PAPER • OPEN ACCESS

Phase diagram and eigenvalue dynamics of stochastic gradient descent in multilayer neural networks

To cite this article: Chanju Park *et al* 2025 *Mach. Learn.: Sci. Technol.* **6** 045048

IOP Publishing Mach. Learn.: Sci. Technol. 6 (2025) 045048 https://doi.org/10.1088/2632-2153/ae1acc

MACHINE LEARNING
Science and Technology

PAPER

Phase diagram and eigenvalue dynamics of stochastic gradient descent in multilayer neural networks

Chanju Park[1,*], **Biagio Lucini**[2] and **Gert Aarts**[1]

[1] Centre for Quantum Fields and Gravity, Department of Physics, Swansea University, Swansea SA2 8PP, United Kingdom
[2] School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: chanju.b.park@gmail.com

Keywords: multilayer perceptron, phase diagram, stochastic gradient descent, dyson brownian motion, hyperparameters

OPEN ACCESS

RECEIVED
5 September 2025

REVISED
20 October 2025

ACCEPTED FOR PUBLICATION
3 November 2025

PUBLISHED
18 November 2025

Abstract

Hyperparameter tuning is one of the essential steps to guarantee the convergence of machine learning models. We argue that intuition about the optimal choice of hyperparameters for stochastic gradient descent can be obtained by studying a neural network's phase diagram, in which each phase is characterised by distinctive dynamics of the singular values of weight matrices. Taking inspiration from disordered systems, we start from the observation that the loss landscape of a multilayer neural network with mean squared error can be interpreted as a disordered system in feature space, where the learnt features are mapped to soft spin degrees of freedom, the initial variance of the weight matrices is interpreted as the strength of the disorder, and temperature is given by the ratio of the learning rate and the batch size. As the model is trained, three phases can be identified, in which the dynamics of weight matrices is qualitatively different. Employing a Langevin equation for stochastic gradient descent, previously derived using Dyson Brownian motion, we demonstrate that the three dynamical regimes can be classified effectively, providing practical guidance for the choice of hyperparameters of the optimiser.

1. Introduction

With the discovery of large machine learning (ML) models and their capability of generalisation [1], significant activity has developed in adopting ML methods in the physical sciences [2], including in quantum physics [3] and (lattice) quantum chromodynamics [4–6]. From the perspective of a physicist, ML systems are intriguing to study in their own right in the context of statistical physics of complex systems [7]. Indeed, a deeper understanding of ML architectures as systems with many fluctuating degrees of freedom, evolving 'out of equilibrium' during the training phase, may shed light on how and why certain systems are highly successful.

A description of learning rooted in statistical physics dates back to (at least) the Hopfield model [8–11], which effectively captures the information storage aspect of the neural network [12], demystifying the scalability and generalisability of large models [13–15]. However, it does not necessarily include practical training settings, which play a crucial role in the performance of a model. In general, it turns out that deep neural networks during training can be thought of as disordered systems in a non-equilibrium setting with non-Gaussian random couplings, for which analytical solutions are not easily accessible. Some recent work has suggested that these systems can be studied by considering an ensemble of models at a fixed training time [16, 17], where the non-Gaussianity can be mitigated by a perturbative approach, or in the limit of infinite width [18–22], where the relevant distributions become Gaussian.

In this paper, we suggest that multilayer neural networks, including the training stage, can be studied by observing the training dynamics of the weight matrices, whose singular values undergo Dyson Brownian motion [23] and follow qualitatively different dynamics depending on the choice of hyperparameters. We start by recalling that the loss function of a multilayer neural network can be interpreted as a disordered Hamiltonian [11, 24], where we further argue that the variance of the weight

matrices can be interpreted as the strength of the disorder, and the stochasticity of the training induces the notion of effective temperature [23, 25–28]. Then, depending on the hyperparameters, three different 'phases' of the training dynamics can be observed, where each phase is characterised by distinctive weight matrix dynamics. Such phase diagrams of ML architectures are expected to provide theoretical insight into the capacity of the model and the choice of hyperparameters [14, 20, 29–34]. Here, we study the trainability of the model depending on the ratio of the learning rate and batch size of stochastic gradient descent, and the initial variance of the weight matrices, which provides guidance to the optimal choice of hyperparameters with regard to the training dynamics.

This paper is structured as follows. In section 2 we define a multilayer neural network and explain how it can be interpreted as a disordered system. We discuss in detail how the degrees of freedom and hyperparameters in the neural network are mapped to 'soft spins' and physical parameters (temperature, disorder) in the disordered system. In section 3 we present the empirical phase diagram of a neural network with two hidden layers and hyperbolic tangent activation functions. In particular, we analyse a range of observables and identify three phases during and after training. The phases depend on the choice of hyperparameters and are identified with an ordered or ferromagnetic phase, in which the network learns well, a disordered or jamming phase, in which the network does not learn, and a paramagnetic phase, in which the dynamics is dominated by fluctuations preventing learning. In section 4, we focus on the singular values of the weight matrices and demonstrate that they follow qualitatively different dynamics in each phase, closely related to the existence of stationary distributions at the end of training. To characterise the phase structure further, we derive expressions for the phase boundaries, using a symmetry-breaking argument as well as a stochastic equation for the average level spacing. In the final section, we show good agreement between these phase boundaries and the regions in the empirical phase diagram obtained by numerical simulations, and discuss the practical implications for successful learning. Some more details on the interpretation of the loss function and on the average level spacing can be found in two appendices.

2. Deep neural networks as disordered systems

2.1. Feed-forward neural network and loss function

To make the connection between neural neural networks and disordered systems, we start by defining the former, following closely the notation of [21]. We consider a neural network with $L+1$ layers, with each layer consisting of n_l nodes ($l=0,\dots,L$). The first layer is the input layer, with the input data given by n_0-dimensional vectors, and the final layer is the output layer, with n_L components. The layers are connected by L weight matrices $W^{(l)}$ ($l=1,\dots,L$) of size $n_l \times n_{l-1}$. We do not include a bias (but this can easily be done). The input data set is indicated as

$$\mathcal{D} = \{x_{i\alpha}\} \qquad i = 1,\dots,n_0, \qquad \alpha = 1,\dots,|\mathcal{D}|, \tag{2.1}$$

where the first index (i) is the component index and the second index (α) labels each sample in the data set. Activation functions, denoted as $\phi(z)$, are applied at the hidden nodes and act component-wise. Here, the z's are the pre-activations, i.e. linear combinations of the components of the previous layer. Explicitly, the first and subsequent pre-activations are then given by

$$z_i^{(l+1)}(x_\alpha) = \sum_{j=1}^{n_l} W_{ij}^{(l+1)} \phi\left(z_j^{(l)}(x_\alpha)\right), \qquad z_i^{(1)}(x_\alpha) = \sum_{j=1}^{n_0} W_{ij}^{(1)} x_{j\alpha}, \tag{2.2}$$

and the result on the final layer defines the neural network function

$$\hat{y}_i(x_\alpha;\theta) \equiv z_i^{(L)}(x_\alpha) = \sum_{j=1}^{n_{L-1}} W_{ij}^{(L)} \phi\left(z_j^{(L-1)}(x_\alpha)\right). \tag{2.3}$$

This function depends on the N_θ learnable parameters, collectively denoted as

$$\theta = \left\{W^{(1)},\dots,W^{(L)}\right\}, \qquad N_\theta = \sum_{l=1}^{L} n_{l-1} n_l. \tag{2.4}$$

The choice of activation function is discussed below. The outputs of the final hidden layer are often called features or representations learnt by the neural network [35], such that the neural network function is a linear combination of these features. Below we indicate the features as

$$\phi_{j\alpha} \equiv \phi\left(z_j^{(L-1)}(x_\alpha)\right). \tag{2.5}$$

In this paper, we consider the mean squared error (MSE) as the loss function, i.e.

$$\mathcal{L}(\theta) \equiv \frac{1}{|\mathcal{D}|}\sum_{\alpha=1}^{|\mathcal{D}|}\ell\left(y(x_\alpha),\hat{y}(x_\alpha;\theta)\right), \qquad \ell(y,\hat{y}) \equiv \frac{1}{2}\sum_{i=1}^{n_L}(y_i-\hat{y}_i)^2, \tag{2.6}$$

where ℓ is the per-sample loss function for input data sample $x_\alpha \in \mathcal{D}$, with target value $y_i(x_\alpha) \equiv y_{i\alpha}$.

To make the connection with disordered systems, we expand the loss function and write

$$\begin{aligned}\mathcal{L}(\theta) &= \frac{1}{2|\mathcal{D}|}\sum_{\alpha=1}^{\mathcal{D}}\sum_{i=1}^{n_L}\left(y_{i\alpha}-\sum_{j=1}^{n_{L-1}}W_{ij}^{(L)}\phi_{j\alpha}\right)^2 \\ &= \frac{1}{2|\mathcal{D}|}\sum_{\alpha=1}^{|\mathcal{D}|}\sum_{i=1}^{n_L}\left(\sum_{j,k=1}^{n_{L-1}}W_{ij}^{(L)}W_{ik}^{(L)}\phi_{j\alpha}\phi_{k\alpha}-2\sum_{j=1}^{n_{L-1}}y_{i\alpha}W_{ij}^{(L)}\phi_{j\alpha}+y_{i\alpha}y_{i\alpha}\right) \\ &= \frac{1}{2|\mathcal{D}|}\sum_{\alpha=1}^{|\mathcal{D}|}\sum_{i,j=1}^{n_{L-1}}J_{ij}\phi_{i\alpha}\phi_{j\alpha}-\frac{1}{|\mathcal{D}|}\sum_{\alpha=1}^{|\mathcal{D}|}\sum_{j=1}^{n_{L-1}}h_{j\alpha}\phi_{j\alpha}+C\end{aligned} \tag{2.7}$$

where to reach the final line, we renamed the indices and introduced

$$J_{ij} \equiv \sum_{k=1}^{n_L}W_{ki}^{(L)}W_{kj}^{(L)}, \qquad h_{j\alpha} \equiv \sum_{i=1}^{n_L}y_{i\alpha}W_{ij}^{(L)}, \qquad C \equiv \frac{1}{2|\mathcal{D}|}\sum_{\alpha=1}^{|\mathcal{D}|}\sum_{i=1}^{n_L}y_{i\alpha}^2. \tag{2.8}$$

Here we intentionally have chosen a notation that resembles the one familiar from disordered systems, with the features ϕ_i playing the role of *spin* degrees of freedom, interacting via a *spin-spin coupling* J_{ij} and with an *external magnetic field* h_j, see also appendix A. The final term is independent of the neural network and will be dropped; hence, from now on, we consider the loss function

$$\mathcal{L}(\theta) = \frac{1}{2|\mathcal{D}|}\sum_{\alpha=1}^{|\mathcal{D}|}\sum_{i,j=1}^{n_{L-1}}J_{ij}\phi_{i\alpha}\phi_{j\alpha}-\frac{1}{|\mathcal{D}|}\sum_{\alpha=1}^{|\mathcal{D}|}\sum_{j=1}^{n_{L-1}}h_{j\alpha}\phi_{j\alpha}. \tag{2.9}$$

Note that this form of MSE is generic for any neural network model whose output is defined as a linear combination of the nodes on the last hidden layer, regardless of the structure of the preceding layers. In the remainder of this section, we further explore the relation to disordered systems. To do so, we remind the reader that the usual spin-glass Hamiltonian is written for binary spins, $s_i = \pm 1$, as [36]

$$\mathcal{H} = -\frac{1}{2}\sum_{i,j}J_{ij}s_is_j+\sum_j h_js_j, \tag{2.10}$$

where J_{ij} is the random coupling and h_j is a random external field, both drawn from a Gaussian distribution. Phase diagrams of spin-glass systems have been studied since the 1970s, and may exhibit ferromagnetic, paramagnetic and spin-glass phases, in the plane spanned by temperature and the strength of the disorder.

