = \frac{\pi}{2} s e^{-\pi s^2/4}$ universal curve

- many RBM training runs, stochasticity due to mini-batches, collect histograms of x_i
- vary learning rate and batch size [no ordering of eigenvalues by hand, induces bias!]

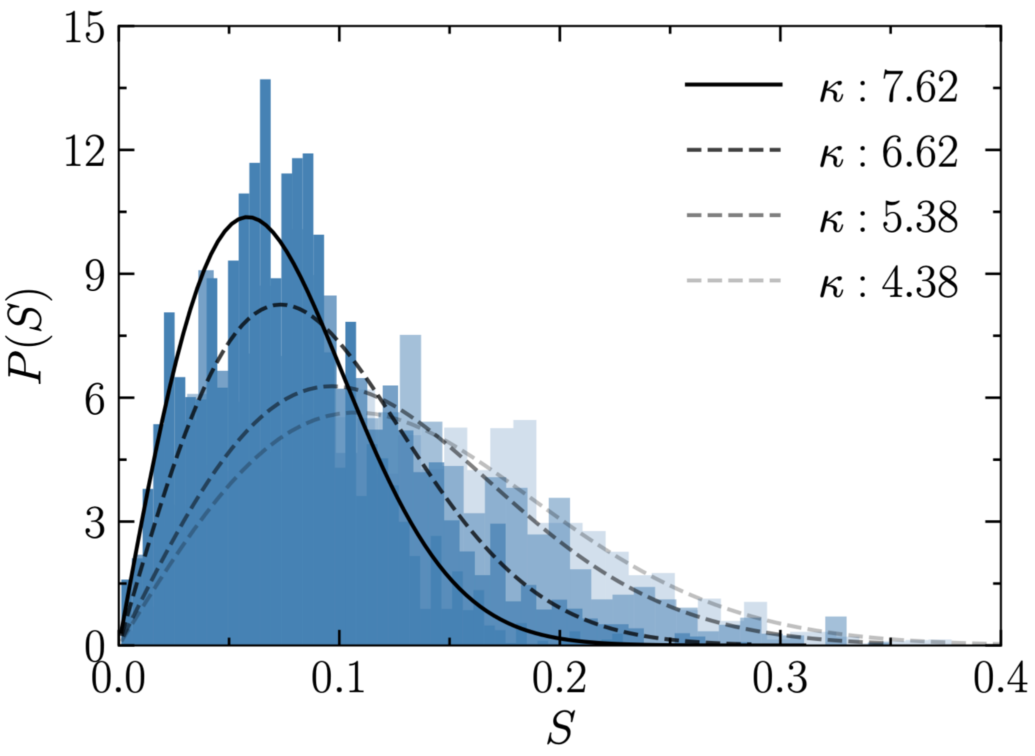


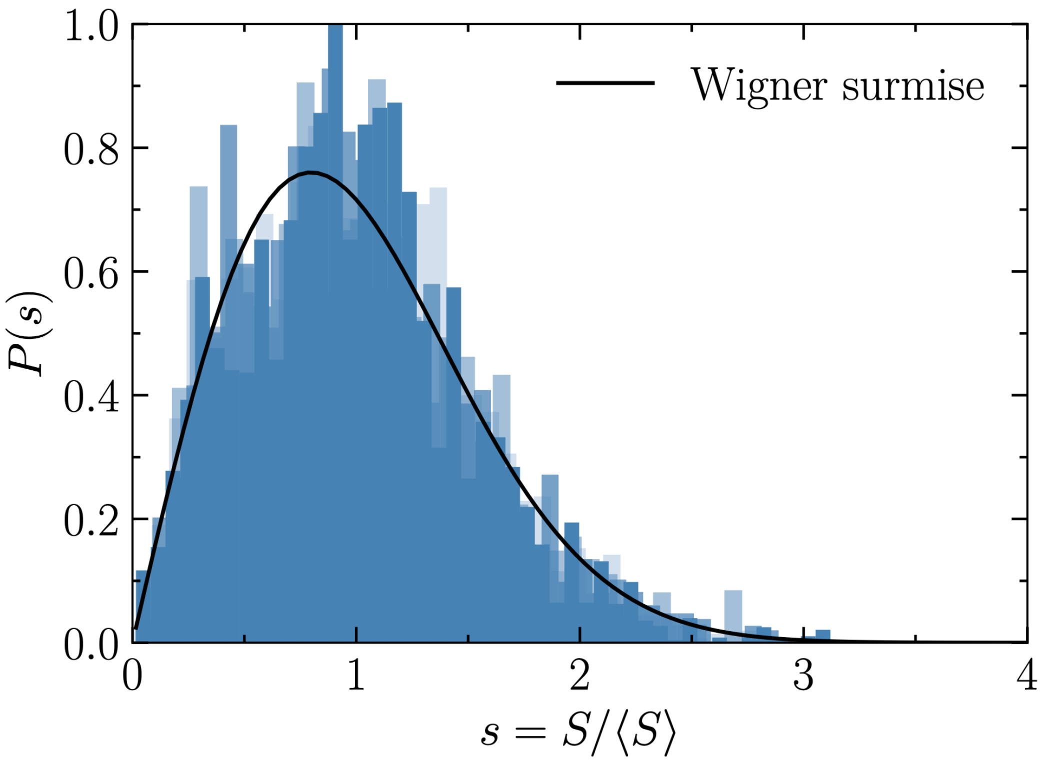
Wigner surmise: 4 degenerate pairs

$$P(S) = \frac{S}{2\sigma^2} e^{-S^2/(4\sigma^2)}$$

$$\langle S \rangle = \sqrt{\pi}\sigma$$

$$P(s) = \frac{\pi}{2} s e^{-\pi s^2/4}$$





data collapse

universality

Wigner surmise: vary learning rate and batch size

- prediction:

$$\sigma_i^2 = (\alpha/|\mathcal{B}|)\ (\tilde{g}_i^2/\Omega_i)$$

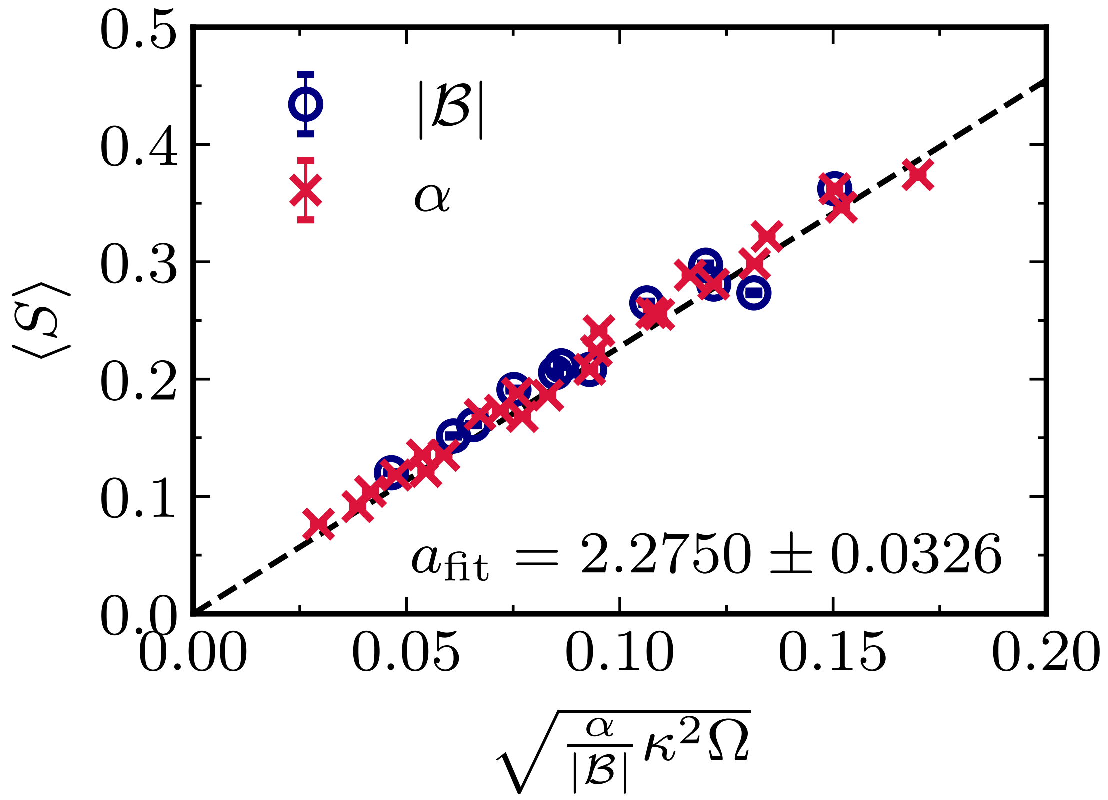
- linear dependence on $(\alpha/|\mathcal{B}|)$

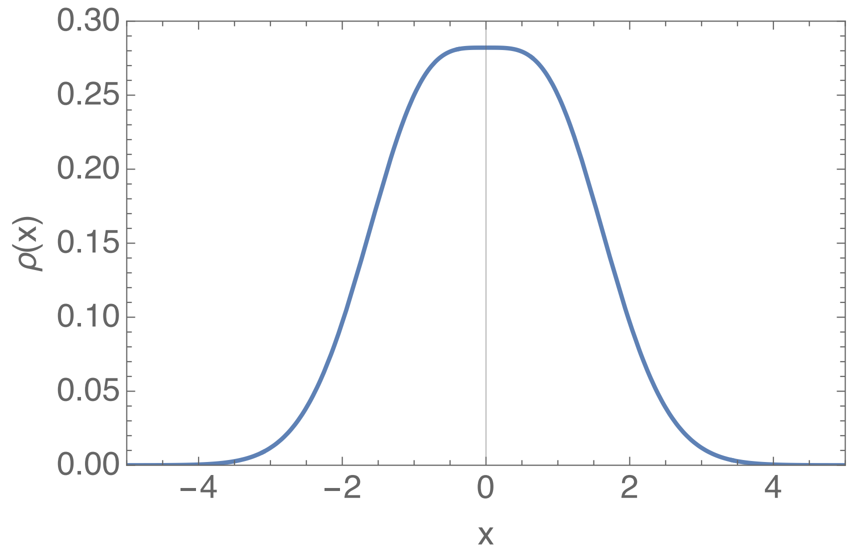
- mean level spacing

$$\langle S_i \rangle = \pi\sqrt{(\alpha/|\mathcal{B}|)(\tilde{g}_i^2/\Omega_i)}$$
$$= a_{\text{fit}}\sqrt{(\alpha/|\mathcal{B}|)(\kappa_i^2\Omega_i)}$$

- fit function includes

  non-universal parameters as well





Wigner semi-circle

- spectral density: $\rho(x) = \left\langle \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i) \right\rangle$

- for two modes:

$$\rho(x) = \frac{e^{-x^2/(2\sigma^2)}}{4\sqrt{\pi}\sigma} \left[2e^{-x^2/(2\sigma^2)} + \sqrt{2\pi}\frac{x}{\sigma}\mathrm{Erf}\left(\frac{x}{\sqrt{2}\sigma}\right) \right]$$

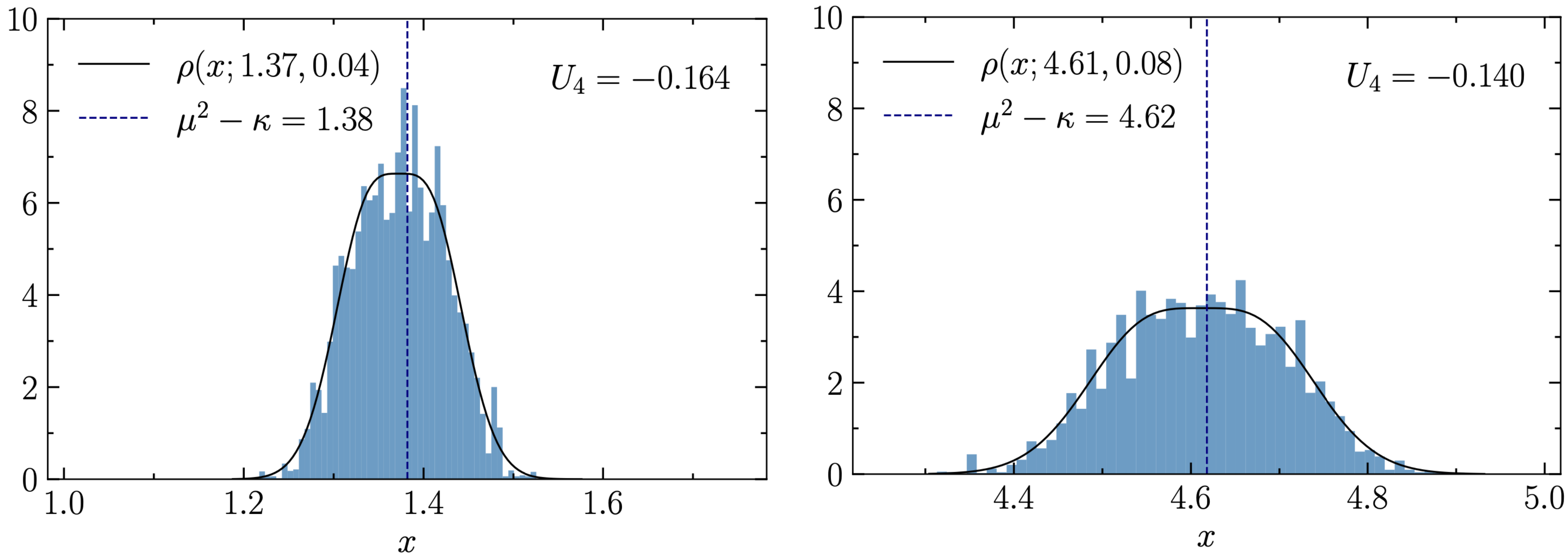
- broadened and flattened Gaussian

- fit σ parameter and position for each doubly degenerate mode

$$\rho(x) = \frac{e^{-x^2/(2\sigma^2)}}{4\sqrt{\pi}\sigma}\left[2e^{-x^2/(2\sigma^2)} + \sqrt{2\pi}\frac{x}{\sigma}\mathrm{Erf}\left(\frac{x}{\sqrt{2}\sigma}\right)\right]$$

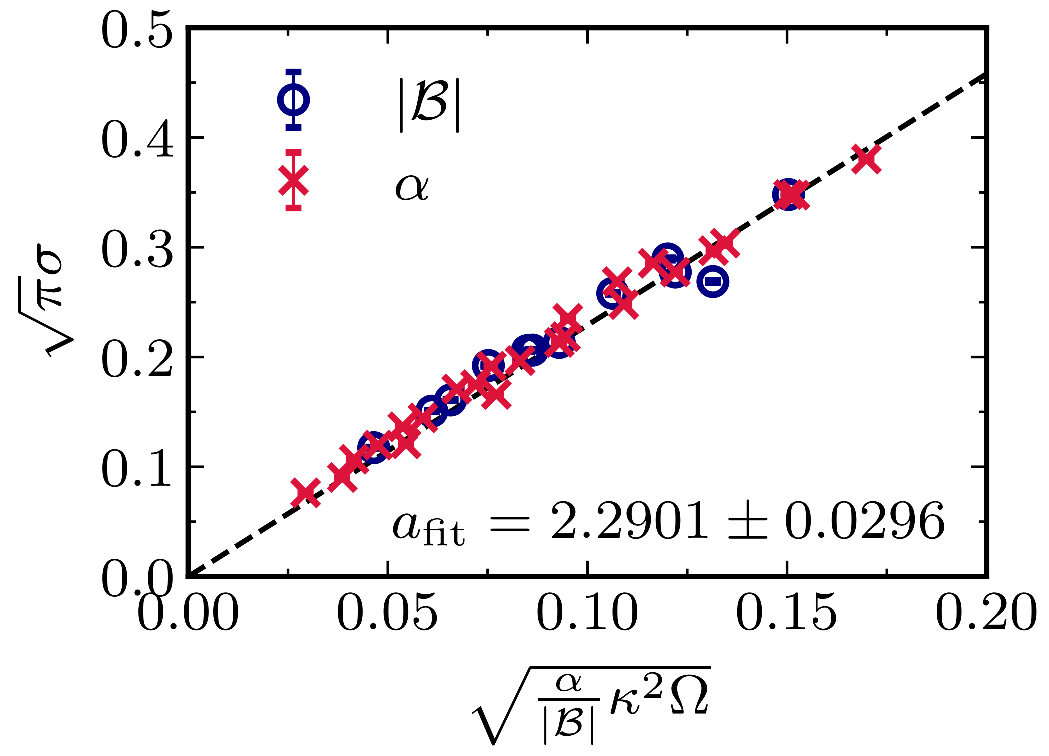
Wigner semi-circle

- fit to semi-circle for two different κ_i values with fixed learning rate and batch size
- Binder cumulant $U_4 = -4/27 \approx -0.148$ for semi-circle (vanishes for Gaussian)

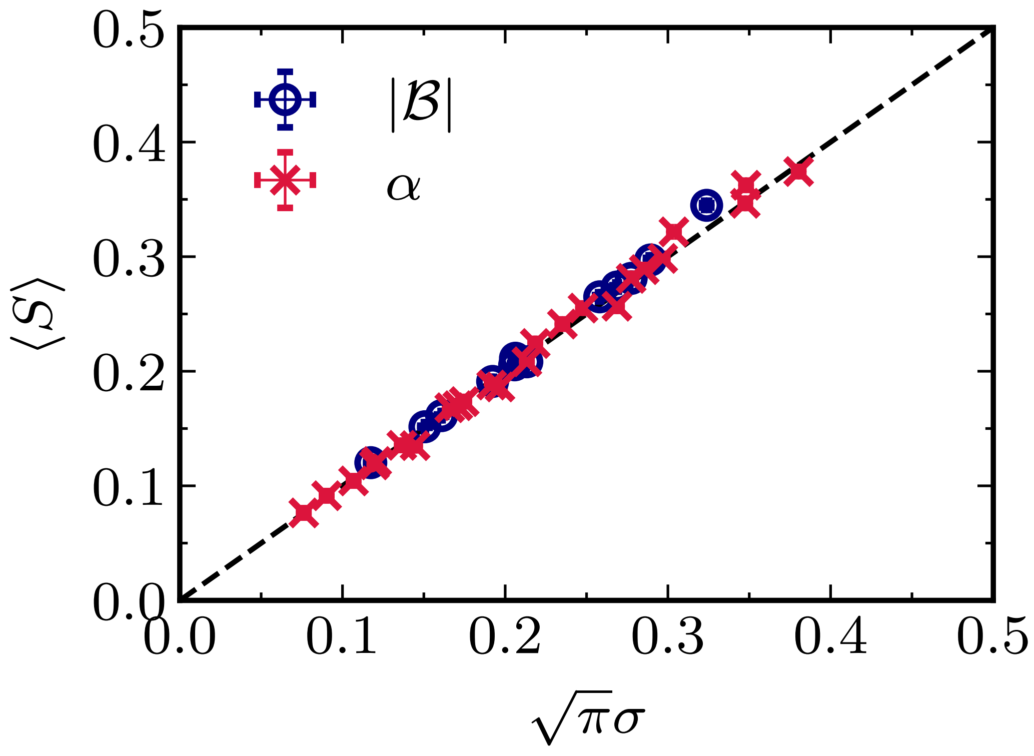


Wigner semi-circle and surmise

semi-circle
dependence on learning rate/batch size



consistency between surmise
and semi-circle fits



Wigner surmise and semi-circle

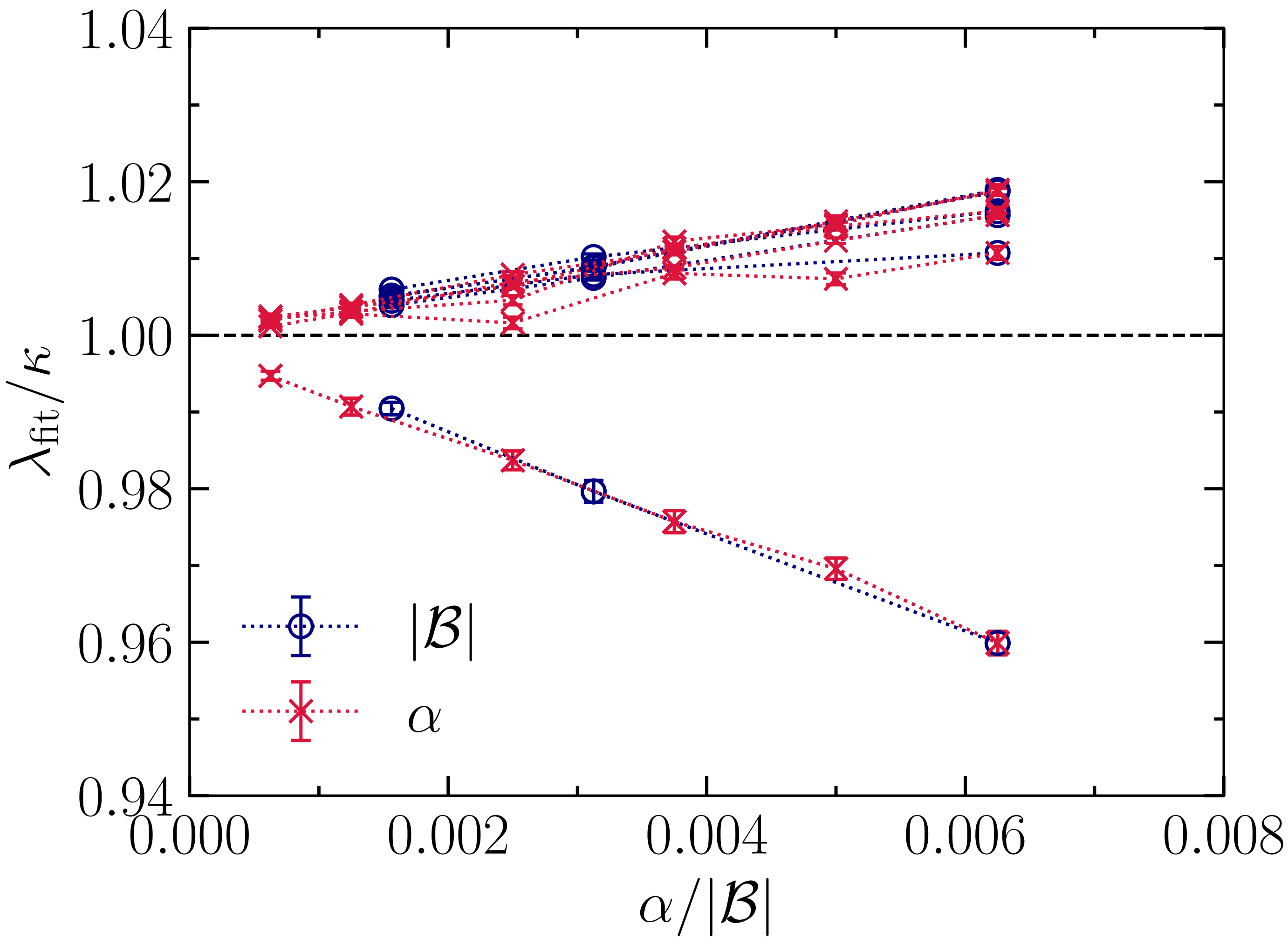
- ✓ parameter σ scales as: $\sigma_i^2 = (\alpha/|\mathcal{B}|)\ \ (\tilde{g}_i^2/\Omega_i)$

  universal scaling   model-dependent

- ✓ stochasticity leads to universal features in trained models
- ✓ derived that learning rate and finite batch size appear as ratio
- ✓ previously observed as empirical linear scaling rule

Eigenvalue repulsion

- Coulomb interaction between all eigenvalues
- learned eigenvalue/target
- repulsion for nonzero learning rate/batch size
- no “perfect learning” unless stochasticity vanishes
- overfitting, generalisation, …



Non-universal dynamics

- distribution $P_s(\{x_i\}) = \frac{1}{Z} \prod_{i<j} |x_i - x_j| \, e^{-\sum_i V_i(x_i)/g_i^2}$
- RBM specific drift $K_i(x_i) = \left(\frac{1}{\kappa_i} - \frac{1}{\mu^2 - x_i} \right) x_i$ determines potential $V_i(x_i)$
- consider this for one mode only (drop the index)

$$V(x) = -\int^x dx' \, K(x') = -\frac{x^2}{2\kappa} - x - \mu^2 \log\left(\mu^2 - x\right)$$

- stationary distribution

$$P_s(x) = \frac{1}{Z} e^{-V(x)/g^2} = \frac{1}{Z} \exp\left[\frac{1}{g^2} \left(\frac{x^2}{2\kappa} + x + \mu^2 \log\left(\mu^2 - x\right) \right) \right]$$

Time-dependent dynamics

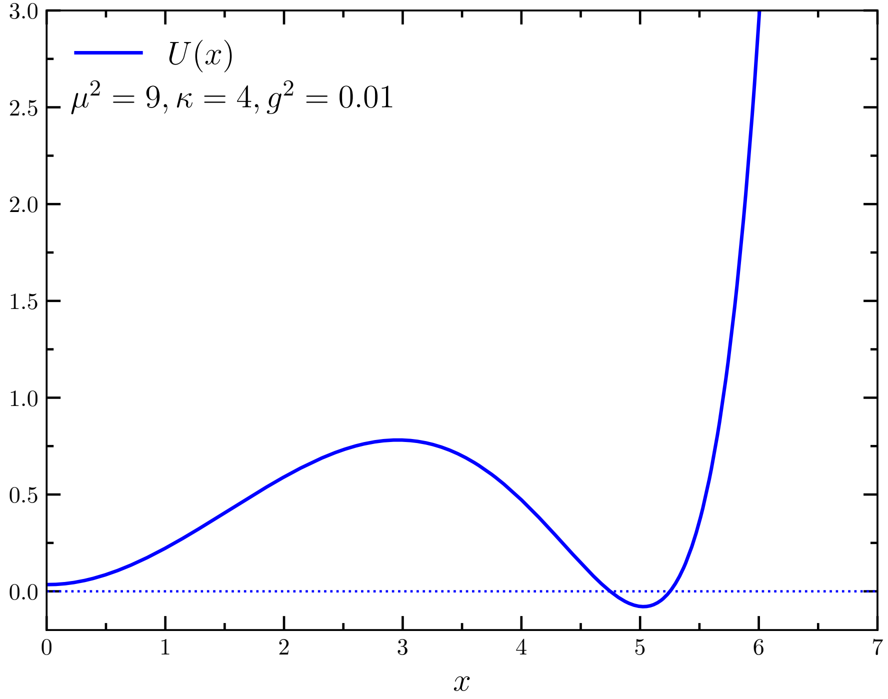
- assume continuous time limit exists
- analyse FPE for one mode: $\partial_\tau P(x,\tau) = \partial_x \left(g^2\partial_x - K(x)\right) P(x,\tau)$
- solve using standard stochastic quantisation/FP methods: $P(x,\tau) = \sqrt{P_s(x)}\psi(x,\tau)$
- evolution: $\partial_\tau \psi(x,\tau) = \left(g^2\partial_x^2 - \frac{1}{4g^2}\left[\partial_x V(x)\right]^2 + \frac{1}{2}\left[\partial_x^2 V(x)\right]\right)\psi(x,\tau) \equiv -2H_{\mathrm{FP}}\psi(x,\tau)$
- Fokker-Planck Hamiltonian: $H_{\mathrm{FP}} = \frac{1}{2}L^\dagger L$

