Linear scaling rule from random matrix theory

Chanju Park,

Robots-in-disguise

G. Aarts, B. Lucini, C. Park, arXiv:2407.16427

Outline

Observation

1. Hyperparameters in the Stochastic Gradient Descent
2. Empirical linear scaling rule

Theory

3. Langevin equation describing the SGD
4. Eigenvalue distribution of weight matrices
5. Linear scaling rule from theory

Stochastic Gradient Descent

Stochasticity is introduced from the finite sample size effect.

$$W_{ij}^{(n+1)} = W_{ij}^{(n)} - \alpha \left\langle \Delta_p \right\rangle_{p \in B},$$

α: Step size

B: Batch

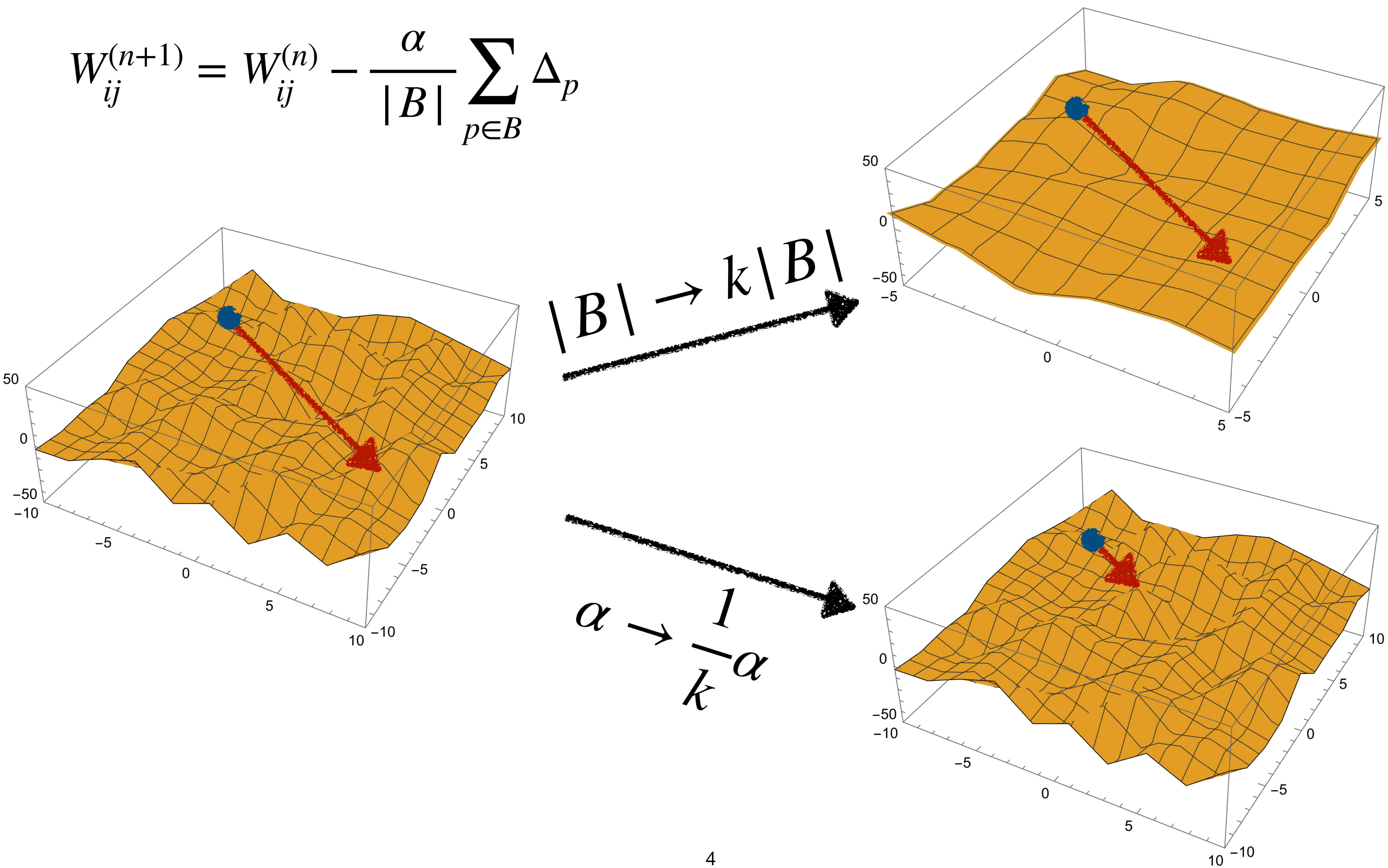
where,

$$\left\langle \Delta_p \right\rangle_{p \in B} \equiv \frac{1}{|B|} \sum_{p \in B} \Delta_p, \quad \Delta_p \equiv \frac{\partial \mathscr{L}}{\partial W_{ij}^{(n)}} \bigg|_p$$

Two hyperparameters in the SGD algorithm, α and $|B|$. How should we choose them?

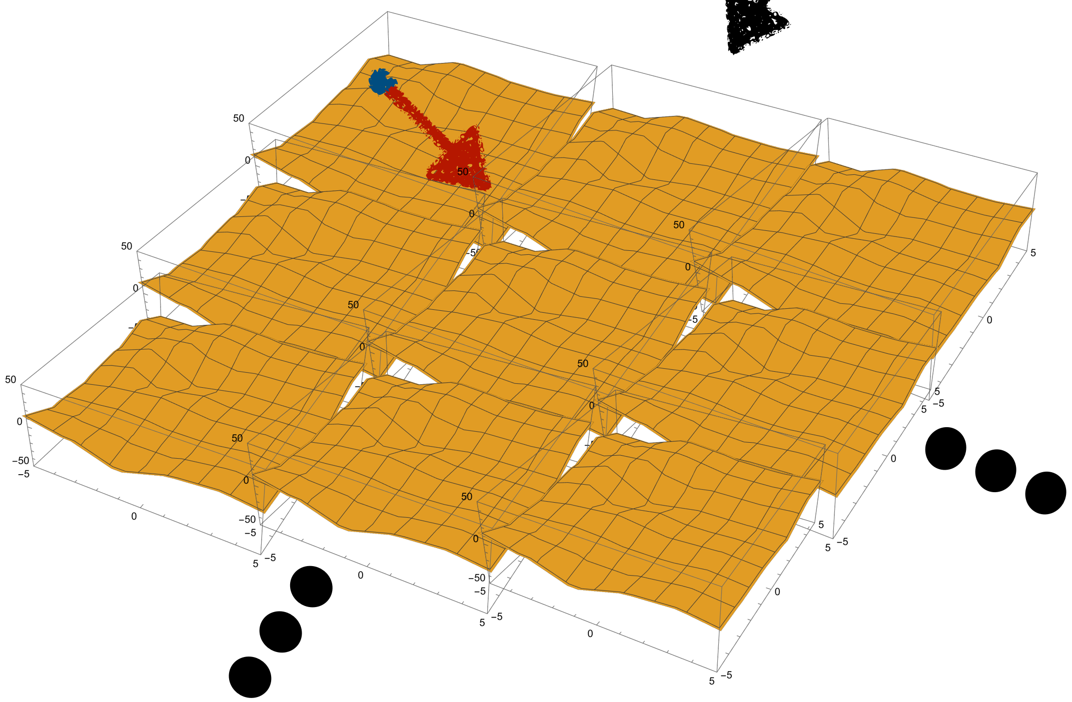
Geometrical intuition on α and B

$$W_{ij}^{(n+1)} = W_{ij}^{(n)} - \frac{\alpha}{|B|} \sum_{p \in B} \Delta_p$$

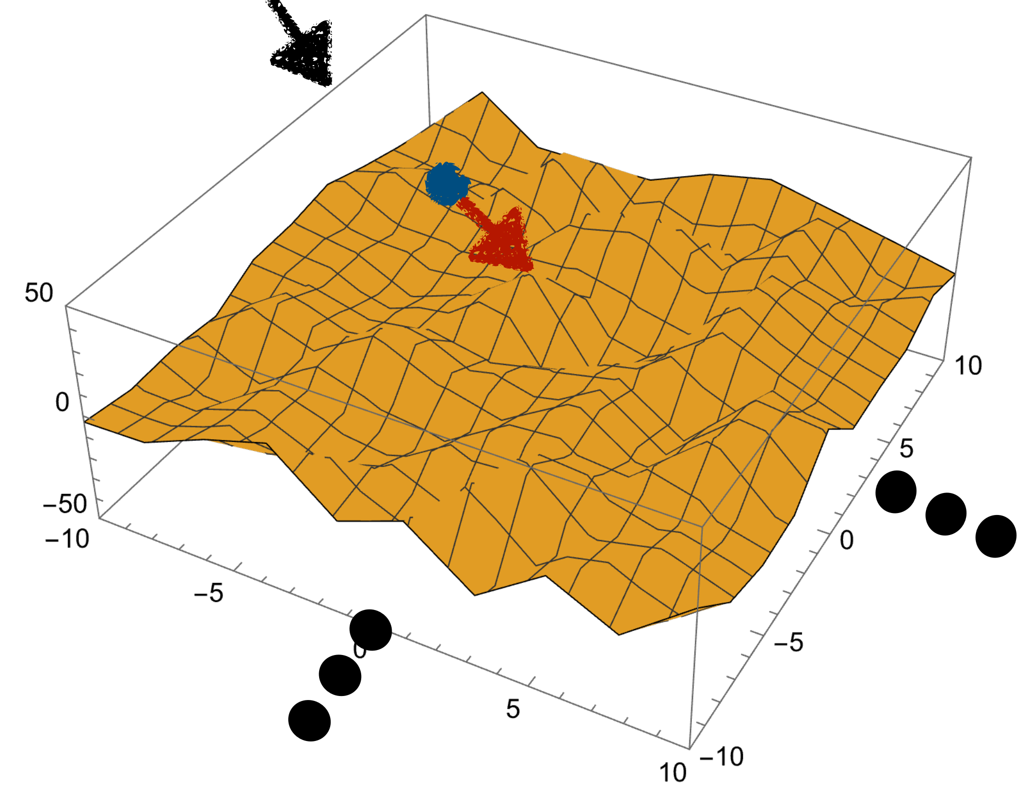


Geometric intuition on α and B

$$W_{ij}^{(n+1)} = W_{ij}^{(n)} - \frac{\alpha}{k|B|}\sum_{p\in B}\Delta_p = W_{ij}^{(n)} - \frac{\alpha/k}{|B|}\sum_{p\in B}\Delta_p$$



$=$



In infinite dataset limit $D \to \infty$ (or practically $|B| \ll D$), reducing α by a factor of k is equivalent to increasing $|B|$ by a same factor.

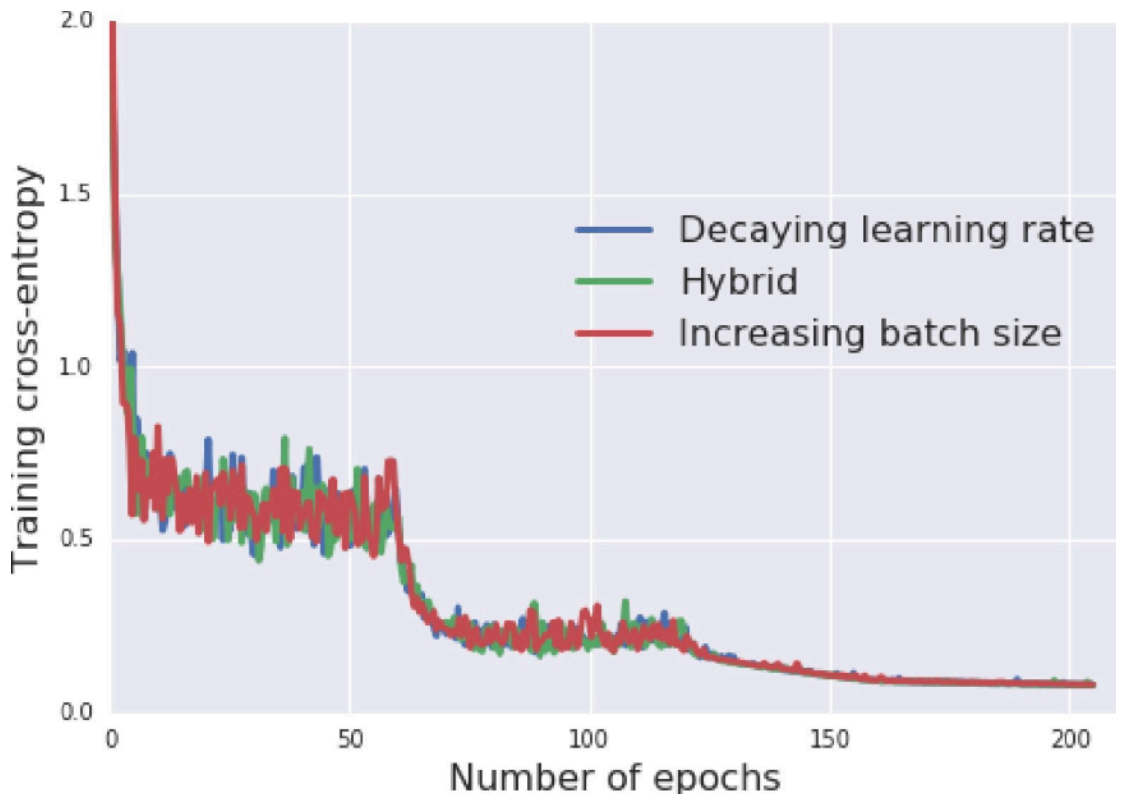
Linear Scaling Rule

Empirical scaling relation between learning rate and batch size.

Linear scaling relation between learning rate and batch size is widely known in practical ML training.

$$\text{Training quality} \propto \frac{\alpha/k}{|B|} = \frac{\alpha}{k|B|}.$$

α: Learning rate, $|B|$: Batch size.



The learning curve is equivalent in both cases. Figure from [1].

[1] S. L. Smith, et. al., *Don't Decay the Learning Rate, Increase the Batch Size,* arXiv:1711.00489
[2] P. Goyal, et. al. *Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour*, arXiv:1706.02677

Langevin dynamics

The Langevin equation can model an object undergoing stochastic motion.

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

$$\frac{dx}{dt} = \underbrace{-K(x;t)}_{\text{Mean drift}} + \underbrace{\sqrt{2}g(x;t)\eta}_{\text{Fluctuation}}$$

We can study the dynamics of the training using the corresponding Langevin equation.

Langevin equation for SGD

$$W_{ij}^{(n+1)} = W_{ij}^{(n)} - \alpha \left\langle \Delta_p \right\rangle_{p\in B},$$

α: Step size
B: Batch

$$\left\langle \Delta_p \right\rangle_{p\in B} \equiv \frac{1}{|B|}\sum_{p\in B} \Delta_p, \quad \Delta_p \equiv \left.\frac{\partial \mathscr{L}}{\partial W_{ij}^{(n)}}\right|_p$$

Assuming training data being i.i.d., the fluctuation can be separated by central limit theorem.

$$\left\langle \Delta_p \right\rangle_{p\in B} \sim \mathscr{N}\left(\mathbb{E}_B\left[\Delta\right], \frac{1}{|B|}\mathbb{V}_B\left[\Delta\right]\right)$$

The Langevin equation describing the SGD update is given as,

$$W_{ij}^{(n+1)} = W_{ij}^{(n)} - \alpha \mathbb{E}_B\left[\Delta_{ij}\right] + \frac{\alpha}{\sqrt{|B|}}\sqrt{\mathbb{V}_B\left[\Delta_{ij}\right]}\eta$$