To analyse our loss function (2.9) in the language of disordered systems and arrive at a phase diagram, we now address the various ingredients, namely the initalisation of the network, the feature degrees of freedom $\phi_{i\alpha}$, disorder and the coupling J_{ij}, the external field $h_{j\alpha}$, and the emergence of an effective temperature when using stochastic gradient descent.

2.2. Initialisation

The network is initialised by sampling the weight matrix elements from normal distributions, according to

$$W_{ij}^{(l)} \sim \mathcal{N}\left(0,\sigma_W^2/n_{l-1}\right), \tag{2.11}$$

where we will refer to σ_W^2 as the weight matrix variance (the factor n_{l-1} is discussed below). In principle, weight matrices connecting different layers can have their own variance, $\sigma_W^2 \to \sigma_W^{(l)2}$, but for notational simplicity we take them identical. Denoting the pre-activations as

$$z_{i\alpha}^{(l)} \equiv z_i^{(l)}(x_\alpha), \tag{2.12}$$

moments of pre-activations at initialisation over the weight matrix distributions can be computed recursively [21, 37]. The first moments vanish for symmetry reasons. The second moments read, for $l > 1$,

$$\begin{aligned}\mathbb{E}_{p(W)}\left[z_{i\alpha}^{(l)} z_{j\beta}^{(l)}\right] &= \sum_{k,k'=1}^{n_{l-1}} \mathbb{E}_{p(W)}\left[W_{ik}^{(l)} W_{jk'}^{(l)}\right] \phi\left(z_{k\alpha}^{(l-1)}\right) \phi\left(z_{k'\beta}^{(l-1)}\right) \\ &= \delta_{ij} \frac{\sigma_W^2}{n_{l-1}} \sum_{k=1}^{n_{l-1}} \phi\left(z_{k\alpha}^{(l-1)}\right) \phi\left(z_{k\beta}^{(l-1)}\right), \end{aligned} \tag{2.13}$$

while at the first layer, one finds

$$\mathbb{E}_{p(W)}\left[z_{i\alpha}^{(1)} z_{j\beta}^{(1)}\right] = \sum_{k,k'=1}^{n_0} \mathbb{E}_{p(W)}\left[W_{ik}^{(1)} W_{jk'}^{(1)}\right] x_{k\alpha} x_{k'\beta} = \delta_{ij} \frac{\sigma_W^2}{n_0} \sum_{k=1}^{n_0} x_{k\alpha} x_{k\beta}. \tag{2.14}$$

Averaging also over the data set, assuming it is standardised, with

$$\mathbb{E}_{\mathcal{D}}\left[x_{i\alpha} x_{j\beta}\right] = \delta_{ij}\delta_{\alpha\beta}, \tag{2.15}$$

then yields

$$\mathbb{E}_{p(W),\mathcal{D}}\left[z_{i\alpha}^{(1)} z_{j\beta}^{(1)}\right] = \delta_{ij}\delta_{\alpha\beta}\sigma_W^2. \tag{2.16}$$

The important observation for us is that all second moments scale with σ_W^2.

As activation function we use the hyperbolic tangent, $\phi(z) = \tanh(z)$, which is bounded between ± 1. We then trivially have

$$\left|\phi\left(z_{k\alpha}^{(l)}\right) \phi\left(z_{k'\beta}^{(l)}\right)\right| \leqslant 1, \tag{2.17}$$

and hence find

$$\left|\mathbb{E}_{p(W)}\left[z_{i\alpha}^{(l)} z_{j\beta}^{(l)}\right]\right| \leqslant \frac{\sigma_W^2}{n_{l-1}} \sum_{k=1}^{n_{l-1}} \left|\phi\left(z_{k\alpha}^{(l-1)}\right) \phi\left(z_{k\beta}^{(l-1)}\right)\right| \leqslant \sigma_W^2. \tag{2.18}$$

At initialisation, the variances of the pre-activations therefore scale with and are bounded by σ_W^2; this is one reason to normalise the weight matrix variance with n_{l-1}. Note that the scaling with σ_W^2 always holds, while the boundedness is due to the boundedness of the activation function.

2.3. Features as 'soft spins'

Disordered systems are usually formulated in terms of binary spins, $s_i = \pm 1$, see equation (2.10). Here we argue that the features, $\phi_{i\alpha}$, play the role of 'soft spins', which take continuous values, but are still bounded between ± 1, due to the choice of the hyperbolic tangent activation function. The distribution of features at initialisation depends on pre-activations, $z_{j\alpha}^{(L-1)}$. As shown above, the variance of the pre-activation scales with (and is bounded by) σ_W^2. Motivated by results in the large width limit of neural networks [21] and the central limit theorem, we assume that

$$z_{i\alpha}^{(L-1)} \sim \mathcal{N}\left(0, \sigma_z^2\right), \qquad \sigma_z^2 \lesssim \sigma_W^2. \tag{2.19}$$

Dropping the indices for simplicity, we write

$$\phi(z) = \tanh(z), \qquad -1 \leqslant \phi(z) \leqslant 1, \tag{2.20}$$

with the distribution

$$p(\phi) \sim p(z(\phi)) \left|\frac{dz}{d\phi}\right| \sim \exp\left[-\frac{1}{2\sigma_z^2} (\operatorname{arctanh}\phi)^2 - \ln\left(1-\phi^2\right)\right] \sim \exp\left[-V(\phi)\right]. \tag{2.21}$$

Expanding the potential around $\phi = 0$, we find

$$V(\phi) = \left(\frac{1}{2\sigma_z^2} - 1\right)\phi^2 + \left(\frac{1}{3\sigma_z^2} - \frac{1}{2}\right)\phi^4 + \ldots, \tag{2.22}$$

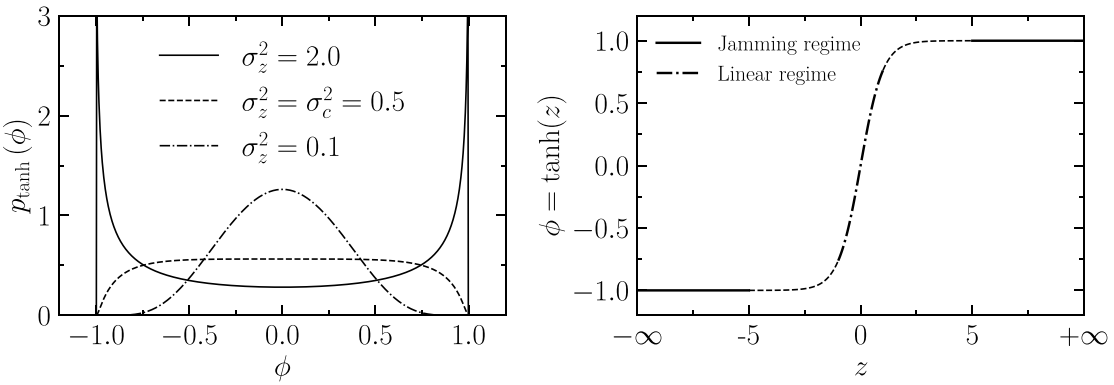

Figure 1. Left: Distribution of post-activation $\phi(z)$, with $z \sim \mathcal{N}(0, \sigma_z^2)$. For small $\sigma_z^2 < 1/2$, the distribution is peaked around zero, while for large $\sigma_z^2 \gg 1/2$, the distribution is sharply peaked towards ± 1. Right: Hyperbolic tangent function $\phi(z) = \tanh(z)$ as activation function, with a linear regime ($|z| \lesssim 1$) and a jamming or vanishing gradient regime ($|z| \gg 1$).

which exhibits a transition from a single well to a double well at $\sigma_z^2 = \sigma_c^2 = 1/2$. This observation also holds without the expansion, as the extrema of the potential are determined by $\phi = \tanh(2\sigma_z^2\phi)$. We conclude that the distribution of features has a single peak centred around 0 when $\sigma_z^2 \leqslant 1/2$ and a double peak when $\sigma_z^2 > 1/2$, as demonstrated in figure 1(left).

The activation function is shown in figure 1(right). We can now draw a first conclusion on how the initial variance of the weight matrices, a hyperparameter, affects the learning efficiency for bounded activation functions. For large $\sigma_z^2 \sim \sigma_W^2$, features ϕ are close to ± 1. The activation function is then mostly in the vanishing gradient regime, and poor learning is expected. On the other hand, for small $\sigma_z^2 \sim \sigma_W^2$, the features lie around 0, which corresponds to the linear regime in the activation function, where good learning is expected. This vanishing gradient transition can be regarded as a jamming transition in disordered systems, where a distribution of pre-activations z, with variance σ_z^2, is placed in a finite box (the activation function) with a vanishing gradient on the boundary, and the dynamics of the distribution starts to get jammed once the size of the distribution exceeds a critical value σ_c^2. Below, we will numerically confirm that indeed a large initial variance of the weight matrices results in poor learning, corresponding to a jammed phase.

2.4. Disorder

Next, we turn to the coupling between the features, J_{ij}, the product of the weight matrices at the final layer,

$$J_{ij} = \sum_{k=1}^{n_L} W_{ki}^{(L)} W_{kj}^{(L)}, \tag{2.23}$$

see equation (2.8). In a standard disordered system with N spins [36], the spin couplings are drawn from a normal distribution, $J_{ij} \sim \mathcal{N}(J_0, J^2)$, and a transition from a ferromagnetic phase to a spin-glass phase is observed as $\sqrt{N}J_0/J$ is reduced, where the scaling with N ensures a proper thermodynamic limit.

In the case considered here, at initialisation the matrix elements of $W^{(L)}$ are drawn from a normal distribution with variance σ_W^2/n_{L-1}. The $n_{L-1} \times n_{L-1}$ coupling matrices J_{ij} are hence symmetric and positive semi-definite, and belong to a Wishart-Laguerre ensemble with Dyson index $\beta = 1$ [38]. In the limit of large matrix size, its spectral density tends to the Marchenko-Pastur distribution defined on a finite support.

At initialisation, the mean and covariance of the J's are given by

$$\mathbb{E}_{p(W)}\left[J_{ij}\right] = \frac{n_L}{n_{L-1}}\sigma_W^2\delta_{ij}, \qquad \mathrm{Cov}\left[J_{ij}J_{kl}\right] = \frac{n_L}{n_{L-1}^2}\sigma_W^4\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right). \tag{2.24}$$

To determine what variable to vary to observe a possible transition between an ordered and a disordered phase, we may follow the choice in the disordered spin system and use mean divided by $\sqrt{\text{variance}}$, J_0/J. However, for the Wishart ensemble, this scales as σ_W^0 and hence it is not a suitable combination. Instead, we will use the scaling with $1/\sqrt{\text{variance}}$ of the initial weight matrices, i.e. $1/\sigma_W$. As we will demonstrate below, this choice indeed provides a useful control parameter.

2.5. External field alignment

The second term in the loss function (2.9) is the coupling between the features $\phi_{j\alpha}$ and the *external magnetic field* $h_{j\alpha}$,

$$\mathcal{L}_h(\theta) = -\frac{1}{|\mathcal{D}|}\sum_{\alpha=1}^{|\mathcal{D}|}\sum_{j=1}^{n_{L-1}} h_{j\alpha}\phi_{j\alpha}, \qquad h_{j\alpha} = \sum_{i=1}^{n_L} y_{i\alpha} W_{ij}^{(L)}, \tag{2.25}$$

where $y_{i\alpha}$ is the target value and both $h_{j\alpha}$ and $\phi_{j\alpha}$ depend on the input data.

Given that the prediction of the network is

$$\hat{y}_{i\alpha} = \sum_{j=1}^{n_{L-1}} W_{ij}^{(L)}\phi_{j\alpha}, \tag{2.26}$$

this term can also be written as

$$\mathcal{L}_h(\theta) = -\frac{1}{|\mathcal{D}|}\sum_{\alpha=1}^{|\mathcal{D}|}\sum_{i=1}^{n_L} y_{i\alpha}\hat{y}_{i\alpha} \tag{2.27}$$

which is minimised when the output vectors are aligned, as expected. This explains the relation between the alignment of the output data on one hand and the features and external field on the other hand.

2.6. Temperature

An important difference between disordered systems and neural networks is that the former are usually considered at a temperature T, whereas the latter evolve during training, from initialisation to potentially being well-trained, and should therefore be considered as non-equilibrium systems. Indeed, weight matrix elements are typically drawn from normal distributions at initialisation and training will induce correlations, leading to characteristic heavy-tailed distributions of weight matrix spectra [39, 40], known from random matrix theory [41, 42].

Nevertheless, it is possible to identify a notion of temperature, characterising fluctuations observed during training. This can be made precise in the case of stochastic gradient descent (SGD), one of the most commonly used training algorithms. Each weight matrix element is updated according to

$$W_{ij}^{(l)\prime} = W_{ij}^{(l)} - \epsilon\Delta_{ij,\mathcal{B}}^{(l)}, \qquad \Delta_{ij,\mathcal{B}}^{(l)} = \frac{1}{|\mathcal{B}|}\sum_{\alpha\in\mathcal{B}}\Delta_{ij,\alpha}^{(l)} \qquad \Delta_{ij,\alpha}^{(l)} = \frac{\partial\ell(x_\alpha)}{\partial W_{ij}^{(l)}}. \tag{2.28}$$

Here $\epsilon > 0$ is the learning rate (or step size), which we take fixed in this analysis, $\Delta_{ij,\alpha}^{(l)}$ is the element-wise gradient of the loss function for data point x_α, and updates are performed using batches of size $|\mathcal{B}|$.

Stochasticity is induced because of the finite batch size. The mean and fluctuations of the gradient can be separated using the central limit theorem, assuming the input data is i.i.d. The evolution of the model parameters can then be described by a discrete Langevin equation,

$$W_{ij}^{(l)\prime} = W_{ij}^{(l)} - \epsilon\mathbb{E}\left[\Delta_{ij}^{(l)}\right] + \sqrt{\frac{\epsilon^2}{|\mathcal{B}|}\mathbb{V}\left[\Delta_{ij}^{(l)}\right]}\,\eta_{ij}, \qquad \eta_{ij}\sim\mathcal{N}(0,1). \tag{2.29}$$

Here $\mathbb{E}[\Delta]$ and $\mathbb{V}[\Delta]$ are the mean and variance of the gradient, respectively, and the scaling with learning rate and batch size has been derived in [23] in the context of Dyson Brownian motion. The scaling with the step size does not follow the standard (Itô) scaling in (discretised) stochastic differential equations (SDEs). In [23] it was demonstrated that the particular scaling in equation (2.29) leads to the linear scaling rule, i.e. a dependence on the ratio $\epsilon/|\mathcal{B}|$ only[3]. Moreover, this ratio combines with the Coulomb potential in the random matrix theory description of Dyson Brownian motion in the same way as a temperature would [23].

An alternative way to arrive at $\epsilon/|\mathcal{B}|$ as an effective temperature is by taking the limit of zero step size in a weak sense. A naive $\epsilon\to 0$ limit would lead to deterministic gradient flow without stochasticity

[3] The dependence on the ratio $\epsilon/|\mathcal{B}|$, and not on the hyperparameters ϵ and $|\mathcal{B}|$ separately, has been observed empirically in SGD optimisation and is known as the linear scaling rule [43, 44]. It implies that decreasing the learning rate by a factor k has the same effect as increasing the batch size by the same factor. This is explained by noting that the effective temperature $T = \epsilon/|\mathcal{B}|$ is invariant under a simultaneous scaling of both.

[25, 26, 45]. Instead, the correct SDE limit is obtained if the ratio $\epsilon/|\mathcal{B}|$ is kept fixed when the learning rate ϵ and the batch size $|\mathcal{B}|$ are taken to zero simultaneously. This yields a Langevin equation in continuous time,

$$\dot{W}_{ij}^{(l)} = -\mathbb{E}\left[\Delta_{ij}^{(l)}\right] + \sqrt{T\mathbb{V}\left[\Delta_{ij}^{(l)}\right]}\,\eta_{ij}, \qquad T \equiv \frac{\epsilon}{|\mathcal{B}|}, \tag{2.30}$$

where the dot indicates the time derivative and the effective temperature T is a measure of the stochasticity, appearing as a temperature in the stationary solution of the corresponding Fokker-Planck equation (assuming that it exists) [23]. Below, we identify $\epsilon/|\mathcal{B}|$ with the temperature axis in the phase diagram, to which we turn now.

3. Empirical phase diagram

In the preceding section, we have argued that a neural network with hyperbolic tangent activation functions and subject to the MSE loss function should be considered as a disordered system, with a temperature set by the ratio of learning rate and batch size, $T = \epsilon/|\mathcal{B}|$, disorder by the variance σ_W^2 of the weight matrices upon initialisation, and features ϕ_i as soft spins. In contrast to a disordered spin system at a given temperature, the learnable parameters in a neural network, i.e. the weight matrix elements, evolve during training, and our aim is to determine how the efficiency and accuracy of training dynamics change depending on the temperature and initial variance. A practical application of this study is that it provides guidance on how to choose the hyperparameters.

As argued above, we consider the phase diagram in the plane spanned by the temperature and the inverse square root of the initial variance, the $T - 1/\sigma_W$ plane. Intuitively, three characteristic phases resembling a disordered system [36, 46] can be expected:

- a high-temperature or paramagnetic phase, in which the model is subject to large fluctuations and the features of the model are randomly distributed;
- a disordered spin-glass or jamming phase at large σ_W, in which the model parameters accumulate near gradient vanishing points and the training dynamics is jammed from initialisation;
- a low-temperature or ferromagnetic phase with little disorder, in which the model is capable to learn the target features and reach an equilibrium state at the end of training.

Traditionally, in statistical mechanics, a phase diagram is studied by identifying symmetries and order parameters, and subsequently calculating the system's free energy. Here, the neural network is evolving during training, which calls for a non-equilibrium treatment, and the identification of order parameters is not obvious. Hence, we probe the phase diagram empirically, using numerical experiments, and consider observables which are both inspired by disordered systems but also relevant in the context of learning, such as the loss, the gradient of the loss, and alignment at the end of training.

3.1. Numerical setup

We consider a neural network with two hidden layers ($L = 3$): the input layer has $n_0 = 3$ input neurons, the hidden layers have $n_1 = 32$ and $n_2 = 16$ neurons respectively, and the output is a scalar, $n_{L=3} = 1$. The task is regression and, as stated, the activation functions are hyperbolic tangents. Specifically, we use a teacher-student setting, where the training data for the network are generated from a teacher neural network with weight matrices $\overline{W}^{(l)}$ ($l = 1, 2, 3$). The matrix elements in the teacher network, as well as the input data, are sampled from a normal distribution,

$$\overline{W}_{ij}^{(l)} \sim \mathcal{N}(0, 1/n_{l-1}), \qquad x_{i\alpha} \sim \mathcal{N}(0, 1). \tag{3.1}$$

Note that the (scaled) variance in the teacher weight matrices is equal to 1. This setup determines the output of the teacher network.

The student network is defined with the same architecture as the teacher network, but with the intermediate weight matrix $W^{(2)}$ set to be trainable and initialised according to equation (2.11). The first and final student weight matrices are identical to the corresponding teacher ones, $W^{(1)} = \overline{W}^{(1)}$, $W^{(3)} = \overline{W}^{(3)}$. This setting can be regarded as a variation of a Random Feature Neural Network [47], where instead of given random features, the linear combinations are given and the features representing the correct target feature space are learnt. From a theoretical perspective, this setting enables us to isolate the dynamics of a single layer and carry out a precise numerical study of the phase diagram and weight matrix dynamics.

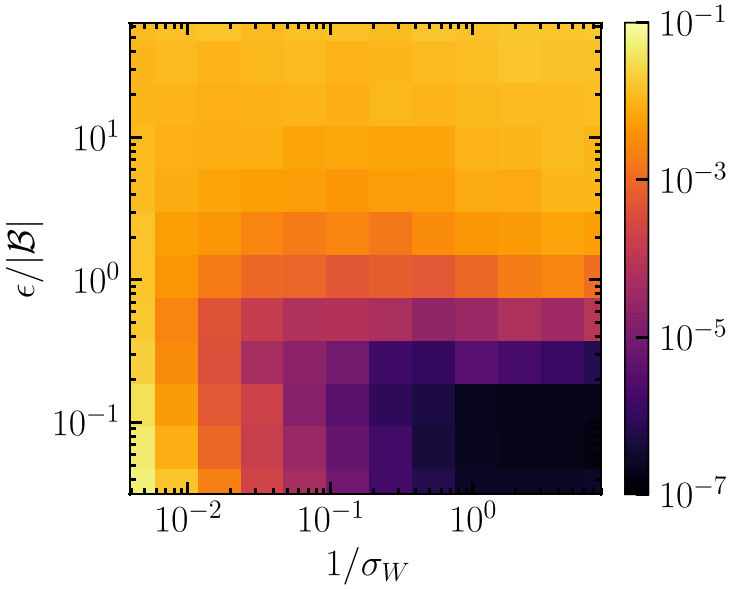

Figure 2. The mean test loss of the trained models in the $T - 1/\sigma_W$ plane for the hyperbolic tangent teacher-student network. A darker colour indicates a lower value.

Moreover, using the teacher-student setting is useful as one has direct control over the target distribution for the student network, allowing one to check the theoretical predictions with experimental outcomes.

In the numerical experiments discussed below, for each combination of hyperparameters, an ensemble $\mathcal{S}$ with a total of $|\mathcal{S}| = 100$ teacher networks is generated, and each student network is trained on a different teacher network. The hyperparameters are selected as follows: the learning rate is varied from $\epsilon = 2^{-3}, \ldots, 2^{8}$ in powers of 2, with a batch size of $|\mathcal{B}| = 4$. Hence the temperature range is $T = \epsilon/|\mathcal{B}| = 2^{-5}, \ldots, 2^{6}$. The inverse square root of the variance of the student weight matrix $W^{(2)}$ is varied as $1/\sigma_W = 2^{-8}, \ldots, 2^{3}$. Each model is trained with the same number of iterations $t = 10^5$. A typical training set consists of $|\mathcal{D}| = 4000$ data points.

3.2. Observables

We now come to the results of the numerical experiments and introduce four observables that capture different aspects of the dynamics of learning across the phase diagram. The training time is denoted with t and the end of the training with $t = t_f$. All observables are evaluated on a test set with $|\mathcal{D}_{\text{test}}| = 100$ data points after training.

First, we measure the mean test loss at the end of the training, defined as an average of the loss (2.6) over the ensemble $\mathcal{S}$ of trained networks at a fixed choice of hyperparameters,

$$\mathbb{E}_{\mathcal{S}}\left[\mathcal{L}\right] = \frac{1}{|\mathcal{S}|}\sum_{s=1}^{|\mathcal{S}|}\mathcal{L}^{s}, \tag{3.2}$$

where $\mathcal{L}^s$ is the final test loss of the s-th model in the ensemble $\mathcal{S}$. A high mean loss indicates that the model has not converged to the correct solution and the training is unsuccessful, while a low mean loss indicates that the model is well trained.

The results for the final mean loss are shown in figure 2, where the mean loss is seen to vary between approximately 10^{-1} and 10^{-7} in the plane spanned by $T = \epsilon/|\mathcal{B}|$ and $1/\sigma_W$. The darker region corresponds to a smaller final loss. At higher temperatures and at larger initial variance, the loss remains large, indicating that the ensembles of models are not close to the target ones. On the other hand, at low temperature and small initial variance, i.e. in the bottom-right corner of this phase diagram, the average loss is substantially smaller, and we may deduce that the ensembles of models have converged with high probability. Below, we identify the bottom-right corner in figure 2 with a ferromagnetic phase.

As a second observable, we consider the mean gradient to quantify whether the training is active or dormant. We define the mean gradient as

$$\mathbb{E}_{\mathcal{S}}\left[\|\nabla\mathcal{L}\|\right] = \frac{1}{|\mathcal{S}|}\sum_{s=1}^{|\mathcal{S}|}\sqrt{\sum_{i=1}^{n_2}\sum_{j=1}^{n_1}\left(\Delta_{ij}^{(2),s}\right)^2}, \tag{3.3}$$

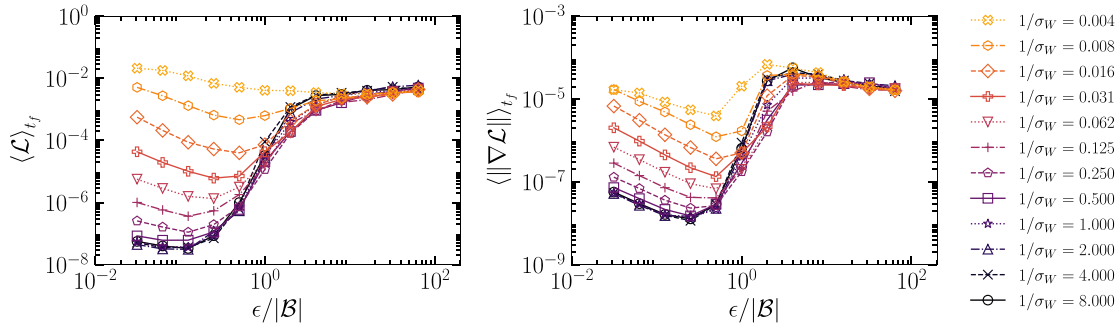

Figure 3. Mean loss (left) and gradient (right) at the end of training as a function of $\epsilon/|\mathcal{B}|$ for various values of $1/\sigma_W$. Error bars are omitted for visibility.

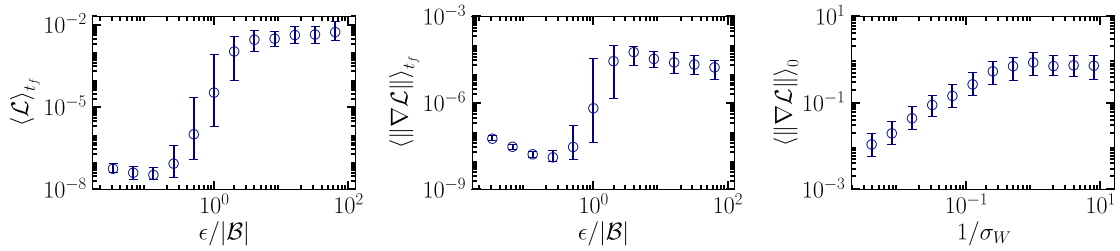

Figure 4. The mean loss (left) and mean gradient (centre) at the end of training along the temperature axis at the smallest value of σ_W considered, $1/\sigma_W = 8$. A transition is visible at $\epsilon/|\mathcal{B}| \sim 1$. Mean gradient at initialisation (right) along the $1/\sigma_W$ axis at the lowest temperature, $\epsilon/|\mathcal{B}| = 2^{-5}$. For small $1/\sigma_W$ (large initial disorder), learning is already inefficient at initialisation.

where $\Delta_{ij}^{(2),s}$ is the gradient of the loss function for the s-th model with respect to $W_{ij}^{(2)}$, averaged over the data set.

In figure 3, we show the mean loss and gradient at the end of training as a function of $\epsilon/|\mathcal{B}|$ for various values of $1/\sigma_W$, i.e. each collection of data points connected by a line corresponds to a vertical slice of the preceding phase diagram. Error bars are omitted for visibility, but are shown in figure 4 for $1/\sigma_W = 8$. We observe different behaviour, which we identify with the phases discussed above:

- If the temperature and initial variance are small, $\epsilon/|\mathcal{B}| \sim 10^{-1}$, $1/\sigma_W \gtrsim 1$, both the final loss and gradient are small, indicating that the models have learnt successfully and that learning is completed. This corresponds to the ordered or ferromagnetic phase.
- Increasing the initial variance at small temperature results in a larger loss, while the gradient remains large as well. This indicates that learning remains active but without making progress. This is identified with a jamming phase with slow dynamics.
- At higher temperature, $\epsilon/|\mathcal{B}| \sim 10^0$, a transition is observed to a regime where both the final loss and final gradient are large, indicating a lack of convergence. Since this dynamics is independent of the matrix initialisation, it signifies a transition to a paramagnetic phase[4].
- At even higher temperature, $\epsilon/|\mathcal{B}| \sim 10^2$, and for all values of $1/\sigma_W$, the dependence on the initialisation has vanished and the models are clearly in the paramagnetic phase, with the dynamics dominated by thermal fluctuations.

In figure 4, we add error bars for the loss (left) and the gradient (middle) as a function of $\epsilon/|\mathcal{B}|$ for the smallest value of σ_W considered, $1/\sigma_W = 8$. We note a rapid increase at $\epsilon/|\mathcal{B}| \sim 1$, in a manner that is not dissimilar to a phase transition, albeit in a finite volume or with a finite number of degrees of freedom. In figure 4(right), we show the gradient of the loss at initialisation, along the $1/\sigma_W$ axis at the lowest temperature considered. We observe that for large initial variance, $1/\sigma_W \sim 10^{-2}$, the mean gradient is already small at initialisation, resulting in slow spin-glass-like dynamics, while with small initial variance, $1/\sigma_W \gtrsim 10^0$, the dynamics is much more efficient. We argue, therefore, that for practical implementations, the optimal values of the learning rate over batch size are located right before the transition

[4] This corresponds to the dotted regions in the phase diagrams analysed in [31], which are considered in the plane spanned by the learning rate and the batch size.

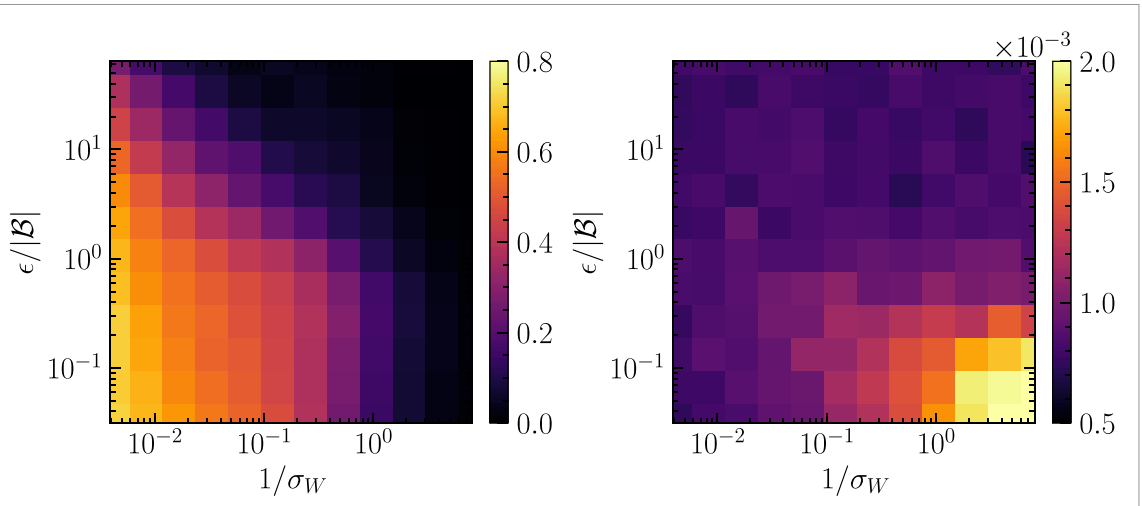

Figure 5. The correlation between features at initial and final time, $G(t_f, 0) = \mathbb{E}[\phi(t_f)\phi(0)]$, (left) and between the features ϕ and the external magnetic field h, $\mathbb{E}[h\phi]$, or alignment between model and target features, (right), in the $T - 1/\sigma_W$ plane for the hyperbolic tangent teacher-student network. A darker colour indicates less correlations.

from the low-temperature to the high-temperature phase, with small initial variance for the weight matrix elements, allowing for fast training while yielding good convergence.

The next observable we consider is the time correlation function of features, defined by

$$G(t,t') \equiv \mathbb{E}_{\mathcal{S}}\left[\phi(t)\,\phi(t')\right] = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\frac{1}{|\mathcal{D}|}\sum_{\alpha\in\mathcal{D}}\frac{1}{n_{L-1}}\sum_{j=1}^{n_{L-1}}\phi^s_{j\alpha}(t)\,\phi^s_{j\alpha}(t')\,, \tag{3.4}$$

where $\phi^s_{j\alpha}(t)$ is a component of a feature at time t in the s-th model. We consider in particular the correlation between the initial features, at $t = 0$, and the features at the end of training, at $t = t_f$, i.e. $G(t_f, 0)$. The results for $G(t_f, 0)$ are shown in figure 5(left). A darker colour corresponds to a lower value and hence less correlations, whereas a lighter colour implies that correlations are preserved. The observed correlations in figure 5 are in agreement with the discussion presented so far. In the high-temperature (paramagnetic) phase and in the well-trained (ferromagnetic) phase, the correlations between features at initialisation and after training are lost, as expected. With a large initial variance, on the other hand, correlations are preserved, which is interpreted as poor learning in the disordered spin-glass phase. Increasing the temperature reduces this correlation, but leads to poor training due to the transition to the paramagnetic phase.

The final observable we consider is the alignment between target and model features, via the correlation between the features $\phi_{j\alpha}$ and the external magnetic field $h_{j\alpha}$, see section 2.5. We define

$$\mathbb{E}_{\mathcal{S}}\left[h\phi\right] = \frac{1}{|\mathcal{S}|}\sum_{s=1}^{|\mathcal{S}|}\frac{h^s\cdot\phi^s}{\|h^s\|\,\|\phi^s\|} \tag{3.5}$$

where

$$h^s\cdot\phi^s = \sum_{\alpha=1}^{|\mathcal{D}|}\sum_{j=1}^{n_{L-1}} h^s_{j\alpha}\phi^s_{j\alpha}, \qquad \|h^s\|^2 = \sum_{\alpha=1}^{|\mathcal{D}|}\sum_{j=1}^{n_{L-1}} h^s_{j\alpha}h^s_{j\alpha}, \tag{3.6}$$

and similar for $\|\phi^s\|^2$. The results are shown in figure 5(right). We observe the strongest alignment in the well-trained region of the phase diagram, giving a clear justification to associate the region with the ferromagnetic phase.

In summary, we have considered several observables in the plane spanned by the hyperparameters. Each observable probes different characteristics of learning, allowing us to identify the various phases. The behaviour of the mean loss gives the first indication of a well-trained phase. The alignment with the external magnetic field emphasises the similarity with a ferromagnetic phase. The difference between the paramagnetic phase and the disordered spin-glass phase is captured by the correlations of features in time, which are preserved in the latter. These findings are consistent with the understanding of different phases in disordered systems, using the mapping between the hyperparameters and the parameters in generic spin-glass models.

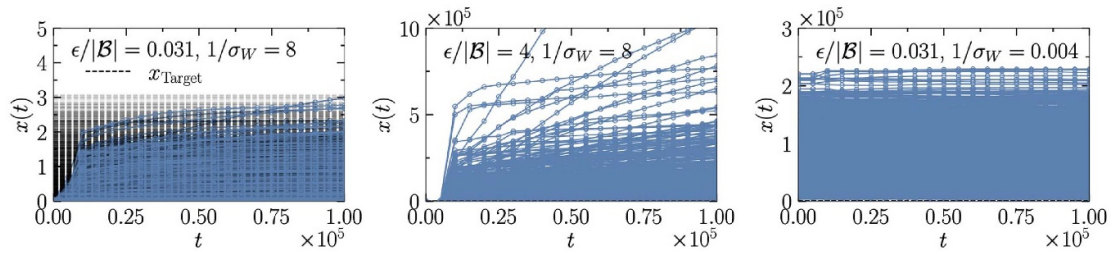

Figure 6. Evolution of the eigenvalues of $X = W^{(2)} W^{(2)\,T}$ during training for ensembles of teacher-student models at three choices of hyperparameters, in the ordered phase (left) with $\epsilon/|\mathcal{B}| = 2^{-5}$, $1/\sigma_W = 8$, in the high-temperature phase (centre) with $\epsilon/|\mathcal{B}| = 4$, $1/\sigma_W = 8$, and in the disordered phase (right) with $\epsilon/|\mathcal{B}| = 2^{-5}$, $1/\sigma_W = 2^{-8}$. Dashed horizontal lines denote the target eigenvalues and are only visible on the left. Note the difference in scale on the vertical axis.

4. Dynamics of training

In this section, we take a complementary approach, i.e. rather than focussing on the features, we study the dynamics of the weight matrices and demonstrate that the various phases can also be observed in the evolution of these. We consider the $n_2 \times n_1$ weight matrix $W^{(2)}$, where in our case $n_2 = 16 < n_1 = 32$. Following [23], we consider the symmetric combination

$$X = W^{(2)} W^{(2)\,T}, \tag{4.1}$$

and denote its eigenvalues with x_i $(i = 1, \ldots, n_2)$. These eigenvalues are semipositive and are the squares of the singular values of $W^{(2)}$.

The evolution of the eigenvalues $x_i(t)$ during training is presented in figure 6 in the ordered phase (left), the high-temperature phase (centre), and the disordered phase (right). Dashed horizontal lines denote the target eigenvalues and are only visible on the left. Note the difference in scale on the vertical axis. What is shown in each figure is the evolution for an ensemble of 100 teacher-student models, hence a total of 1600 eigenvalues.

At initialisation, with $W^{(2)}_{ij} \sim \mathcal{N}(0, \sigma_W^2/n_1)$, the distribution of eigenvalues of X is given by the Marchenko-Pastur distribution,

$$P_{\text{MP}}(x) = \frac{1}{2\pi\sigma_W^2 r x}\sqrt{(x_+ - x)(x - x_-)}, \qquad x_- < x < x_+, \tag{4.2}$$

where $r = n_2/n_1 = 1/2$, $x_\pm = \sigma_W^2(1 \pm \sqrt{r})^2$. The upper limit is hence given by $x_+ = 0.046$ for $1/\sigma_W = 8$ and $x_+ = 1.9 \times 10^5$ for $1/\sigma_W = 2^{-8}$. This explains the difference in the initial range between the figure on the right and the two other examples.

We observe dynamics characteristic of all three phases. In the ordered phase (left), the distribution of eigenvalues flows towards the target distribution. In the high-temperature phase (centre), stochastic fluctuations are so strong that the distribution quickly widens and the eigenvalues grow until they potentially reach a gradient vanishing point. In the jamming phase (right), the initial distribution is already very broad, and eigenvalues evolve slowly without ever converging.

4.1. Stochastic dynamics of the average level spacing

The dynamics of the eigenvalues can in principle be modelled by generalised Dyson Brownian motion [23], derived from the stochastic equation for the weight matrix, with discrete updates (2.29) or in continuous time (2.30). Dyson Brownian motion for the eigenvalues of X in continuous time, with $T = \epsilon/|\mathcal{B}|$, is given by [23, 48, 49],

$$\dot{x}_i = K_i + T\sum_{j\neq i}\frac{V_{ij}}{x_i - x_j} + \sqrt{T V_{ii}}\,\eta_i, \qquad \eta_i \sim \mathcal{N}(0,1). \tag{4.3}$$

Here $K_i = K_{ii}$ and V_{ii} are the diagonal elements of the mean and variance of the gradient after diagonalisation of the former, and $V_{i\neq j}$ are the off-diagonal components, with

$$K_{ij} = -\mathbb{E}\left[\Delta^X_{ij}\right], \qquad V_{ij} = \mathbb{V}\left[\Delta^X_{ij}\right], \qquad \Delta^X_{ij} = \sum_{k=1}^{n_1}\left(W^{(2)}_{ik}\Delta^{(2)}_{kj} + \Delta^{(2)}_{ik}W^{(2)}_{kj}\right), \tag{4.4}$$

where the final equation follows from the update for $W^{(2)}$, see equation (2.28). The second term in equation (4.3) is the Coulomb term, leading to eigenvalue repulsion.

Solving the coupled set (4.3) of SDEs would give the time evolution of the eigenvalues, but it is not straightforward to do so, as the drift, Coulomb term, and diffusion coefficients are complicated non-linear and time-dependent functions. To simplify the problem, we introduce a 'one-particle theory', see also [50], by focussing on the average level spacing. Let x_i be the i-th eigenvalue of X and assume that the eigenvalues are ordered, $0 \leqslant x_1 < x_2 < \cdots < x_{n_2}$. We consider the level spacing $S_i = x_{i+1} - x_i$, and its average value,

$$S = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2 - 1} S_i = \frac{1}{n_2 - 1} \left(x_{n_2} - x_1 \right). \tag{4.5}$$

As a first Ansatz, we assume that the average level spacing effectively evolves with a local rate $\lambda(t)$, i.e.

$$\dot{S}(t) = \lambda(t) S(t) \qquad \Rightarrow \qquad S(t) = S(0)\, e^{\int_0^t ds\, \lambda(s)}. \tag{4.6}$$

We can then define a time-averaged rate,

$$\overline{\lambda}(t) = \frac{1}{t} \int_0^t \mathrm{d}s\, \lambda(s) = \frac{1}{t} \log \frac{S(t)}{S(0)}. \tag{4.7}$$

The computed values at the end of learning, $\overline{\lambda}(t_f)$, are shown in figure 7. We observe that the averaged rate is smallest in the jammed phase and largest in the paramagnetic phase, consistent with the time correlation of the features shown in figure 5(left)[5].

To improve on this, we construct a stochastic equation for S, by subtracting the equation for the smallest eigenvalue x_1 from the one for x_{n_2}, see equation (4.5). A detailed derivation is given in appendix B. This results in

$$\dot{S} = K_S + V_S \frac{T}{S} + \sqrt{2TD_S}\,\eta, \qquad \eta \sim \mathcal{N}(0,1). \tag{4.8}$$

Expressions for K_S, V_S, and D_S in terms of K_{ij} and V_{ij} are given in equations (B.7) and (B.9). In the following, we use equation (4.8) to derive a relation for the phase boundary between the high-temperature phase and the other phases.

4.2. Phase boundary from a stability analysis

In the high-temperature phase, the level spacing keeps increasing, see figure 6(centre), unlike in the two other phases in which it approximately stabilises after the initial stage, albeit for different reasons. We now derive a condition of convergence using a stability analysis of the combined force term in equation (4.8). A restoring force is required to balance the effects of fluctuations due to the noise, c.f. the fluctuation-dissipation theorem.

