A statistical mechanics framework for Bayesian deep neural networks beyond the infinite-width limit

R. Pacelli, S. Ariosto, M. Pastore, F. Ginelli, M. Gherardi, P. Rotondo

A statistical mechanics framework for Bayesian deep neural networks beyond the infinite-width limit

R. Pacelli,[1,2] S. Ariosto,[3,4] M. Pastore,[5,6] F. Ginelli,[3,4] M. Gherardi,[7,4] and P. Rotondo[8,4]

[1] *Dipartimento di Scienza Applicata e Tecnologia, Politecnico di Torino, 10129 Torino, Italy*
[2] *Artificial Intelligence Lab, Bocconi University, 20136 Milano, Italy*
[3] *Dipartimento di Scienza e Alta Tecnologia and Center for Nonlinear and Complex Systems, Università degli Studi dell'Insubria, Via Valleggio 11, 22100 Como, Italy*
[4] *I.N.F.N. Sezione di Milano, Via Celoria 16, 20133 Milano, Italy*
[5] *Université Paris-Saclay, CNRS, LPTMS, 91405 Orsay, France*
[6] *Laboratoire de physique de l'École normale supérieure, CNRS, PSL University, Sorbonne University, Université Paris-Cité, 24 rue Lhomond, 75005 Paris, France*
[7] *Università degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy*
[8] *Dipartimento di Scienze Matematiche, Fisiche e Informatiche, Università degli Studi di Parma, Parco Area delle Scienze, 7/A 43124 Parma, Italy*

Despite the practical success of deep neural networks, a comprehensive theoretical framework that can predict practically relevant scores, such as the test accuracy, from knowledge of the training data is currently lacking. Huge simplifications arise in the infinite-width limit, where the number of units N_ℓ in each hidden layer ($\ell = 1, \dots, L$, being L the depth of the network) far exceeds the number P of training examples. This idealisation, however, blatantly departs from the reality of deep learning practice. Here, we use the toolset of statistical mechanics to overcome these limitations and derive an approximate partition function for fully-connected deep neural architectures, which encodes information about the trained models. The computation holds in the "thermodynamic limit" where both N_ℓ and P are large and their ratio $\alpha_\ell = P/N_\ell$ is finite. This advance allows us to obtain (i) a closed formula for the generalisation error associated to a regression task in a one-hidden layer network with finite α_1; (ii) an approximate expression of the partition function for deep architectures (via an "effective action" that depends on a finite number of "order parameters"); (iii) a link between deep neural networks in the proportional asymptotic limit and Student's t processes.

I. INTRODUCTION

The rise of deep learning, driven by advances in computing technology and foreshadowed by decades of research, has outpaced our ability to develop a solid theoretical foundation [1, 2]. Filling the gaps in our understanding of deep learning on a fundamental level is a long-time collective effort involving several communities. Statistical physics achieved far-reaching results in this regard, and remains a wellspring of fresh perspectives and breakthroughs [3–11]. One notable recent advance was obtained by considering the infinite-width limit, where the number of training data P is fixed and the size of the hidden layers is taken to infinity. The observation that such deep models are equivalent to Gaussian processes (GPs) [12–20] established a connection between deep learning and kernel methods [21], and provided a statistical physics description of this regime [7, 22, 23].

However, there is agreement that a more complete theory should address deep learning beyond the infinite-width limit [24–29]: in fact, realistic neural networks operate in a qualitatively different regime, where the number of training examples exceeds the width of the largest layer. Modeling the finite-width regime in the thermodynamic limit, where the number of degrees of freedom diverges and the tools of statistical mechanics are most effective, amounts to taking the asymptotic limit where both the size of the training set P and the number of units in each hidden layer N_ℓ are taken to infinity with their ratios fixed, as we consider in the present work:

$$P, N_\ell \to \infty, \quad \alpha_\ell = \frac{P}{N_\ell} \text{ finite} \qquad \forall \ell = 1, \dots, L \tag{1}$$

with L being the (finite) depth of the network (the scaling of P with the input size N_0, which deserves special care, is discussed in section III F of the Methods). This choice guarantees that such networks work in the over-parametrised regime.

Another fruitful line of research, in the direction of overcoming the limitations of the infinite-width limit, sacrifices the non-linear nature of the network by considering a deep *linear* input-output mapping: even if the resulting architecture lacks the expressive power [30–37] of the same model with non-linearities, the multi-layer structure maintains the learning problem non-convex, while amenable to analytical investigation [38, 39]. Very recently, Li and Sompolinsky [5] proposed a method to analytically evaluate properties of deep finite-width linear networks (e.g., their generalisation error) trained on a generic fixed training set. However, the more relevant case of generic non-linear DNNs remains an open problem, despite some recent notable attempts to address it [3, 26–29, 40–42].

In statistical mechanics, the partition function is the central object encoding the properties of the system in the thermodynamic limit. In this work we address the analytical computation of the partition function of a fully-connected, multi-layer, non-linear neural network, as a

function of the training set in the asymptotic limit defined in (1). Technically, the computation amounts to integrating out an extensive number of degrees of freedom (the weights of the network), thus landing on an expression that involves only a finite number (proportional to the depth L) of integrals, to be evaluated by the saddle-point method. In the one-hidden-layer (1HL) case, the only key approximation is justified by a generalised central limit theorem due to Bardet and Surgailis [43] (which belongs to a class of results known as Breuer-Major (BM) theorems [44] from the seminal paper [45]).

In the general case of an architecture with L hidden fully-connected layers, we show that the distribution of the pre-activations at each layer ℓ is a mixture of Gaussians that depends on ℓ parameters. Notably, the back-propagating integration performed in [5] is not a viable option as soon as non-linearities are added to the model. We introduce a forward-propagating method to carry out nested integrations starting from the input layer. This result depends on an assumption that is similar, at least in spirit, to the Gaussian equivalence principle employed for random and generic feature models [46–52].

From these developments, we are able to obtain quantitative predictions for the generalisation error of the network below the interpolation threshold. Moreover, our results have an intriguing interpretation from the point of view of stochastic processes: we show that in the case of finite α_ℓ the GP arising in the infinite-width limit of Bayesian neural networks [15] should be generalised to a Student's t stochastic process [53].

As a first application of the theory, we establish a simple criterion (equivalent to the one found in the linear case [5, 54] and in finite-P perturbation theory [55, 56]) to predict whether it is convenient, in terms of generalisation performance, to employ a finite-width deep neural network over its infinite-width version.

Problem setting - We consider a supervised learning problem with training set $\mathcal{T}_P = \{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^P$, where each $\mathbf{x}^\mu \in \mathbb{R}^{N_0}$ and the corresponding labels $y^\mu \in \mathbb{R}$. The architecture is a deep neural network $f_{\mathrm{DNN}}(\mathbf{x})$ with $(L-1)$ fully-connected hidden (FC) layers and a final linear readout layer as defined in (23). We analyse regression problems with a quadratic loss function:

$$\mathcal{L} = \frac{1}{2}\sum_{\mu=1}^{P}[y^\mu - f_{\mathrm{DNN}}(\mathbf{x}^\mu)]^2 + \mathcal{L}_{\mathrm{reg}}\,, \tag{2}$$

$$\mathcal{L}_{\mathrm{reg}} = \frac{\lambda_L}{2\beta}\sum_{i_L=1}^{N_L} v_{i_L}^2 + \frac{1}{2\beta}\sum_{\ell=0}^{L-1}\lambda^{(\ell)}\|W^{(\ell)}\|^2\,, \tag{3}$$

where L^2 regularisations have been added for each layer to the loss function, $\|\cdot\|$ is the standard Frobenius norm defined for the weights matrices $W^{(\ell)}$, and $\beta = 1/T$ is the inverse temperature parameter.

As a standard practice in statistical mechanics of deep learning, we define the partition function of the problem as:

$$Z = \int \mathcal{D}\theta\, e^{-\beta\mathcal{L}(\theta)}\,. \tag{4}$$

where the symbol $\int \mathcal{D}\theta$ indicates the collective integration over the weights of the network, $\theta = \{W^{(\ell)}, v\}$. This choice enforces minimization of the training error for $\beta \to \infty$. We notice that scaling $\mathcal{L}_{\mathrm{reg}}$ by $1/\beta$ has a natural Bayesian learning interpretation: the Gibbs probability $P_\beta(\theta) = Z^{-1}e^{-\beta\mathcal{L}(\theta)}$ associated with the partition function in equation (4) is the posterior distribution of the weights after training, whereas the Gaussian regularization is a prior equivalent to assuming that weights at initialization have been drawn from a Gaussian distribution

In this framework, the average test error over a new (unseen) example $(\mathbf{x}^0, y^0)$ is given by:

$$\langle \epsilon_{\mathrm{g}}(\mathbf{x}^0, y^0)\rangle = \int \mathcal{D}\theta\,[y^0 - f_{\mathrm{DNN}}(\mathbf{x}^0)]^2\frac{e^{-\beta\mathcal{L}(\theta)}}{Z}\,. \tag{5}$$

II. RESULTS

A. Asymptotic effective action for one-hidden-layer neural networks in the Bayesian setting

In the case of 1HL architectures, we are able to reduce the partition function (4) to the following two-variables integral in the thermodynamic limit described in (1):

$$Z = \int \mathrm{d}Q \int \mathrm{d}\bar{Q}\, \exp\left[-\frac{N_1}{2}S(Q,\bar{Q})\right] \tag{6}$$

where we have defined an effective action S given by:

$$\begin{aligned} S =& -Q\bar{Q} + \log(1+Q) + \frac{\alpha_1}{P}\,\mathrm{Tr}\log\beta\left[\frac{\mathbb{1}}{\beta} + \frac{\bar{Q}K}{\lambda_1}\right] \\ &+ \frac{\alpha_1}{P}y^\top\left[\frac{\mathbb{1}}{\beta} + \frac{\bar{Q}K}{\lambda_1}\right]^{-1} y \end{aligned} \tag{7}$$

and we have introduced a vectorial notation for the output $y^\top = (y^1, y^2, \ldots, y^P)$. The $P \times P$, input-dependent kernel K/λ_1 is the neural network Gaussian process (NNGP) kernel [15] arising in the infinite-width limit and its precise definition in terms of the input covariance matrix (rescaled by the Gaussian prior of the first layer λ_0) $C_{\mu\nu} = \mathbf{x}^\mu \cdot \mathbf{x}^\nu/(\lambda_0 N_0)$ is given in the Methods, equation (45). Note also that equation (7) holds for zero-mean activation functions, that is functions whose average over a centered Gaussian is zero (see equation (44); an effective action for the generic finite-mean case is reported in the supplemental material [57], Sec. IV) and that for many reasonable non-linearities and input data distributions the derivation goes through at least in the regime $P = O(N_0)$ (we discuss this key technical point in the Methods). This is the first main result of our work

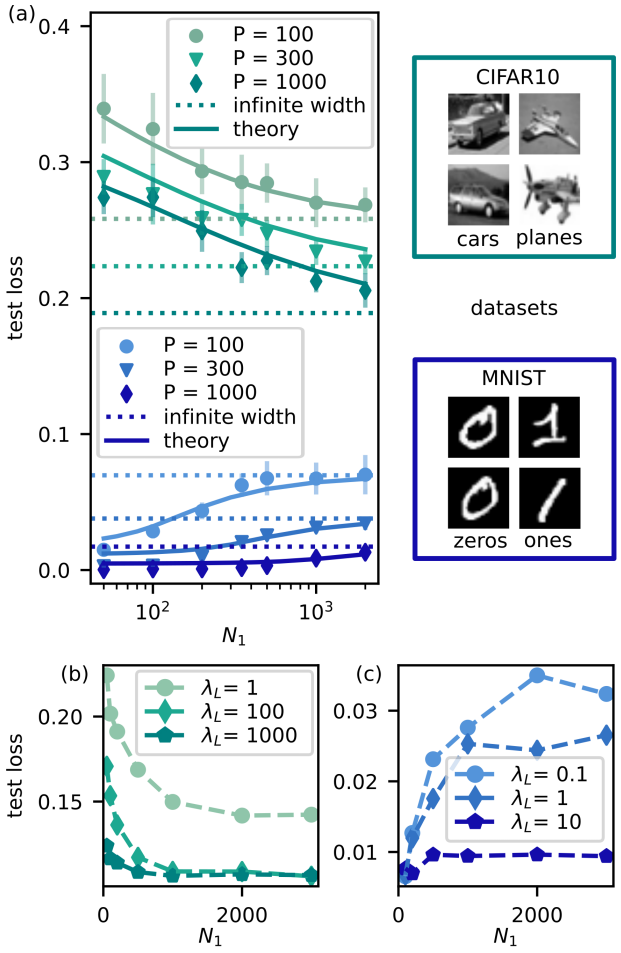

FIG. 1. (a) Learning curves of 1HL architectures with Erf activation (trained with a discretised Langevin dynamics, see also Methods) as a function of the hidden layer size N_1 for two regression tasks on the CIFAR10 (above) and MNIST (below) datasets. Zero/one labels have been chosen in both cases and the images of the CIFAR10 dataset have been gray-scaled and down-scaled to $N_0 = 28 \times 28$. The experimental test loss at different values of the trainset size P (points with error bars indicating one standard deviation) are compared with the theory computed from equation (8) (solid lines). The bar centres are computed as the average over an ensemble of $S = 450$ equilibrium configurations. Samples are taken every 10^4 Langevin steps (after thermalisation of the dynamics). The error bar represents one standard deviations from the average. (b,c) Experimental learning curves as a function of N_1 for increasing values of the Gaussian prior of the last layer λ_1. Error bars are within points, and dashed lines connecting the points are shown to guide the eye. The nets are trained on $P = 3000$ examples from the CIFAR10 dataset in (b) and $P = 500$ examples from MNIST in (c). Two qualitative predictions of the theory at zero temperature are checked: (i) the generalisation loss should decrease for any N_1 when λ_1 grows; (ii) the dependence of the learning curves on N_1 disappears in the large-λ_1 limit, since the bias is constant (see also main text).

and we conjecture it is exact since the only key Gaussian approximation that we perform is justified by the extension of the Breuer-Major theorem [43], as argued in the Methods.

In the supplemental material [57], we obtain a number of additional results that did not enter here for space limitations: (i) a re-derivation of the effective action in the case of linear activation function, valid at fixed P, N_1, N_0, together with a comparison with the results given in [5, 58]; (ii) a specific derivation of the effective action for quadratic activation function, which makes no use of the Breuer-Major theorem; (iii) the generalisation of the effective action in equation (7) to the case of multiple (but finite) outputs.

We can now solve equation (7) using the saddle-point method, since $N_1 \to \infty$, which amounts to finding the solutions Q^*, $\bar{Q}^*$ of the system of equations $\partial_Q S = 0$, $\partial_{\bar{Q}} S = 0$ (the infinite-width limit is re-obtained for $\alpha_1 \to 0$ and corresponds to the particular solution $Q^* = 0$, $\bar{Q}^* = 1$). In the zero-temperature limit, we can find the analytical solution of the saddle-point equations (see Methods). A straightforward computation shows that the generalisation error is given in terms of the usual bias-variance decomposition:

$$\begin{aligned} \langle \epsilon_g(\mathbf{x}^0, y^0) \rangle &= \left(y^0 - \Gamma_1\right)^2 + \sigma_1^2 \,, \\ \Gamma_1 &= \sum_{\mu,\nu} \kappa_\mu(\mathbf{x}^0) K^{-1}_{\mu\nu}\, y_\nu \,, \\ \sigma_1^2 &= \frac{\bar{Q}^*}{\lambda_1} \left[\kappa_0(\mathbf{x}^0) - \sum_{\mu,\nu} \kappa_\mu(\mathbf{x}^0) K^{-1}_{\mu\nu}\, \kappa_\nu(\mathbf{x}^0) \right] \,, \end{aligned} \tag{8}$$

where $\kappa_\mu(\mathbf{x}^0)$, $\kappa_0(\mathbf{x}^0)$ can be computed from the functional definition of the NNGP kernel using the new unseen input $\mathbf{x}^0$, as shown in the Methods.

We can directly employ equation (8) to obtain testable predictions for the generalisation error of finite-width 1HL architectures trained in the Bayesian learning setting, as we do in panel (a) of Fig. 1 for two specific regression tasks defined on the CIFAR10 and MNIST datasets (details on the numerical experiments are provided in the Methods section III H and in Sec. V of the supplemental material [57]). It turns out that the generalisation curves for the two regression tasks are monotonically increasing (decreasing) as a function of N_1 depending on the fact that the observable $y^\top (K/\lambda_1)^{-1} y/P$ is smaller (larger) than one. The importance of this quantity in controlling the generalisation performance has been already noted in linear networks [5, 54] as well as in direct perturbation theory at finite P for non-linear networks [55, 56].

We also point out two semi-quantitative predictions for the general behavior of the generalisation error, just by looking at the dependence of equation (8) on the size of the hidden layer N_1 and on the Gaussian prior of the last layer λ_1. At $T = 0$, the bias is constant as a function of N_1 (as explicitly observed also in the linear case in Ref. [5]) and of λ_1. On the contrary, the variance depends on N_1 and decreases as $1/\sqrt{\lambda_1}$ in the large-λ_1 limit. These observations lead to the following

two testable predictions: (i) increasing the magnitude of the Gaussian prior λ_1 should systematically improve the generalisation performance at any N_1; (ii) for large λ_1 the dependence on N_1 of the generalisation error should disappear (see also the numerical experiments performed in panel (b) in Fig. 1).

B. Link between Student's t-processes and shallow neural networks in the proportional limit

In obtaining the results reported in Sec. II A, our theory can be formulated as a statement on the probability distribution of the output variables

$$s^\mu \equiv \frac{1}{\sqrt{N_1}} \sum_{i_1=1}^{N_1} v_{i_1} \sigma(h^\mu_{i_1}) \,, \tag{9}$$

where $h \sim \mathcal{N}(0, C \otimes \mathbb{1}_{N_1})$, $v \sim \mathcal{N}(0, \lambda_1^{-1} \mathbb{1}_{N_1})$. Proceeding as in the derivation of the partition function presented in Methods, the p.d.f. of these variables can be written as a re-weighted Fourier transform,

$$P(s|\mathcal{T}_P) = \frac{e^{-\frac{\beta}{2}\sum_\mu (y^\mu - s^\mu)^2}}{Z} \int \prod_\mu \frac{\mathrm{d}\bar{s}^\mu}{2\pi} \, e^{i\bar{s}^\top s} \Xi(\bar{s}) \,, \tag{10}$$

of the function

$$\Xi(\bar{s}) = \left(1 + \frac{1}{\lambda_1 N_1} \sum_{\mu,\nu}^{P} \bar{s}^\mu K_{\mu\nu}(C) \bar{s}^\nu \right)^{-\frac{N_1}{2}} . \tag{11}$$

It is straightforward to notice that as long as $N_1 \to \infty$ and $N_1 \gg P$, the dependence on N_1 disappears and we get:

$$\Xi(\bar{s}) \to e^{-\frac{1}{2\lambda_1}\sum_{\mu,\nu}^{P} \bar{s}^\mu K_{\mu\nu}(C) \bar{s}^\nu} . \tag{12}$$

This quantity has a very natural interpretation in view of the NNGP literature. Indeed, for N_1 large and P finite, the variables (9) are jointly multivariate Gaussian distributed according to the central limit theorem, as noted for example in [15]: this limit corresponds indeed to the RHS of our equation (12) and is the cornerstone of the mapping of an infinite-width Bayesian neural network to a GP. This is however no more the case when P is comparable to N_1: equation (11), derived exploiting the Gaussian equivalence based on the BM theorem in the proportional asymptotic limit $P/N_1 \sim O(1)$, is suggesting that the variables $\bar{s}^\mu$ are distributed according to a multivariate Student's t-distribution [53, 59–61].

The need of considering Student's t-processes as a generalisation of NNGPs has been noted already in the case of different priors on the distribution of the last layer's weights [62]. Non-Gaussianity of the posterior in a form similar to that of Eq. (11) has appeared also in [63–65]. The reason why this kind of process arises in the case we are considering here can be understood with an heuristic argument: when N_1 and P are of the same order, we cannot take the limit $N_1 \to \infty$ before $P \to \infty$, and so we need to use the empirical covariance of the output variables s^μ instead of their true one in estimating their probability distribution. A more precise characterization of these neural network Student's t-processes (NNTPs) and the regime where they arise represent interesting topics for future work.

C. Asymptotic effective action for deep neural networks in the Bayesian setting

In the generic case of a deep fully-connected architecture with a finite number of layers L and zero-mean activation function, we express the partition function in terms of a $2L$-dimensional integral (see Methods):

$$Z_{\mathrm{DNN}} = \int \prod_{\ell=1}^{L} \mathrm{d}Q_\ell \, \mathrm{d}\bar{Q}_\ell e^{-\frac{N_L}{2} S_{\mathrm{DNN}}(\{Q_\ell, \bar{Q}_\ell\})} \,, \tag{13}$$

where the effective action is given by:

$$\begin{aligned} S_{\mathrm{DNN}} = &\sum_{\ell=1}^{L} \frac{\alpha_L}{\alpha_\ell} \left[-Q_\ell \bar{Q}_\ell + \log(1 + Q_\ell) \right] \\ &+ \frac{\alpha_L}{P} \mathrm{Tr} \log \beta \left(\frac{\mathbb{1}}{\beta} + K_L^{(R)}(\{\bar{Q}_\ell\}) \right) \\ &+ \frac{\alpha_L}{P} y^T \left(\frac{\mathbb{1}}{\beta} + K_L^{(R)}(\{\bar{Q}_\ell\}) \right)^{-1} y \end{aligned} \tag{14}$$

and we have introduced a renormalised kernel $K^{(R)}$ that generalises the the recurrence relation for the L-layer NNGP kernel as:

$$K_\ell^{(R)}(\{\bar{Q}_\ell\}) = \bar{Q}_\ell / \lambda_\ell K \circ \left[K_{\ell-1}^{(R)}(\{\bar{Q}_\ell\}) \right] \,, \quad K_0^{(R)} = C \,, \tag{15}$$

where C is the covariance matrix of the inputs defined above and we stress that each $K_\ell^{(R)}$ depends on the variables $\bar{Q}_1, \ldots, \bar{Q}_{\ell-1}$ only. For completeness, we notice that the recurrence relation for the infinite-width kernel K_L is given by equation (15) with $\bar{Q}_\ell = 1 \; \forall \ell = 1, \ldots, L$.

This action shares the same structure as the one found in section II A for the special case of 1HL, with the difference that for L hidden layers, the recursive nature of the derivation introduces additional order parameters that are nested in the definition of the kernel K_L. Furthermore, since our derivation applies to layers of arbitrary size N_ℓ, the action also depends on the aspect ratios α_ℓ. In the supplemental material [57], we derive a series of additional results: (i) we generalise this effective action for finite-mean activation functions; (ii) we show how to recover the linear case in the isotropic limit $\alpha_\ell = \alpha$ $\forall \ell = 1, \ldots, L$; (iii) using (i) we show how to correct the heuristic theory for ReLU activation presented in Ref. [5].

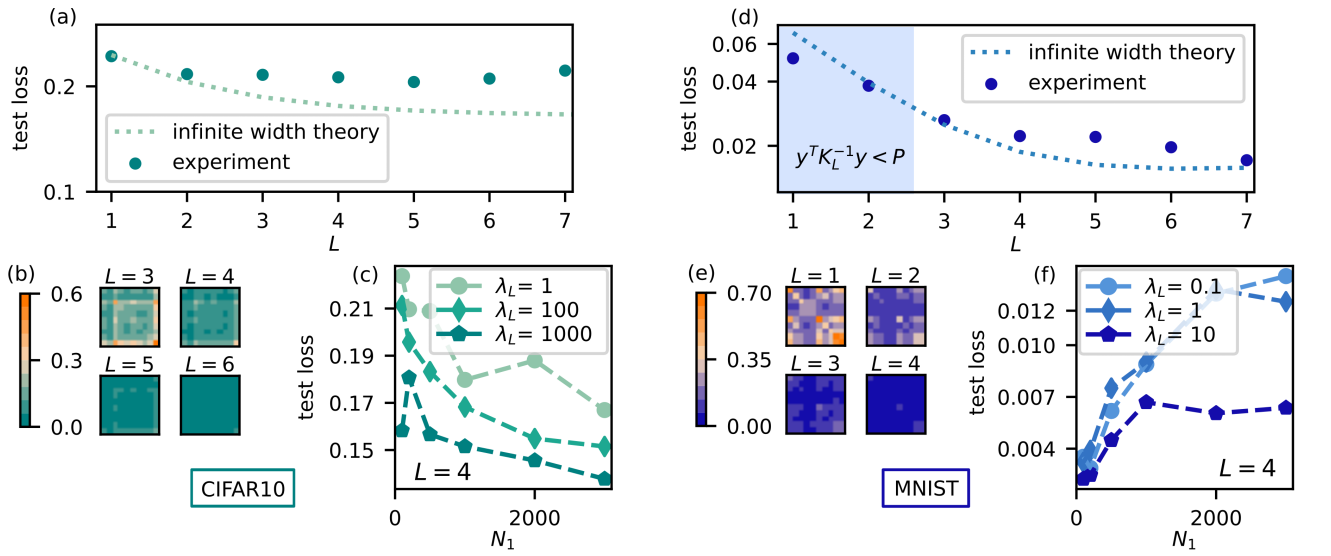

FIG. 2. (a,d) Test loss of a L-HL neural network with ReLU activation, as a function of the depth L, for $P = 100$. The net is trained on a regression task in the small α regime ($\alpha = 0.1$), close to the infinite-width limit. The finite-width network can outperform the infinite-width prediction only when $s_L < 1$ (shaded area), i.e. only for the MNIST task and for depth $L < 3$. (b,e) Visualisation of the entries of the infinite-width NNGP kernel at different layers of the network. The ReLU NNGP kernel converges to zero after repeated iterations. This generates almost vanishing eigenvalues that makes s_L eventually always larger than one. (c,f) Test loss of a 4-HL network trained on $P = 1000$ examples with different regularisation strengths (with $N_\ell = N = 1000$). While increasing the magnitude of the Gaussian prior of the last layer still improves generalisation for all N, it is not clear anymore (as it was for 1HL networks) that the curve at large λ_L is a constant as a function of N. The dashed line is shown to guide the eye. In all panels, error bars lie within points.

The computation of the generalisation error over a new example $(\mathbf{x}^0, y^0)$ gives:

$$\langle \epsilon_{\rm g}(\mathbf{x}^0, y^0)\rangle = (y^0 - \Gamma_L)^2 + \sigma_L^2 \tag{16}$$

where

$$\Gamma_L = \sum_{\mu\nu} \kappa^{(R)}_{L\mu} \left(\frac{\mathbb{1}}{\beta} + K^{(R)}_L(\{\bar{Q}_\ell\})\right)^{-1}_{\mu\nu} y_\nu, \tag{17}$$

$$\sigma_L^2 = \kappa^{(R)}_{L0} - \sum_{\mu\nu} \kappa^{(R)}_{L\mu} \left(\frac{\mathbb{1}}{\beta} + K^{(R)}_L(\{\bar{Q}_\ell\})\right)^{-1}_{\mu\nu} \kappa^{(R)}_{L\nu} \tag{18}$$

and $\kappa^{(R)}_{L\mu}$, $\kappa^{(R)}_{L0}$ are recursive kernels computed from the recurrence given in equation (15) using the input $\mathbf{x}^0$ in the initial conditions.

Note that also in this case we can perform the same scaling analysis of the dependence of the generalisation error on the Gaussian prior in the last layer λ_L (in the zero temperature limit). It turns out that the bias does not depend on it, whereas the variance σ_L^2 approaches zero as $1/\sqrt{\lambda_L}$ as λ_L is taken to infinity. This means that also in the case of finite depth $L > 1$, training at large values of the Gaussian prior of the last layer should improve generalisation at any aspect ratio of the network. We confirm this general observation with numerical experiments in panels (c) and (f) of Fig. 2. However, differently from the 1HL case, we observe that the bias does depend on the aspect ratio even in the zero-temperature limit and we cannot expect anymore that the dependence on the aspect ratios of the networks α_ℓ disappears in the $\lambda_L \to \infty$ limit.

We can obtain another prediction of the theory at L layers (that again confirms previous results on linear networks and perturbative calculations for non-linear networks [5, 54–56]) by considering the effective action for ReLU activation. A straightforward Taylor expansion around the infinite-width limit $\alpha_\ell = \alpha = 0\ \forall \ell = 1, \dots, L$ shows that the first correction to the test loss $\Delta\epsilon_{\rm g}$ is proportional to:

$$\Delta\epsilon_{\rm g} \propto \alpha \left(\frac{1}{P} y^T K_L^{-1} y - 1\right). \tag{19}$$

where K_L is the solution of recurrence in equation (15) for $\bar{Q}_\ell = 1\ \forall \ell = 1, \dots, L$ and ReLU activation. This means that there exists a simple scalar observable that determines whether the finite-width deep neural network will outperform its infinite-width counterpart that generalises the one found at 1HL:

$$s_L = \frac{1}{P} y^T K_L^{-1} y\,. \tag{20}$$

In particular, we expect the finite-width network to outperform its infinite-width counterpart whenever $s_L < 1$. In panel (a) and (c) of Fig. 2 we check this prediction for deep architectures with ReLU activation on the same regression tasks employed in the 1HL case. Notice that s_L

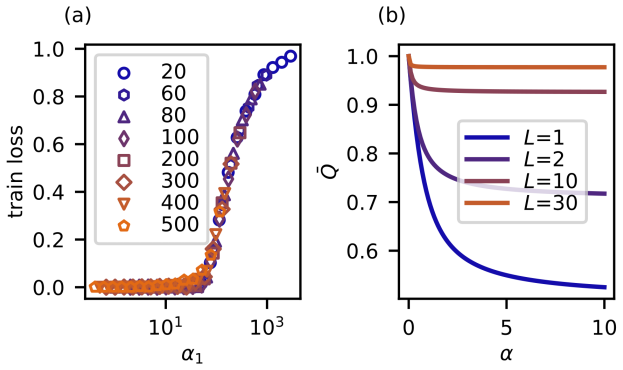

FIG. 3. (Left panel) Training loss of different one-hidden layer architectures trained on a completely random task (i.e. both the inputs $\mathbf{x} \in \mathbb{R}^{N_0}$ with $N_0 = 50$ and the scalar outputs y are i.i.d. random variables sampled from a normal distribution with zero mean and unit variance) as a function of α_1. At the moment we can not capture this universal phenomenon with our theory, which only describes the overparametrised limit where the training error is exactly zero. (Right panel) The numerical evaluation of the solution $\bar{Q}$ is shown in the case of ReLU activation function and isotropic network $\alpha_\ell = \alpha \ \forall \ell$, for different depths L. As L grows ($L \sim 30$), the parameter $\bar{Q}$ quickly approaches 1 for all α, suggesting that also DNNs in the asymptotic regime converge to a kernel limit in the sequential limit where the depth L is taken to infinity after P, N.

quickly diverges to infinity as the number of hidden layers L grows. The reason for this is simply that the ReLU NNGP kernel K_L develops at least one zero eigenvalue as $L \to \infty$. This ultimately occurs because each element of the matrix K_L converges to zero as L grows (see panel (b) and (c) of Fig. 2), as one can easily check by looking at the explicit recurrence relation for the NNGP ReLU kernel. [15, 66]. We note that this singularity can be equivalently thought as the fixed point of the discrete dynamical map defined by the recurrence relation for the NNGP kernel and therefore it might be worth investigating the relation between the generalisation performance in our asymptotic limit and the line of work on the edge of chaos in random neural networks [67, 68].

Equation (19) provides an additional link with Student's t inference. In fact, the same criterion has been found by Tracey and Wolpert [69] in the study of Bayesian optimization with Student's t-processes. Here the authors show that the value of s_L determines whether the Student's t-process they consider has a larger/smaller variance than the corresponding GP with the same kernel.

III. DISCUSSION

In our work we have described a strategy to investigate the statistical mechanics of deep neural networks beyond the infinite-width limit, that is in the finite asymptotic regime $P, N_\ell \to \infty$ at $\alpha_\ell = P/N_\ell > 0$ as opposed to the infinite-width $\alpha_\ell = 0$. In the 1HL case, we conjecture that our evaluation is exact in the above thermodynamic limit. As such, we do not expect any additional corrections to the result, at least in the asymptotic regime. In particular, we have found a closed expression for the generalisation error that in principle provides a Bayesian estimator of the generalisation capabilities of fully-connected architectures for any given empirical dataset, provided that the chosen architecture is capable of perfectly fitting the trainset.

For the case of finite depth $L > 1$ networks, it should be possible, at least in principle, to take systematically into account non-Gaussian corrections to the saddle-point action to check whether these are relevant or not for the theory at finite width, since the assumptions we made in deriving the results are clear [70] (see also Methods).

From the mathematical perspective, we find the link with Student's t-processes very promising. The precise characterization of this mapping and its limits of validity represent a research line for future investigation.

Notably, our theory predicts that a kernel limit should also appear in the asymptotic regime as the depth L approaches infinity. This could be checked, for instance, considering the isotropic limit $\alpha_\ell = \alpha \ \forall \ell$ and ReLU activation. Here one can numerically solve the saddle-point equation for $\bar{Q}$ at large L and verify that $\bar{Q} \to 1$ for all α, as shown in panel (b) Fig. 3. As such, also in this limit we expect an equivalence with a kernel theory with kernel given by $K_\infty(C)$. Note that from our framework it is clear that we are taking the depth L infinity only after P, N. As such we are not making claims about the challenging simultaneous limit $L, N \to \infty$ at fixed L/N, as done for instance in Refs. [27, 28, 58, 70].

It is fair to stress that our theory only describes the equilibrium regime of zero train loss, so that our analysis should not apply in the regime $P/N_1 \gg 1$. Interestingly, numerical simulations performed with 1HL architectures of varying width and random training labels show that the train loss follows a universal behavior [54, 71] w.r.t. α_1 (see Fig. 3, panel (a)) also in the regime where the DNN is not capable to perfectly fit the data. It would be desirable to develop a theory that also captures this phase.

Another interesting aspect to understand is the degree to which this mean-field static analysis can be extended beyond equilibrium in order to assess the full training dynamics; such a theory would indeed make it possible to investigate the performance of the many (often heuristics) learning algorithms employed to train deep neural networks.

We conclude by pointing out that it would be interesting to compare our theory at fixed data with the data-averaged cases studied in [54, 72] and to extend our results to convolutional layers, as done in the infinite-width case in Ref. [17].

METHODS

A. Setting of the learning problem and notation

We consider deep neural networks $f_{\rm DNN}(\mathbf{x})$ with L fully-connected hidden layers, where the pre-activations of each layer $h_{i_\ell}^{(\ell)}$ ($i_\ell = 1, \dots, N_\ell$; $\ell = 1, \dots, L$) are given recursively as a non-linear function of the pre-activations at the previous layer $h_{i_{\ell-1}}^{(\ell-1)}$ ($i_{\ell-1} = 1, \dots, N_{\ell-1}$):

$$h_{i_\ell}^{(\ell)} = \frac{1}{\sqrt{N_{\ell-1}}} \sum_{i_{\ell-1}=1}^{N_{\ell-1}} W_{i_\ell i_{\ell-1}}^{(\ell)} \sigma\left(h_{i_{\ell-1}}^{(\ell-1)}\right) + b_{i_\ell}^{(\ell)} , \qquad (21)$$

$$h_{i_1}^{(1)} = \frac{1}{\sqrt{N_0}} \sum_{i_0=1}^{N_0} W_{i_1 i_0}^{(1)} x_{i_0} + b_{i_1}^{(1)} \qquad (22)$$

where $W^{(\ell)}$ and $b^{(\ell)}$ are respectively the weights and the biases of the ℓ-th layer, whereas the input layer has dimension N_0 (the input data dimension). σ is a non-linear activation function and it is common to each layer. We add one last readout layer and we define the function implemented by the deep neural network as:

$$f_{\rm DNN}(\mathbf{x}) = \frac{1}{\sqrt{N_L}} \sum_{i_L=1}^{N_L} v_{i_L} \sigma\left[h_{i_L}^{(L)}(\mathbf{x})\right] , \qquad (23)$$

where $\mathbf{v}$ is the vector of weights of the last layer.

The average training error at a given inverse temperature β is given by:

$$\langle \epsilon_{\rm t} \rangle = \frac{1}{P} \int \mathcal{D}\theta \, [\mathcal{L}(\theta) - \mathcal{L}_{\rm reg}(\theta)] \frac{e^{-\beta \mathcal{L}(\theta)}}{Z} , \qquad (24)$$

Training and test errors (as defined in equation (5)) represent two special observables, but more in general, for an arbitrary observable O we have:

$$\langle O \rangle = \int \mathcal{D}\theta \, O(\theta) \frac{e^{-\beta \mathcal{L}(\theta)}}{Z} . \qquad (25)$$

B. The Breuer-Major theorem as a justification for the Gaussian equivalence in shallow networks

The Breuer-Major theorem and its extensions deal with the following sequence of random variables:

$$S_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} c_i F(x_i) \quad N \geq 1 . \qquad (26)$$

Clearly, if the distribution of the vector $\mathbf{x} = (x_1, \dots, x_N)$ is factorized over its coordinates, i.e. $p(\mathbf{x}) = \prod_i p(x_i)$ and $F(x) = x$, the random variable $S = \lim_{N\to\infty} S_N$ is normal distributed as long as the mean $\mathbb{E}(x_i) = 0$, the variance $\mathbb{E}(x_i^2)$ is finite and the c_i's satisfy the so-called Lindeberg's condition. This is also true whenever F is a well-behaved non-linearity.

The Breuer-Major theorem essentially extends this result to generic GPs, providing sufficient conditions on the covariance matrix of the GP and on the non-linearity F that guarantee convergence of S_N to the normal distribution. We report here the modern statement of the theorem given in Ref. [44].

We first consider a stationary (unidimensional) GP $x = (x_k)_{k\in\mathbb{Z}}$. Stationarity –which is not essential and will be replaced by a weaker condition in the following– amounts to require that the covariance of the process $C_{ij} = \mathbb{E}(x_i x_j)$ is a function of the difference $i - j$, i.e. $C_{ij} = C(i - j)$. The only technical condition to be imposed on the non-linear function F is to have well-defined *Hermite rank* R. The Hermite rank is the smallest positive integer that appears in the decomposition of F over the Hermite polynomials:

$$F(x) = \sum_{k=R}^{\infty} f_k \, {\rm He}_k(x) , \qquad (27)$$

where ${\rm He}_k(x)$ is the k-th Hermite polynomial and f_k the coefficient of the expansion. For many reasonable activation functions F, $R = 1$.

Theorem 1 (Breuer and Major, 1983) *Let $x = (x_k)_{k\in\mathbb{Z}}$ be a stationary unidimensional GP with covariance $C(i - j)$. Let $\mathbb{E}[F(x_1)] = 0$ and $\mathbb{E}\left[F^2(x_1)\right] < \infty$ and assume that the function F has Hermite rank $R \geq 1$. Suppose that:*

$$\sum_{j\in\mathbb{Z}} |C_{1j}|^R < \infty . \qquad (28)$$

Then $\sigma^2 := \mathbb{E}\left[F(x_1)^2\right] + 2\sum_{j=1}^{\infty} \mathbb{E}\left[F(x_1)F(x_j)\right]$ is finite. Moreover, one has that the sequence of random variables

$$S_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} F(x_i) \quad N \geq 1 \qquad (29)$$

converges in distribution to $\mathcal{N}(0, \sigma^2)$, i.e. to a Gaussian distribution with zero mean and variance σ^2.

For our scopes we will need a slightly stronger statement than the one just mentioned: (i) in our calculation the covariance will not be stationary and (ii) we will need to consider a more general sequence of nonlinear functions $c_i F(x_i)$, such that each term of the sum (29) is weighted by a factor $c_i \neq 1$.

It has been shown, already in the original reference [45], that the hypothesis of stationarity can be weakened and replaced with a requirement of uniform convergence of the elements of the covariance, namely:

$$\sum_{j\in\mathbb{Z}} |C_{ij}|^R < B_0 \quad \forall i \in \mathbb{Z} , \qquad (30)$$

where B_0 is a positive finite constant. Extensions (i) and (ii) have been addressed more recently by Bardet and

Surgailis in [43], as we report in the following. Let $\mathbf{x}^N$ be an N-dimensional Gaussian vector, such that $\mathbb{E}[x_i^N] = 0$, $\mathbb{E}[(x_i^N)^2] = 1$. Now define $C_{ij}^N = \mathbb{E}[x_i^N x_j^N]$. For a given integer $m \geq 1$, assume

$$\sup_{N\geq 1} \max_{1\leq j\leq N} \sum_{i=1}^{N} |C_{ij}^N|^m < \infty\,, \tag{31}$$

$$\sup_{N\geq 1} \frac{1}{N} \sum_{\substack{1\leq i,j\leq N \\ |i-j|>K}} |C_{ij}^N|^m \underset{K\to\infty}{\longrightarrow} 0\,. \tag{32}$$

Take also $\mathbb{L}_0^2(x) = \left\{f : \mathbb{E}\, f(x) = 0, \mathbb{E}\, f^2(x) < \infty\right\}$, where x is a standard normal variable. Then

Theorem 2 (Bardet and Surgailis [43], 1.ii) *Assume (31), (32). Let $f_i^N \in \mathbb{L}_0^2(x)$ $(N \geq 1,\ 1 \leq i \leq N)$ be a sequence of functions all having Hermite rank m at least one. Assume that there exist a $\mathbb{L}_0^2(x)$-valued continuous function ϕ_τ, $\tau \in [0,1]$, such that*

$$\sup_{\tau\in(0,1]} \mathbb{E}[f_{[\tau N]}^N(x) - \phi_\tau(x)]^2 \underset{N\to\infty}{\longrightarrow} 0\,. \tag{33}$$

Moreover, let

$$(\sigma^N)^2 = \mathbb{E}\left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} f_i^N(x_i^N)\right]^2 \underset{N\to\infty}{\longrightarrow} \sigma^2\,, \tag{34}$$

where $\sigma^2 > 0$. Then

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} f_i^N(x_i^N) \underset{N\to\infty}{\overset{d}{\longrightarrow}} \mathcal{N}(0,\sigma^2)\,, \tag{35}$$

where $\overset{d}{\underset{N\to\infty}{\longrightarrow}}$ denotes convergence in distribution.

The hypotheses of this theorem should be taken as conditions on the activation function σ, on the rescaled input covariance matrix $C_{\mu\nu}$ and on the dominant configurations $\bar{s}$ in the Fourier integral (10) in order to justify our Gaussian ansatz (see below, Eq. (42)).

C. Sketch of the calculation of the effective action in the Bayesian setup for one-hidden layer fully-connected neural networks

We now discuss the salient aspects of the calculation. The starting point is the following partition function:

$$Z = \int \prod_{i_1}^{N_1} \mathrm{d}v_{i_1} \prod_{i_1,i_0}^{N_1,N_0} \mathrm{d}w_{i_1 i_0} \exp\left\{-\frac{\lambda_1}{2}\sum_{i_1} v_{i_1}^2 - \frac{\lambda_0}{2}\|w\|^2 - \frac{\beta}{2}\sum_{\mu}^{P}\left[y^\mu - \frac{1}{\sqrt{N_1}}\sum_{i_1} v_{i_1}\sigma\left(\sum_{i_0}\frac{w_{i_1,i_0}x_{i_0}^\mu}{\sqrt{N_0}}\right)\right]^2\right\}. \tag{36}$$

where $w = W^{(1)}$ and we took $b^{(1)} = 0$ without loss generality [1]. The first step is to decouple the weights of the different layers in the loss function. This can be done including standard identities built over two families of Dirac deltas, one for the pre-activations of the hidden layer and one for the output of the network:

$$1 = \int \prod_{\mu}^{P} ds^\mu \delta\left[s^\mu - \frac{1}{\sqrt{N_1}}\sum_{i}^{N_1} v_{i_1}\sigma(h_{i_1}^\mu)\right]\,, \tag{37}$$

$$1 = \int \prod_{\mu}^{P}\prod_{i_1}^{N_1} dh_{i_1}^\mu \delta\left(h_{i_1}^\mu - \frac{1}{\sqrt{N_0}}\sum_{i_0}^{N_0} w_{i_1 i_0}x_{i_0}^\mu\right)\,. \tag{38}$$

By using a standard Fourier representation of these deltas, which introduces the conjugate variables $\bar{h}_{i_1}^\mu$ and $\bar{s}^\mu$, we can perform the gaussian integrals on the internal and external weights:

$$Z = \int \prod_{\mu}^{P} \frac{\mathrm{d}s^\mu\,\mathrm{d}\bar{s}^\mu}{2\pi} e^{-\frac{\beta}{2}\sum_\mu (y^\mu - s^\mu)^2 + i\sum_\mu^P s^\mu \bar{s}^\mu} \times \left\{\int \prod_{\mu}^{P} \frac{\mathrm{d}h^\mu\,\mathrm{d}\bar{h}^\mu}{2\pi} e^{i\sum_\mu^P h^\mu \bar{h}^\mu - \frac{1}{2\lambda_1 N_1}\left[\sum_\mu^P \bar{s}^\mu \sigma(h^\mu)\right]^2} \times e^{-\frac{1}{2\lambda_0 N_0}\sum_{i_0}^{N_0}\left(\sum_\mu^P \bar{h}^\mu x_{i_0}^\mu\right)^2}\right\}^{N_1}, \tag{39}$$

where we used the fact that the integrals on $h_{i_1}^\mu$ and $\bar{h}_{i_1}^\mu$ can be factorized on the index i_1. The integral over the $\bar{h}^\mu$ is Gaussian and can be solved:

$$\int \prod_{\mu}^{P} \frac{\mathrm{d}\bar{h}^\mu}{2\pi} e^{i\sum_\mu^P h^\mu \bar{h}^\mu - \frac{1}{2\lambda_0 N_0}\sum_{i_0}^{N_0}\left(\sum_\mu^P \bar{h}^\mu x_{i_0}^\mu\right)^2} = P_1(\{h^\mu\})\,, \tag{40}$$

where

$$P_1(\{h^\mu\}) = \frac{e^{-\frac{1}{2}\sum_{\mu,\nu}^P h^\mu C_{\mu,\nu}^{-1} h^\mu}}{\sqrt{(2\pi)^P \det C}}\,,\quad C_{\mu\nu} = \frac{1}{\lambda_0 N_0}\sum_{i_0}^{N_0} x_{i_0}^\mu x_{i_0}^\nu\,. \tag{41}$$

This last step requires the covariance matrix C to be invertible. Note that this is false as soon as $P > N_0$, but adding a small diagonal term to C solves the issue. One can explicitly check that the final result does not depend on this extra regularization.

To deal with the integral over h^μ we can include a further Dirac delta identity for the random variable $q = 1/\sqrt{\lambda_1 N_1}\sum_\mu \bar{s}^\mu \sigma\left(h^\mu\right)$. This leaves us with the problem of finding the probability density $P(q)$. In the limit defined in (1), this is exactly the same setting of the

[1] One can map a system with non-zero biases in a zero-bias one increasing by one the dimensions of the input and of the activations at each layer. The original biases are then trivially mapped in the extra weights of the augmented system.

Breuer-Major theorems [43–45]. As such, it is sufficient that both the (regularized) covariance C and the activation function σ satisfy the hypotheses of the theorem to guarantee that the probability distribution $P(q)$ converges in distribution to a Gaussian:

$$P(q) = \int \mathrm{d}^P h\, P_1(\{h^\mu\})\delta\left[q - \frac{1}{\sqrt{\lambda_1 N_1}}\sum_\mu \bar{s}^\mu \sigma\left(h^\mu\right)\right],$$
$$P(q) \to \mathcal{N}_q(0, Q)\,. \tag{42}$$

with variance

$$Q(\bar{s}, C) = \frac{1}{\lambda_1 N_1}\sum_{\mu,\nu}^{P} \bar{s}^\mu \left[\int \mathrm{d}^P h\, P_1(\{h^\rho\})\sigma(h^\mu)\sigma(h^\nu)\right]\bar{s}^\nu$$
$$= \frac{1}{\lambda_1 N_1}\sum_{\mu,\nu}^{P} \bar{s}^\mu K_{\mu\nu}(C)\bar{s}^\nu\,. \tag{43}$$

One can show that there exist special configurations $\bar{s}$ in the domain of integration for which we are not allowed to invoke a Gaussian equivalence (see for instance our discussion at the end of Sec. II in the supplemental material [57]). In our derivation, we are assuming that the contribution of these special configurations to the effective action is negligible in the thermodynamic limit. Here we have also assumed that the variable q has zero mean, a condition verified as long as

$$\int \mathrm{d}^P h\, P_1(\{h^\mu\})\sigma(h^\nu) = 0\,, \tag{44}$$

that is whenever σ is zero-mean; for a more general derivation, relevant for finite-mean activation functions such as ReLU, see the supplemental material [57], Sec. IV.

Each element of the kernel matrix $K_{\mu\nu}(C)$ can be easily reduced from a P-dimensional integral to a simpler two-dimensional one:

$$K_{\mu\nu}(C) = \int \frac{\mathrm{d}t_1\,\mathrm{d}t_2}{\sqrt{(2\pi)^2 \det \tilde{C}}} e^{-\frac{1}{2}\mathbf{t}^T\tilde{C}^{-1}\mathbf{t}}\sigma(t_1)\sigma(t_2)\,, \tag{45}$$

where $\mathbf{t} = (t_1, t_2)^T$ and

$$\tilde{C} = \begin{pmatrix} C_{\mu\mu} & C_{\mu\nu} \\ C_{\mu\nu} & C_{\nu\nu}\end{pmatrix}. \tag{46}$$

is the reduced 2×2 input covariance matrix. It is worth pointing out that the kernel we find here is the so-called *neural network Gaussian process* (NNGP) kernel. It differs from the neural tangent kernel (NTK) that is found in the infinite-width limit of networks trained under gradient descent [73]. The fact that the infinite-width limit of a Bayesian neural network differs from the one obtained from gradient descent is indeed known and discussed in literature [20].

Now we can integrate over the variable q and obtain:

$$\left[\int \frac{\mathrm{d}q\, e^{-\frac{q^2}{2} - \frac{q^2}{2Q(\bar{s},C)}}}{\sqrt{2\pi Q(\bar{s}, C)}}\right]^{\frac{N_1}{2}} = [Q(\bar{s}, C) + 1]^{-\frac{N_1}{2}}\,. \tag{47}$$

In the general case of finite $\alpha_1 = P/N_1$, we are only left with the integrals in s^μ and $\bar{s}^\mu$. To solve them it is convenient to introduce one final Dirac delta identity:

$$1 = \int dQ\,\delta\left[Q - \frac{1}{\lambda_1 N_1}\sum_{\mu,\nu}\bar{s}^\mu K(C)_{\mu\nu}\bar{s}^\nu\right], \tag{48}$$

where $Q \geq -1$ is now an integration variable and not a function of $\bar{s}$, so that we have removed the explicit dependence on $\sqrt{Q(\bar{s}, C) + 1}$ in the partition function. Finally, the integrals in s^μ and $\bar{s}^\mu$ are Gaussian once another integral representation of the delta via a conjugate variable $\bar{Q}$ is inserted. This allows us to get the final effective action obtained in equation (7).

D. Exact solution of the saddle-point equations in the zero temperature limit

The saddle-point equations obtained from (7) considerably simplify in the zero temperature limit ($\beta \to \infty$). In particular, using the fact that the kernel K has only positive eigenvalues (in the asymptotic regime α_1, α_0 finite), we get:

$$\bar{Q} = \frac{1}{1+Q}\,, \qquad Q = +\frac{\alpha_1}{\bar{Q}} - \frac{\alpha_1}{\bar{Q}^2}\frac{1}{P}y^T\left(\frac{K}{\lambda_1}\right)^{-1} y\,. \tag{49}$$

Given the condition $Q \geq -1$, the unique exact solution for $\bar{Q}$ is positive and reads:

$$\bar{Q}^* = \frac{\sqrt{(\alpha_1 - 1)^2 + 4\alpha_1\frac{1}{P}y^T\left(\frac{K}{\lambda_1}\right)^{-1} y} - (\alpha_1 - 1)}{2}\,. \tag{50}$$

E. Predictors statistics

The main observable we are interested in is the generalisation error (5). We can proceed along the same lines of the calculation performed in Sec. III C introducing, other than the variables s^μ, h_i^μ defined by (37), (38), additional variables s^0, h_i^0 that describe output and pre-activations

of the new test example. We thus get:

$$\langle \epsilon_{\rm g}(\mathbf{x}^0, y^0)\rangle = \frac{1}{Z}\int \frac{\mathrm{d}s^0\,\mathrm{d}\bar{s}^0}{2\pi}\int\prod_{\mu=1}^{P}\frac{\mathrm{d}s^\mu\,\mathrm{d}\bar{s}^\mu}{2\pi}\,(y^0-s^0)^2 \times e^{-\frac{\beta}{2}\sum_{\mu=1}^{P}(y^\mu-s^\mu)^2+i\sum_{\mu=1}^{P}s^\mu\bar{s}^\mu+is^0\bar{s}^0} \times\left[1+\frac{1}{\lambda_1 N_1}\left(\sum_{\mu,\nu=1}^{P}\bar{s}^\mu K_{\mu\nu}\bar{s}^\nu + 2\bar{s}^0\sum_{\mu=1}^{P}\bar{s}^\mu\kappa_\mu(\mathbf{x}^0)+(\bar{s}^0)^2\kappa_0(\mathbf{x}^0)\right)\right]^{-\frac{N_1}{2}}, \tag{51}$$

where κ_μ and κ_0 are respectively the train-test and the test-test kernel integrals defined as in (43) when the covariance matrix involves the test input, namely:

$$\kappa_\mu = \int \frac{dt_1 dt_2}{\sqrt{(2\pi)^2\det\tilde{C}_\mu}} e^{-\frac{1}{2}\mathbf{t}^T\tilde{C}_\mu^{-1}\mathbf{t}}\sigma(t_1)\sigma(t_2)\,, \tag{52}$$

$$\kappa_0 = \int \frac{dt}{\sqrt{2\pi C_{00}}} e^{-\frac{t^2}{2C_{00}}}\sigma(t)^2\,, \tag{53}$$

where

$$\tilde{C}_\mu = \begin{pmatrix} C_{\mu\mu} & C_{\mu 0}\\ C_{\mu 0} & C_{00}\end{pmatrix}\,,\quad C_{\mu 0} = \frac{1}{\lambda_0 N_0}\sum_{i_0}^{N_0} x^\mu_{i_0}x^0_{i_0}\,, \qquad C_{00} = \frac{1}{\lambda_0 N_0}\sum_{i_0}^{N_0}(x^0_{i_0})^2\,. \tag{54}$$

Now we can introduce the order parameters Q and $\bar{Q}$ via equation (48) and their Fourier representation and perform the integration over all the $s^\mu, \bar{s}^\mu$ and over the $\bar{s}^0$. Doing so yields a single integral in s^0 and integrals on Q and $\bar{Q}$.

$$\langle \epsilon_{\rm g}(\mathbf{x}^0, y^0)\rangle = \frac{1}{Z}\int \frac{\mathrm{d}Q\,\mathrm{d}\bar{Q}}{2\pi} e^{-\frac{N_1}{2}S(Q,\bar{Q})} \times\int \frac{\mathrm{d}s^0 (y^0-s^0)^2}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s^0+\Gamma_1)^2}{2\sigma_1^2}}, \tag{55}$$

with

$$\Gamma_1 = \frac{\bar{Q}}{\lambda_1}\sum_{\mu\nu}\kappa_\mu(\mathbf{x}^0)\left(\frac{\mathbb{1}}{\beta}+\frac{\bar{Q}}{\lambda_1}K\right)^{-1}_{\mu\nu} y_\nu,$$
$$\sigma_1^2 = \frac{\bar{Q}}{\lambda_1}\left[\kappa_0(\mathbf{x}^0) - \frac{\bar{Q}}{\lambda_1}\sum_{\mu\nu}\kappa_\mu(\mathbf{x}^0)\left(\frac{\mathbb{1}}{\beta}+\frac{\bar{Q}}{\lambda_1}K\right)^{-1}_{\mu\nu}\kappa_\nu(\mathbf{x}^0)\right] \tag{56}$$

We can then unfold the easy integrals in s^0 and evaluate the result on the saddle point solution. The generalisation error is expressed in terms of Γ_1 and σ_1^2 as in equation (8). Taking the $\beta\to\infty$ limit in equations (56) yields the expressions in (8).

F. Constraints on the scaling of the size of the dataset P with the input dimension N_0

In this section we address the additional constraints to the thermodynamic scaling ($P, N_1 \to \infty$ with $\alpha_1 = P/N_1$ finite) that may come from the hypotheses of the Breuer-Major on the covariance matrix C. The only stringent condition to verify is equation (30), that is

$$\sum_{\mu=1}^{P}|C_{\mu\nu}|^R < B_0 \qquad \forall\,\nu = 1,\dots,P\,, \tag{57}$$

where B_0 is a given finite constant and R the Hermite rank of the activation function σ. In the case of inputs $\mathbf{x}$ with i.i.d. standard Gaussian coordinates, $C_{\mu\nu}$ is a Wishart random matrix with off-diagonal entries of order $1/\sqrt{N_0}$ and random signs: after taking the absolute value, the sum in Eq (57) is of order $P(N_0)^{-R/2}$. Note that this provides an infinite class of activation functions (those with Hermite rank $R \geq 2$) where we can safely work at least at finite $\alpha_0 = P/N_0$. For activation functions with Hermite rank $R = 1$ (such as Erf or ReLU) we cannot provide such a guarantee by only looking at the hypothesis of the Breuer-Major theorem. It is also worth stressing that, given any odd (non-odd) activation function $\sigma(x)$ with Hermite rank $R = 1$, it is easy to engineer a new reasonable activation function with Hermite rank $R = 3$ ($R = 2$), just by replacing the old activation function with a new one $\sigma_1(x) = \sigma(x) - g_1 x$, where the coefficient $g_1 = \langle\sigma(x)\,\mathrm{He}_1(x)\rangle$ and the average is over a normal distribution of zero mean and unit variance.

We observe that there is at least one case of activation function with $R = 1$ where the derivation goes through at finite α_0, i.e. the linear function $\sigma(x) = x$ (in this case we can obtain the result at finite P, N_1, N_0, as done also in Ref. [58]). In the supplemental material [57], Sec. II, we examine the specific case of quadratic activation $\sigma(x) = x+x^2$ (that has $R = 1$), deriving the final effective action without employing the BM theorem. As in the linear case, this derivation goes through at finite α_0. We are thus led to think that the scaling $P = O(\sqrt{N_0})$ suggested for $R = 1$ is overly-pessimistic.

G. Generalisation to deep neural networks with a finite number of hidden layers $L > 1$ and zero-mean activation

In the same spirit of the 1HL calculation, we introduce L sets of auxiliary variables $h^\mu_{i_\ell}$ (where $i_\ell = 1,\dots,N_\ell$) that are equal to the pre-activations at each layer. The strategy to perform the calculation is to show that the probability distribution of the preactivations at each layer $P_\ell(\{h^\mu_{i_\ell}\})$ can be computed recursively, starting from the input layer. We notice that this is conceptually different from the backpropagating kernel renormalisation group introduced in Ref. [5]. It is still a kernel renormalisation group, but forward-propagating, and

represents a generalisation to NNTPs of the kernel recurrence arising in NNGPs [15]. In practice, our approach amounts to a systematic, layer-by-layer description of the pre-activation statistics by the Student's t distribution that we have shown to appear in the 1HL case. This can be seen as a quantitative correction to the standard Gaussian statistics that is recovered in the infinite width limit. At the moment we are not able to re-derive the same result using the backpropagating method introduced in [5].

Let us start by integrating the weights of the first layer. This defines a probability distribution over the pre-activations of the first layer via:

$$P_1(\{h^\mu_{i_1}\}) = \int \mathcal{D}W^{(1)} \prod_{i_1,\mu} \delta\left(h^\mu_{i_1} - \frac{1}{\sqrt{N_0}} \sum_{i_0=1}^{N_0} W^{(1)}_{i_1 i_0} x^\mu_{i_0}\right) = \prod_{i_1=1}^{N_1} \frac{e^{-\frac{1}{2}\sum_{\mu\nu} h^\mu_{i_1} C^{-1}_{\mu\nu} h^\nu_{i_1}}}{\sqrt{(2\pi)^P \det C}} \,. \tag{58}$$

where C is defined in (41). This result is straightforward and it is valid for any N_0, P and N_1, since the prior for the weights is gaussian. At the second layer we have:

$$P_2(\{h^\mu_{i_2}\}) = \int \mathcal{D}W^{(2)} \mathcal{D}h_1 P_1(\{h^\mu_{i_1}\}) \times \prod_{i_2,\mu} \delta\left(h^\mu_{i_2} - \frac{1}{\sqrt{N_1}} \sum_{i_1=1}^{N_1} W^{(2)}_{i_2 i_1} \sigma(h^\mu_{i_1})\right) \tag{59}$$

We now introduce conjugate variables $\bar{h}^\mu_{i_2}$ to the activation of the second layer and the calculation proceeds as in the case of 1HL architectures. To make analytical progress we need to make two fundamental approximations: (i) assuming that the set of random variables $q_{i_2} = 1/(\sqrt{N_1}\lambda_1)\sum_\mu \bar{h}^\mu_{i_2}\sigma(h^\mu)$, where $\mathbf{h} \sim \mathcal{N}(0, C)$, is Gaussian-distributed; (ii) neglecting correlations between different pre-activations of the second hidden layer. In conclusion we get:

$$P_2(\{h^\mu_{i_2}\}) = \int \mathrm{d}Q_1 \,\mathrm{d}\bar{Q}_1 e^{-\frac{N_1}{2}(-Q_1\bar{Q}_1 + \log(1+Q_1))} \times \prod_{i_2=1}^{N_2} \frac{e^{-\frac{1}{2}\sum_{\mu\nu} h^\mu_{i_2} \left(\bar{Q}_1 K(C)/\lambda_1\right)^{-1}_{\mu\nu} h^\nu_{i_2}}}{\sqrt{(2\pi)^P \det(\bar{Q}_1 K(C)/\lambda_1)}} \,, \tag{60}$$

where $K(C)$ is defined by equation (43). Notice that except for the integration over the two variables Q_1 and $\bar{Q}_1$, this is the same as the probability distribution of the 1HL system (41) if we replace C with $\bar{Q}_1 K(C)/\lambda_1$. This reasoning can be easily iterated across layers and gives:

$$P_L(\{h^\mu_{i_L}\}) = \int \prod_{\ell=1}^{L-1} \mathrm{d}Q_\ell \,\mathrm{d}\bar{Q}_\ell e^{-\sum_{\ell=1}^{L-1} \frac{N_\ell}{2}\left[-Q_\ell\bar{Q}_\ell + \log(1+Q_\ell)\right]} \times \prod_{i_L=1}^{N_L} \frac{e^{-\frac{1}{2}\sum_{\mu\nu} h^\mu_{i_L} \left(K^{(R)}_{L-1}(\{\bar{Q}_\ell\})\right)^{-1}_{\mu\nu} h^\nu_{i_L}}}{\sqrt{(2\pi)^P \det(K^{(R)}_{L-1}(\{\bar{Q}_\ell\}))}} \,, \tag{61}$$

where $K^{(R)}_\ell(\{\bar{Q}_\ell\})$ is a renormalised kernel that satisfies the recurrence relation in equation (15).

The computation of the generalisation error over a new example $(\mathbf{x}^0, y^0)$ gives:

$$\langle \epsilon_g(\mathbf{x}^0, y^0)\rangle = (y^0 - \Gamma_L)^2 + \sigma^2_L \tag{62}$$

where Γ_L and σ^2_L are defined respectively in Eqs. (17) and (18). Note that $\kappa^{(R)}_{L\mu}$, $\kappa^{(R)}_{L0}$ are recursive kernels that generalise the train-test and test-test kernels (52)-(53). They are defined starting from equation (15) where the kernel K is now evaluated with the covariance matrix C involving train-test or test-test points. Note that L-hidden layers generalisation error is found replacing the 1HL kernel with its recursive generalisation (15).

H. Numerical experiments

1. *Network architectures*

We perform numerical experiments with deep fully-connected architectures trained on two regression tasks in computer vision. In particular we use the 0 and 1 classes of the MNIST and CIFAR10 datasets, which for the latter correspond to the labels "cars" and "planes". Examples from CIFAR10 are coarse grained to $N_0 = 28 \times 28$ pixels and converted to grayscale.

To test our theory in the zero-mean activation function case, we used the Erf function, for which the NNGP kernel can be computed analytically [74]:

$$K^{\mathrm{Erf}}_{\mu\nu}(C) = \frac{2}{\pi} \arcsin\left(\frac{2C_{\mu\nu}}{\sqrt{(1+2C_{\mu\mu})\,(1+2C_{\nu\nu})}}\right) \,. \tag{63}$$

In Fig. 2 we train networks with $\sigma = \mathrm{ReLU}$. The kernel can be computed analytically also in this case [66] and reads:

$$K^{\mathrm{ReLU}}_{\mu\nu}(C) = \sqrt{C_{\mu\mu}C_{\nu\nu}}\,\kappa\left(\frac{C_{\mu\nu}}{\sqrt{C_{\mu\mu}C_{\nu\nu}}}\right) \,, \tag{64}$$

$$\kappa(x) = \frac{1}{2\pi}\left[x(\pi - \arccos(x)) + \sqrt{1-x^2}\right] \,.$$

2. *Sampling from the Bayesian posterior*

To ensure convergence of the posterior weights distribution to the Gibbs ensemble, we train our networks using a discretised Langevin dynamics, similarly to what is done in [3, 5]. At each training step t the parameters $\theta = \{W^\ell, v\}$ are updated according to:

$$\theta(t+1) = \theta(t) - \eta\nabla_\theta \mathcal{L}(\theta(t)) + \sqrt{2T\eta}\epsilon(t) \tag{65}$$

where $T = 1/\beta$ is the temperature, η is the learning rate, $\epsilon(t)$ is a white Gaussian noise vector with entries

drawn from a standard normal distribution, and the loss is the one defined in equation (3). We employ $T = \eta = 10^{-3}$ throughout all the experiments. This is sufficient to approximate the $T = 0$ dynamics in the regime we are considering. This dynamics requires $10^5/10^6$ steps to reach thermalisation, depending on the sizes of the dataset and network. We extract the generalisation loss within a single run: after the train error has reached its minimum and the test loss is thermalised, we average test loss values every $10^3/10^4$ epochs (depending again on the magnitude of P, N_ℓ). For the sake of completeness, we report the best test accuracy achieved on both datasets by 1HL architectures: 0.86 on CIFAR10 with $P = 3000$ and $\lambda_1 = 1000$, 0.999 on MNIST with the same Gaussian prior and $P = 1000$. The train accuracy is always 1. Additional comments on the technical issues encountered in simulating the Bayesian dynamics are discussed in the Supplemental material [57] in Sec. V.

DATA AVAILABILITY

The CIFAR10 and MNIST datasets that we used for all our experiments are publicly available online, respectively at https://www.cs.toronto.edu/~kriz/cifar.html and http://yann.lecun.com/exdb/mnist/.

CODE AVAILABILITY

The code used to perform experiments, compute theory predictions and analyze data is available at: https://github.com/rpacelli/FC_deep_bayesian_networks [75].

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016).
[2] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, 2001).
[3] I. Seroussi, G. Naveh, and Z. Ringel, Separation of scales and a thermodynamic description of feature learning in some cnns, Nature Communications **14**, 908 (2023).
[4] A. J. Wakhloo, T. J. Sussman, and S. Chung, Linear classification of neural manifolds with correlated variability, Phys. Rev. Lett. **131**, 027301 (2023).
[5] Q. Li and H. Sompolinsky, Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization, Phys. Rev. X **11**, 031059 (2021).
[6] C. Baldassi, C. Lauditi, E. M. Malatesta, R. Pacelli, G. Perugini, and R. Zecchina, Learning through atypical phase transitions in overparameterized neural networks, Phys. Rev. E **106**, 014116 (2022).
[7] A. Canatar, B. Bordelon, and C. Pehlevan, Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks, Nature communications **12**, 1 (2021).
[8] A. Mozeika, B. Li, and D. Saad, Space of functions computed by deep-layered machines, Phys. Rev. Lett. **125**, 168301 (2020).
[9] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, Modeling the influence of data structure on learning in neural networks: The hidden manifold model, Phys. Rev. X **10**, 041044 (2020).
[10] 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**, 501 (2020).
[11] B. Li and D. Saad, Exploring the function space of deep-learning machines, Phys. Rev. Lett. **120**, 248301 (2018).
[12] R. M. Neal, