Data generated using SCW cluster

AIMLAC
UKRI CENTRE FOR DOCTORAL TRAINING IN ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND ADVANCED COMPUTING

UNDEB EWROPEAIDD
EUROPEAN UNION
Llywodraeth Cymru
Welsh Government
Cronfa Datblygu Rhanbarthol Ewrop
European Regional Development Fund

Random Matrix Theory in Stochastic Optimisation

Chanju Park
With Gert Aarts and Biagio Lucini

Swansea University
Prifysgol Abertawe

Aarts, Lucini, Park, arXiv:2407.16427

Our result

- Stochastic optimisation of a matrix has universal features described by Random Matrix Theory.

- The stochasticity of the system scales as $\alpha/|\mathcal{B}|$.

α : Step size $|\mathcal{B}|$: Batch size

⋯ Linear Scaling Rule [Goyal et al, arXiv:1706.02677]

Outline

1. Stochastic Optimisation

2. Random Matrix Theory

3. Random Matrix Theory in Stochastic Gradient Descent

4. Experiments

Stochastic Optimisation

Robbins - Monro algorithm

Imagine we have a stochastic variable x and a function $f(x)$ with an extremum at $x = x^*$. We can find the value x^* by following iterative algorithm.

$$x_{n+1} = x_n - \alpha_n \left(\mathbb{E}\left[f(x)\right] - f(x^*) \right)$$

For α_n satisfying $\sum_{n=0}^{\infty} \alpha_n = \infty, \quad \sum_{n=0}^{\infty} \alpha_n^2 < \infty$

Eg) Stochastic Gradient Descent

$$x = W_{ij}, \quad f\left(W_{ij}\right) = \frac{\partial \mathscr{L}}{\partial W_{ij}}, \quad f\left(W_{ij}^*\right) = 0$$

Langevin Dynamics

Stochastic optimisation

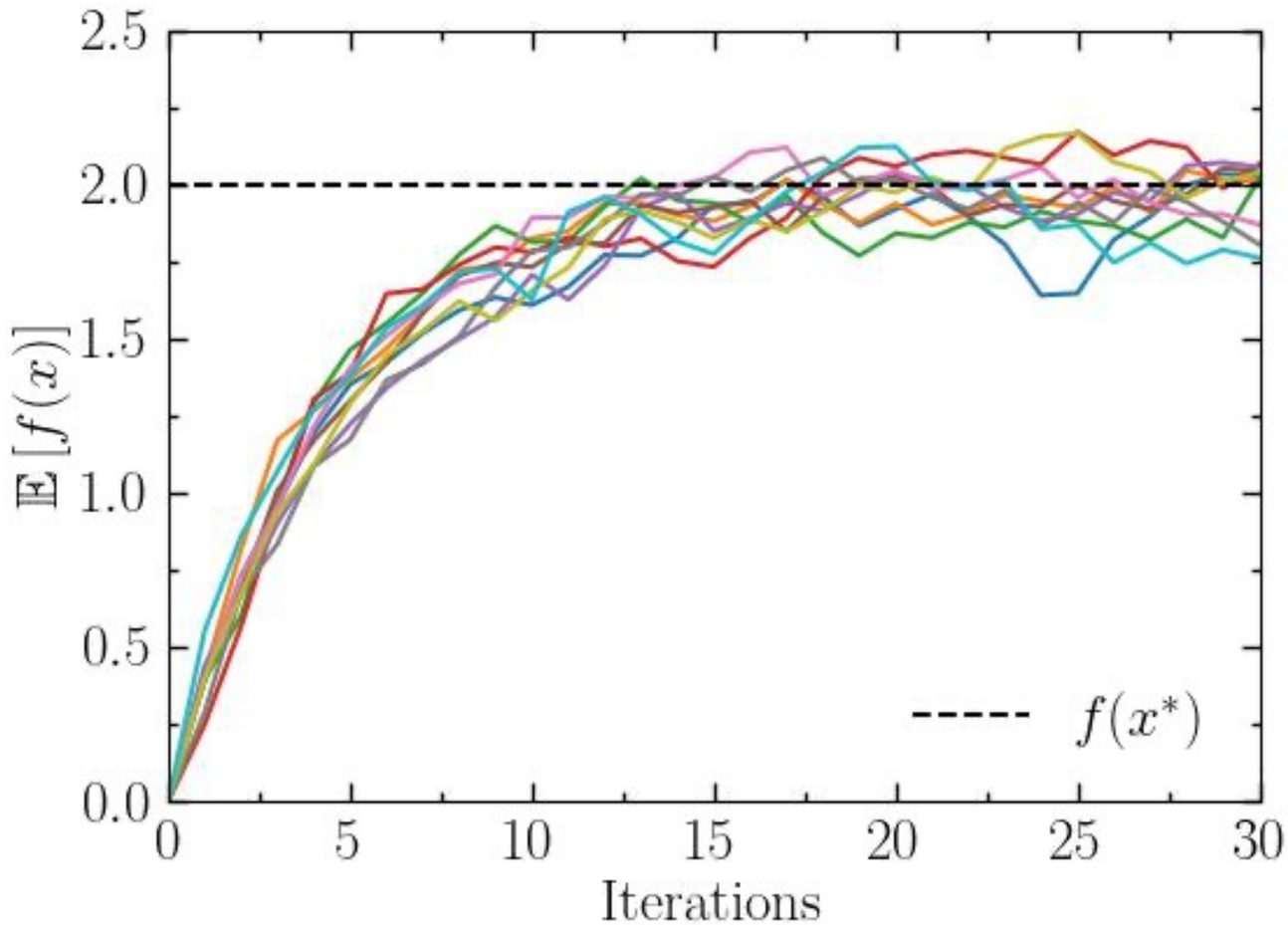
Langevin equation

$$x_{n+1} = x_n - \alpha\left(\mathbb{E}\left[f(x_n)\right] - f(x^*)\right) \quad \longleftrightarrow \quad \frac{dx}{dt} = -K(x;t) + \sqrt{2}g(x;t)\,\eta$$

$$f(x^*) = 2, \quad P\left(f(x_n)\right) \sim \mathcal{N}\left(\mu_n, \sigma_n\right)$$

$$K(x;t) = \alpha\left(\mu(t) - f(x^*)\right), \quad g(x;t) = \alpha\frac{\sigma(t)}{\sqrt{2}}$$

Simulation



Theory

Stochastic optimisation can be described by Langevin dynamics.

Random Matrix Theory

Gaussian Orthogonal Ensemble

Imagine we have an ensemble of symmetric matrices, where the elements are Gaussian random variables.

$$M_{ij} \sim \mathcal{N}(0,1)$$

$P\left(M_{ij}\right) \propto e^{-\frac{1}{2}TrM_{ij}^2}$ invariant under $M' = O^T M O, \quad O^T O = I$

Because of the symmetry, there is degeneracy in the representation and it is convenient to choose eigenvalues as the basis.

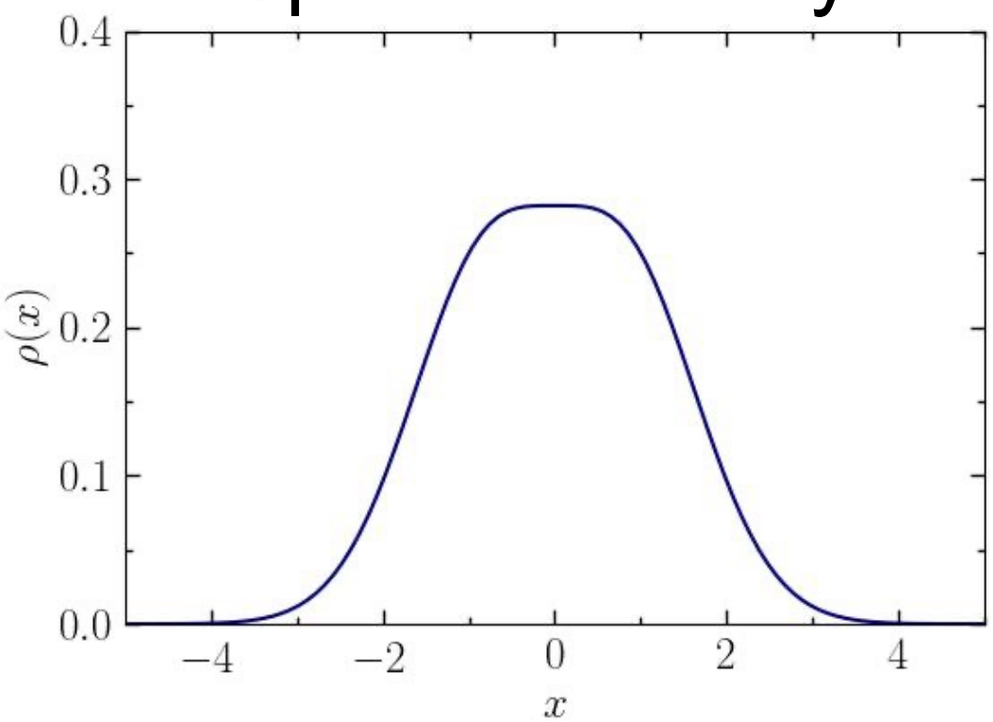
$$P\left(M_{ij}\right) \quad \Rightarrow \quad P\left(x_i\right) \propto \prod_{i<j} |x_i - x_j| e^{-\frac{1}{2}\sum_i x_i^2}$$

Jacobian

Properties of Random Matrix eigenvalues

Some useful statistical properties of the eigenvalues are known for these types of matrices.

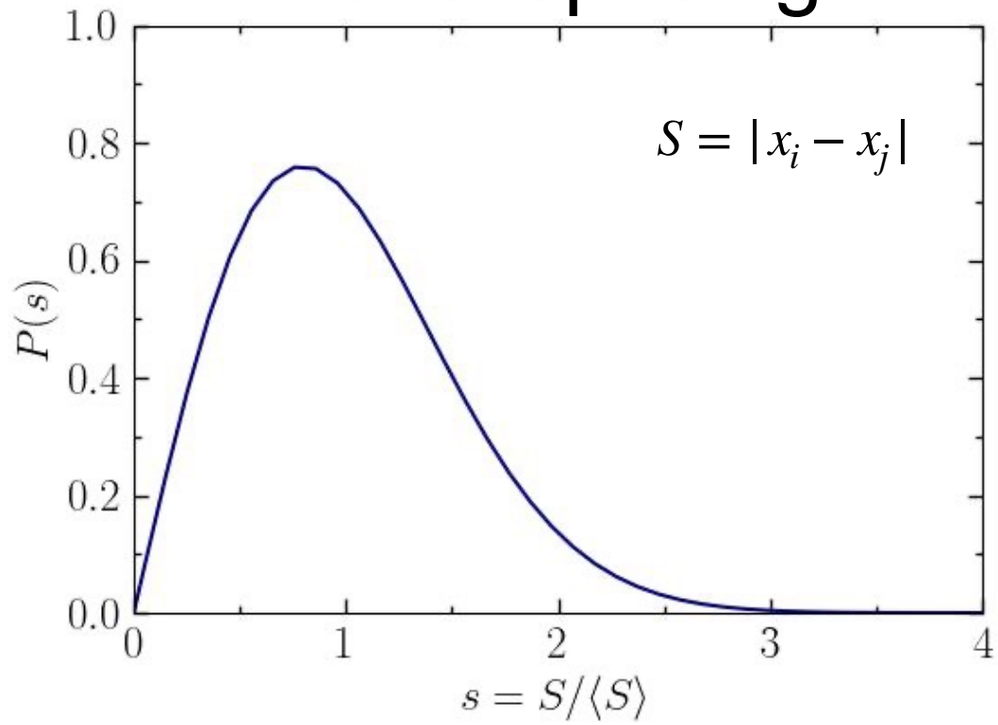
Spectral density



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

"Wigner semi-circle"

Level spacing

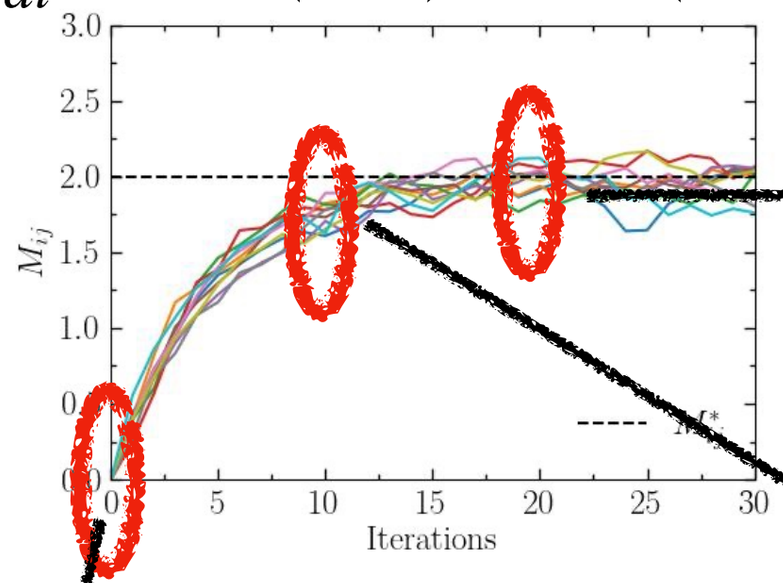


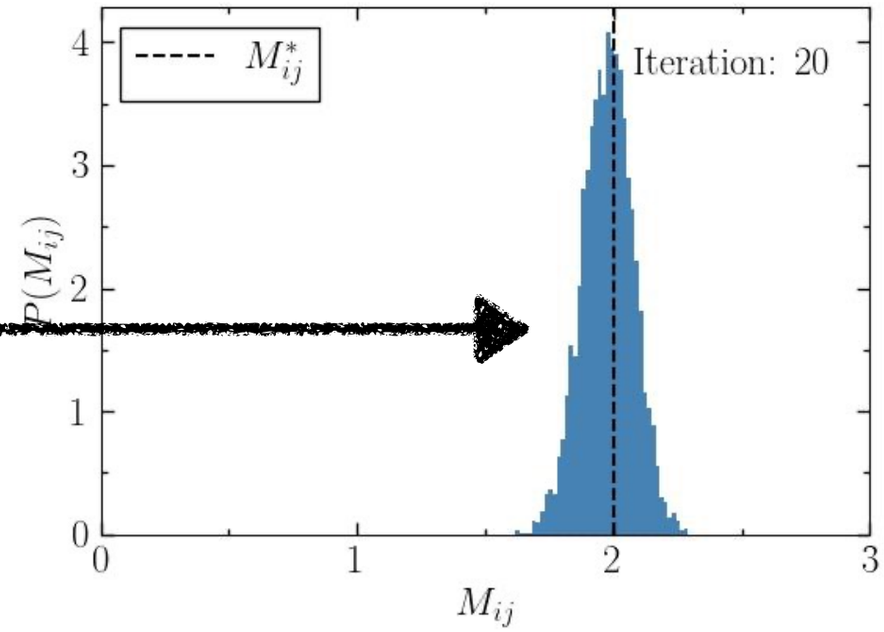
$$P(s) = \frac{\pi}{2} s e^{-\pi s^2/4}$$