We consider the deterministic part of equation (4.8),

$$\dot{S} = K_S + V_S \frac{T}{S}, \tag{4.9}$$

where in general, S will converge if the RHS is negative and diverge if it is positive. Note that $K_S < 0$ close to a (local) minimum, when the loss function is minimised. We assume $V_S > 0$. Hence, the convergence condition is given by the inequality,

$$\frac{T}{S} \leqslant -\frac{K_S}{V_S} \quad \Leftrightarrow \quad \text{converge}, \qquad \frac{T}{S} > -\frac{K_S}{V_S} \quad \Leftrightarrow \quad \text{diverge}. \tag{4.10}$$

The quantity on the RHS represents the signal-to-noise ratio of the gradient of the loss function. The LHS can be related to the initial hyperparameters, as follows. If the average level spacing keeps increasing, we have

$$S(t_f) > S(t) > S(0) \qquad \text{or} \qquad \frac{1}{S(0)} > \frac{1}{S(t)} > \frac{1}{S(t_f)}. \tag{4.11}$$

[5] In the paramagnetic phase, at the highest temperature and smallest variance and at $t_f \sim 10^5$, see figure 6(centre), we find $(n_2 - 1)S(t_f) \sim 10^6$, $(n_2 - 1)S(0) = x_+ - x_- \sim 0.04$, and hence $\overline{\lambda}(t_f) \sim 10^{-5} \log(10^6/0.04) \sim 2 \times 10^{-4}$, in agreement with figure 7.

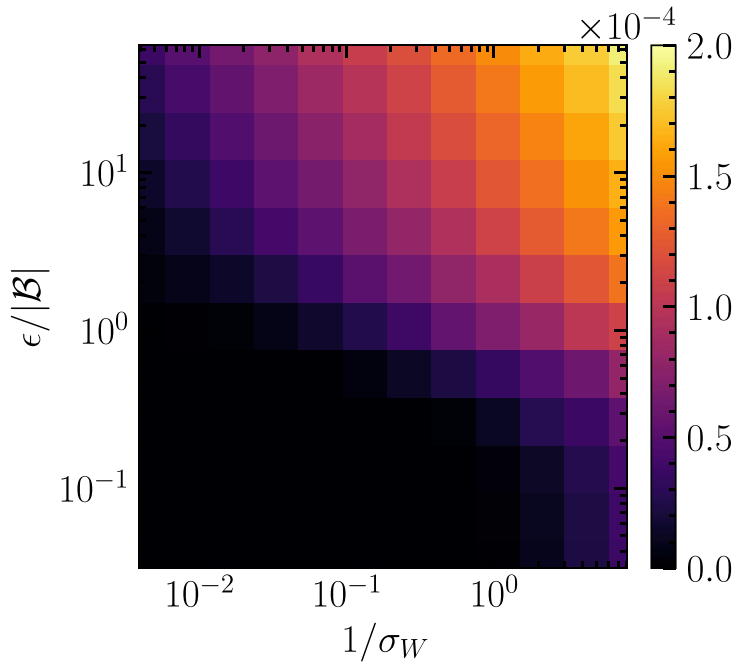

Figure 7. Time-averaged rate at the end of learning, $\overline{\lambda}(t_f)$, in the $T - 1/\sigma_W$ plane. A darker colour means a smaller value.

Combining this with equation (4.10) in the case of divergent dynamics then yields the following inequality during training,

$$\frac{T}{S(0)} > \frac{T}{S(t)} > -\frac{K_S(t)}{V_S(t)}. \tag{4.12}$$

On the other hand, a conservative estimate for converging dynamics is given by

$$\frac{T}{S(0)} < -\frac{K_S(t_f)}{V_S(t_f)}. \tag{4.13}$$

The initial eigenvalue distribution of X is given by the Marchenko-Pastur distribution (4.2). Hence, the average level spacing at initialisation is bounded by

$$S(0) \leqslant \frac{1}{n_2 - 1}(x_+ - x_-) = \frac{1}{n_2 - 1} 4\sigma_W^2 \sqrt{r}, \qquad r = \frac{n_2}{n_1}. \tag{4.14}$$

The boundary for divergent dynamics is then approximately given by

$$\frac{n_2 - 1}{4\sqrt{r}} \frac{T}{\sigma_W^2} > -\frac{K_S(t_f)}{V_S(t_f)}. \tag{4.15}$$

The quantities on the LHS are variables determining the network setup, initialisation, and details of the SGD ($n_1, n_2, \sigma_W, T = \epsilon/|\mathcal{B}|$), whereas the quantity on the RHS indicates the signal-to-noise ratio of the gradient at the end of training. Equation (4.15) therefore establishes that the average level spacing converges to a stationary value only when the signal-to-noise ratio of the gradient on the RHS is larger than the stochasticity on the LHS.

To apply equation (4.15) in practice, we compute the gradient and variance of the loss function at the end of training and estimate the RHS for each run, see appendix B for details. We then compare both sides of equation (4.15) to determine whether, according to this criterion, the dynamics has converged or not. Each point in the phase diagram is assessed using an ensemble of $\mathcal{S} = 100$ models, and we determine whether 1σ (68%), 2σ (95%), and 3σ (99.7%) of the models satisfy the inequality (4.15). These boundaries are shown in figure 8. Figure 8(left) shows the 1σ boundary for the smallest σ_W: the vertical line separates the ordered phase and the high-temperature phase, indicating the temperature of the transition. In figure 8(centre and right) three lines are shown, indicating that 68%, 95%, and 99.7% of the models satisfy inequality (4.15) at the end of the training. In the upper region, above the 1σ line,

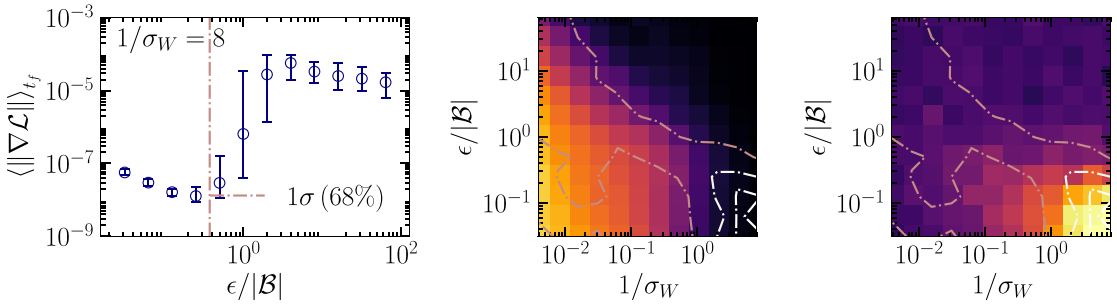

Figure 8. Convergence boundaries obtained from equation (4.15). Left: the vertical dashed-dotted line is the 1σ boundary at fixed $1/\sigma_W = 8$, separating the ordered and the high-temperature phase. Three convergence boundaries on top of the time correlation (centre) and feature alignment (right). The dimmest line indicates that 1σ (68%) of the models have converged, and the brighter two lines indicate that 2σ (95%) and 3σ (99.7%) of the models have converged.

the dynamics has not converged due to thermal fluctuations; this is the high-temperature phase. This boundary tracks the transition region especially well for the time correlation function, shown in the centre. In the lower region towards the left, below the 1σ line, the dynamics has not converged due to disorder; this is the jamming phase in which the dynamics is slow but not completely static. Increasing the requirement for convergence, we observe that the 2σ and 3σ boundaries capture the ordered phase in the lower-right region very well: here essentially all models converge due to the strong alignment of features.

5. Discussion and outlook

Starting from the observation that the loss function of a multilayer neural network can be interpreted as a Hamiltonian of a disordered system, we analysed in some detail its training dynamics under stochastic gradient descent in the case of a teacher-student model. We have shown that three phases can be identified, depending on the choice of hyperparameters. These findings can be conveniently summarised in a phase diagram, in which the ratio of the learning rate over batch size, $T = \epsilon/|\mathcal{B}|$, determines the 'temperature axis', while the initial variance σ_W^2 of the weight matrix elements sets the 'disorder axis'.

Figure 9 contains a summary of our findings, with four observables probing different aspects of the phase structure. Superimposed are the estimates of the phase boundaries between the three phases. Starting in the lower-left corner, the dynamics is dominated by the large initial variance, leading to a jamming phase familiar from spin-glass like systems, in which training is extremely slow. This results in a large loss, poor feature alignment, and persistent memory of the initial state of the network. Reducing the initial variance while keeping the temperature low leads to a transition to a well-trained—or ferromagnetic—phase, in which the loss is small, the features are aligned with the target data, and memory of the initial state is successfully erased. The phase boundary is given by the vertical dashed line, which is derived from a symmetry-breaking argument, directly linked to our choice of hyperbolic tangents as activation functions, see section 2.3. Increasing the level of stochasticity brings us to the high-temperature—or paramagnetic—phase, in which the loss is again large and feature alignment is absent, but due to the fluctuations, the memory of the initial state is erased rapidly. The phase boundary, indicated by the dashed-dotted line, is derived using the framework of Dyson Brownian motion for stochastic weight matrix dynamics and follows essentially a signal-to-noise argument, with the high-temperature phase dominated by noise.

A practical implication of this study is that the phase boundaries can guide the choice of optimal hyperparameters: selecting the largest temperature in the ordered phase leads to fast training. It is noted that to determine the phase boundary to the high-temperature phase, one has to compute the signal-to-noise ratio of the gradient at the end of the training, but having a grasp of what the phase diagram looks like will help in finding optimal hyperparameters effectively, even without explicit knowledge of this boundary. More generally, the significance of this work lies in the fact that it incorporates stochastic training dynamics in the analysis, proposes a direct interpretation of hyperparameters as physical quantities, and provides intuition from theory on how hyperparameters affect the training dynamics.