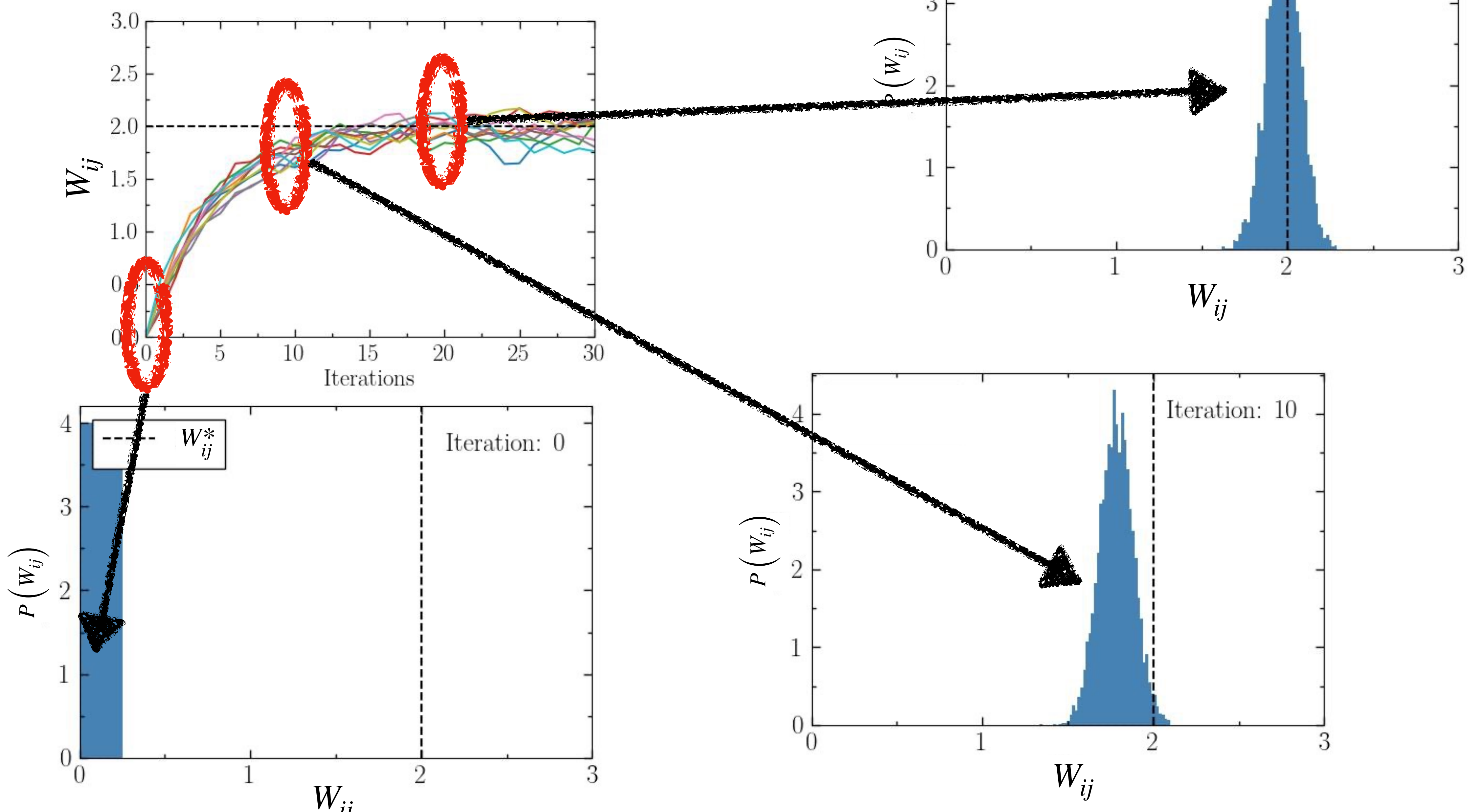
Drift

Fluctuation

$\eta \sim \mathscr{N}(0,1)$

Weight matrix = Random matrix

$$W_{ij}^{(n+1)} = W_{ij}^{(n)} - \alpha \mathbb{E}_B\left[\Delta_{ij}\right] + \frac{\alpha}{\sqrt{|B|}}\sqrt{\mathbb{V}_B\left[\Delta_{ij}\right]}\,\eta$$

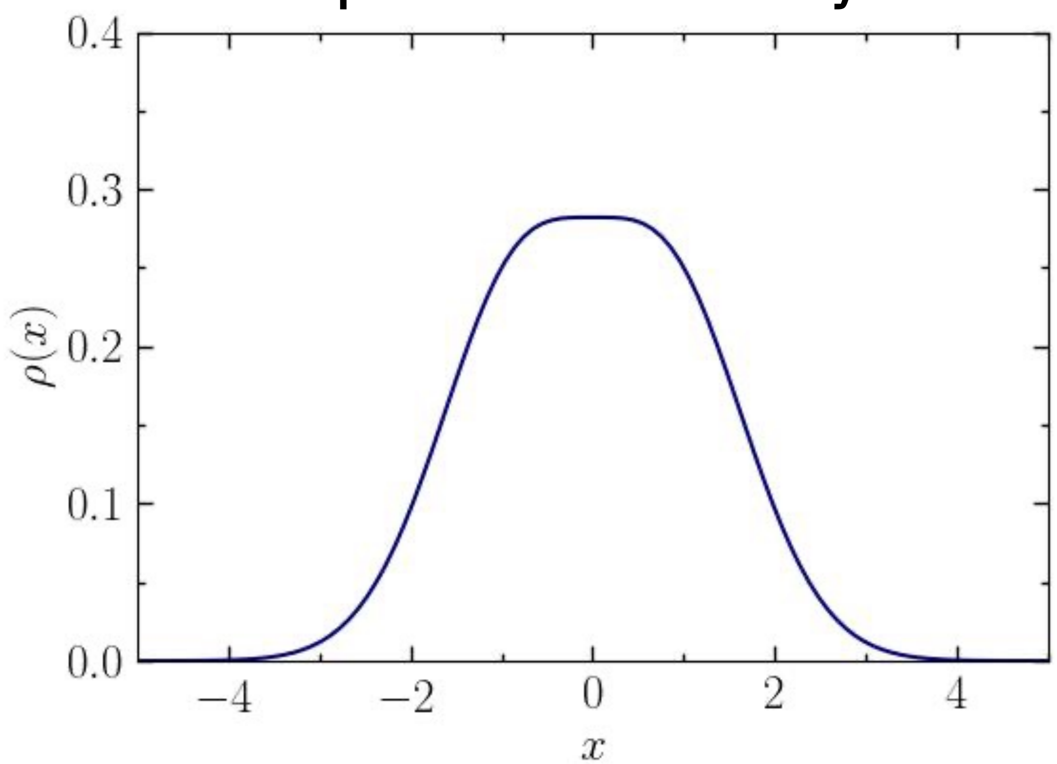


At each time slice, matrix elements are randomly distributed => Random Matrix!

Random Matrix Theory

Some useful spectral properties are known for random matrices.

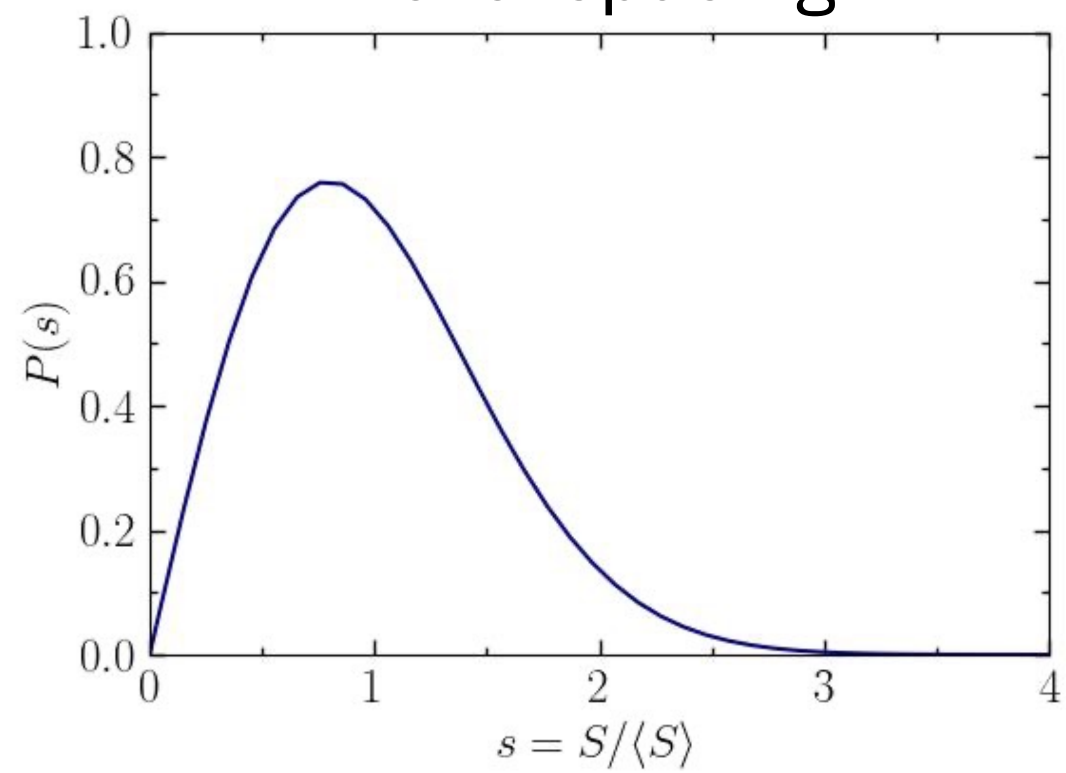
Spectral density



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

"Wigner semi-circle"

Level spacing



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

"Wigner surmise"

Dyson Brownian motion

The Langevin equation of eigenvalues can be derived from the random matrix theory.

$$x_i^{(n+1)} = x_i^{(n)} - \alpha \mathbb{E}_B \left[\Delta_{ii}\right] + \frac{\alpha^2}{|B|} \sum_{j \neq i} \frac{\mathbb{V}_B \left[\Delta_{ij}\right]}{x_i - x_j} + \frac{\alpha}{\sqrt{|B|}} \sqrt{\mathbb{V}_B \left[\Delta_{ii}\right]} \eta_i$$

An additional repulsion term is introduced from the Jacobian determinant.

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

Matrix elements | Eigenvalues

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

Distribution of eigenvalues

The distribution of the eigenvalues can be obtained by solving the associated Fokker-Planck equation.

$$x_i^{(n+1)} = x_i^{(n)} \underbrace{- \alpha \mathbb{E}_B \left[\Delta_{ii}\right] + \frac{\alpha^2}{|B|} \sum_{j \neq i} \frac{\mathbb{V}_B \left[\Delta_{ij}\right]}{x_i - x_j}}_{\equiv K_{ii}(x)} + \frac{\alpha}{\sqrt{|B|}} \sqrt{\underbrace{\mathbb{V}_B \left[\Delta_{ii}\right]}_{\equiv g_{ii}^2(x)}} \eta_i$$

Simplify the notation $\equiv K_{ii}(x)$ $\equiv g_{ii}^2(x)$

The Fokker-Planck equation is given by,

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

Linear Scaling Rule again

The linear scaling rule is obtained starting from the SGD equation.

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

Stationary limit solution: Coulomb gas distribution

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

The stationary distribution scales with a combination of scaling factors coming from the optimiser and the model architecture.

Width of the distribution

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

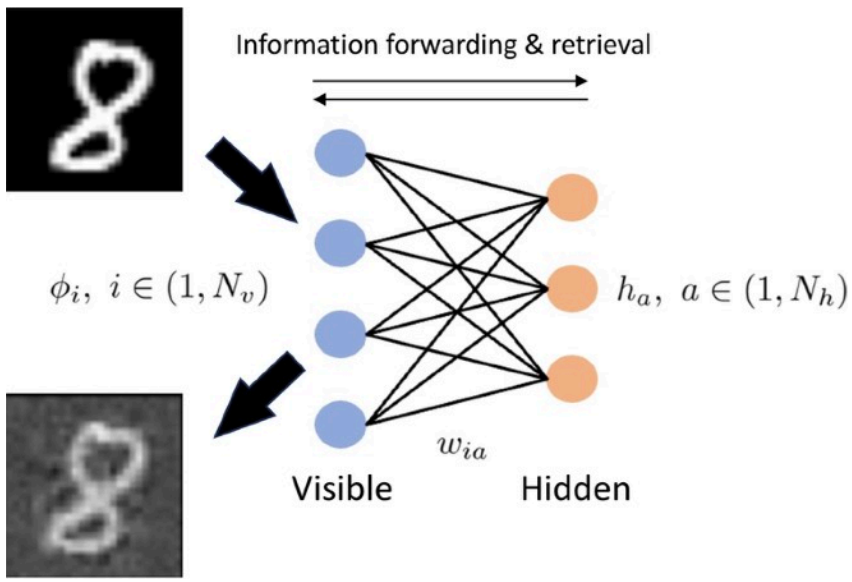
Linear Scaling Rule

(Optimiser)

Model-specific scaling

(Loss function, architecture, etc.)

Gaussian Restricted Boltzmann Machine



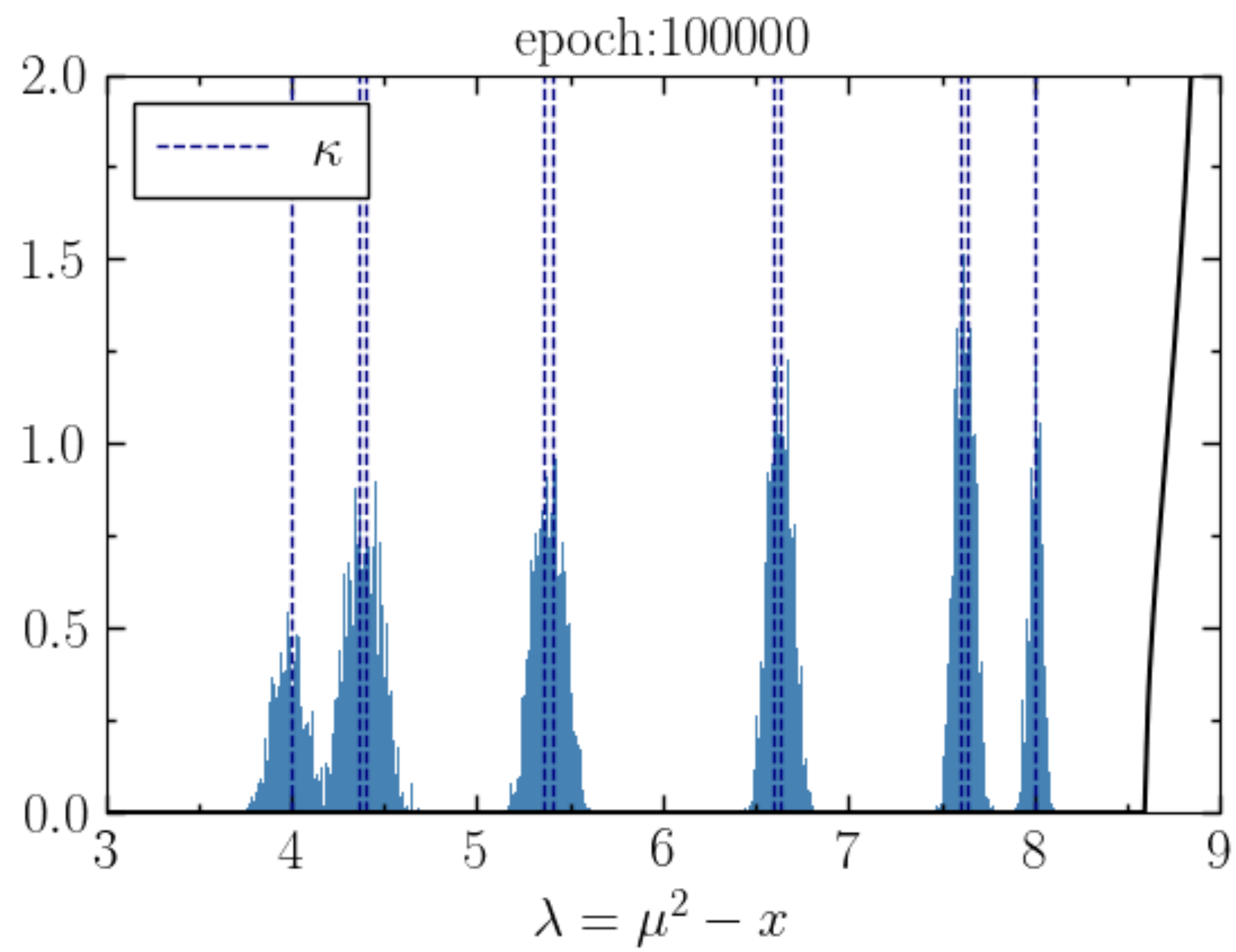
Gaussian RBM is an analytically solvable model and we can test the scaling law with the analytic calculation.

Target eigenvalues:

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

Gradient (drift) of Scalar field RBM:

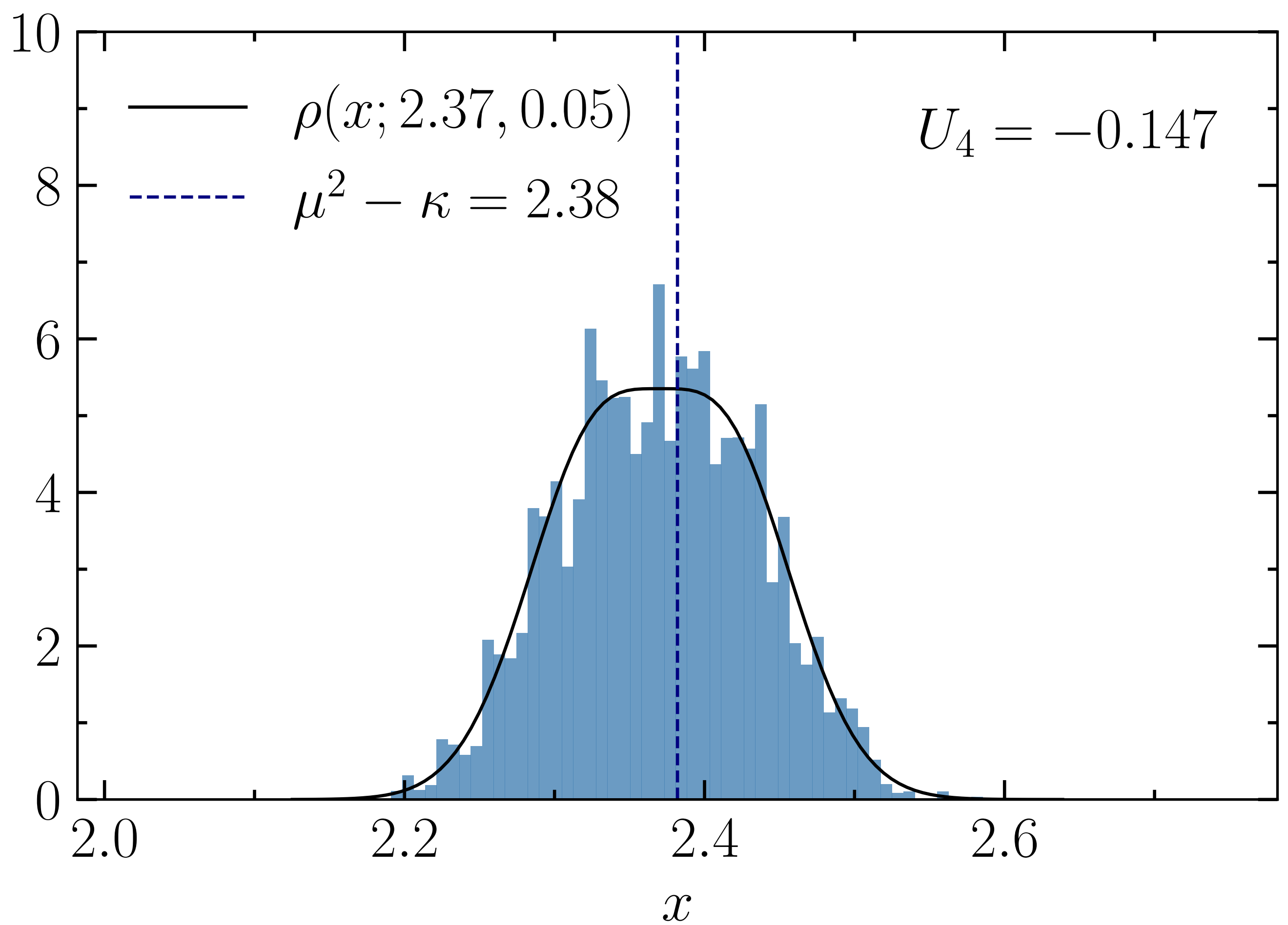
$$\frac{\partial \mathscr{L}}{\partial W_{ii}} \Rightarrow K_i\left(x_i\right) = \left(\frac{1}{\kappa_i} - \frac{1}{\mu^2 - x_i}\right) x_i$$