$$L^\dagger = -g\partial_x + \frac{1}{2g}\partial_x V(x)$$

$$L = +g\partial_x + \frac{1}{2g}\partial_x V(x)$$

Quantum-mechanical bound state problem

- $H_{\mathrm{FP}} = \frac{1}{2} L^\dagger L$ eigenvalue problem: $H_{\mathrm{FP}} \psi_n(x) = E_n \psi_n(x)$
- explicit form: $H_{\mathrm{FP}} = -\frac{g^2}{2} \partial_x^2 + U(x)$



$$U(x) = \frac{1}{g^2} \left[U_0(x) + g^2 U_1(x) \right]$$

$$U_0(x) = \frac{1}{8} \left[\partial_x V(x) \right]^2$$

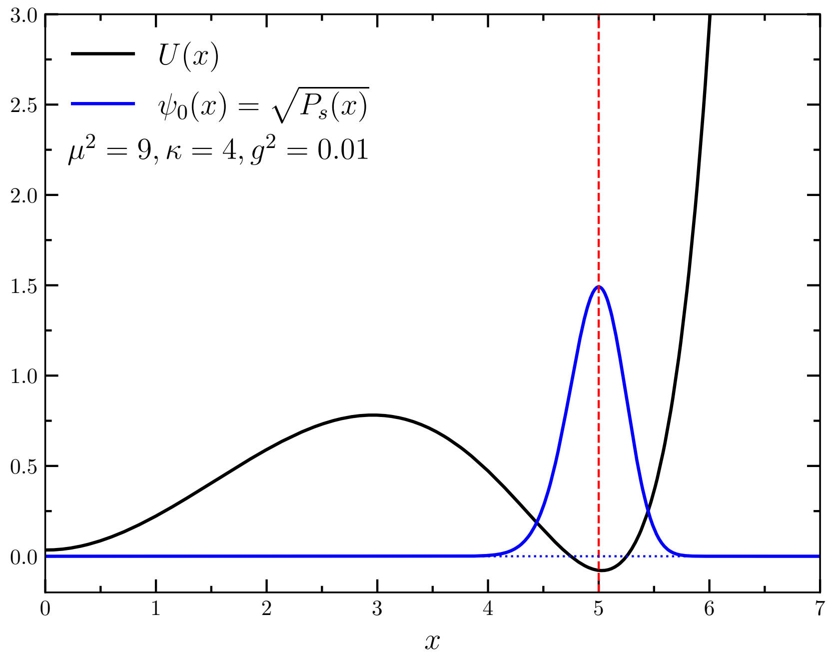
$$U_1(x) = -\frac{1}{4} \partial_x^2 V(x)$$

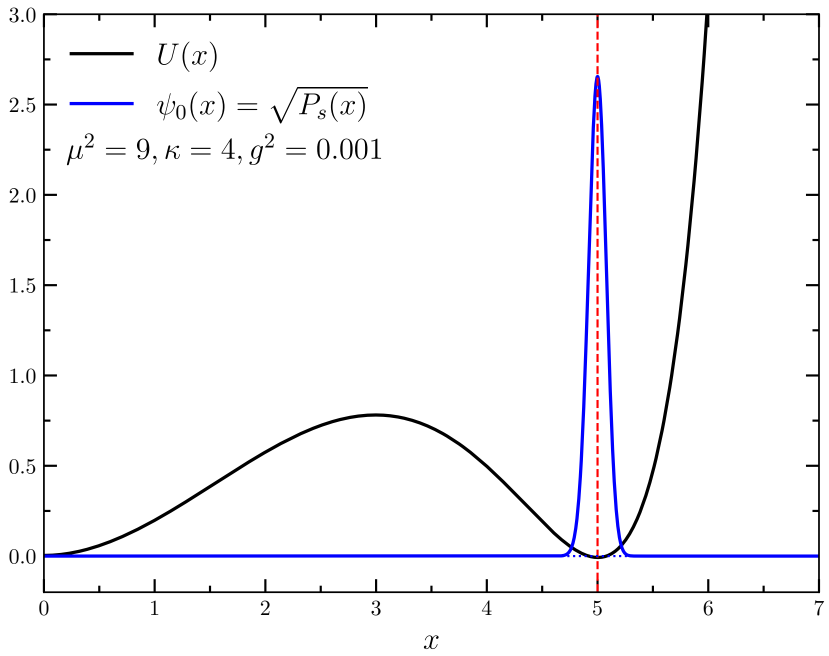
double well potential on interval $0 \leq x \leq \mu^2$

Quantum-mechanical bound state problem

- $H_{\text{FP}}\psi_n(x) = E_n\psi_n(x)$ ground state exactly known: $\psi_0(x) = \sqrt{P_s(x)}$
- width of solution depends on strength of the noise g^2 : better description of target κ

$g^2 = 0.01$

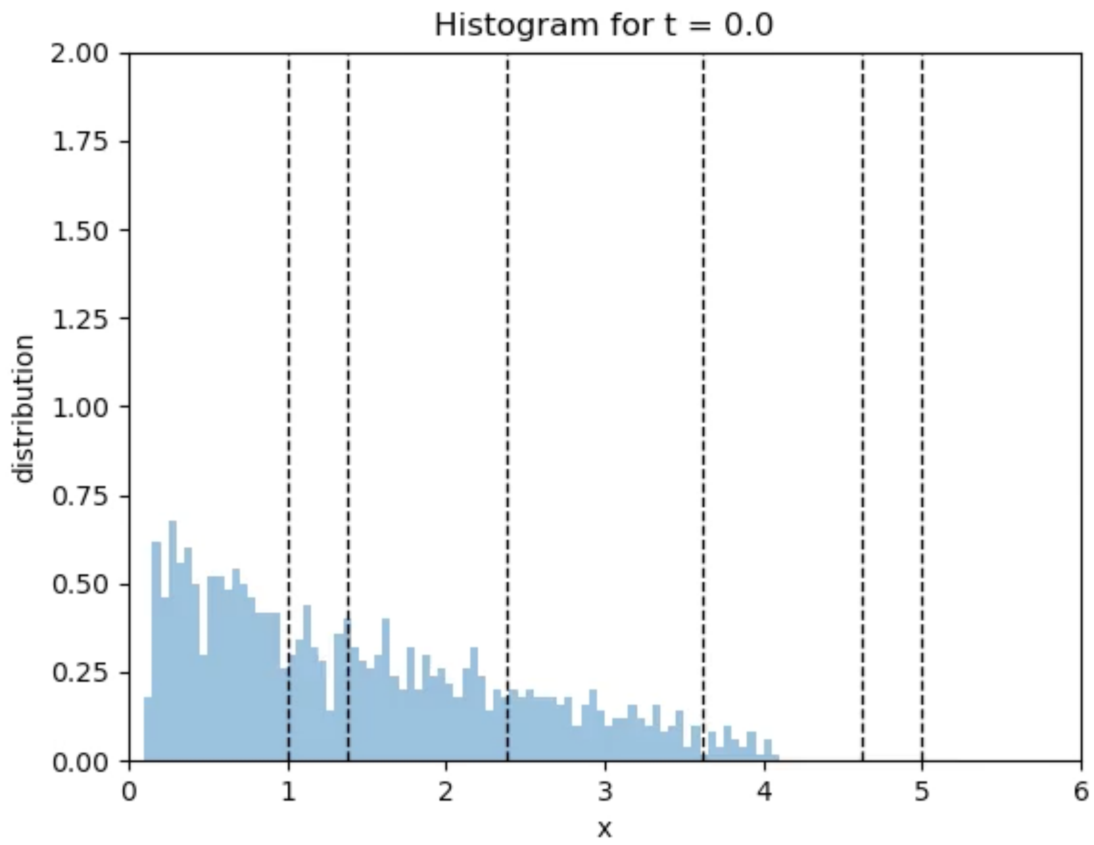




$g^2 = 0.001$

Full time-dependent dynamics: learning

- combine Coulomb repulsion and drift
- from Marchenko-Pastur distribution
  to stochastic target distribution
- 10 modes, 4 doubly degenerate ones
- dynamics of $P(\{x_i\}, t)$ described by FPE



- effective description of learning dynamics in terms of eigenvalues

Second application: Transformers

- Gaussian RBM has one weight matrix, target spectrum is known, essentially solvable

in more advanced architectures:

- many weight matrices, target spectra not known, do they even exist?

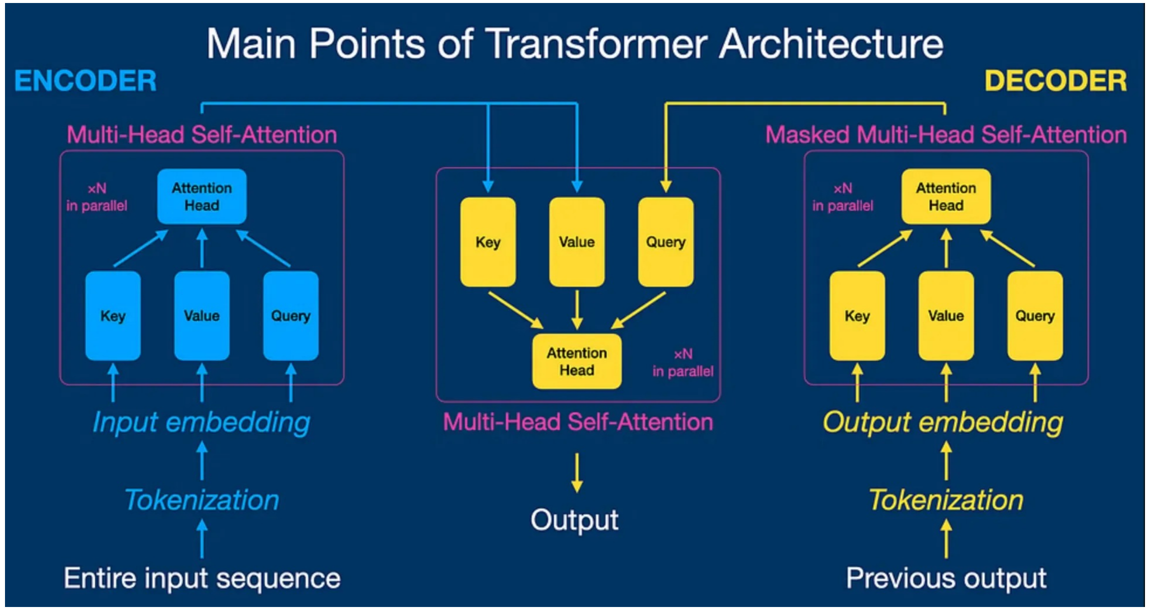
- what is the loss function landscape? localised minima, flat directions, … ?

- empirical study following dynamics of eigenvalues of $X = W^T W$

GA, O Hajizadeh, B Lucini, C Park, NeurIPS 2024 workshop *ML and the Physical Sciences*, 2411.13512 [cond-mat.dis-nn]

Transformer: nano-GPT

- four attention blocks with each four attention heads
- each attention head: one key (K), one query (Q) and one value (V) matrix
- matrix sizes: $M \times N = 64 \times 16$
- about 2.1×10^5 parameters
- use AdamW optimiser
  (highly adaptive stepsize during training)
- trained on opus of Shakespeare

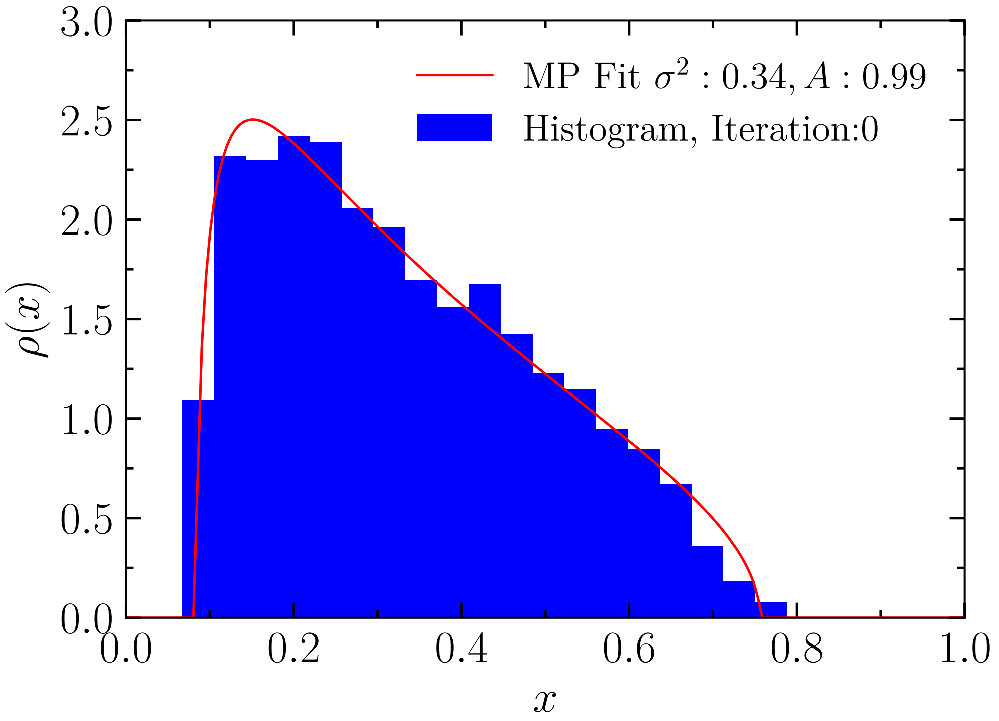


https://botpenguin.com/glossary/transformer-architecture

Transformer: empirical analysis

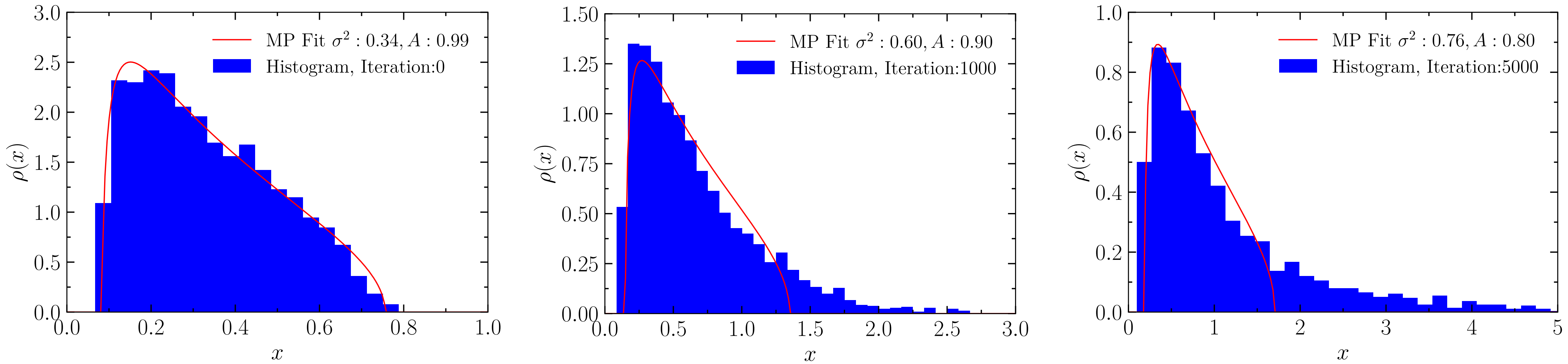
- initialisation: eigenvalues of $X = W^T W$ follow Marchenko-Pastur distribution

$$P_{\mathrm{MP}}(x; \sigma^2, A) = \frac{A}{2\pi\sigma^2 r x}\sqrt{(x_+ - x)(x - x_-)}\,\theta(x_+ - x)\theta(x - x_-)$$



Transformer: empirical analysis

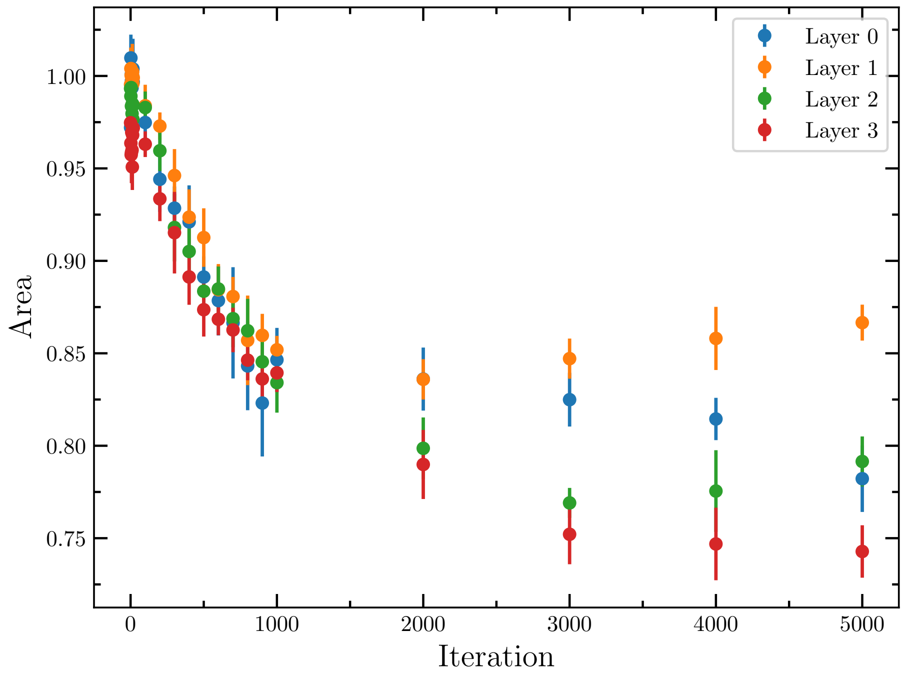
- appearance of tail in distribution (shown K matrix of layer 1)
- part of spectrum described by Marchenko-Pastur distribution is reduced, $A < 1$
- use area A and width σ^2 as fit parameters during evolution



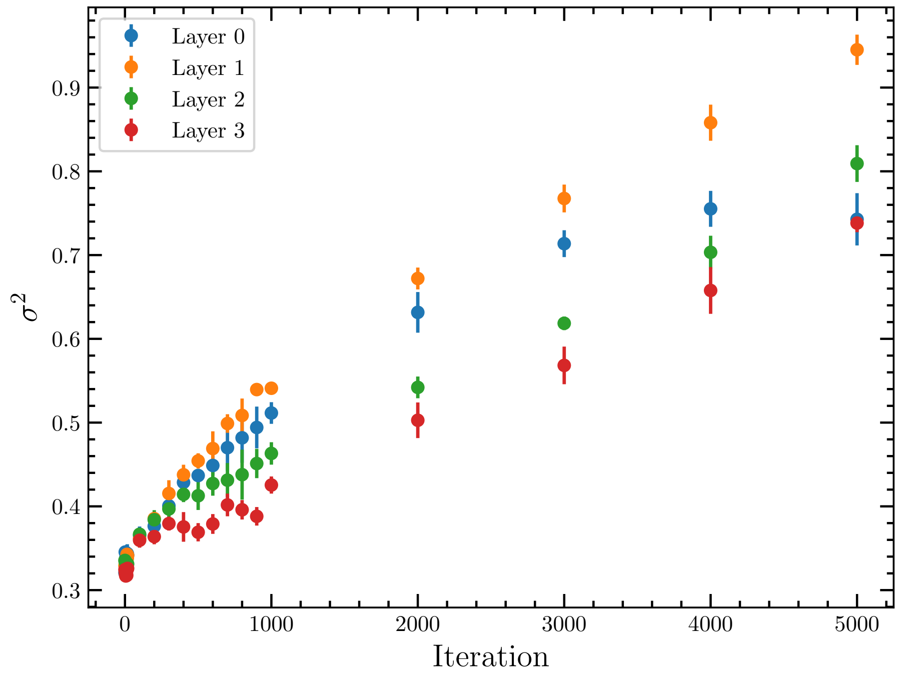
iteration 0 iteration 1000 iteration 5000