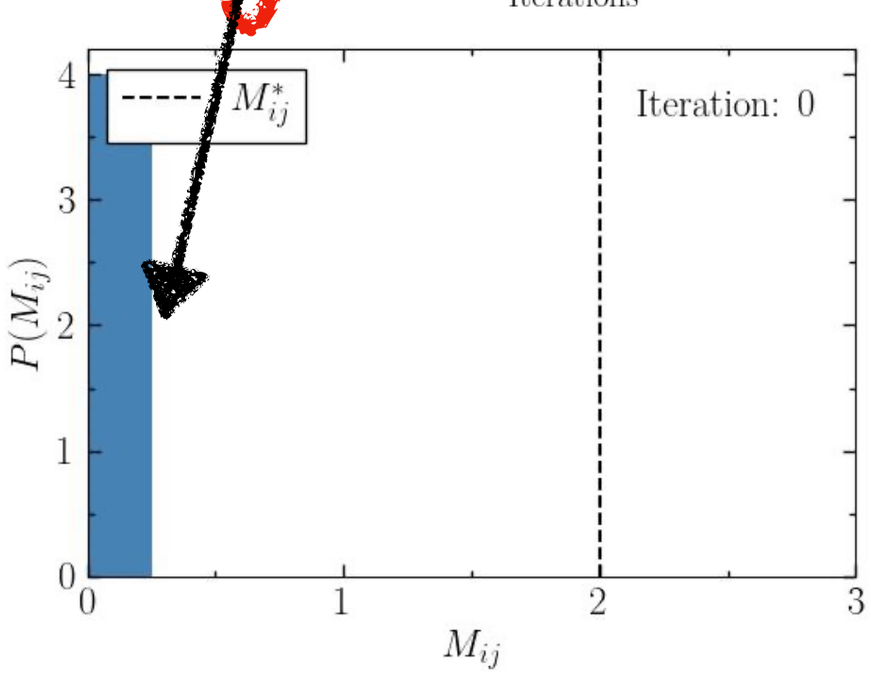
"Wigner surmise"

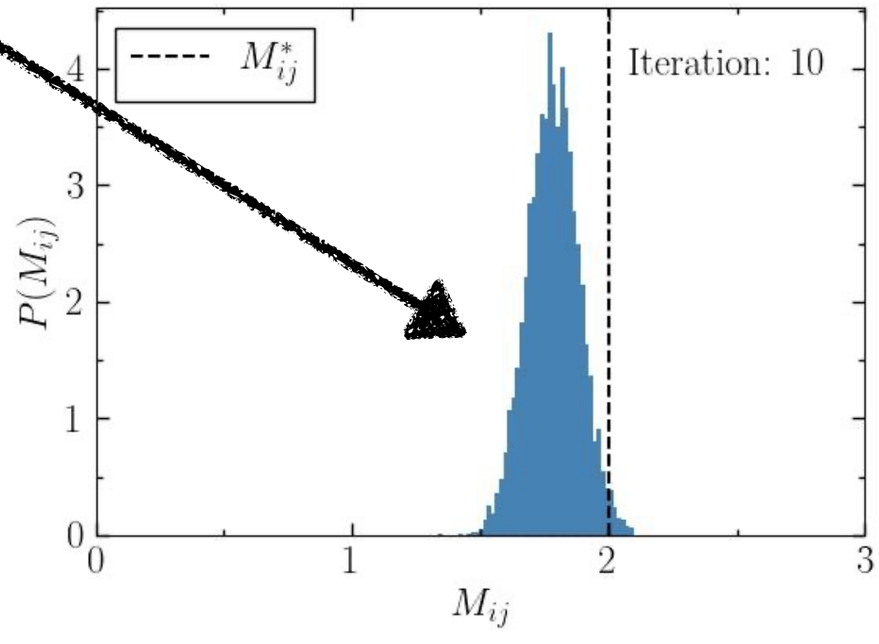
Stochastic optimisation of Matrices

$$\frac{dM_{ij}}{dt} = -K\left(M_{ij};t\right) + \sqrt{2}g\left(M_{ij};t\right)\eta_{ij}$$









At each time slice, matrix elements are randomly distributed!

=> Random Matrix Theory!

Dyson, J. Math. Phys. 3 (1962) 1191

Dynamics of eigenvalue

Dyson-Brownian motion

When a matrix is stochastically optimised,

$$\frac{dM_{ij}}{dt} = -K_{ij} + g_{ij}\eta_{ij}, \quad K_{ij} = \frac{\partial \mathscr{L}}{\partial M_{ij}}$$

Change of variable introduces Jacobian (Vandermonde determinant),

$$P\left(M_{ij}\right) \Rightarrow P\left(x_i\right) \propto \prod_{i<j} |x_i - x_j| e^{-\mathscr{L}} = e^{-\mathscr{L} + \sum_{i<j} \log\left(x_i - x_j\right)}$$

Dynamics of eigenvalues can be derived.

$$\frac{dx_i}{dt} = -K_{ii} + \sum_{j \neq i} \frac{g_{ij}^2}{x_i - x_j} + \sqrt{2} g_{ii}\eta_{ii}$$

Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent is one of the Stochastic optimisation algorithms.

$$M_{n+1} = M_n + \alpha \left\langle \Delta_p \right\rangle_{p \in \mathscr{B}}$$

$$\left\langle \Delta_p \right\rangle_{p \in \mathscr{B}} \equiv \frac{1}{|\mathscr{B}|} \sum_{p \in \mathscr{B}} \Delta_p, \quad \Delta_p \equiv \left. \frac{\partial \mathscr{L}}{\partial M_n} \right|_p$$

Stochasticity is introduced from the finite batch size, and its strength can be extracted using the CLT.

$$\mathbb{V}\left[\Delta_p\right]_{p \in \mathscr{B}} \propto \frac{1}{|\mathscr{B}|} \tilde{g}^2, \quad \tilde{g}^2 \equiv \mathbb{V}\left[\frac{\partial \mathscr{L}}{\partial M_n}\right]$$

Where $\tilde{g}$ is a variance of the gradient.

Langevin equation for SGD

Corresponding Langevin equation for stochastic gradient descent can be written as,

$$M'_{ij} = M_{ij} - \underbrace{\alpha \mathbb{E}_{ij}\left[\Delta_p\right]}_{= K_{ij}(M)} + \frac{\alpha}{\sqrt{|\mathcal{B}|}}\sqrt{\tilde{g}_{ij}^2}\eta_{ij}$$

$$\eta_{ij} \sim \mathcal{N}(0,1)$$

Dynamic equation for eigenvalue is given as,

$$x'_i = x_i + \underbrace{\alpha \mathbb{E}_{ii}\left[\Delta_p\right] + \frac{\alpha^2}{|\mathcal{B}|}\sum_{j\neq i}\frac{\tilde{g}_{ij}^2}{x_i - x_j}}_{= K_{ii}^{(\mathrm{eff})}(x)} + \frac{\alpha}{\sqrt{|\mathcal{B}|}}\sqrt{2\tilde{g}_{ii}^2}\eta_i$$

$$\eta_i \sim \mathcal{N}(0,1)$$

Universal scaling factor

Stationary limit distribution

The distribution of the eigenvalues is obtained by solving the Fokker-Planck equation.

$$\partial_t P\left(\{x_i\},t\right) = \sum_{i=1}^{N} \partial_{x_i}\left[\left(\frac{\alpha^2}{|\mathscr{B}|}\tilde{g}_i^2\partial_{x_i} - K_{ii}^{(\mathrm{eff})}\right)\right] P\left(\{x_i\},t\right)$$

Stationary limit solution: Coulomb gas distribution

$$P\left(\{x_i\}\right) = \frac{1}{Z}\prod_{i<j}|x_i - x_j|\, e^{-\sum_i V_i(x_i)/\sigma_i^2}, \quad K_{ii}\left(x_i\right) = -\alpha\frac{dV_i\left(x_i\right)}{dx_i}$$

··· details in Matteo's poster

and $\sigma_i^2 = \dfrac{\alpha}{|\mathscr{B}|}\dfrac{\tilde{g}_i^2}{\Omega_i}$

Universal scaling
(SGD)

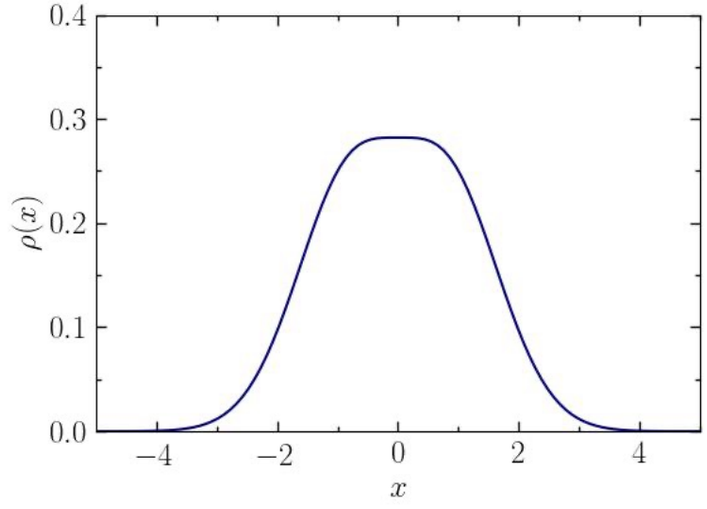
Model-specific factor
(Loss function, architecture, etc.)

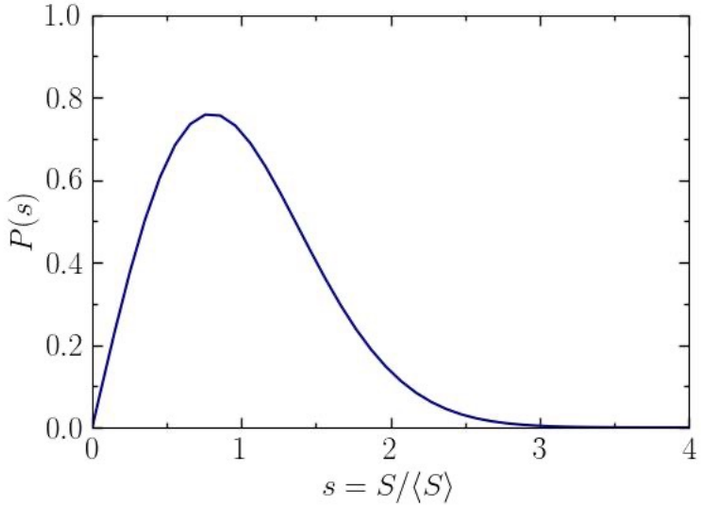
Experiment

$$\sigma_i^2 = \frac{\alpha}{|\mathscr{B}|}\frac{\tilde{g}_i^2}{\Omega_i}$$

1. Ensembles of the model are trained with different values of α and $|\mathscr{B}|$.

2. Random matrix behaviour of the stochastic optimisation is checked by comparing eigenvalue distributions to Wigner surmise and semi-circle.



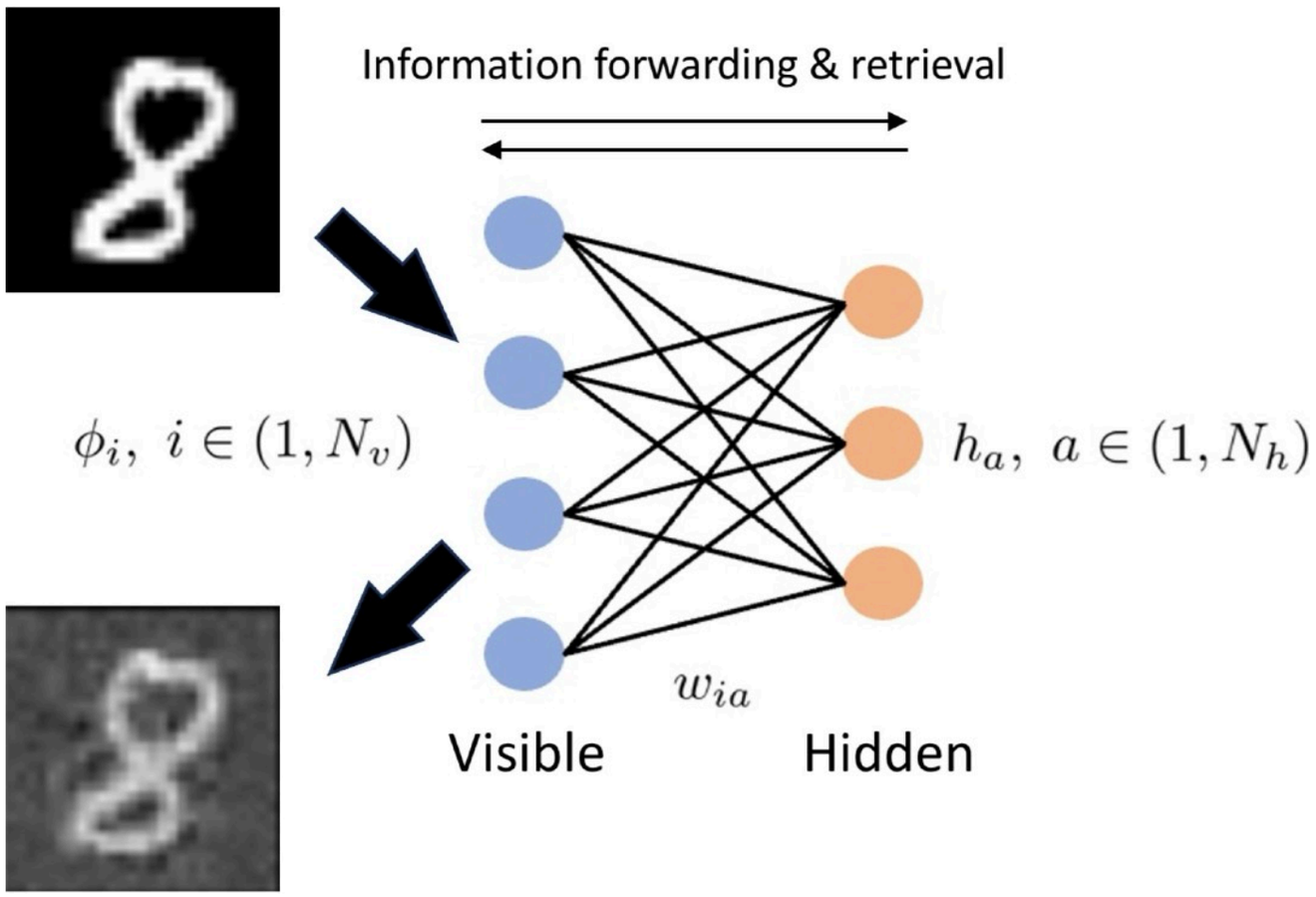


3. Universal scaling factor is checked by observing the scaling behaviour of Wigner surmise and semi-circle according to different values of α and $|\mathscr{B}|$.

Aarts, Lucini, Park, PRD 109, (2024) 3, 034521

Scalar field RBM

Gaussian Restricted Boltzmann Machine



- RBM is an energy-based generative model.
- Each layer is sampled based on the energy function.

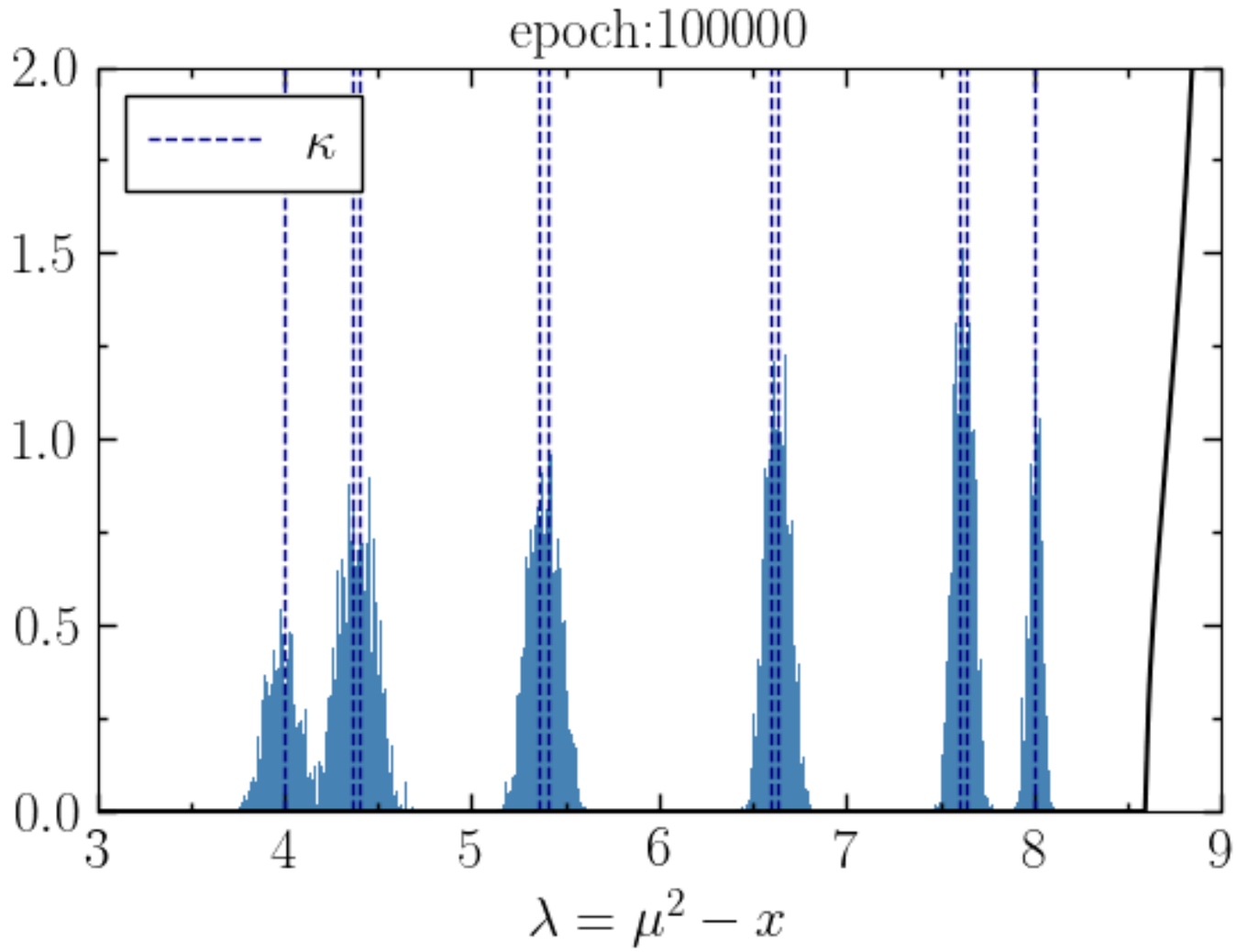
$$p\left(\phi, h\right) = \frac{1}{Z} e^{-H(\phi, h)}$$ with an energy function $H\left(\phi, h\right)$.

Learning scalar field theory

RBM is trained to learn a distribution representing a lattice scalar field theory.

Target distribution eigenvalue: $\kappa_i = m^2 + 2 - 2\cos\left(\frac{2\pi i}{N}\right)$

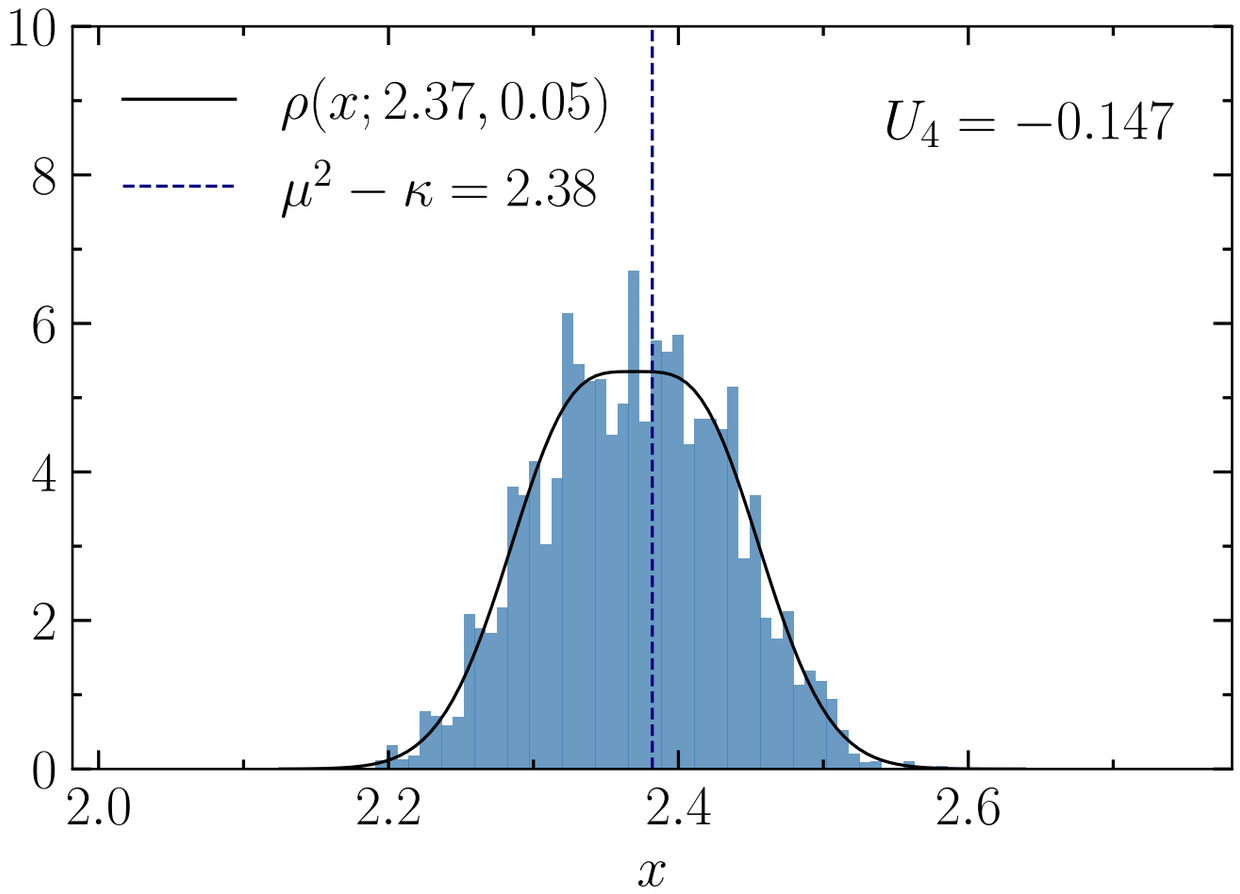
RBM eigenvalue distribution:

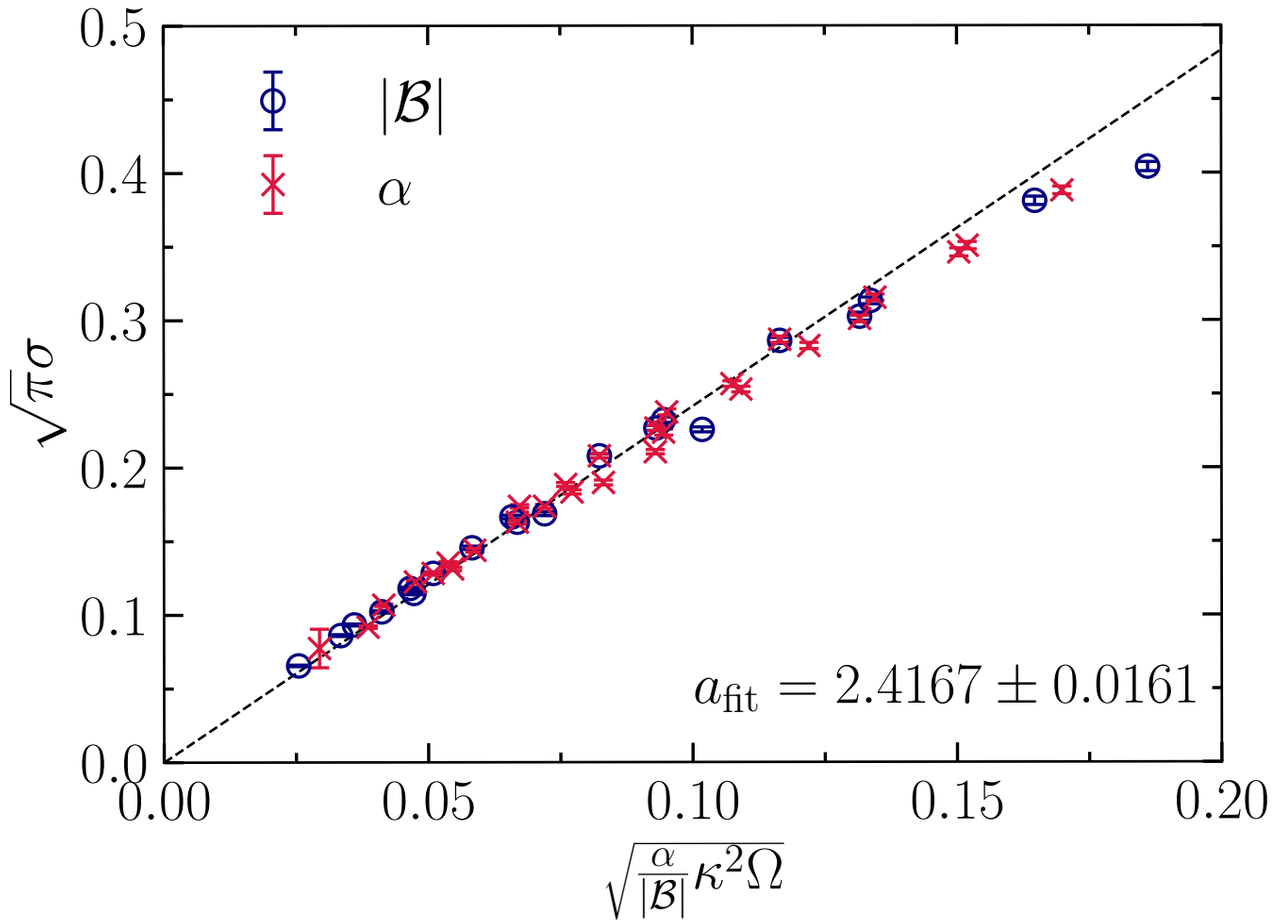


Gradient (drift) of Scalar field RBM:

$$\frac{\partial \mathcal{L}}{\partial M_{ii}} = K_i(x_i) = \left(\frac{1}{\kappa_i} - \frac{1}{\mu^2 - x_i}\right) x_i$$