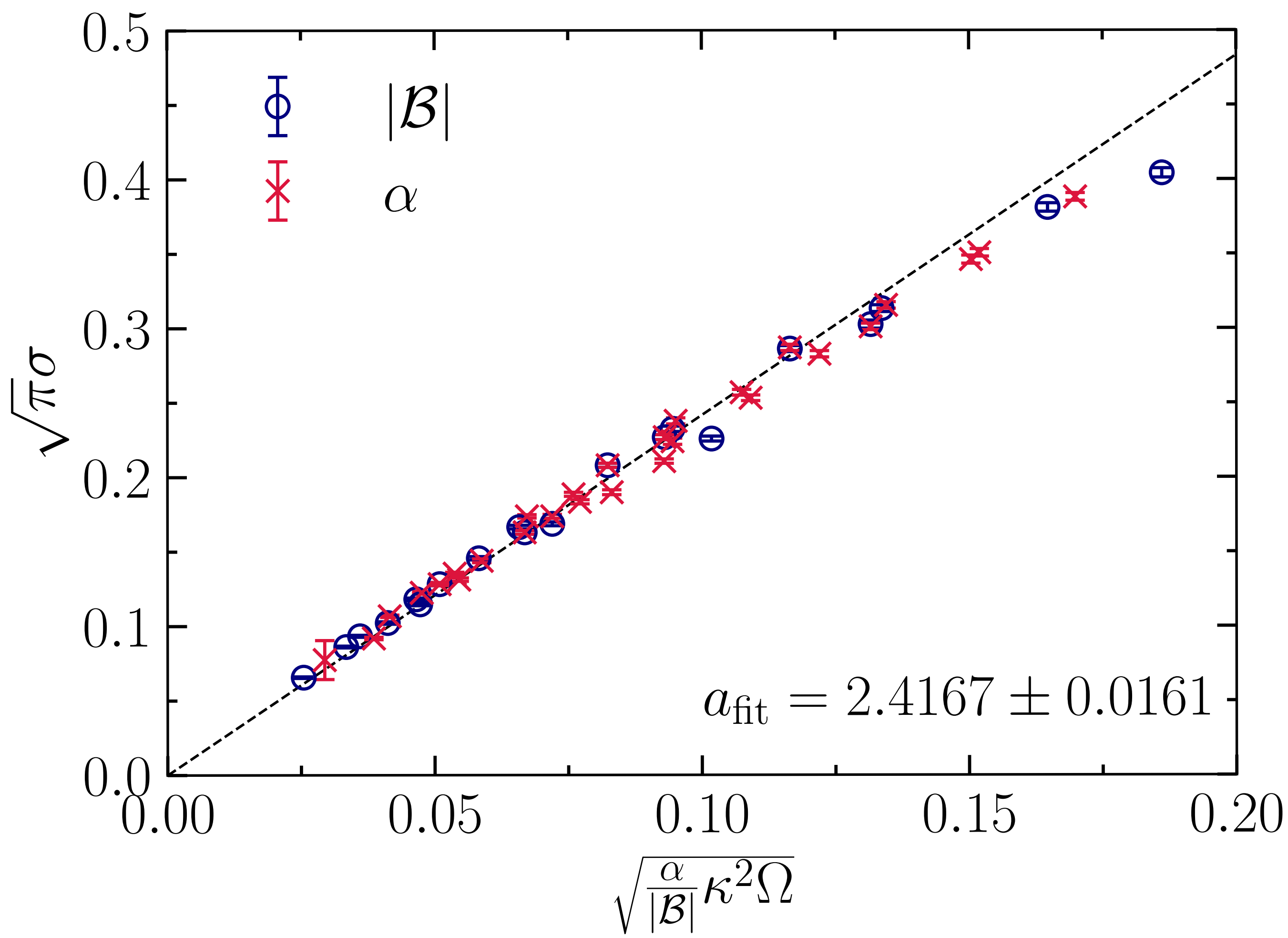


Trained eigenvalue of Gaussian RBM

G. Aarts, B. Lucini, C. Park, *Scalar field Restricted Boltzmann Machine as an ultraviolet regulator*, arXiv:2309.15002

Spectral density



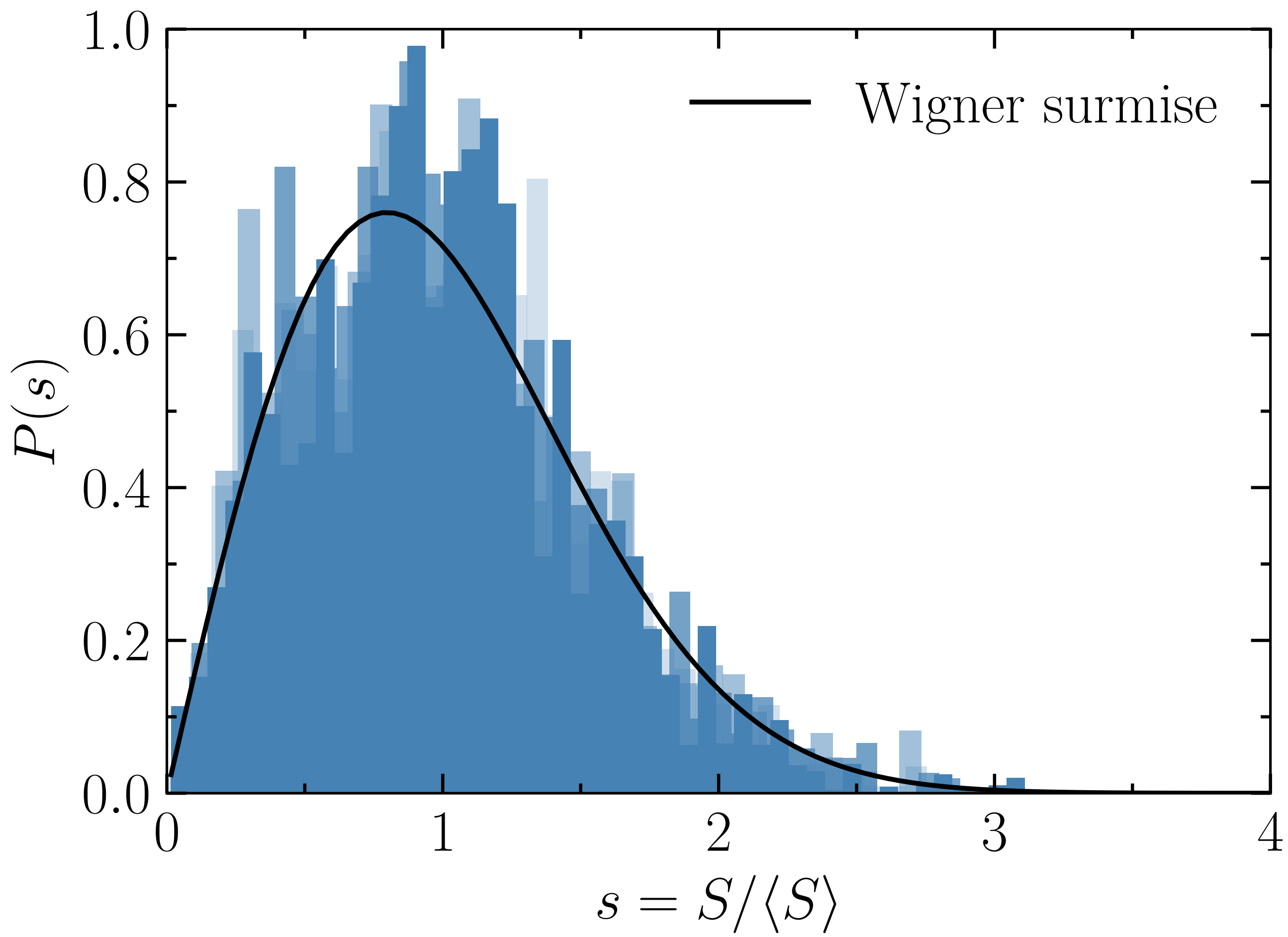


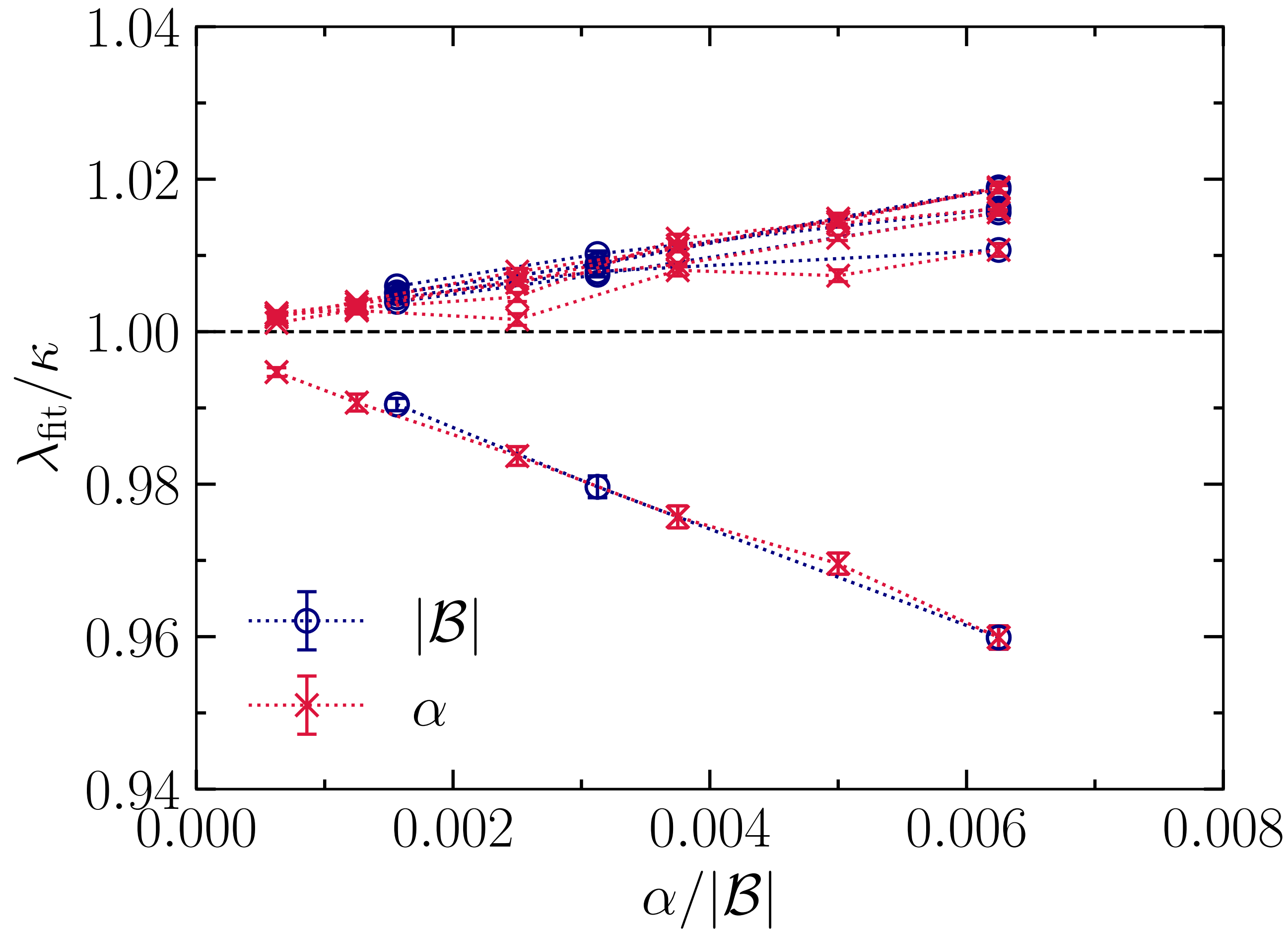
Eigenvalue distribution follows the Wigner semi-circle.

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

The width of the distribution follows the linear scaling rule $\alpha/|B|$.

Level spacing





Mean level spacing collapses into the universal curve.

Correct eigenvalues are retrieved only in the $\alpha/|B| \to 0$ limit.

Summary and Outlook

- Linear scaling relation between the learning rate α and the batch size $|B|$ has been empirically observed.
- The training dynamics of SGD can be described using Langevin dynamics.
- The linear scaling rule can be derived analytically from the random matrix theory.

Thank you!