

# Weight matrix dynamics and Dyson Brownian motion

Gert Aarts

with **Chanju Park** and Biagio Lucini

[2407.16427](#) [cond-mat.dis-nn]



# Motivation

- further understanding learning: weight matrix dynamics
- make connection with random matrix theory (RMT)
- identify universal behaviour during and after training
- describe universal dependence on and fundamental limitations for finite learning rate and batch size in terms of RMT

# 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)
- summary and outlook

# Stochastic weight matrix dynamics

- consider some  $M \times N$  weight matrix  $W$
- update (e.g. stochastic gradient descent):  $W \rightarrow W' = W + \delta W$  with  $\delta W = -\alpha \frac{\delta \mathcal{L}}{\delta W}$
- obtained from loss function  $\mathcal{L}[W]$ , learning rate  $\alpha$
- $\delta W$  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):  $\frac{1}{|\mathcal{B}|} \text{Var}(\delta W)$

# Stochastic weight matrix dynamics

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

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

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

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

# 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:  $\xi_i \quad (i = 1 \dots N)$  [take  $N \leq M$  without loss of generality]

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

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

$$D = \Xi^T \Xi = \text{diag}(\xi_1^2, \dots, \xi_N^2) = \text{diag}(x_1, \dots, x_N)$$

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

# Initialisation: Marchenko-Pastur distribution

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

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

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

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

$$P_{\text{MP}}(x) = \frac{1}{2\pi r x} \sqrt{(x_+ - x)(x - x_-)} \quad 0 \leq x_- \leq x \leq x_+ \leq 4 \quad x_{\pm} = (1 \pm \sqrt{r})^2$$

# 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

$$\begin{aligned}\frac{dx_i}{dt} &= K_i(x_i) + \sum_{j \neq i} \frac{g_i^2}{x_i - x_j} + \sqrt{2}g_i\eta_i \\ &\equiv K_i^{(\text{eff})}(x_i) + \sqrt{2}g_i\eta_i\end{aligned}$$

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

# 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: 
$$\partial_t P(\{x_i\}, t) = \sum_{i=1}^N \partial_{x_i} \left[ \left( g_i^2 \partial_{x_i} - K_i^{(\text{eff})}(x_i) \right) \right] P(\{x_i\}, t)$$
- 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: 
$$Z = \int dx_1 \dots dx_N P_s(\{x_i\})$$
- and provided drift can be derived from a potential 
$$K_i(x_i) = -\frac{dV_i(x_i)}{dx_i}$$
- known as Coulomb gas, describes universal features of random matrices

# Back to weight matrix dynamics

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

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

# Back to weight matrix dynamics

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

- insert learning rate and batch size dependence:

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

- no usual scaling of drift and noise with learning rate (Ito calculus:  $\epsilon, \sqrt{\epsilon}$ ):  
no obvious continuous time limit (SDE)
- known issue: from SGD to SDE [see e.g. Yaida, ArXiv:1810.00004] but is in fact blessing (see below)

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

# 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

$$g_i = \frac{\alpha}{\sqrt{|\mathcal{B}|}} \tilde{g}_i$$

$$V_i(x_i) = \alpha \tilde{V}_i(x_i)$$

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

- if drift vanishes at  $x_i = x_i^s$ , expand potential 
$$\tilde{V}_i(x_i) = \tilde{V}_i(x_i^s) + \frac{1}{2} \Omega_i (x_i - x_i^s)^2 + \dots$$

- 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

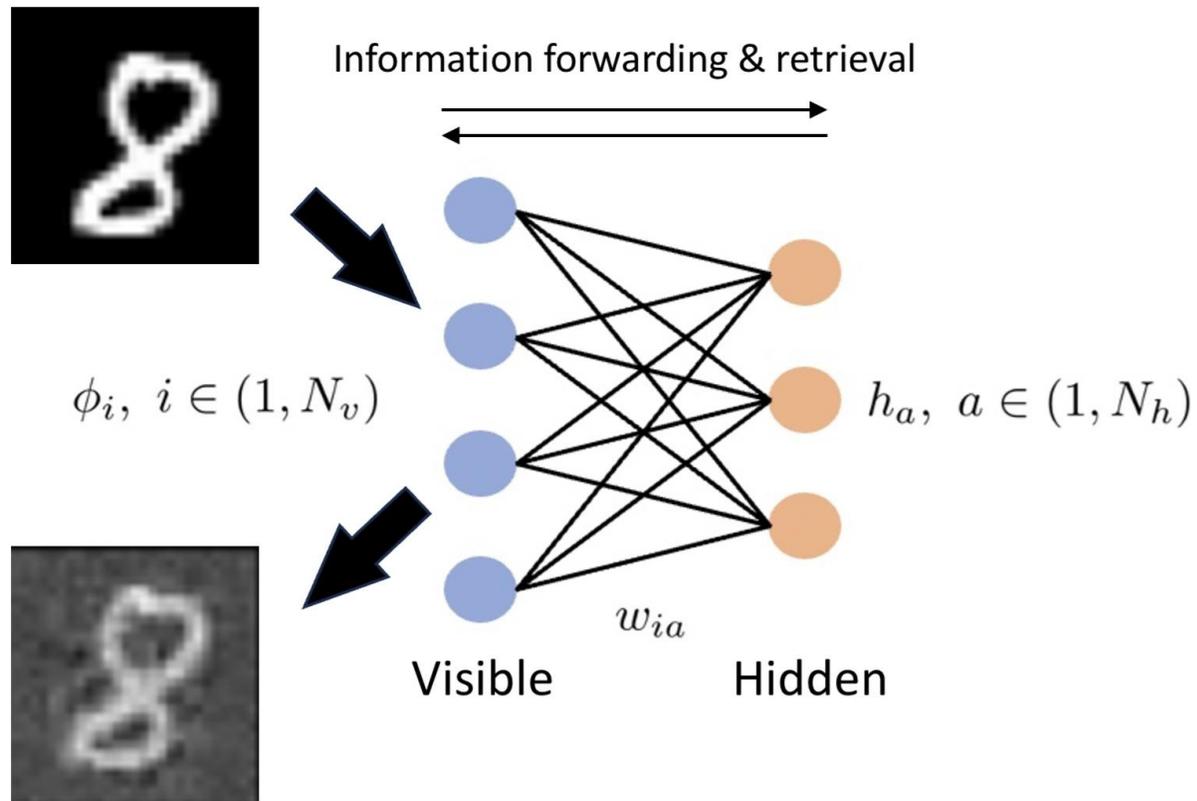
- so far, the derivation is general: prediction of eigenvalue distribution after learning
- apply to actual ML models to observe universal features and support derivation
  - teacher-student model
  - Gaussian Restricted Boltzmann Machine

building on previous analysis of RBM:

Chanju Park, Biagio Lucini, GA, Phys. Rev. D 109 (2024) 034521 [[2309.15002](#)] [hep-lat]

current paper: Chanju Park, Biagio Lucini, GA, [2407.16427](#) [cond-mat.dis-nn]

# Restricted Boltzmann Machine: generative network

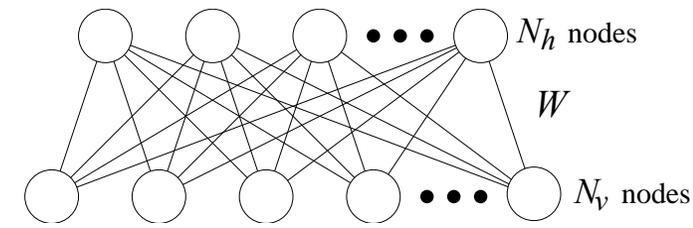


- energy-based method
- probability distribution
- binary or continuous d.o.f.

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

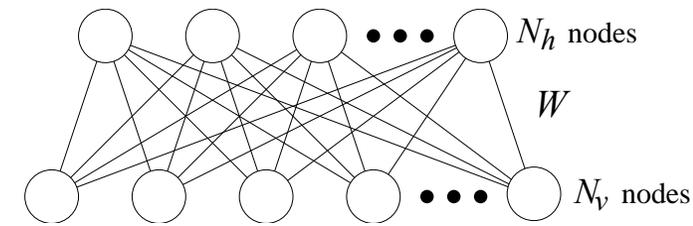
$$Z = \int D\phi D h e^{-S(\phi, h)}$$

# Scalar field RBM



- distribution:  $p(\phi, h) = \frac{1}{Z} e^{-S(\phi, h)}$        $S(\phi, h) = \frac{1}{2} \mu^2 \phi^T \phi + \frac{1}{2\sigma_h^2} (h - \eta)^T (h - \eta) - \phi^T W h$
- $M \times N = N_v \times N_h$  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  $K = \mu^2 \mathbb{1} - \sigma_h^2 W W^T = \mu^2 \mathbb{1} - \sigma_h^2 U \Xi \Xi^T U^T = U [\mu^2 \mathbb{1} - \sigma_h^2 \Xi \Xi^T] U^T \equiv U D_K U^T$
- eigenvalues  $D_K = \text{diag} \left( \underbrace{\mu^2 - \sigma_h^2 \xi_1^2, \mu^2 - \sigma_h^2 \xi_2^2, \dots, \mu^2 - \sigma_h^2 \xi_N^2}_N, \underbrace{\mu^2, \dots, \mu^2}_{M-N} \right)$

# Scalar field RBM



- weight matrix is updated using persistent contrastive divergence (PCD)

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

- target distribution has eigenvalues  $\kappa_i$ , eigenvalues of  $X = W^T W$  denoted as  $x_i$

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

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

# Scalar field RBM

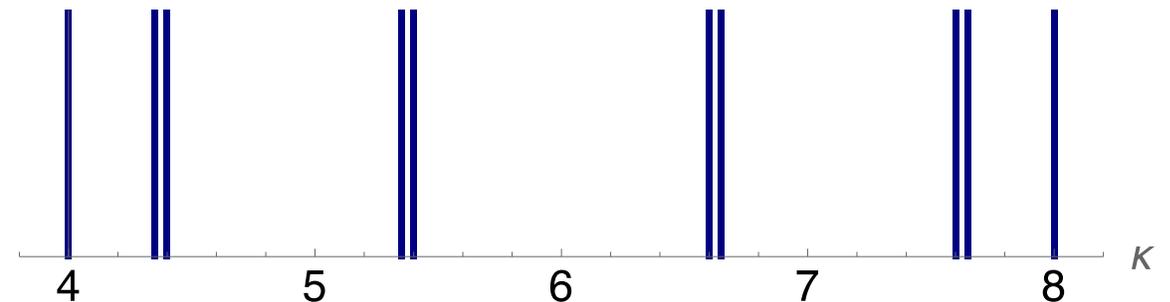
- implement for simple target distribution: scalar field in LFT in 1 dimension

- spectrum is free dispersion relation:  $\kappa_k = m^2 + p_{\text{lat},k}^2 = m^2 + 2 - 2 \cos\left(\frac{2\pi k}{N_v}\right)$

- each mode is doubly degenerate, except lowest and highest one

- example for 10 modes

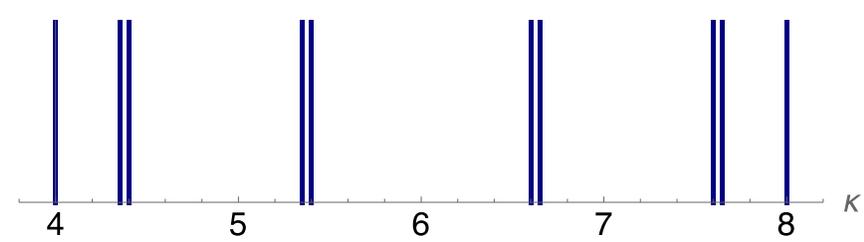
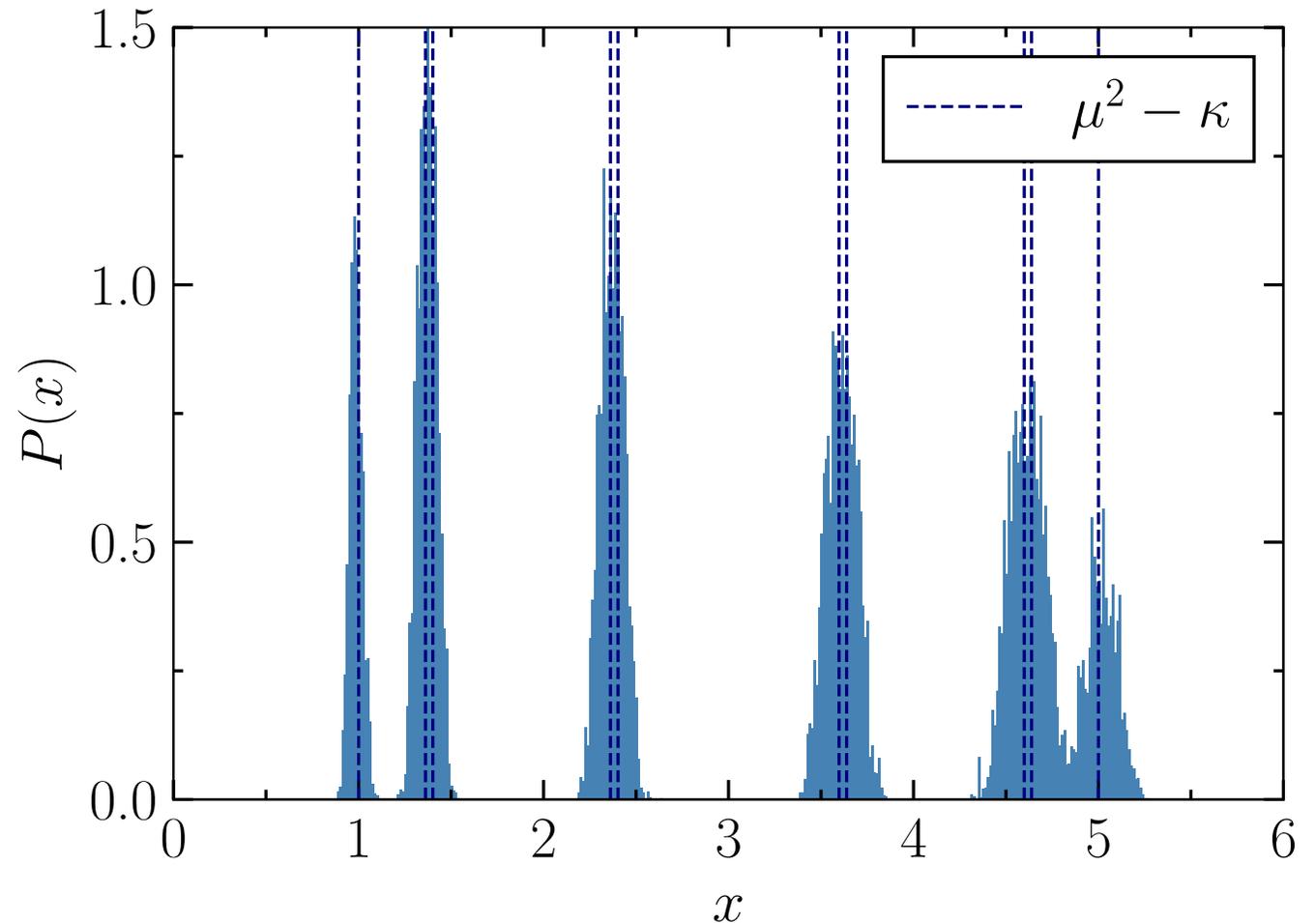
- degenerate modes split for clarity



# RBM evolution

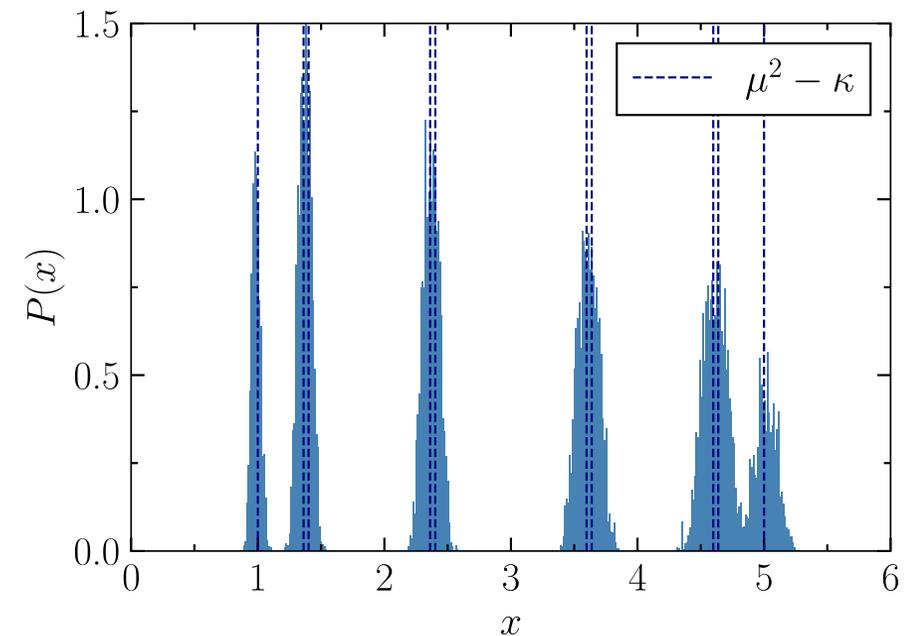
weight matrix updates using  
persistent contrastive divergence  
with mini-batches

no sharp lines:  
distributions around  
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

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

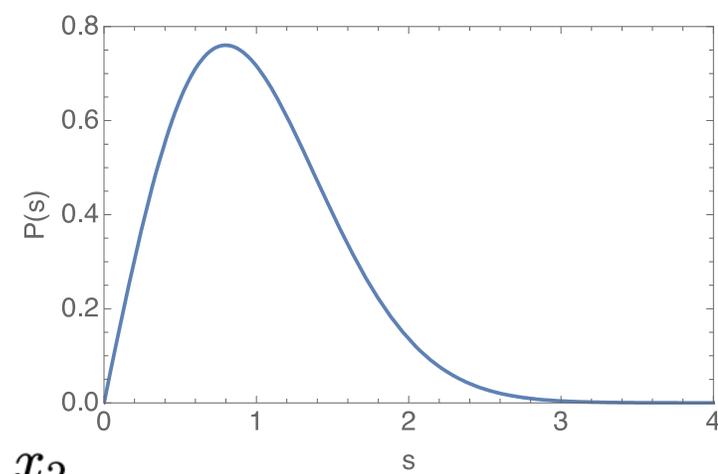
- eigenvalues  $x_1, x_2$  cannot both be equal to  $\kappa$  due to Coulomb repulsion

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

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

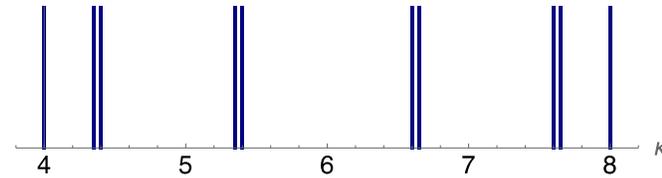
- mean level spacing  $\langle S \rangle = \int_0^\infty dS S P(S) = \sqrt{\pi}\sigma$   $s = S/\langle S \rangle$

# 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!]

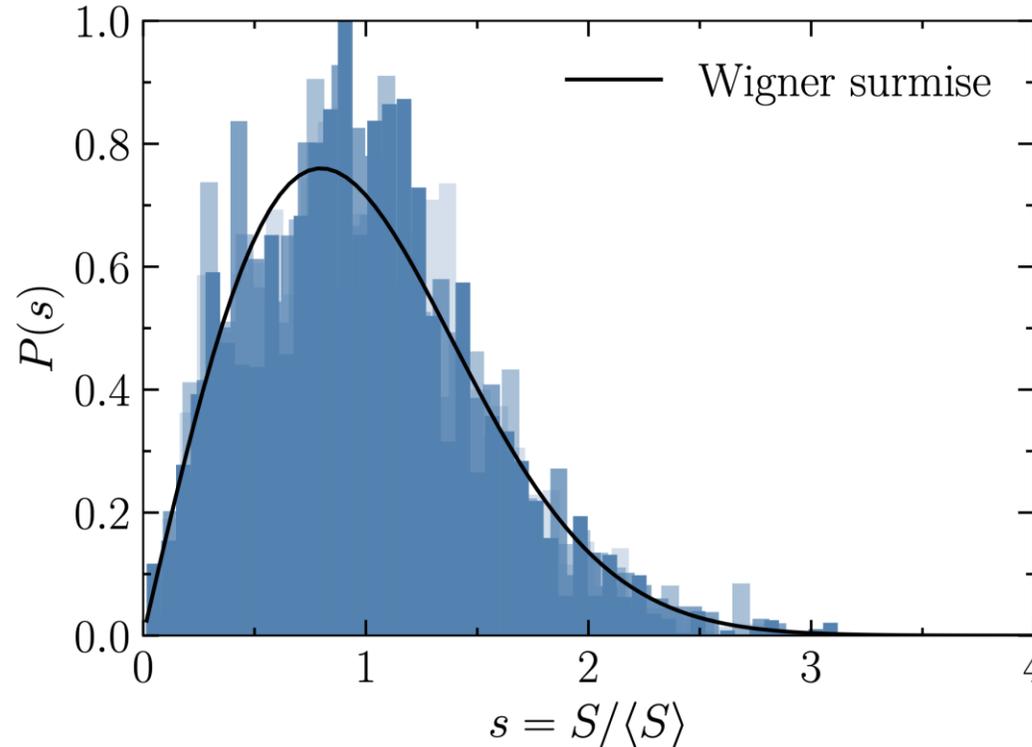
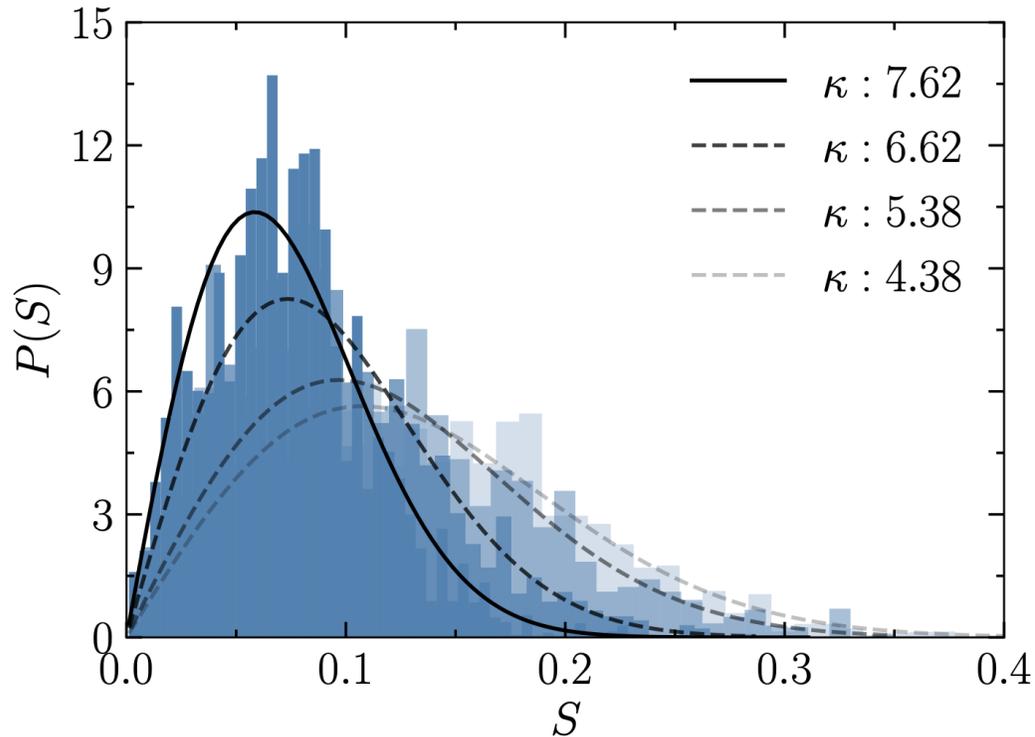
# Wigner surmise: 4 degenerate pairs



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

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

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



data collapse

universality

# Wigner surmise: vary learning rate and batch size

- prediction:

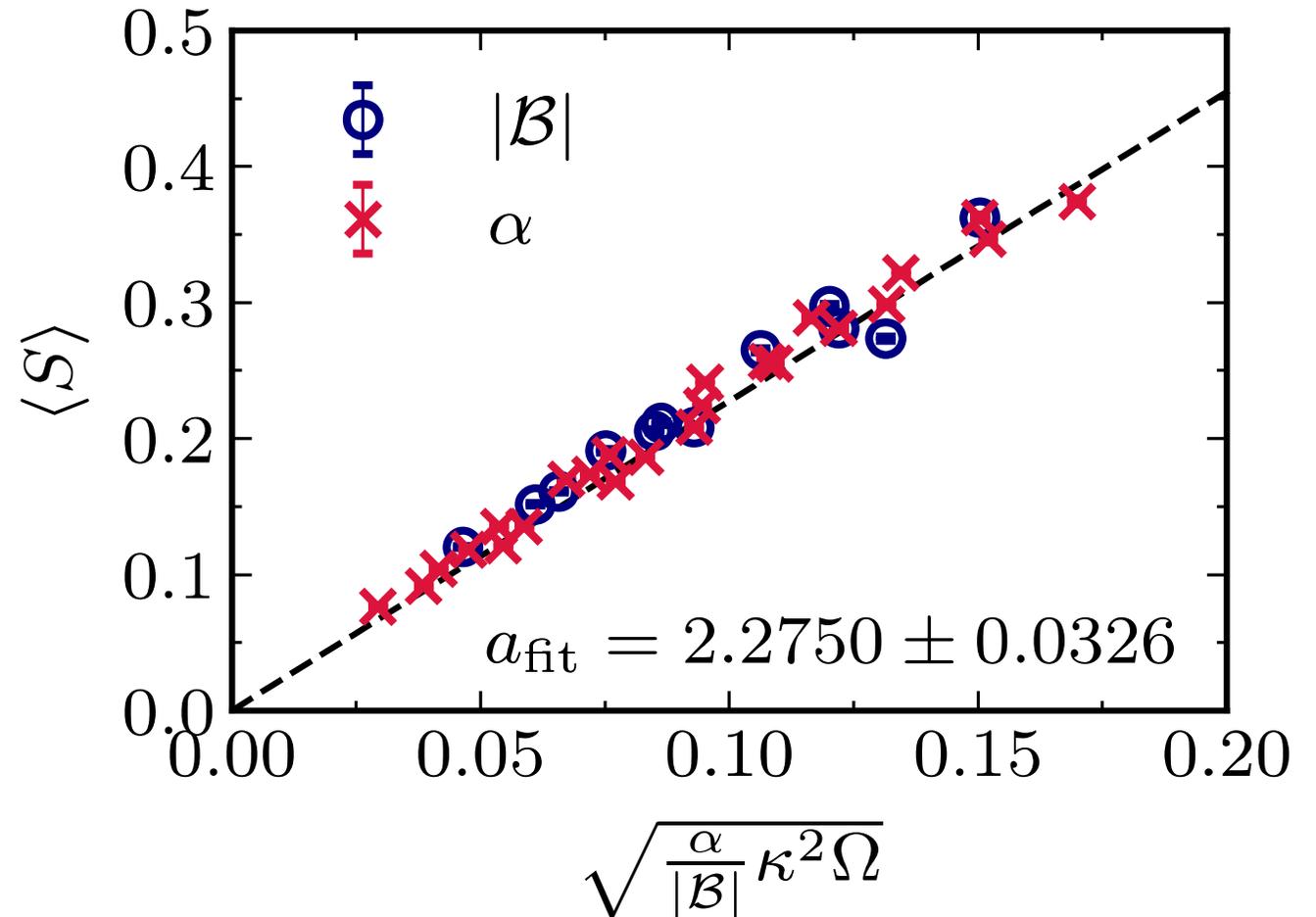
$$\sigma_i^2 = (\alpha/|\mathcal{B}|) (\tilde{g}_i^2/\Omega_i)$$

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

- mean level spacing

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

- fit function includes non-universal parameters as well



# Wigner semi-circle

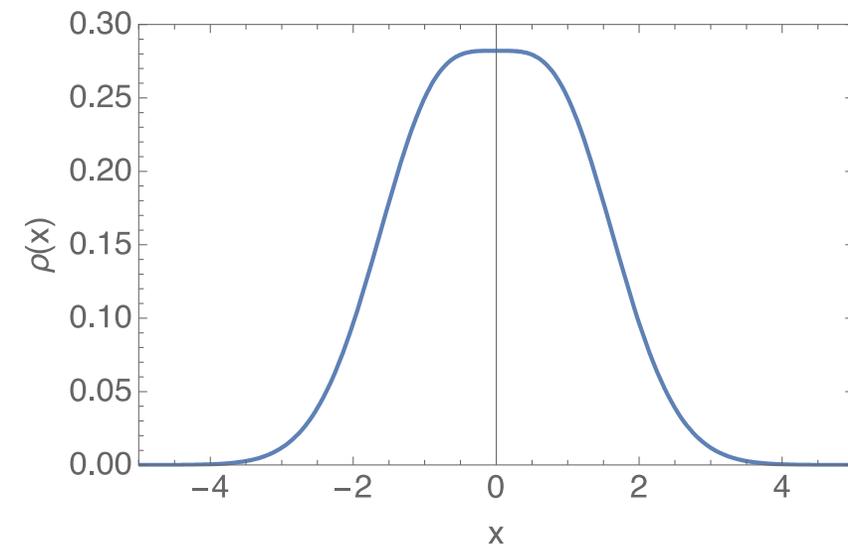
- spectral density:  $\rho(x) = \left\langle \frac{1}{N} \sum_{i=1}^N \delta(x - x_i) \right\rangle$

- for two modes:

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

- broadened and flattened Gaussian

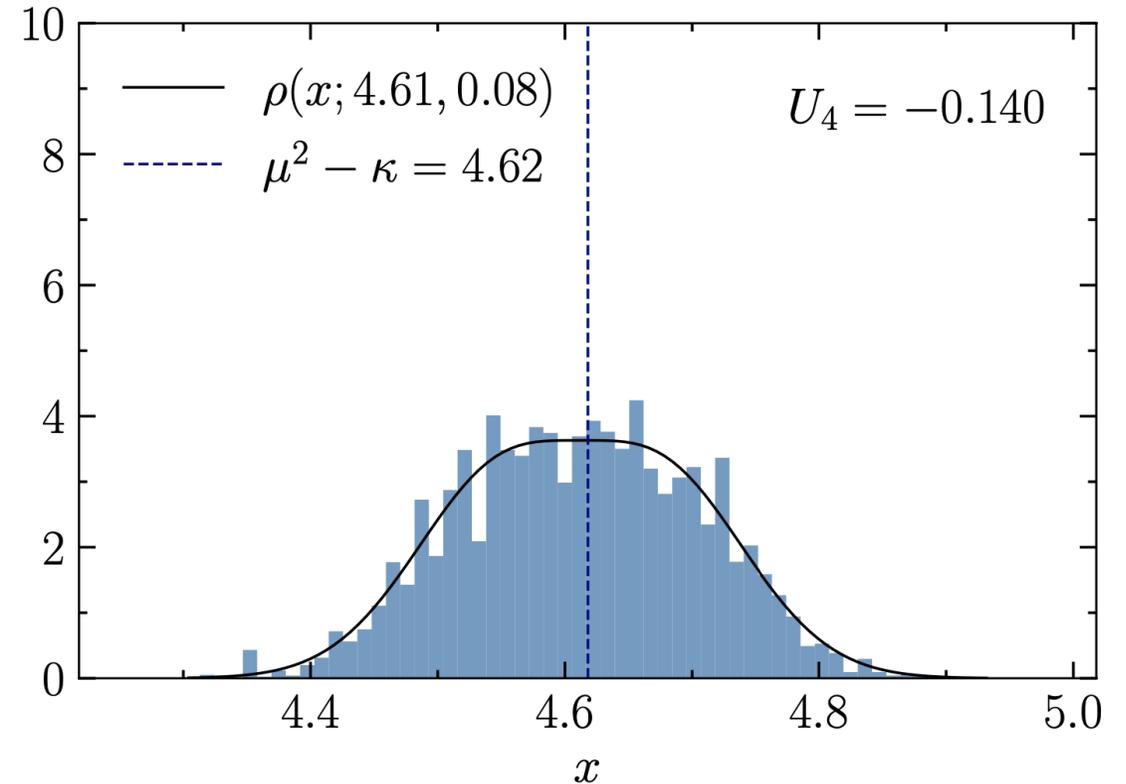
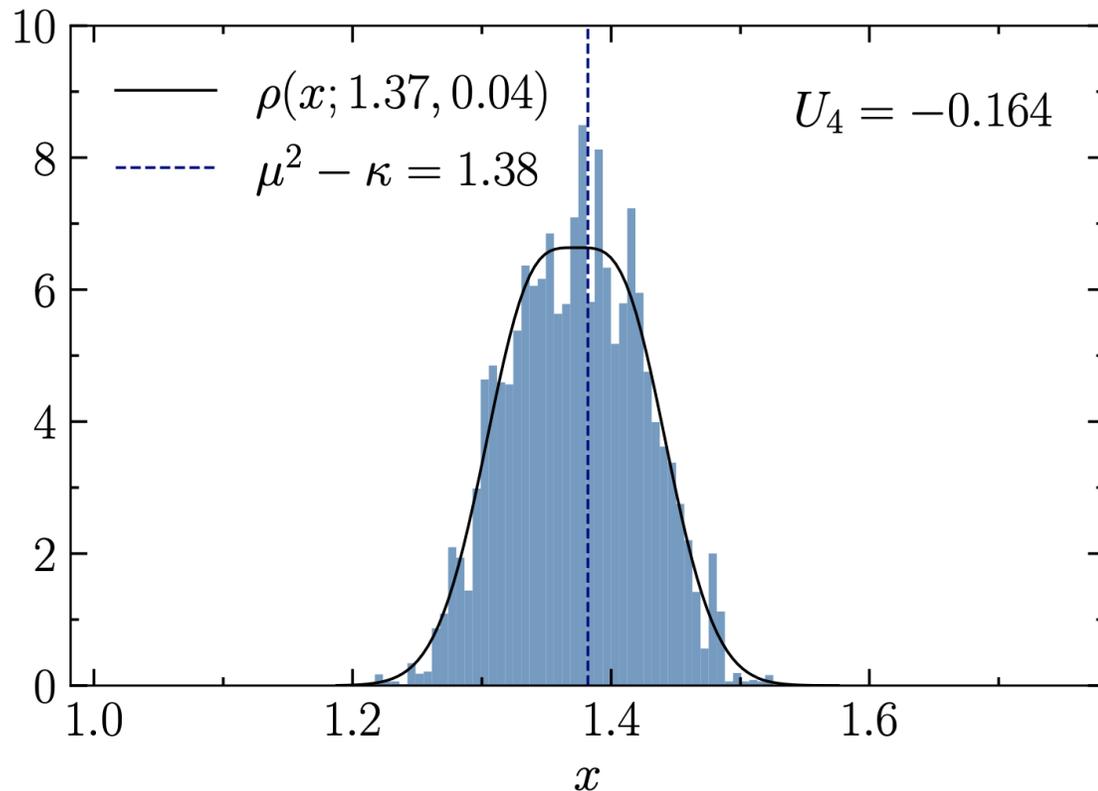
- fit  $\sigma$  parameter and position for each doubly degenerate mode