Spectral density



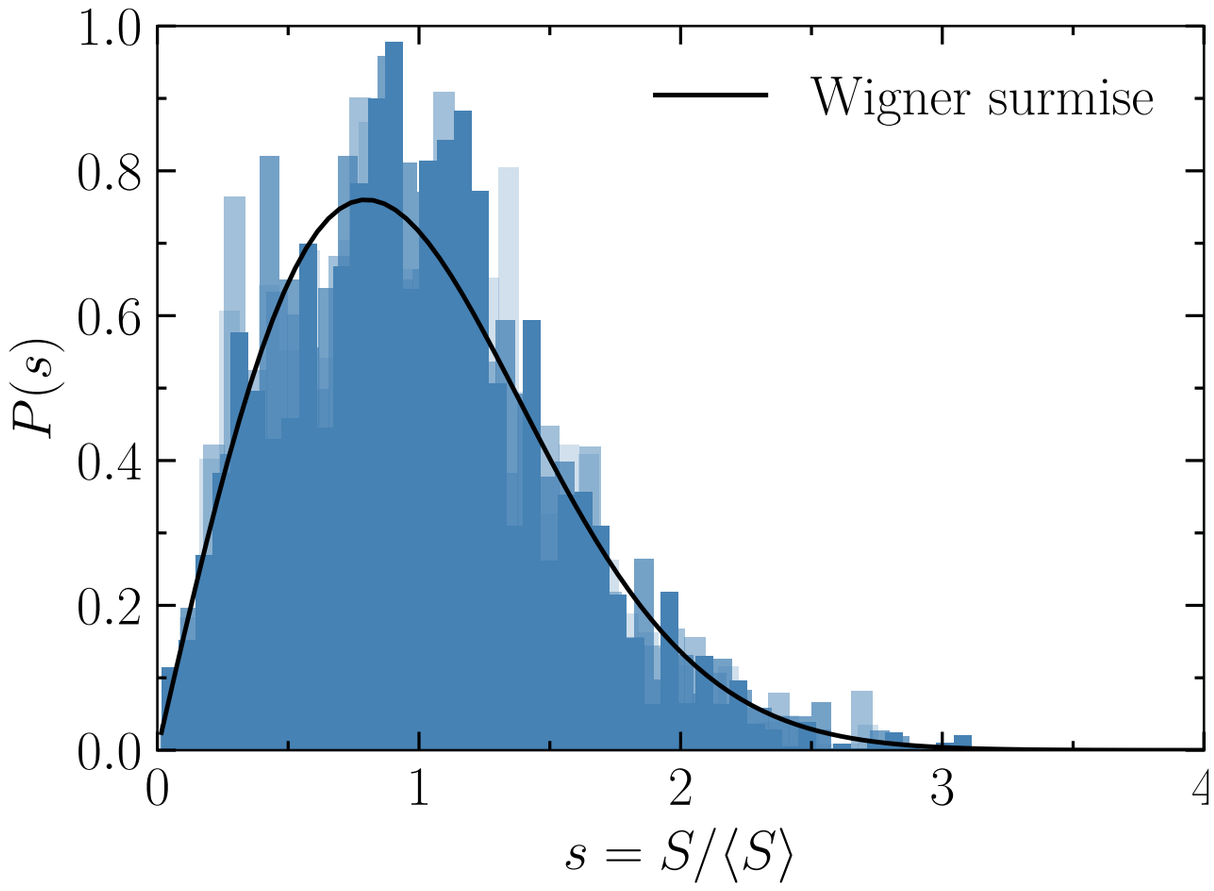


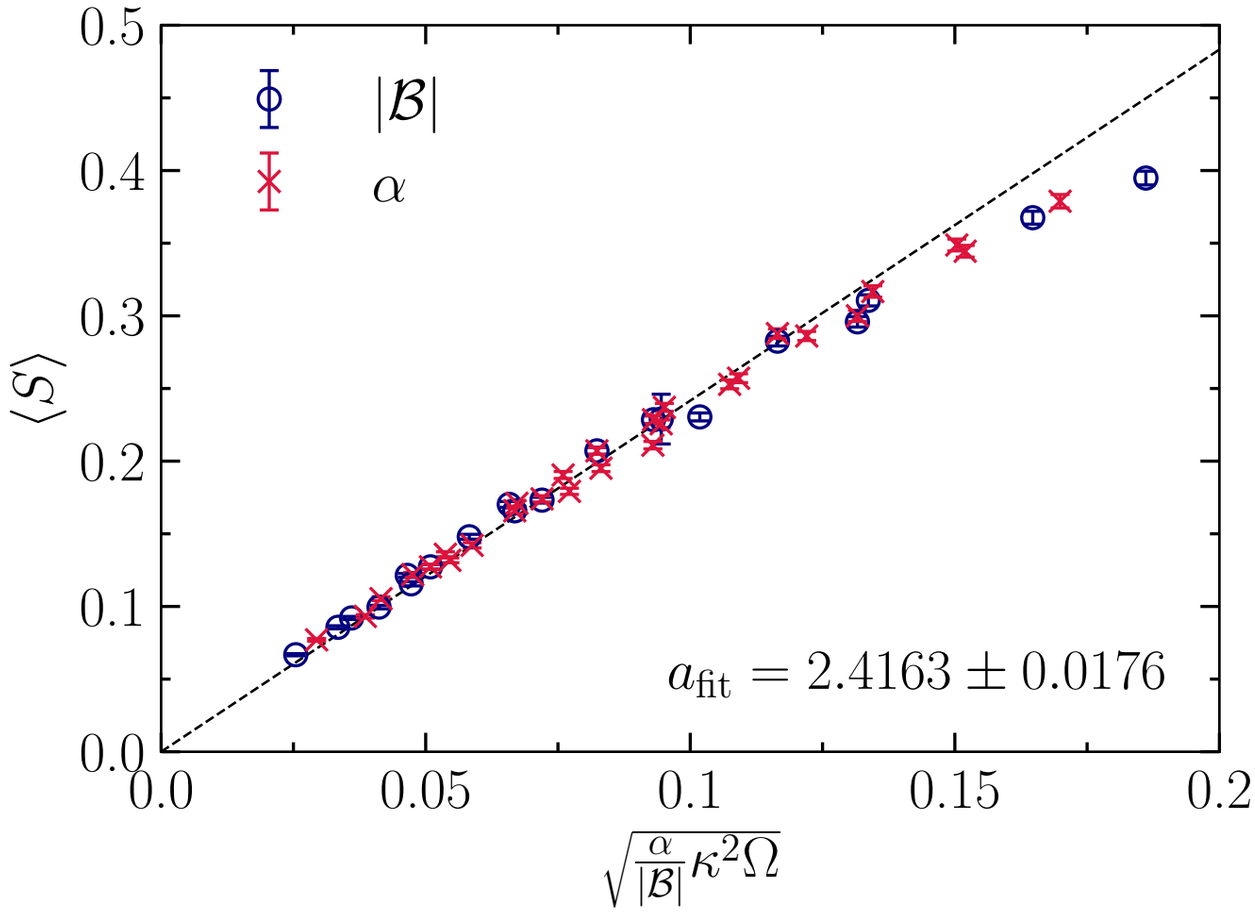
Eigenvalue distribution follows the Wigner semi-circle.

$$U_4 \equiv \frac{\langle \delta x^4 \rangle}{3 \langle \delta x^2 \rangle^2} - 1 = -\frac{4}{27} \approx -0.147\ldots$$

The width of the distribution scales with the universal scaling factor $\alpha/|\mathcal{B}|$.

Level spacing





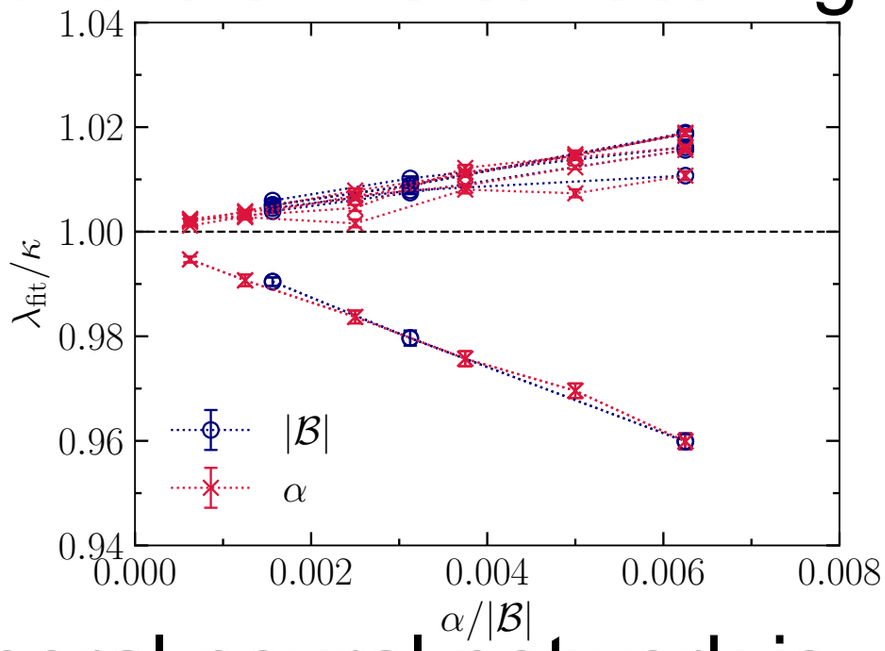
Mean level spacing collapses into the universal curve.

Level spacing scales with the universal scaling factor $\alpha/|\mathcal{B}|$.

Summary and Outlook

- The Linear Scaling Rule of the stochastic gradient descent can be derived from the random matrix theory.

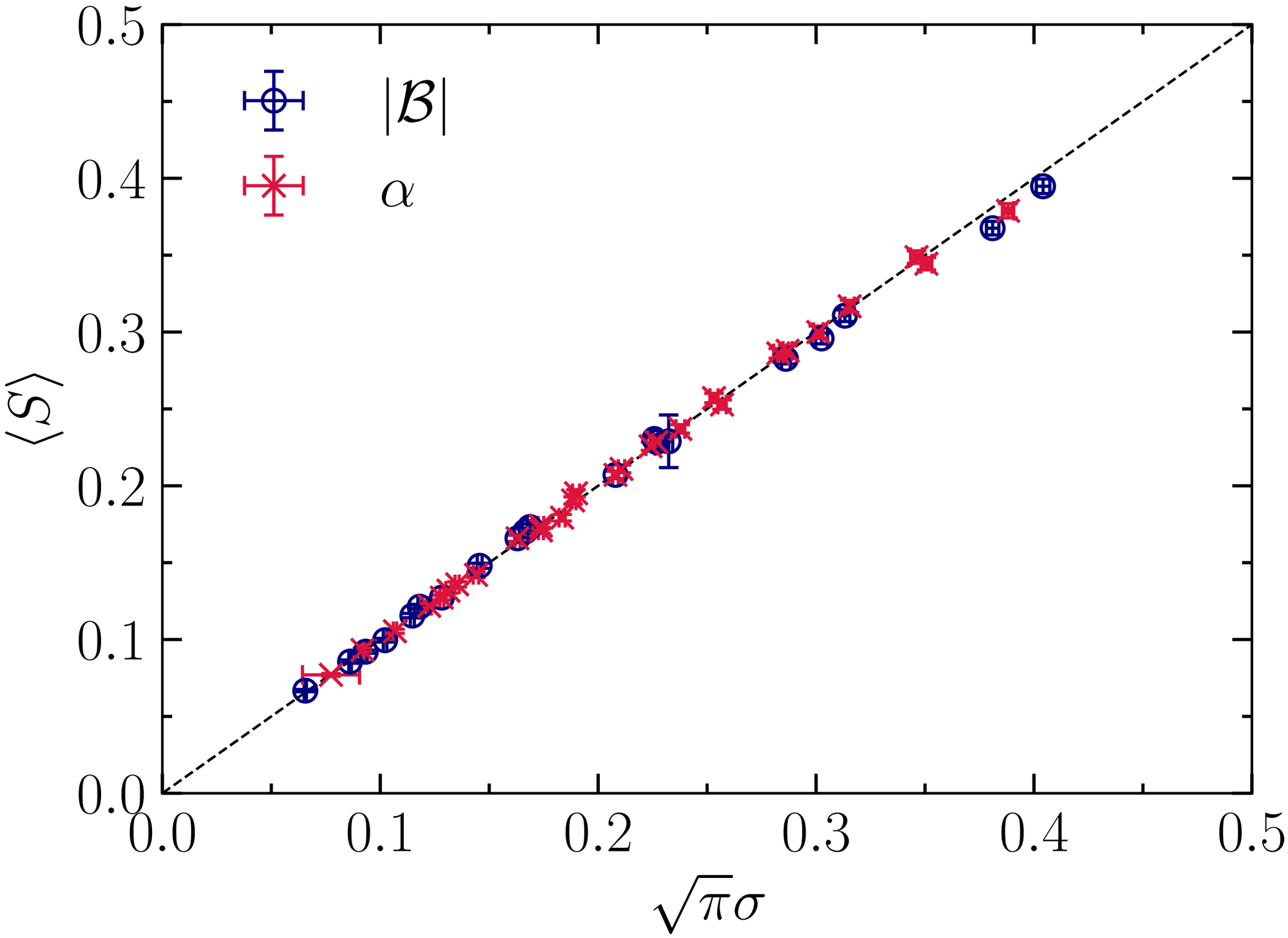
- Stochasticity of the model scales with the universal scaling factor $\alpha/|\mathcal{B}|$.

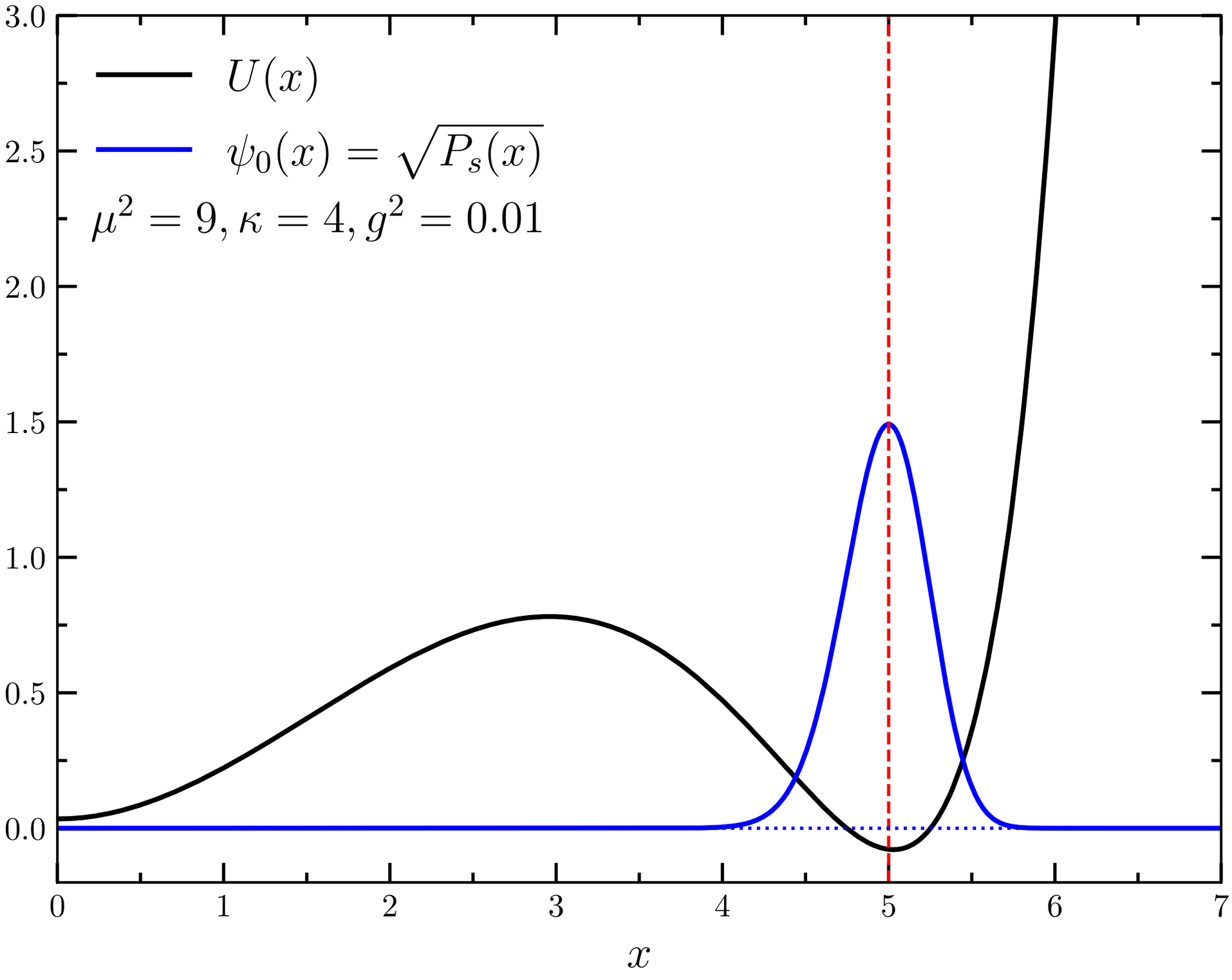


- Training error (precision) is bounded below by the fluctuation.

- Extending the experiment to the general neural network is in process. ⋯ see Matteo's poster

More nice plots


0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
0.4
0.5
$|\mathcal{B}|$
α
$\langle S \rangle$
$\sqrt{\pi}\sigma$


$U(x)$
$\psi_0(x) = \sqrt{P_s(x)}$
$\mu^2 = 9, \kappa = 4, g^2 = 0.01$
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
6
7
x