PROCEEDINGS
OF SCIENCE

# Random Matrix Theory for Stochastic Gradient Descent

**Chanju Park,**[a,*] **Matteo Favoni,**[a,*] **Biagio Lucini**[b] **and Gert Aarts**[a]

[a]*Department of Physics, Swansea University,*
*Swansea, SA2 8PP, United Kingdom*

[b]*Department of Mathematics, Swansea University (Bay Campus),*
*Swansea, SA1 8EN, United Kingdom*

*E-mail:* chanju.b.park@gmail.com, matteo.favoni@swansea.ac.uk,
b.lucini@swansea.ac.uk, g.aarts@swansea.ac.uk

Investigating the dynamics of learning in machine learning algorithms is of paramount importance for understanding how and why an approach may be successful. The tools of physics and statistics provide a robust setting for such investigations. Here, we apply concepts from random matrix theory to describe stochastic weight matrix dynamics, using the framework of Dyson Brownian motion. We derive the linear scaling rule between the learning rate (step size) and the batch size, and identify universal and non-universal aspects of weight matrix dynamics. We test our findings in the (near-)solvable case of the Gaussian Restricted Boltzmann Machine and in a linear one-hidden-layer neural network.

*The 41st International Symposium on Lattice Field Theory (LATTICE2024)*
*28 July - 3 August 2024*
*Liverpool, UK*

*∗Speaker*

## 1. Introduction

Machine learning (ML) and artificial intelligence (AI) can provide powerful tools for the scientific community, as demonstrated by the recent Nobel Prize in Chemistry. Reversely, insights from traditional physics theories also contribute to a deeper understanding of the mechanism of learning. Ref. [1] contains a broad overview of the successful cross-fertilisation between ML and the physical sciences, covering a number of domains. One way to mitigate against possible scepticism with regard to using ML as a "black box" is by unveiling the dynamics of training (or learning) and explaining how the relevant information is engraved in the model during the training stage.

To further develop this programme, we study here the dynamics of first-order stochastic gradient descent as applied to weight matrices, reporting and expanding on the work presented in Ref. [2]. When training ML models, weight matrices are commonly updated by one of the variants of the stochastic gradient descent algorithm. The dynamics can then be decomposed into a drift and a fluctuating term, and such a system can be described by a discrete Langevin equation. The dynamics of stochastic matrix updates is richer than the dynamics for vector or scalar quantities, as captured by Dyson Brownian motion and random matrix theory (RMT), with the appearance of universal features for the eigenvalues [3–9]. Earlier descriptions of the statistical properties of weight matrices in terms of RMT can be found in e.g. Refs. [10, 11], but here we specifically focus on RMT effects for stochastic gradient dynamics via Dyson Brownian motion, which leads to additional Coulomb-type repulsion between eigenvalues due to the Vandermonde determinant. Importantly, we have shown that in this framework a specific combination of hyperparameters of the optimiser, namely the ratio of the learning rate (or step size) and the batch size, naturally arises as a scaling factor determining the strength of the fluctuations in the process [2]. In fact, this specific combination had already been observed at an empirical level in practical ML training and dubbed the *linear scaling rule* [12–15]. We derived this relation from first-principle matrix dynamics [2].

In this contribution, we first summarise the results of Ref. [2] and then present some new results for a simple linear neural network with one hidden layer. Related results for a nano-GPT can be found in Ref. [16].

## 2. Langevin equation for stochastic gradient descent

Stochastic gradient descent (SGD) is one of the most commonly used first-order gradient optimisation algorithms in the ML community. Given an objective function $\mathcal{L}(W)$ depending on a (collection of) weight matrices $W$, the optimal state that minimises the objective function is found by searching for the stationary point of the first-order equation,

$$W_{n+1} = W_n - \alpha \left\langle \Delta_P \right\rangle_{p \in \mathcal{B}}, \tag{1}$$

where

$$\left\langle \Delta_P \right\rangle_{p \in \mathcal{B}} \equiv \frac{1}{|\mathcal{B}|} \sum_{p \in \mathcal{B}} \Delta_p, \qquad \Delta_p \equiv \left. \frac{\partial \mathcal{L}}{\partial W_n} \right|_p. \tag{2}$$

Here $\alpha$ is the learning rate and the gradient $\Delta_p$ of the objective function $\mathcal{L}(W)$, depending on the current state $W_n$, is calculated for each data point $p$ and then averaged over the data points within a (mini-)batch $\mathcal{B}$.

Unlike in standard gradient descent, the input data is split into a number of small mini-batches and stochasticity is introduced due to the effect of having finite sample sizes. Assuming that the input data is well standardised, each gradient within a mini-batch is an i.i.d. random variable as each data point in the mini-batch is randomly sampled from the total dataset. As the measured batch gradient is the average of i.i.d. variables, we can use the central limit theorem to write $\langle \Delta_p \rangle$ in terms of the mean gradient of the batch and its fluctuation,

$$\langle \Delta_p \rangle_{p \in \mathcal{B}} = \mathbb{E}_{\mathcal{B}}[\Delta_p] + \frac{1}{\sqrt{|\mathcal{B}|}} \sqrt{\mathbb{V}_{\mathcal{B}}[\Delta_p]} \, \eta, \qquad \eta \sim \mathcal{N}(0, 1), \tag{3}$$

where $\mathbb{E}_{\mathcal{B}}[\Delta_p]$ and $\mathbb{V}_{\mathcal{B}}[\Delta_p]$ are the mean and the variance of the gradient distribution of the batch respectively, and $\eta$ is Gaussian noise. Rewriting Eq. (1) in terms of the mean drift and the fluctuation then yields

$$W_{n+1} = W_n - \alpha \mathbb{E}_{\mathcal{B}}[\Delta_p] + \frac{\alpha}{\sqrt{|\mathcal{B}|}} \sqrt{\mathbb{V}_{\mathcal{B}}[\Delta_p]} \, \eta, \tag{4}$$

i.e., a discrete Langevin equation for SGD.

Note that the learning rate does not have a natural interpretation as a 'physical' step size, as naively sending $\alpha \to 0$ does not give a correct stochastic differential equation [17–19]. Instead, to obtain a continuous time limit that satisfies Itô calculus, one must consider a combination of parameters that jointly acquire the dimension of time, which we plan to discuss in the future.

## 3. Dynamics of eigenvalues, random matrix theory and the Coulomb gas

To follow the dynamics during learning, it is convenient to work with singular or eigenvalues, as some statistical properties of those are well known in RMT. RMT is usually defined for square symmetric or hermitian matrices. Since weight matrices in ML are typically rectangular, of size $M \times N$, we consider the symmetric combination $X = W^T W$. The update for $X$ follows directly from the update for $W$ given above, using $\delta X = W^T \delta W + \delta W^T W$. We denote the eigenvalues of $X$ with $x_i$ $(i = 1, \ldots, N)$, where we assume $N \leq M$ (if not, swap $W$ and $W^T$). Note that the eigenvalues $x_i$ are real and non-negative.

To obtain a discrete Langevin equation for the eigenvalues, we would have to write the matrix as a product of rotations and a diagonal matrix, and separate the dynamics of the eigenvalues from the dynamics of the angles, which is non-trivial in general. A key result from Dyson Brownian motion [8, 9] is that the equation satisfied by the eigenvalues of $X$ can be written down directly in terms of the drift and fluctuations of $X$, as well as a Coulomb term, namely [2]

$$x_i' = x_i + \alpha \tilde{K}_i + \frac{\alpha^2}{|\mathcal{B}|} \sum_{j \neq i} \frac{\tilde{g}_i^2}{x_i - x_j} + \frac{\alpha}{\sqrt{|\mathcal{B}|}} \sqrt{2} \tilde{g}_i \eta_i. \tag{5}$$

where $\tilde{K}_i = -\mathbb{E}_{\mathcal{B}} [\Delta_p]_{ii}$ and $2\tilde{g}_i^2 = \mathbb{V}_{\mathcal{B}} [\Delta_p]_{ii} = 2\mathbb{V}_{\mathcal{B}} [\Delta_p]_{i \neq j}$. Quantities with a tilde are independent of learning rate and batch size at leading order. The additional Coulomb-type interaction arises from the Jacobian determinant of the change of variables from matrix elements to eigenvalues, cf. the Vandermonde determinant. Intuitively, the interaction term originates from the fact that the

orthogonal transformation that diagonalises $X$ does not necessarily diagonalise the noise matrix $\eta_{ij}$, as in the equivalent of Eq. (4) for $X$.

After having obtained the Langevin equation for the eigenvalues, we can proceed and solve the associated Fokker-Planck equation to study the stationary distribution. The Fokker-Planck equation reads (using continuous time here, we are mostly interested in the stationary solution)

$$\partial_t P\left(\{x_i\}, t\right) = \sum_{i=1}^{N} \partial_{x_i} \left[ \frac{\alpha^2}{|\mathcal{B}|} \tilde{g}_i^2 \partial_{x_i} - \alpha \tilde{K}_i - \frac{\alpha^2}{|\mathcal{B}|} \sum_{j \neq i} \frac{\tilde{g}_i^2}{x_i - x_j} \right] P\left(\{x_i\}, t\right). \tag{6}$$

The stationary distribution, $\partial_t P\left(\{x_i\}, t\right) = 0$, is solved using the Coulomb gas description [8, 9]

$$P_s(\{x_i\}) = \frac{1}{Z} \prod_{i<j} |x_i - x_j| \exp\left[ -\sum_i \frac{1}{\alpha/|\mathcal{B}|} \frac{\tilde{V}_i(x_i)}{\tilde{g}_i^2} \right], \qquad \tilde{K}_i(x_i) = -\frac{d\tilde{V}_i(x_i)}{dx_i}, \tag{7}$$

where the potential $\tilde{V}_i(x_i)$ of the Coulomb gas is defined via the drift. If one assumes there is a unique minimum $x_i = x_i^s$ for each eigenvalue, such that the potential can be expanded as

$$\tilde{V}_i(x_i) = \tilde{V}_i(x_i^s) + \frac{1}{2}\Omega_i \left(x_i - x_i^s\right)^2 + \ldots, \tag{8}$$

with $\Omega_i$ the curvature around the minimum, the Coulomb gas potential becomes a sum of Gaussians centred at $x_i = x_i^s$ with variance

$$\sigma_i^2 = \frac{\alpha}{|\mathcal{B}|} \frac{\tilde{g}_i^2}{\Omega_i} = \frac{\alpha}{|\mathcal{B}|} \frac{\mathbb{V}_{\mathcal{B}}\left[\Delta_P\right]_{ii}}{2\Omega_i}, \tag{9}$$

The beauty of expression (9) is that the contributions to the fluctuations in the system are clearly separated into two factors from different sources, with the first part $\alpha/|\mathcal{B}|$ solely coming from the stochasticity of the optimiser, which leads to the linear scaling rule [12], and the second part $\mathbb{V}_{\mathcal{B}}\left[\Delta_P\right]_{ii}/\Omega_i$ only depending on the specific profile of the model [2].