Transformer: empirical analysis

- evolution of area A and width σ^2 during evolution (shown K matrix for all four layers)



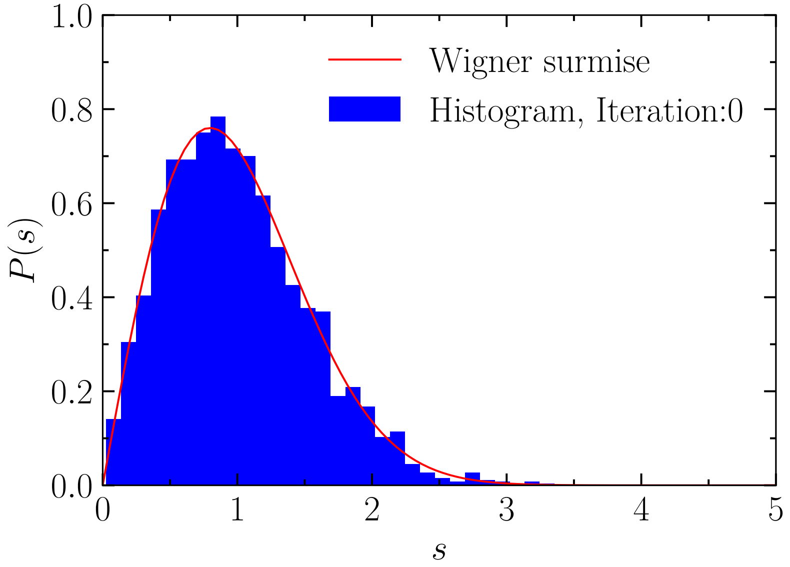
15-25% of spectral weight moves to the tail



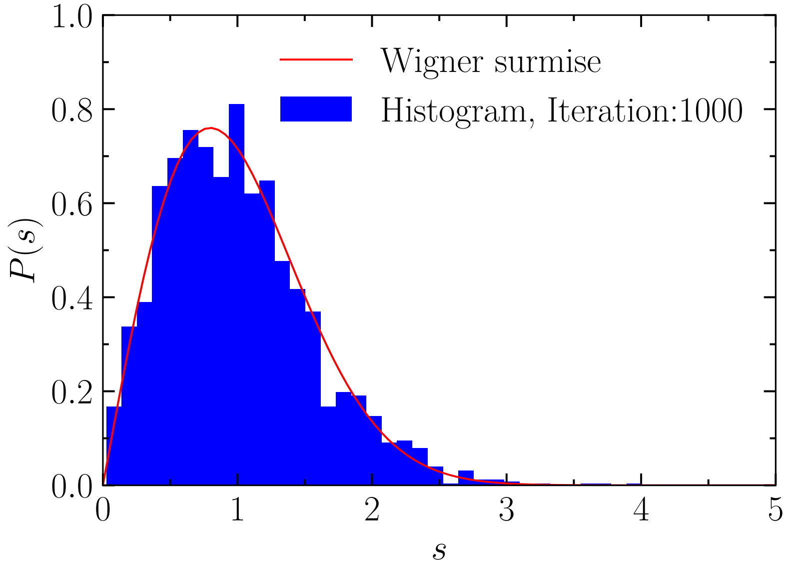
MP distribution broadens due to Brownian motion

Transformer: Wigner surmise

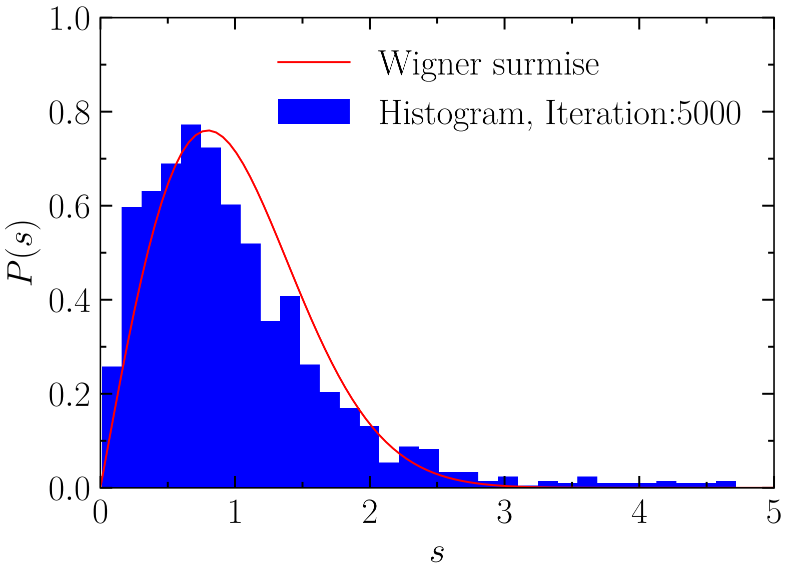
- short-distance fluctuations: level spacing described by Wigner surmise
- remains approximately described by RMT for real, symmetric matrices



iteration 0



iteration 1000



iteration 5000

Transformer: empirical analysis

requires further understanding:

- what is the “final” target spectrum? does it even exist?
- tail drops as a power, what does this imply? can the power be understood?

significant part of the spectrum remains MP: random matrix elements

- how relevant is this part of the spectrum? remove? sparse weight matrices?

see also CH Martin, MW Mahoney, *Traditional and Heavy-Tailed Self Regularization in Neural Network Models*, 1901.08276

Summary

- stochastic weight matrix dynamics has universal features described by RMT
- manifests in eigenvalue repulsion, quantified by Wigner surmise and semi-circle
- fundamental limitation of learning for finite learning rate and batch size
- stochasticity controlled by learning rate/batch size: reduce ratio to improve agreement with target distribution, but stochasticity allows for generalisation

Outlook

- Dyson Brownian motion is present at “microscopic” level
- how does it manifest itself for more advanced architectures?
- is there universality beyond level repulsion (power law tails)?
- what are the practical implications? description of learning, algorithmic advances?