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

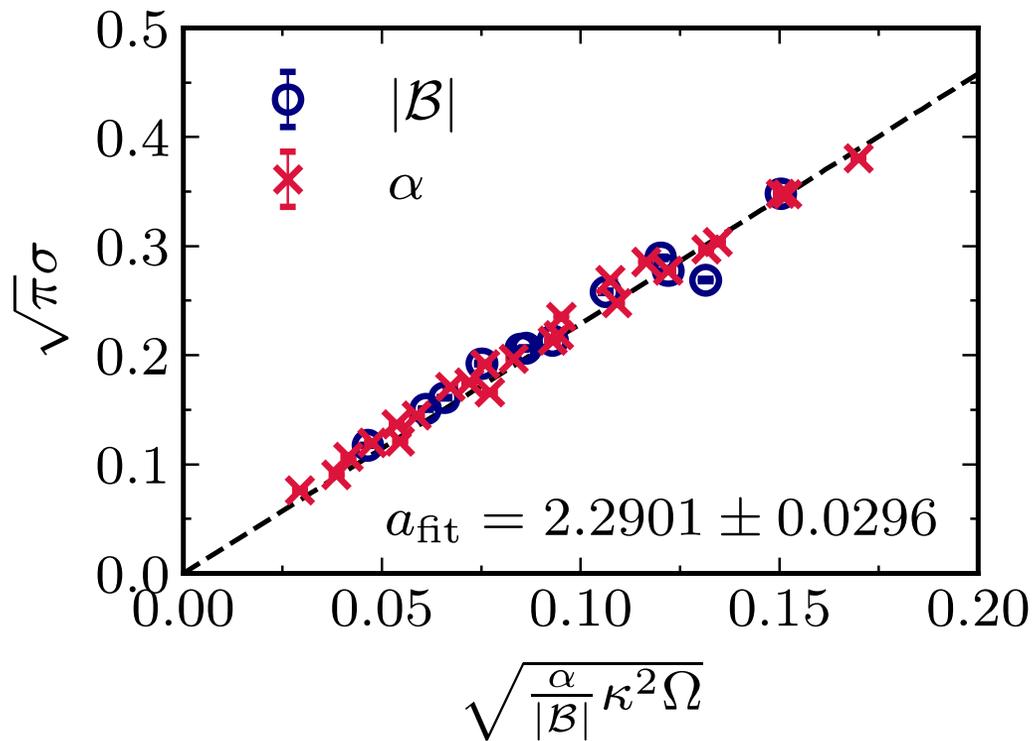
# Wigner semi-circle

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

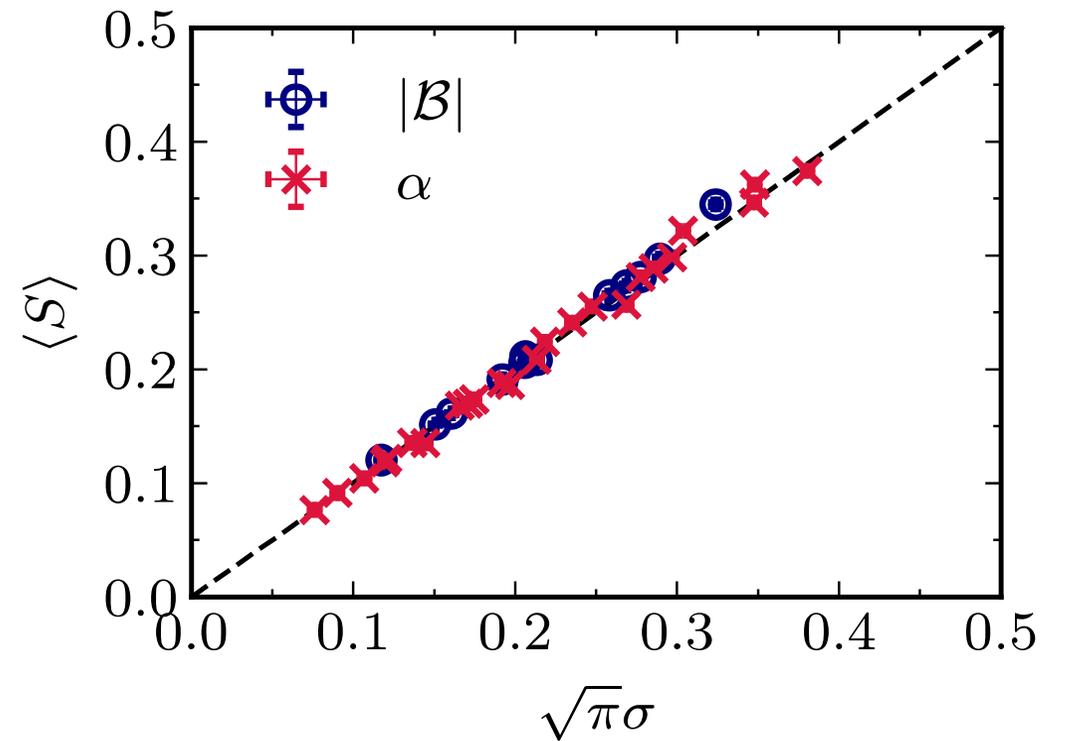


# Wigner semi-circle and surmise

semi-circle  
dependence on learning rate/batch size



consistency between surmise  
and semi-circle fits



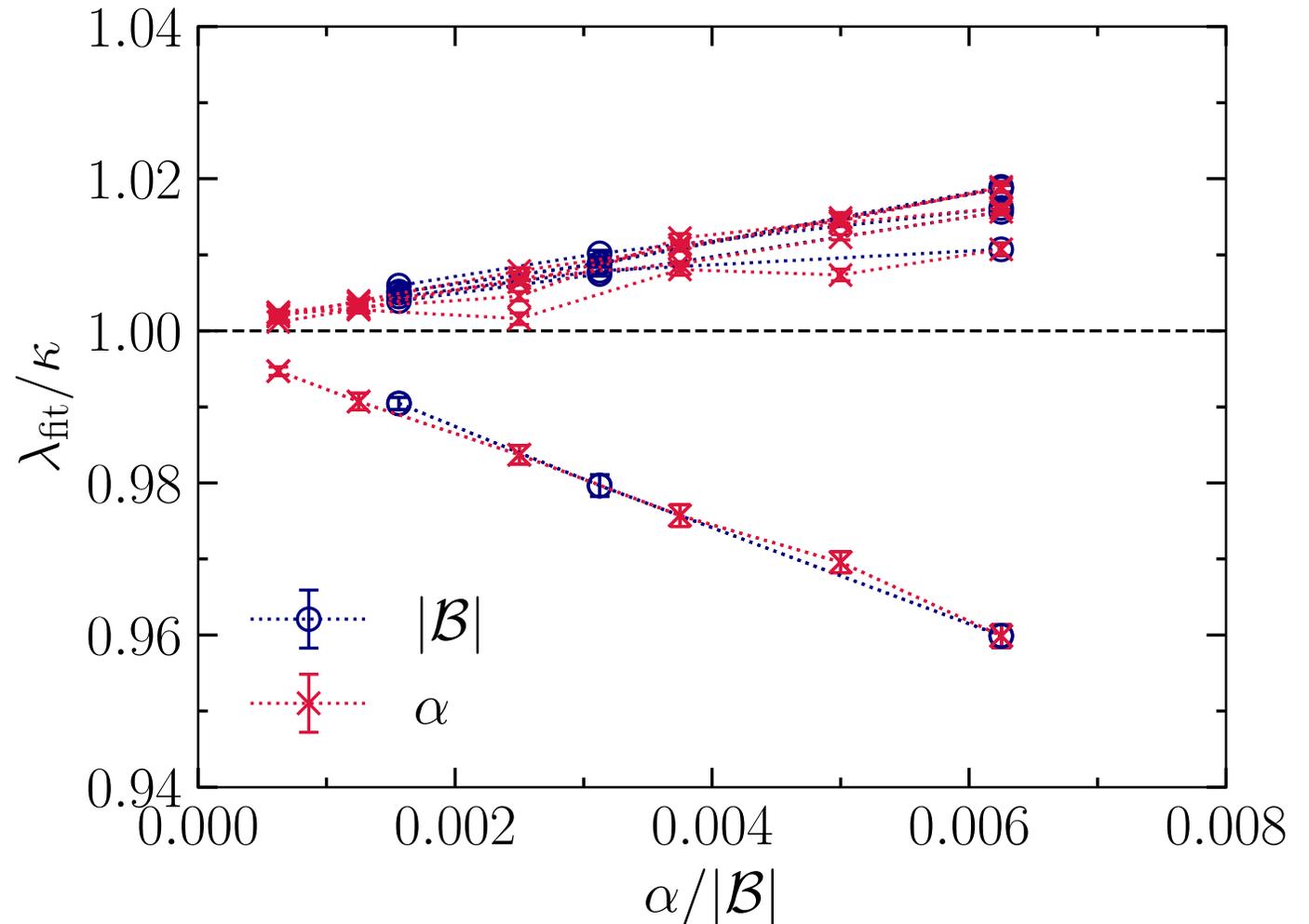
# Wigner surmise and semi-circle

- ✓ parameter  $\sigma$  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

Goyal et al, 1706.02677, S.L. Smith & Q.V. Le, 1710.06451, S.L. Smith, P. Kindermans & Q.V. Le, 1711.00489

# 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)

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

- stationary distribution

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

# Time-dependent dynamics

- assume continuous time limit exists

- analyse FPE for one mode:  $\partial_\tau P(x, \tau) = \partial_x (g^2 \partial_x - K(x)) P(x, \tau)$

- solve using standard stochastic quantisation/FP methods:  $P(x, \tau) = \sqrt{P_s(x)} \psi(x, \tau)$

- evolution:  $\partial_\tau \psi(x, \tau) = \left( g^2 \partial_x^2 - \frac{1}{4g^2} [\partial_x V(x)]^2 + \frac{1}{2} [\partial_x^2 V(x)] \right) \psi(x, \tau) \equiv -2H_{\text{FP}} \psi(x, \tau)$

- Fokker-Planck Hamiltonian:  $H_{\text{FP}} = \frac{1}{2} L^\dagger L$

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

# Quantum-mechanical bound state problem

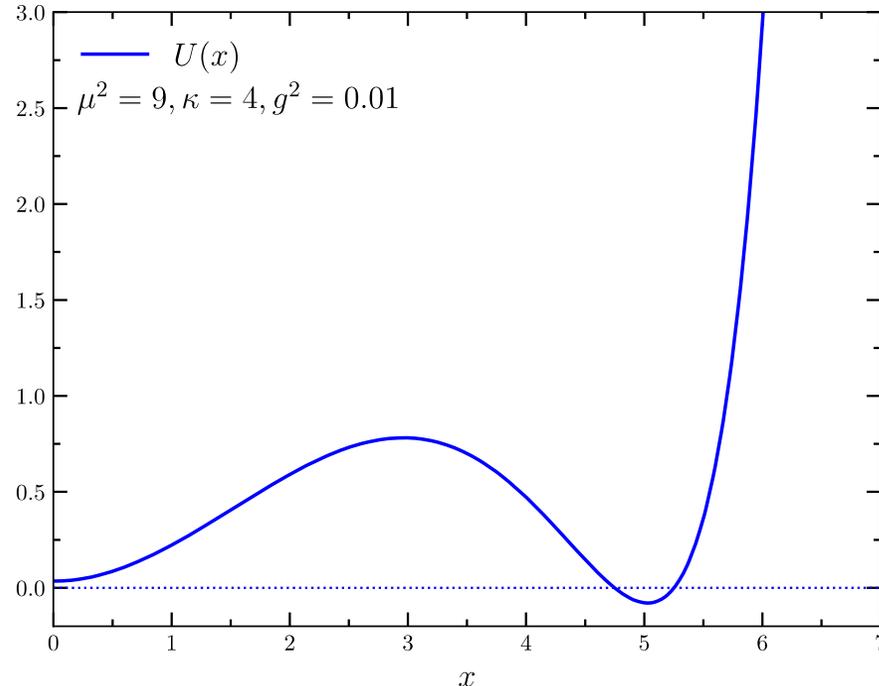
○  $H_{\text{FP}} = \frac{1}{2}L^\dagger L$       eigenvalue problem:  $H_{\text{FP}}\psi_n(x) = E_n\psi_n(x)$

○ explicit form:  $H_{\text{FP}} = -\frac{g^2}{2}\partial_x^2 + U(x)$

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

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

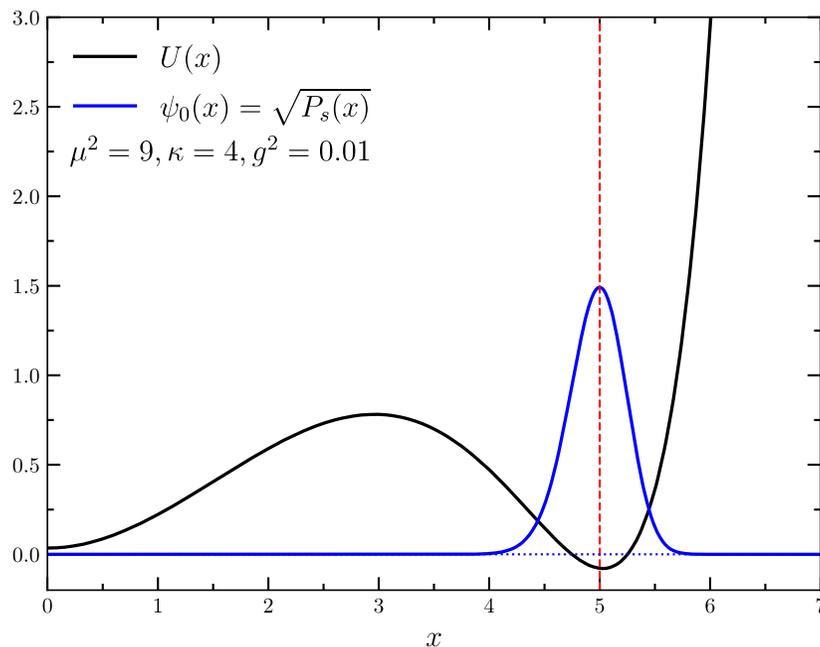
$$U_1(x) = -\frac{1}{4}\partial_x^2 V(x)$$



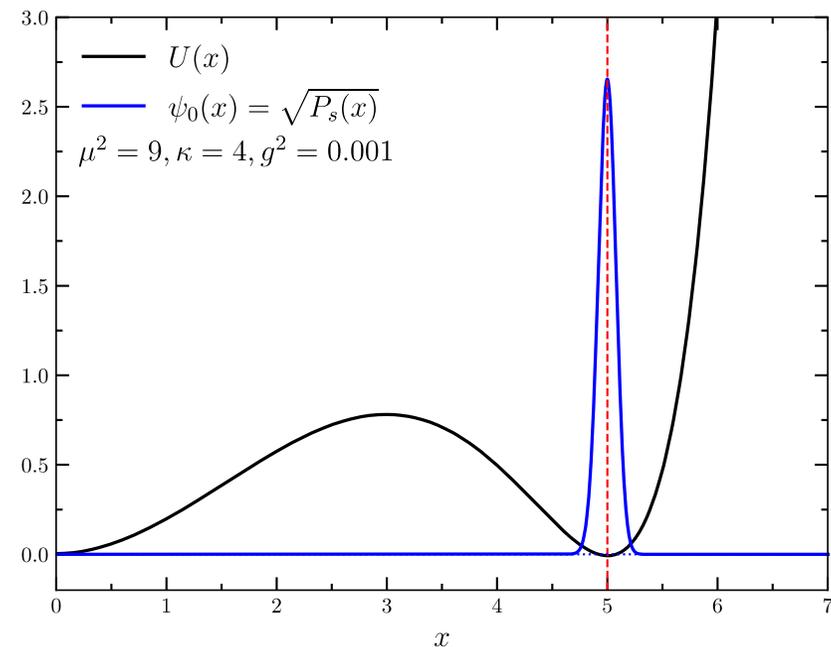
double well potential on interval  $0 \leq x \leq \mu^2$

# 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  $\kappa$



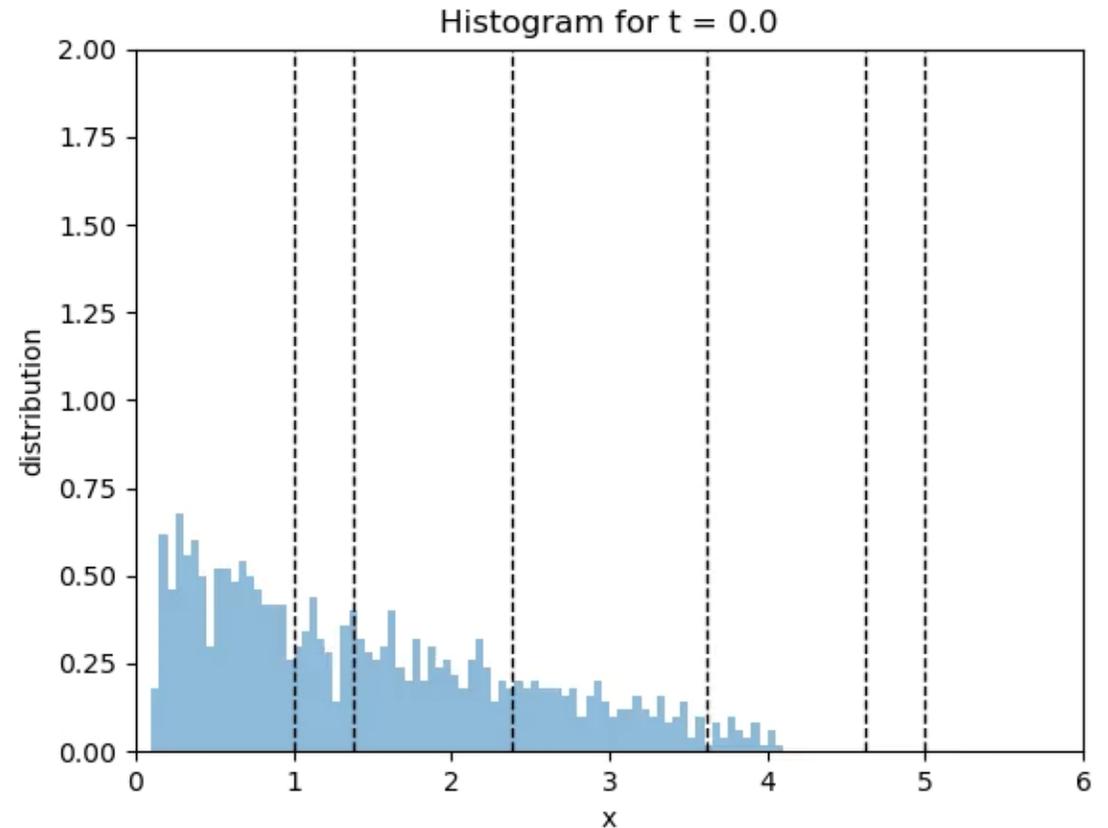
$g^2 = 0.01$



$g^2 = 0.001$

# 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



# 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

- no obvious continuous time limit, i.e. learning rate  $\rightarrow 0$  does not yield SDE
- advantageous since it gives handle on controlling stochasticity
- alternative routes to continuous time
- extending analysis to neural networks, transformers, adaptive stepsize, etc
- next talk: Matteo on neural networks