## 4. Applications

In this section, we validate the RMT description of matrix-valued SGD and the linear scaling rule by observing the eigenvalue distribution of a Gaussian Restricted Boltzmann Machine and explore a dense linear neural network within the teacher-student setting.

### 4.1 Gaussian Restricted Boltzmann Machine

Restricted Boltzmann Machines (RBMs) are generative energy-based models consisting of two layers as shown in Fig. 1 [20–22]. When a sample is fed into the visible layer, the value of the hidden layer is sampled from a conditional probability distribution obtained by marginalising visible degrees of freedom of the given energy function of the model. Subsequently the output of the model is sampled from a conditional probability distribution obtained by marginalising the hidden degrees of freedom. The model is stochastically updated by maximising the log-likelihood between the model distribution and the target distribution. If both the visible and hidden degrees of
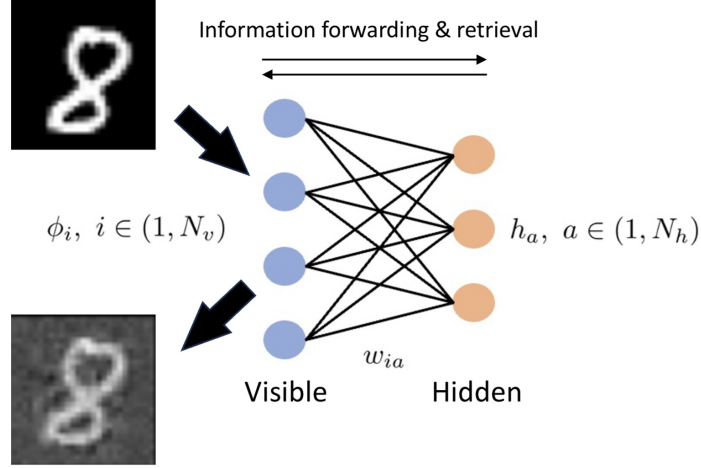
**Figure 1:** General structure of a Restricted Boltzmann Machine, with $N_v$ ($N_h$) visible (hidden) nodes.

freedom are Gaussian fields, interacting via a bilinear coupling, $\phi_i W_{ia} h_a$, one obtains a Gaussian RBM, with a probability distribution $p(\phi, h) \sim \exp[-S(\phi, h)]$ and the "action",

$$S(\phi, h) = \frac{1}{2}\mu^2 \phi^T \phi + \frac{1}{2\sigma_h^2}(h - \eta)^T (h - \eta) - \phi^T W h. \tag{10}$$

Here $\mu^2$ and $\sigma_h^2$ are hyperparameters and we will put the bias $\eta = 0$. A full analysis of this model using the language of lattice field theory (LFT) can be found in Ref. [23].

We train the Gaussian RBM to learn a distribution representing a one-dimensional non-interacting lattice scalar field theory, i.e., the eigenvalues of the target distribution are given as $\kappa_n = m^2 + 2 - 2\cos(2\pi n/N)$, where $m$ is the mass of the scalar field, $N = N_v$ is the size of the lattice and $-N/2 < n \le N/2$. The smallest and largest eigenvalues are non-degenerate, while all the intermediate ones are doubly degenerate. We denote the learnt RBM eigenvalues as $\lambda_i = \mu^2 - x_i$, where $x_i = \sigma_h^2 \xi_i^2$, with $\xi_i$ the singular values of $W$ [2, 23].

After training, the eigenvalues of the RBM flow towards the target eigenvalues, but rather than learning the exact values, they form a distribution around the target values, as shown in Fig. 2 (left). The precise structure of the distributions can be analysed using RMT [2]. In Fig. 2 (right) we show one of the distributions (note that $\mu^2 = 9$ and hence $x = 9 - \lambda$). Using the Coulomb gas description, the predicted spectral density, a.k.a. the Wigner semi-circle, for each pair of doubly-degenerate eigenvalues, is

$$\rho(x; x_m, \sigma) = \frac{1}{N}\sum_{i=1}^{N}\langle\delta(x - x_i)\rangle \overset{N=2}{=} \frac{e^{-\delta x^2/(2\sigma^2)}}{4\sqrt{\pi}\sigma}\left[2e^{-\delta x^2/(2\sigma^2)} + \sqrt{2\pi}\frac{\delta x}{\sigma}\text{Erf}\left(\frac{\delta x}{\sqrt{2}\sigma}\right)\right], \tag{11}$$

where $\delta x = x - x_m$, with $x_m$ the centre of the distribution, and $\sigma \sim \sqrt{\alpha/|\mathcal{B}|}$ is given by Eq. (9) and used as a fit parameter. Fig. 2 (right) shows the fit of this spectral density to one of the histograms. To illustrate that the histogram is indeed not a Gaussian we also compute the Binder cumulant $U_4$,