There are several directions to explore in the future. The phase boundary between the ordered and disordered phases at low temperature is closely linked to hyperbolic tangent activation functions being bounded. In contrast, activation functions of the ReLU type are only bounded on one side, which will affect the phase structure. This is worth exploring further. In this study, we used a teacher-student

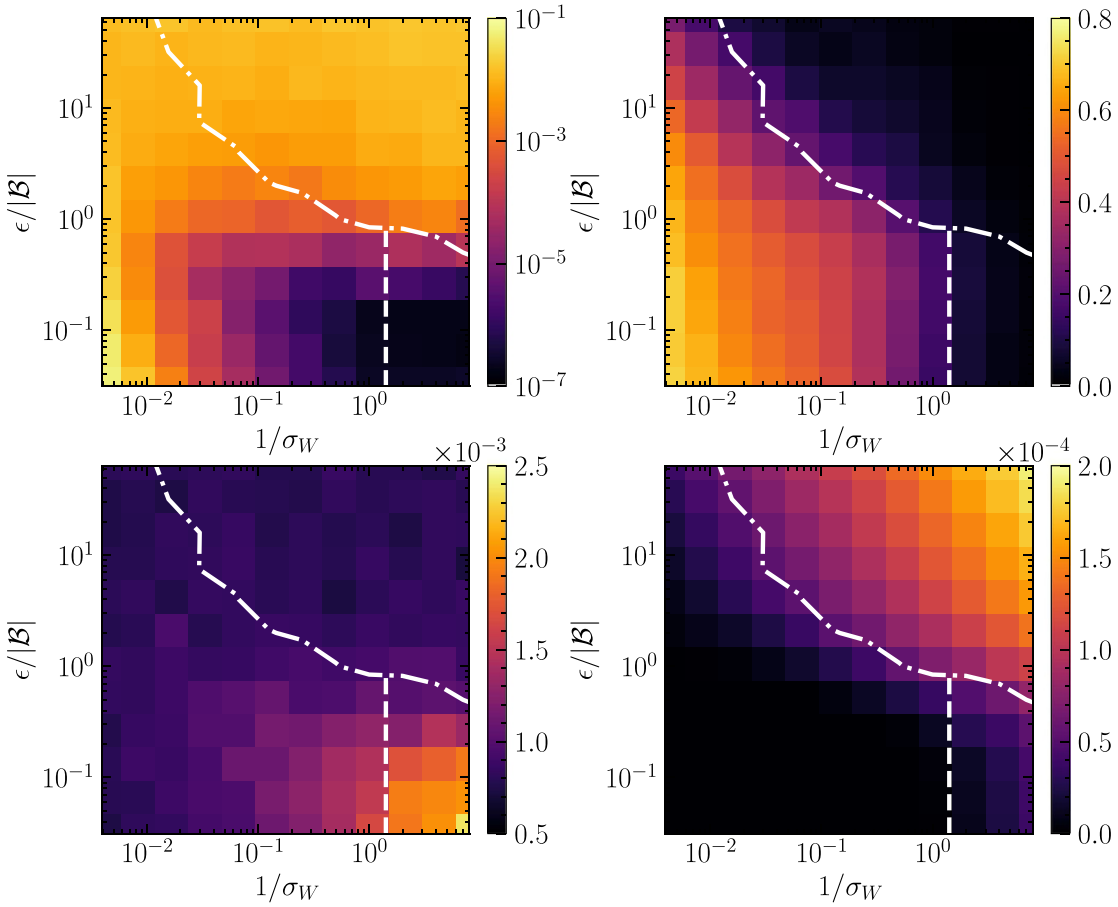

Figure 9. Four observables probing different aspects of the phase structure: final test loss (top left), time correlation of features (top right), feature alignment (lower left), and time-averaged rate of the average level spacing (lower right) in the $T - 1/\sigma_W$ plane, with superimposed the estimates of the phase boundary between the jamming and ferromagnetic phase (dashed vertical line) and between the paramagnetic and the other phases (dashed dotted line).

model and focused on one of the weight matrices in the neural network. It is of interest to extend this to include realistic data sets such as MNIST and analyse the dynamics of all weight matrices. Our neural network is relatively small; considering deeper and especially wider architectures may lead to more pronounced phase boundaries and potentially the use of finite-size scaling techniques to analyse the transitions at a more quantitative level. To achieve this, one first needs to define the correct thermodynamic limit, which preserves the dynamical properties of the optimisation, as naive large width or depth limits lead to lazy training, where the norm of the gradient vanishes [51].

While we followed the evolution of the singular values of the weight matrices during training, we did not study the evolution of their spectral density [23, 52]. Empirical evidence shows that correlations between the weight matrix elements are induced during training, such that the ensemble of weight matrices at the end of training deviates from the Wishart–Laguerre distribution at initialisation, leading, e.g. to heavy tails [39, 40]. It would be interesting to analyse these distributions using techniques developed in the random matrix community [41, 42].

Once the spectral properties of the trained network and a consistent thermodynamic limit are obtained, a theoretical calculation of the phase boundaries might be feasible using methods developed for disordered systems [53, 54]. An ambitious long-term objective would then be to study universality properties of different neural network architectures, potentially leading to insights on designing efficient architectures and predicting scaling behaviour from first principles.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.17046571 [55, 56].

Acknowledgments

We thank Ouraman Hajizadeh and Matteo Favoni for discussion. The main part of this work was carried out when CP was on an Enrichment Placement at The Alan Turing Institute. CP thanks the members of the Turing and fellow Enrichment students for a stimulating experience. CP also thanks the participants of the 2025 Beg Rohu summer school of physics, in particular Elena Agliari, for discussion, and LPENS for support. CP is further supported by the UKRI AIMLAC CDT EP/S023992/1. GA and BL are supported by STFC Consolidated Grant ST/T000813/1. BL is further supported by the UKRI EPSRC ExCALIBUR ExaTEPP Project EP/X017168/1.

We acknowledge the support of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via Welsh Government.

Appendix A. Duality in feature space

In section 2, we claimed that the output of a neural network is a linear combination of learned features. Here, we develop this viewpoint somewhat further.

Using the notation of section 2,

$$\hat{y}_{i\alpha} = \sum_{j=1}^{n_{L-1}} W_{ij}^{(L)} \phi_{j\alpha}, \qquad \phi_{j\alpha} = \phi\left(z_j^{(L-1)}(x_\alpha)\right), \tag{A.1}$$

we note that the $\phi_{j\alpha}$'s are $n_{L-1} \times |\mathcal{D}|$-dimensional matrices. These can be interpreted in two ways, namely as a set of n_{L-1} vectors of size $|\mathcal{D}|$ or as a set of $|\mathcal{D}|$ vectors of size n_{L-1}, i.e.

$$\phi_j \in \left\{ \phi_i \,\middle|\, \phi_i \in \mathbb{R}^{|D|}, i \in [1, n_{L-1}] \right\} \qquad \text{or} \qquad \phi_\alpha \in \left\{ \phi_\beta \middle| \phi_\beta \in \mathbb{R}^{n_{L-1}}, \beta \in [1, |\mathcal{D}|] \right\}. \tag{A.2}$$

In the first interpretation, the output of a neural network is a linear combination of features,

$$\hat{y}_i = \sum_{j=1}^{n_{L-1}} W_{ij}^{(L)} \phi_j, \qquad \hat{y}_i \in \mathbb{R}^{|\mathcal{D}|}, \tag{A.3}$$

whereas in the second interpretation, the output $\hat{y}_\alpha \in \mathbb{R}^{n_L}$ is a vector in the image given by the map $W^{(L)}$, which is not necessarily bijective, depending on the sizes of n_L and n_{L-1}.

We may carry this distinction through to the loss function (2.9), since the sums over the data index and node indices are independent. In the first case, we can write

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{\alpha=1}^{|\mathcal{D}|} \mathcal{H}_\alpha, \qquad \mathcal{H}_\alpha \equiv \frac{1}{2} \sum_{i,j=1}^{n_{L-1}} J_{ij} \phi_{i\alpha} \phi_{j\alpha} - \sum_{i=1}^{n_{L-1}} h_{i\alpha} \phi_{i\alpha}, \tag{A.4}$$

where the term $\mathcal{H}_\alpha$ is a random field spin-glass Hamiltonian of n_{L-1} soft spins. In the second case, we can define a local Hamiltonian along the node index direction, for fixed indices i, j,

$$\mathcal{L} = \sum_{i,j=1}^{n_{L-1}} \mathcal{H}_{ij}, \qquad \mathcal{H}_{ij} \equiv \frac{1}{|\mathcal{D}|} \sum_{\alpha=1}^{|\mathcal{D}|} \left(\frac{1}{2} J_{ij} \phi_{i\alpha} \phi_{j\alpha} - h_{i\alpha} \phi_{j\alpha} \delta_{ij} \right) \tag{A.5}$$

which is a local Hamiltonian of a random field with $n = |\mathcal{D}|$ components and constant J_{ij} for fixed i and j. Similarly, we can define the overlap or alignment $\langle h\phi \rangle$ by contracting over either the data index or the node index, and these two interpretations capture different physics.

In the main part of the paper, we follow the first interpretation and consider the output as a linear combination of features ϕ_j.

Appendix B. Stochastic equation for the average level spacing

Let x_i be an eigenvalue of a symmetric positive definite $N \times N$ matrix X subject to stochastic dynamics,

$$\dot{x}_i = K_i + T\sum_{j\neq i}\frac{V_{ij}}{x_i - x_j} + \sqrt{TV_{ii}}\,\eta_i, \qquad \eta_i \sim \mathcal{N}(0,1). \tag{B.1}$$

We assume the eigenvalues are ordered, $0 \leqslant x_1 < x_2 < \ldots < x_N$. We are interested in the eigenvalue spacing, $S_i = x_{i+1} - x_i$, and its average value for each realisation,

$$S = \frac{1}{N-1}\sum_{i=1}^{N-1} S_i = \frac{1}{N-1}(x_N - x_1). \tag{B.2}$$

In addition, one may average over an ensemble of trajectories.

Our aim is to introduce a 'one-particle theory' for S, in the spirit of [50]. Subtracting the equation for the smallest eigenvalue x_1 from the one for x_N, one obtains

$$\begin{aligned}\dot{x}_N - \dot{x}_1 = (N-1)\,\dot{S} = K_N - K_1 + T\left(\frac{V_{N1}}{x_N - x_1} - \frac{V_{1N}}{x_1 - x_N}\right)\\ + T\sum_{j=2}^{N-1}\left(\frac{V_{Nj}}{x_N - x_j} - \frac{V_{1j}}{x_1 - x_j}\right) + \sqrt{TV_{NN}}\,\eta_N - \sqrt{TV_{11}}\,\eta_1.\end{aligned} \tag{B.3}$$

The noise term is the difference of two independent Gaussian random variables with zero mean and variances TV_{NN} and TV_{11}, respectively. These can be combined as

$$\sqrt{TV_{NN}}\,\eta_N - \sqrt{TV_{11}}\,\eta_1 = \sqrt{T(V_{NN} + V_{11})}\,\eta, \qquad \eta \sim \mathcal{N}(0,1). \tag{B.4}$$

The Coulomb terms are non-trivial. We propose to replace the level spacings by the appropriately scaled average level spacing, i.e.

$$x_N - x_j = (N-j)\,S, \qquad x_1 - x_j = -(j-1)\,S. \tag{B.5}$$

We then write the combination of Coulomb terms as

$$\frac{T}{N-1}\left(\frac{V_{N1}}{x_N - x_1} - \frac{V_{1N}}{x_1 - x_N}\right) + \frac{T}{N-1}\sum_{j=2}^{N-1}\left(\frac{V_{Nj}}{x_N - x_j} - \frac{V_{1j}}{x_1 - x_j}\right) = V_S\frac{T}{S}, \tag{B.6}$$

with

$$V_S = \frac{2V_{N1}}{(N-1)^2} + \frac{1}{N-1}\sum_{k=1}^{N-2}\frac{1}{k}\left(V_{N,N-k} + V_{1,k+1}\right). \tag{B.7}$$

Here we used that V_{ij} is symmetric and relabelled the summation index $(N-j=k, j-1=k)$ in the two sums. The equation for $\dot{S}$ then takes the elegant form

$$\dot{S} = K_S + V_S\frac{T}{S} + \sqrt{2TD_S}\,\eta, \tag{B.8}$$

where we also introduced the drift and the diffusion coefficient,

$$K_S = \frac{1}{N-1}(K_N - K_1), \qquad D_S = \frac{V_{11} + V_{NN}}{2(N-1)^2}. \tag{B.9}$$

In the main text, it is shown that the phase boundary between the high-temperature phase and the other phases depends on the ratio K_S/V_S. We determine this ratio as follows. We start with the gradient after training, K_{ij} at $t = t_f$, for a given training run. This matrix is diagonalised, yielding the eigenvalues K_i. From this, we immediately obtain K_S as the normalised difference between the largest and smallest eigenvalue, see equation (B.9). Using the same transformation, we also rotate $V_{ij} = \mathbb{V}[K_{ij}]$ to compute V_S and D_S, see equations (B.7) and (B.9). Subsequently, we compute K_S/V_S and compare this to LHS of equation (4.15), to test the inequality.

ORCID iDs

Chanju Park 0009-0009-2750-6080
Biagio Lucini 0000-0001-8974-8266
Gert Aarts 0000-0002-6038-3782

References

[1] Zhang C, Bengio S, Hardt M, Recht B and Vinyals O 2017 Understanding deep learning requires rethinking generalization *5th Int. Conf. on Learning Representations (ICLR)* (arXiv:1611.03530)
[2] Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N Vogt-Maranto L and Zdeborová L 2019 Machine learning and the physical sciences *Rev. Mod. Phys.* **91** 045002
[3] Dawid A *et al* Modern applications of machine learning in quantum sciences (arXiv:2204.04198)
[4] Boyda D *et al* 2022 Applications of machine learning to lattice quantum field theory *Snowmass 2021* (arXiv:2202.05838)
[5] Cranmer K, Kanwar G, Racanière S, Rezende D J and Shanahan P E 2023 Advances in machine-learning-based sampling motivated by lattice quantum chromodynamics *Nat. Rev. Phys.* **5** 526–35
[6] Aarts G *et al* 2025 Physics-driven learning for inverse problems in quantum chromodynamics *Nat. Rev. Phys.* **7** 154–63
[7] Zdeborová L 2020 Understanding deep learning is also a job for physicists *Nat. Phys.* **16** 602–4
[8] Amari S-I 1972 Learning patterns and pattern sequences by self-organizing nets of threshold elements *IEEE Trans. Comput.* **C-21** 1197–206
[9] Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities *Proc. Natl Acad. Sci.* **79** 2554–8
[10] Hopfield J J 1984 Neurons with graded response have collective computational properties like those of two-state neurons *Proc. Natl Acad. Sci.* **81** 3088–92
[11] Amit D J, Gutfreund H and Sompolinsky H 1985 Spin-glass models of neural networks *Phys. Rev. A* **32** 1007–18
[12] Gardner E and Derrida B 1988 Optimal storage properties of neural network models *J. Phys. A: Math. Gen.* **21** 271
[13] Nakanishi K and Takayama H 1997 Mean-field theory for a spin-glass model of neural networks: TAP free energy and the paramagnetic to spin-glass transition *J. Phys. A: Math. Gen.* **30** 8085
[14] Barra A, Genovese G, Sollich P and Tantari D 2017 Phase transitions in restricted Boltzmann machines with generic priors *Phys. Rev. E* **96** 042156
[15] Albanese L, Barra A, Bianco P, Durante F and Pallara D 2024 Hebbian learning from first principles *J. Math. Phys.* **65** 113302
[16] Erbin H, Lahoche V and Samary D O 2022 Non-perturbative renormalization for the neural network-QFT correspondence *Mach. Learn.: Sci. Technol.* **3** 015027
[17] Demirtas M, Halverson J, Maiti A, Schwartz M D and Stoner K 2024 Neural network field theories: non-Gaussianity, actions and locality *Mach. Learn.: Sci. Technol.* **5** 015002
[18] Lee J, Sohl-dickstein J, Pennington J, Novak R, Schoenholz S and Bahri Y 2018 Deep neural networks as Gaussian processes *Int. Conf. on Learning Representations* (arXiv:1711.00165)
[19] Jacot A, Gabriel F and Hongler C 2018 Neural tangent kernel: convergence and generalization in neural networks *Advances in Neural Information Processing Systems* vol 31 (arXiv:1806.07572)
[20] Luo T, Xu Z-Q J, Ma Z and Zhang Y 2021 Phase diagram for two-layer relu neural networks at infinite-width limit *J. Mach. Learn. Res.* **22** 1–47 (available at: https://arxiv.org/abs/2007.07497)
[21] Roberts D A, Yaida S and Hanin B 2022 *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks* (Cambridge University Press) (http://dx.doi.org/10.1017/9781009023405) (arXiv:2106.10165)
[22] Hanin B 2023 Random neural networks in the infinite width limit as Gaussian processes *Ann. Appl. Probab.* **33** 4798–819
[23] Aarts G, Lucini B and Park C 2025 Stochastic weight matrix dynamics during learning and Dyson Brownian motion *Phys. Rev. E* **111** 015303
[24] Decelle A and Furtlehner C 2021 Restricted Boltzmann machine: recent advances and mean-field theory *Chin. Phys. B* **30** 040202
[25] Yaida S 2019 Fluctuation-dissipation relations for stochastic gradient descent *Int. Conf. on Learning Representations* (arXiv:1810.00004)
[26] Mandt S, Hoffman M D and Blei D M 2017 Stochastic gradient descent as approximate Bayesian inference *J. Mach. Learn. Res.* **18** 1–35 (available at: www.jmlr.org/papers/v18/17-214.html)
[27] Granziol D, Zohren S and Roberts S J 2020 Learning rates as a function of batch size: a random matrix theory approach to neural network training *J. Mach. Learn. Res.* **23** 173 (available at: https://www.jmlr.org/papers/v23/20-1258.html)
[28] Mandt S, Blei D M and Hoffman M D 2015 Continuous-time limit of stochastic gradient descent revisited *8th NIPS Workshop on Optimization for Machine Learning* (available at: https://opt-ml.org/oldopt/papers/OPT2015_paper_8.pdf)
[29] Schoenholz S S, Gilmer J, Ganguli S and Sohl-Dickstein J 2017 Deep information propagation *Int. Conf. on Learning Representations* (arXiv:1611.01232)
[30] Bassi A, Albert C, Lucchi A, Baity-Jesi M and Francazi E 2025 When the left foot leads to the right path: bridging initial prejudice and trainability (arXiv:2505.12096)
[31] Sclocchi A and Wyart M 2024 On the different regimes of stochastic gradient descent *Proc. Natl Acad. Sci.* **121** e2316301121
[32] Ghio D, Dandi Y, Krzakala F and Zdeborová L 2024 Sampling with flows, diffusion and autoregressive neural networks from a spin-glass perspective *Proc. Natl Acad. Sci.* **121** e2311810121
[33] D'Amico F, Rossi S, del Bono L M and Negri M 2025 Pseudo-likelihood produces associative memories able to generalize, even for asymmetric couplings *New Frontiers in Associative Memories* (arXiv:2507.05147)
[34] Achilli B, Ambrogioni L, Lucibello C, Mézard M and Ventura E 2025 The capacity of modern hopfield networks under the data manifold hypothesis *New Frontiers in Associative Memories* (arXiv:2503.09518)
[35] Bengio Y, Courville A and Vincent P 2013 Representation learning: a review and new perspectives *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 1798–828
[36] Sherrington D and Kirkpatrick S 1975 Solvable model of a spin-glass *Phys. Rev. Lett.* **35** 1792–6
[37] Poole B, Lahiri S, Raghu M, Sohl-Dickstein J and Ganguli S 2016 Exponential expressivity in deep neural networks through transient chaos *Advances in Neural Information Processing Systems* vol 29 (arXiv:1606.05340)

[38] Livan G, Novaes M and Vivo P 2018 *Introduction to Random Matrices—Theory and Practice* vol 26 (Springer) (http://dx.doi.org/10.1007/978-3-319-70885-0) (arXiv:1712.07903)
[39] Martin C H and Mahoney M W 2021 Implicit self-regularization in deep neural networks: evidence from random matrix theory and implications for learning *J. Mach. Learn. Res.* **22** 1–73 (available at: https://jmlr.org/papers/v22/20-410.html)
[40] Mahoney M and Martin C 2019 Traditional and heavy tailed self regularization in neural network models *Proc. 36th Int. Conf. on Machine Learning* vol 97 pp 4284–93(arXiv:1901.08276)
[41] Saberi A A, Saber S and Moessner R 2024 Interaction-correlated random matrices *Phys. Rev. B* **110** L180102
[42] Akemann G and Vivo P 2008 Power-law deformation of Wishart-Laguerre ensembles of random matrices *J. Stat. Mech.* **2008** P09002
[43] Smith S L, Kindermans P-J and Le Q V 2018 Don't decay the learning rate, increase the batch size *Int. Conf. on Learning Representations* (arXiv:1711.00489)
[44] Goyal P *et al* 2018 Accurate, large minibatch SGD: training ImageNet in 1 hour (arXiv:1706.02677)
[45] Ziyin L, Wang M, Li H and Wu L 2024 Parameter symmetry and noise equilibrium of stochastic gradient descent *Advances in Neural Information Processing Systems* vol 37 pp 93874–906
[46] Mezard M, Parisi G and Virasoro M 1986 *Spin Glass Theory and Beyond* (WORLD SCIENTIFIC) (http://dx.doi.org/10.1142/027110.1142/0271)
[47] Rahimi A and Recht B 2007 Random features for large-scale kernel machines *Advances in Neural Information Processing Systems* vol 20 (Curran Associates, Inc.)
[48] Dyson F J 1962 A Brownian-motion model for the eigenvalues of a random matrix *J. Math. Phys.* **3** 1191–8
[49] Mehta M 1967 *Random Matrices and the Statistical Theory of Energy Levels* (Academic)
[50] Pimpinelli A, Gebremariam H and Einstein T L 2005 Evolution of terrace-width distributions on vicinal surfaces: Fokker-Planck derivation of the generalized Wigner surmise *Phys. Rev. Lett.* **95** 246101
[51] Chizat L, Oyallon E and Bach F 2019 On lazy training in differentiable programming *Advances in Neural Information Processing Systems* vol 32 (arXiv:1812.07956)
[52] Aarts G, Hajizadeh O, Lucini B and Park C 2024 Dyson Brownian motion and random matrix dynamics of weight matrices during learning *38th Conf. on Neural Information Processing Systems* (arXiv:2411.13512)
[53] Guerra F 2003 Broken replica symmetry bounds in the mean field spin glass model *Commun. Math. Phys.* **233** 1–12
[54] Agliari E, Alemanno F, Barra A and Fachechi A A novel derivation of the Marchenko-Pastur law through analog bipartite spin-glasses (arXiv:1811.08298)
[55] Park C, Aarts G and Lucini B 2025 Phase diagram and eigenvalue dynamics of stochastic gradient descent in multilayer neural networks—data release *Zenodo* (https://doi.org/10.5281/zenodo.17046571)
[56] Bennett E 2025 The TELOS Collaboration approach to reproducibility and open science (arXiv:2504.01876)