$$U_4 \equiv \frac{\langle\delta x^4\rangle}{3\langle\delta x^2\rangle^2} - 1 = -\frac{4}{27} \approx -0.148\ldots \quad \text{for the Wigner semi-circle,} \tag{12}$$
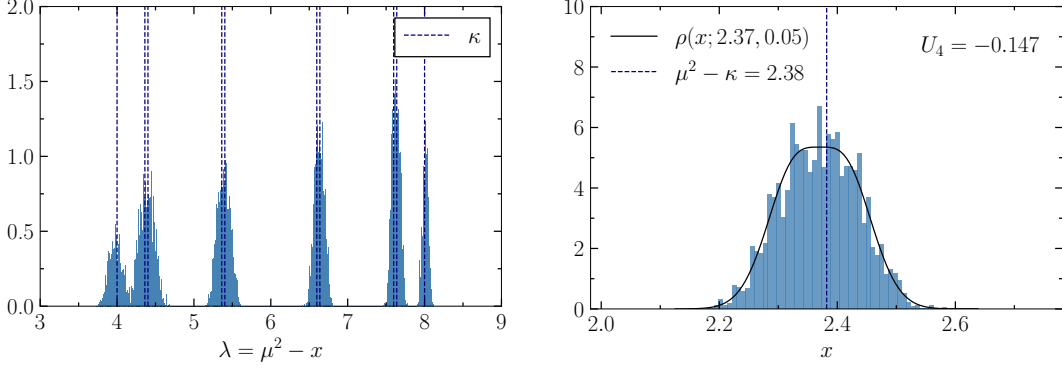
**Figure 2:** (Left) Target eigenvalues (dashed lines) and model eigenvalues (histograms) after training. The middle 8 target eigenvalues are doubly degenerate due to periodic boundary conditions. (Right) Close-up of one of the peaks: the learnt eigenvalue distribution of the RBM follows the Wigner semi-circle (solid line).
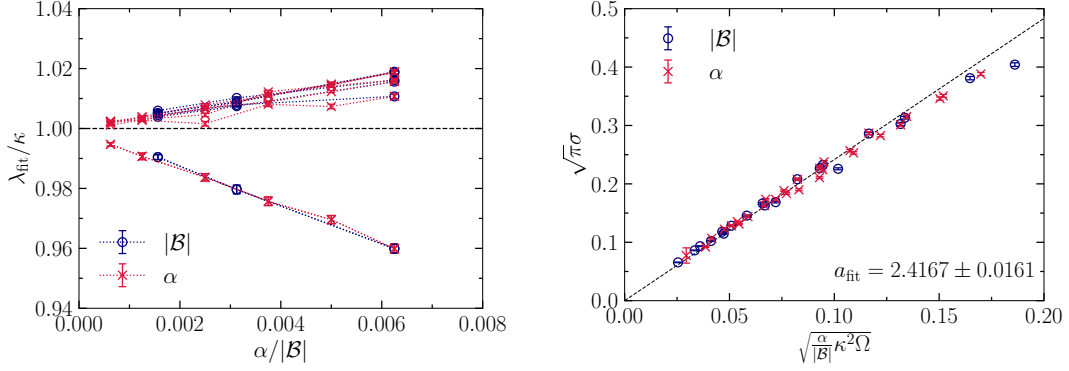


**Figure 3:** (Left) Deviation of the fitted centres of the model eigenvalue distributions from the target ones due to the inter-mode Coulomb interaction. The offset decreases as stochasticity in the model decreases. (Right) The width of eigenvalue distributions scales with the predicted universal scaling factor $\sqrt{\alpha/|\mathcal{B}|}$. In both cases $\alpha$ and $|\mathcal{B}|$ are varied independently.

and find excellent agreement. One may note in Fig. 2 (right) that the centre of the peak is slightly displaced from the target value. This is due to the Coulomb repulsion between all the modes, as demonstrated in Fig. 3 (left), where the ratio of the fitted peak centres and the target values is shown. Only in the limit of vanishing stochasticity ($\alpha/|\mathcal{B}| \to 0$) is the spectrum learnt exactly.

The universal appearance of $\alpha/|\mathcal{B}|$ is also demonstrated in Fig. 3 (right), where the dependence of the width $\sigma$ of the spectral density on $\sqrt{\alpha/|\mathcal{B}|}$ is shown, including the non-universal (model-dependent) factor $\kappa^2\Omega$. Finally, the spacing between nearest eigenvalues follows the Wigner surmise, $P(s) = (\pi/2)s \exp(-\pi s^2/4)$, where $s = S/\langle S \rangle$ and $\langle S \rangle = \sqrt{\pi}\sigma \sim \sqrt{\alpha/|\mathcal{B}|}$, see Ref. [2] for details.

### 4.2 Neural network in teacher-student setting

To extend the analysis to more general architectures, we consider here the simplest case of a neural network with one hidden layer in the teacher-student setting, with the activation function put equal to the identity. Even though this is a linear network, it introduces a new feature in the

spectral density, which is already interesting to understand, since it indicates how the architecture may interplay with the RMT structure seen above.

Teacher-student models are widely used in the statistical mechanics of learning, see e.g. Refs. [24, 25]. They are formulated in terms of a teacher and a student network, with weight matrices $W_t$ and $W_s$ respectively. The input data $\mathbf{x}$ is drawn from a normal distribution with unit variance. The teacher network has fixed weights and it is the task of the student to determine these weights. Denoting the output of the network with $\mathbf{y}_p = \mathbf{f}(\mathbf{x}_p; W)$ for each data point $\mathbf{x}_p$, this can be summarised as

$$\mathcal{L}(W_s) = \frac{1}{2P} \sum_{p=1}^{P} \left| \mathbf{y}_p^{(t)} - \mathbf{y}_p^{(s)} \right|^2, \qquad \mathbf{y}_p^{(t)} = \mathbf{f}(\mathbf{x}_p; W_t), \qquad \mathbf{y}_p^{(s)} = \mathbf{f}(\mathbf{x}_p; W_s), \quad (13)$$

where the sum in the loss function is over the $P$ data points.

The one-hidden-layer network function we use here can be written as

$$\mathbf{f}(\mathbf{x}; W) = Z \, \mathbf{a}(W\mathbf{x}). \tag{14}$$

Bold-faced quantities are vectors; $W$ and $Z$ are rectangular matrices in general. The input and output dimensions do not have to be the same. As stated, we replace the activation function $\mathbf{a}(\cdot)$ with the identity and we will not write it from now on. The $Z$ matrix is the same for the teacher and the student, such that only $W$ needs to be learnt.

The gradient of the loss function for a data point $\mathbf{x}$ with components $x_i$ is given by

$$\frac{\partial \mathcal{L}(W_s)}{\partial W_{s,i'i}} = - \sum_{j',j} (Z^T Z)_{i'j'} \, (W_t - W_s)_{j'j} \, x_j x_i, \tag{15}$$

where primed and unprimed indices may have a different range, reflecting that the matrices are typically rectangular. Averaging over a mini-batch, we can write

$$\frac{1}{|\mathcal{B}|} \sum_{p \in \mathcal{B}} x_{i,p} x_{j,p} \simeq \delta_{ij}. \tag{16}$$

This approximation ignores some stochasticity due to mini-batch sampling, but if the batches are not too small, we found that this can safely be ignored. The gradient of the loss function, averaged over a mini-batch, then reads

$$\left. \frac{\partial \mathcal{L}(W_s)}{\partial W_{s,i'i}} \right|_{\mathcal{B}} = - \sum_{j'} (Z^T Z)_{i'j'} \, (W_t - W_s)_{j'i}. \tag{17}$$

To analyse the dynamics analytically, we assume that some continuous time limit exists, while keeping $\alpha$ as an explicit learning rate, and consider the equation

$$W_s' = W_s - \alpha \left. \frac{\partial \mathcal{L}(W_s)}{\partial W_s} \right|_{\mathcal{B}} \qquad \Rightarrow \qquad \dot{W}_s = \alpha (Z^T Z)(W_t - W_s). \tag{18}$$

Here the dot indicates the time derivative and we no longer write the indices explicitly.

We can study this dynamics in more detail by introducing a singular value decomposition of both the teacher and the student weight matrix, writing

$$W_s = U_s \Xi_s V_s^T, \qquad W_t = U_t \Xi_t V_t^T, \tag{19}$$

where $U_{s,t}$ and $V_{s,t}$ are orthogonal matrices and $\Xi_{s,t}$ are diagonal matrices containing the singular values $\xi_{s,i}$ and $\xi_{t,i}$. Closely following Ref. [23], we take the time derivative of $W_s W_s^T$ and conjugate the resulting expression with $U_s^T$ and $U_s$. This yields

$$U_s^T \frac{d}{dt} (W_s W_s^T) U_s = \frac{d}{dt} (\Xi_s \Xi_s^T) + (U_s^T \dot{U}_s)(\Xi_s \Xi_s^T) + (\Xi_s \Xi_s^T)(\dot{U}_s^T U_s)$$
$$= \alpha (U_s^T Z^T Z U_s) \left[ U_s^T U_t \Xi_t V_t^T V_s \Xi_s^T - \Xi_s \Xi_s^T \right] + \alpha \left[ \Xi_s V_s^T V_t \Xi_t^T U_t^T U_s - \Xi_s \Xi_s^T \right] (U_s^T Z^T Z U_s). \tag{20}$$

A similar expression is obtained starting from $W_s^T W_s$, but with $V_s^T \dot{V}_s$ instead of $U_s^T \dot{U}_s$, etc. From these expressions we can clearly see the process of learning: the RHS of the equation vanishes when the singular values of $W_s$ and $W_t$ agree, $\Xi_s \to \Xi_t$, as well as the left and right basis, $U_s \to U_t, V_s \to V_t$. The rate of learning is determined by both $\alpha$ and the combination $U_s^T Z^T Z U_s$. Note that the final two terms on the first line form a symmetric matrix with zeroes on the diagonal, whereas the first term is purely diagonal.

We focus here on the diagonal terms, i.e. the eigenvalues of $\Xi_s \Xi_s^T$, or equivalently the square of the singular values of $\Xi_s$. We denote the eigenvalues of the student matrix with $x_i = \xi_{s,i}^2$, as in the RBM (there should be no confusion with the data points **x** discussed above). Assuming that $U_s \sim U_t, V_s \sim V_t$, the diagonal part of Eq. (20) can then be reduced to

$$\dot{x}_i = 2a_i \left( \sqrt{\kappa_i x_i} - x_i \right), \tag{21}$$

where $a_i > 0$ depends on $\alpha$ times the diagonal components of $U_s^T Z^T Z U_s$, and $\sqrt{\kappa_i}$ is the $i^{\text{th}}$ singular value of the teacher matrix $\Xi_t$. This equation is solved as

$$x_i(t) = \left[ \sqrt{\kappa_i} + \left( \sqrt{x_{i,0}} - \sqrt{\kappa_i} \right) e^{-a_i t} \right]^2, \tag{22}$$

with $x_i(t \to \infty) = \kappa_i$, as expected for learning in the absence of stochasticity. Initialising $W_s$ from a normal distribution with variance 1, the initial average value of $x_i$ equals $\langle x_{i,0} \rangle = 1$. Note that the dynamics is linear for the singular values,

$$\dot{\xi}_{s,i} = a_i \left( \sqrt{\kappa_i} - \xi_{s,i} \right), \qquad \xi_{s,i}(t) = \sqrt{\kappa_i} + \left( \xi_{s,i}(0) - \sqrt{\kappa_i} \right) e^{-a_i t}. \tag{23}$$

For reference to the Coulomb gas description, we note that the drift, i.e., the RHS of Eq. (21), can be obtained from a potential,
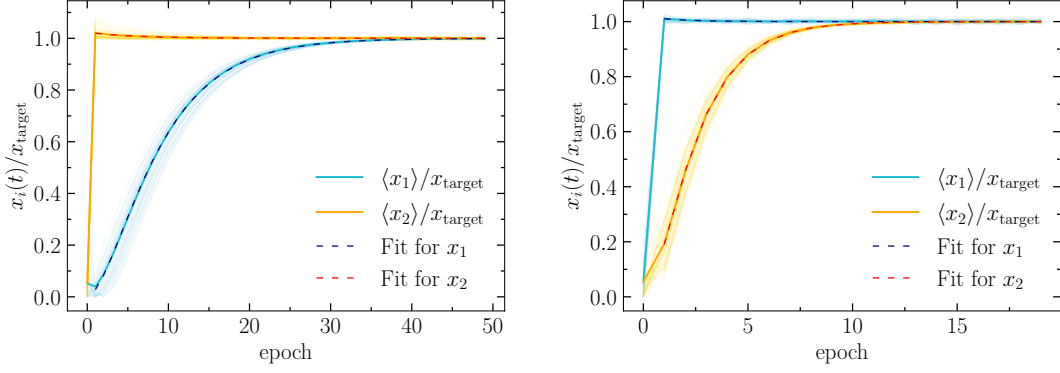
$$V(x_i) = a_i \left( x_i^2 - \frac{4}{3} \sqrt{\kappa_i} x_i^{3/2} \right). \tag{24}$$

Expanding this potential around the minimum at $x_i = \kappa_i$ yields

$$V(x_i) = -\frac{1}{3} a_i \kappa_i^2 + \frac{a_i}{2} (x_i - \kappa_i)^2 + \ldots, \tag{25}$$

i.e., the width of the potential is proportional to the rate of learning $a_i$.

**Figure 4:** Training dynamics of the square of the singular values of a $2 \times 2$ student matrix, given a teacher matrix with doubly degenerate eigenvalues, using $Z$ as in Eq. (26) (left) and Eq. (27) (right). The presence of $Z$ affects the rate of convergence. Shown are an ensemble of 20 networks (with opaque lines), the evolution averaged over an ensemble of 500 networks (with solid blue and orange lines), and fits to Eq. (22), starting from epoch $t = 2$ (with dashed lines), agreeing with the averaged evolution.

We solved the teacher-student model numerically using SGD for the case of two-dimensional input and output, and with $Z$ and $W$ $2 \times 2$ matrices. To study the case of (near-)degeneracy, the eigenvalues of $W_t$ are taken identical. The modes mix due to the presence of $Z$. We study two examples, with

$$Z = \begin{pmatrix} -0.1727 & -0.8341 \\ -0.0106 & -0.7750 \end{pmatrix}, \qquad Z^T Z = \begin{pmatrix} 0.0299 & 0.1522 \\ 0.1522 & 1.2964 \end{pmatrix}, \tag{26}$$

and with

$$Z = \begin{pmatrix} -1.8198 & 0.5861 \\ -0.6574 & 0.0473 \end{pmatrix}, \qquad Z^T Z = \begin{pmatrix} 3.7437 & -1.0977 \\ -1.0977 & 0.3458 \end{pmatrix}. \tag{27}$$

We add a noise term in the learning process, i.e., the update rule for $W$ is modified as

$$W_s' = W_s + \alpha (Z^T Z)(W_t - W_s - \eta), \qquad \eta \sim \mathcal{N}(0, 0.01). \tag{28}$$

This is required, as the algorithm by itself is not noisy enough, unlike in the RBM discussed above. The respective eigenvalue dynamics is shown in Fig. 4. We used here a finite learning rate and batch size, and associate epoch with time. The eigenvalue converging faster is associated with the larger value on the diagonal of $Z^T Z$ and the fit obtained using Eq. (22) aligns with the ensemble average of the eigenvalue dynamics. When $Z = \mathbb{1}$, the rate of convergence is the same for all eigenvalues.

To analyse these results in terms of RMT and the Coulomb potential, we note that the additional layer in the network leads to different rates of learning and curvatures in the potential (24), parametrised by $a_i$. We incorporate this by extending the RMT description to a two-component Coulomb gas, where particles of each species are characterised by a different mass (or variance). We focus on the case of doubly degenerate eigenvalues, with $\kappa_1 = \kappa_2 = \kappa$. By shifting $x_{1,2}$ by $\kappa$ and extending the integration boundaries to $-\infty$, we arrive at the Coulomb gas partition function

$$Z = \frac{1}{N_0} \int dx_1 dx_2 \, |x_1 - x_2| e^{-V(x_1, x_2)}, \qquad V(x_1, x_2) = \frac{x_1^2}{2\sigma_1^2} + \frac{x_2^2}{2\sigma_2^2}, \tag{29}$$
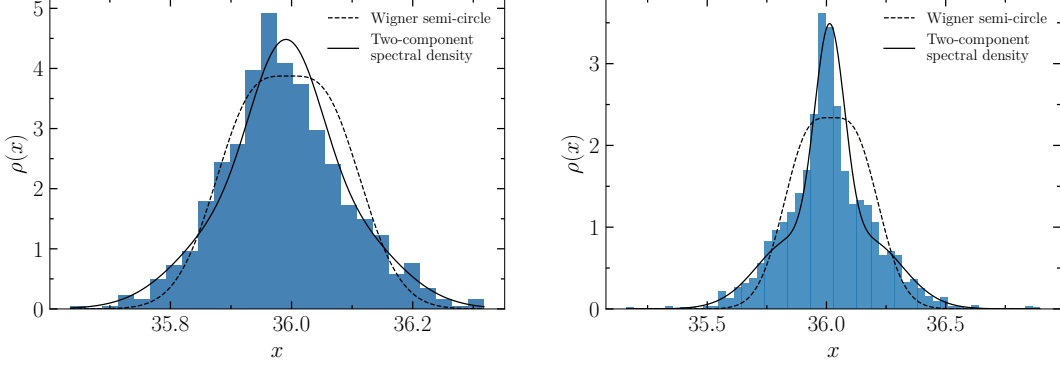
**Figure 5:** Histogram of the spectral density $\rho(x)$ in the presence of a hidden layer, with $Z$ as in Eq. (26) (left) and Eq. (27) (right). Also shown are fits to the standard Wigner semi-circle (11) (dashed line) and the generalised Wigner semi-circle (32) for two species (solid line). The generalised Wigner semi-circle better captures the histogram's peak and wider tails, as seen in particular on the right.

where the two modes have different variances $\sigma_{1,2}^2$. The normalisation constant is $N_0 = 4\sqrt{\pi}\sigma_1\sigma_2\sigma_m$, with

$$\sigma_m^2 = \frac{1}{2}\left(\sigma_1^2 + \sigma_2^2\right). \tag{30}$$

We follow the same procedure as in the RBM. For the Wigner surmise, we write $S = x_1 - x_2$, $x_c = \sigma_2/(2\sigma_1)x_1 + \sigma_1/(2\sigma_2)x_2$, such that $V(x_1, x_2) = S^2/(4\sigma_m^2) + x_c^2/\sigma_m^2$. The Wigner surmise and the average level spacing are then

$$P(S) = \frac{S}{2\sigma_m^2}\exp[-S^2/(4\sigma_m^2)], \qquad \langle S\rangle = \sqrt{\pi}\sigma_m. \tag{31}$$

The spectral density is given by

$$\rho(x; x_c, \sigma_1, \sigma_2) = \frac{e^{-\sigma_m^2\delta x^2/(\sigma_1^2\sigma_2^2)}}{8\sqrt{\pi}\sigma_1\sigma_2\sigma_m}\sum_{i=1,2}\left[2\sigma_i^2 + e^{\delta x^2/(2\sigma_i^2)}\sqrt{2\pi}\delta x\sigma_i\mathrm{Erf}\left(\frac{\delta x}{\sqrt{2}\sigma_i}\right)\right], \tag{32}$$

where $\delta x = x - x_c$ and $\sigma_m$ is related to $\sigma_{1,2}$ using Eq. (30). The presence of the additional layer affects the spectral density, which becomes a generalised version of the Wigner semi-circle, whereas the Wigner surmise is unchanged, albeit with a modified factor for the average level spacing.

We have verified this Coulomb gas description for both the Wigner surmise (not shown here) and the generalised Wigner semi-circle. In Fig. 5, we show the histogram of the eigenvalues and a comparison between the fits performed with the standard Wigner semi-circle (11) and its generalised version (32). It is evident that the standard expression is not able to reproduce the high peak and the wider tails, but that the generalised version is. Finally we note that inserting the values of $\sigma_{1,2}$ into Eq. (30), we obtain a result for $\sigma_m$ compatible with the value obtained directly from a fit of the surmise to Eq. (31).

## 5. Summary

To further develop our understanding of machine learning algorithms, we have formulated stochastic gradient descent in terms of Dyson Brownian motion and the Coulomb gas. In the stationary limit the statistical properties of singular/eigenvalues of weight matrices then follow predictions from random matrix theory. In particular, we have shown that the width of the Coulomb potential around a learnt target value scales proportionally to a specific combination of two hyperparameters of the optimiser, namely the learning rate over batch size, and hence derived the linear scaling rule.

We have verified this behaviour in the Gaussian Restricted Boltzmann Machine, in which the spectral density takes the form of the Wigner semi-circle and the level spacing follows the Wigner surmise, and the predicted scaling of the eigenvalue distribution and level spacing with the learning rate and batch size is observed. Subsequently we have extended the analysis into a more general scenario by considering a linear neural network with one hidden layer in a teacher-student setting. Interestingly, the additional layer modifies the width of the potential for each eigenvalue, resulting in a Coulomb gas with multiple species and a generalised Wigner semi-circle.

For the future we plan to consider larger non-linear neural networks, in which the spectral density is expected to be more intricate, as seen in e.g. Refs. [10, 11, 16].

## References

[1] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby et al., *Machine learning and the physical sciences*, *Rev. Mod. Phys.* **91** (2019) 045002 [1903.10563].

[2] G. Aarts, B. Lucini and C. Park, *Stochastic weight matrix dynamics during learning and dyson brownian motion*, *Phys. Rev. E* **111** (2025) 015303 [2407.16427].

[3] E.P. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, *Annals of Mathematics* **62** (1955) 548.

[4] E.P. Wigner, *Conference on Neutron Physics by Time-of-Flight*, p. 67, 1956.

[5] F.J. Dyson, *Statistical theory of the energy levels of complex systems. I*, *J. Math. Phys.* **3** (1962) 140.

[6] F.J. Dyson, *Statistical theory of the energy levels of complex systems. II*, *J. of Math. Phys.* **3** (1962) 157.

[7] F.J. Dyson, *Statistical Theory of the Energy Levels of Complex Systems. III*, *J. Math. Phys.* **3** (1962) 166.

[8] F.J. Dyson, *A Brownian-Motion Model for the Eigenvalues of a Random Matrix*, *J. Math. Phys.* **3** (1962) 1191.

[9] M.L. Mehta, *Random Matrices*, Academic Press, New York, 3rd ed. (2004).

[10] C.H. Martin and M.W. Mahoney, *Traditional and Heavy-Tailed Self Regularization in Neural Network Models*, 1901.08276.

[11] N.P. Baskerville, D. Granziol and J.P. Keating, *Appearance of random matrix theory in deep learning*, *Physica A* **590** (2022) 126742 [2102.06740].

[12] 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.

[13] S.L. Smith and Q.V. Le, *A bayesian perspective on generalization and stochastic gradient descent*, in *International Conference on Learning Representations*, 2018 [1710.06451].

[14] S.L. Smith, P.-J. Kindermans and Q.V. Le, *Don't decay the learning rate, increase the batch size*, in *International Conference on Learning Representations*, 2018 [1711.00489].

[15] S.L. Smith, D. Duckworth, Q.V. Le and J. Sohl-Dickstein, *Stochastic natural gradient descent draws posterior samples in function space*, 1806.09597.

[16] G. Aarts, O. Hajizadeh, B. Lucini and C. Park, *Dyson Brownian motion and random matrix dynamics of weight matrices during learning*, in *38th conference on Neural Information Processing Systems*, 2024 [2411.13512].

[17] S. Mandt, M.D. Hoffman and D.M. Blei, *Continuous-time limit of stochastic gradient descent revisited*, in *8th NIPS Workshop on Optimization for Machine Learning*, 2015.

[18] Q. Li, C. Tai and W. E, *Stochastic modified equations and adaptive stochastic gradient algorithms*, in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 2101–2110, 2017 [1511.06251].

[19] S. Yaida, *Fluctuation-dissipation relations for stochastic gradient descent*, in *International Conference on Learning Representations*, 2019 [1810.00004].

[20] P. Smolensky, *Chapter 6: Information processing in dynamical systems: Foundations of harmony theory*, in *Parallel Distributed Processing: Volume 1*, D. Rumelhart and J. McLelland, eds., pp. 194–281, MIT Press, 1986.

[21] G.E. Hinton, *Training Products of Experts by Minimizing Contrastive Divergence*, *Neural Comput.* **14** (2002) 1771.

[22] A. Decelle and C. Furtlehner, *Restricted Boltzmann machine: Recent advances and mean-field theory*, *Chinese Phys. B* **30** (2021) 040202 [2011.11307].

[23] G. Aarts, B. Lucini and C. Park, *Scalar field restricted Boltzmann machine as an ultraviolet regulator*, *Phys. Rev. D* **109** (2024) 034521 [`2309.15002`].

[24] Y. Bahri, J. Kadmon, J. Pennington, S.S. Schoenholz, J. Sohl-Dickstein and S. Ganguli, *Statistical Mechanics of Deep Learning*, *Annual Review of Condensed Matter Physics* **11** (2020) 501.

[25] S. Goldt, M.S. Advani, A.M. Saxe, F. Krzakala and L. Zdeborová, *Dynamics of stochastic gradient descent for two-layer neural networks in the teacher–student setup*, *J. Stat. Mech.* **2020** (2020) 124010.

[26] C. Park, G. Aarts and B. Lucini, *chanjure/Stochastic_weight_matrix_dynamics_during_ learning_and_Dyson_Brownian_motion-data_release: v1.0.1*, 2024. 10.5281/zenodo.13310439.

PoS(LATTICE2024)031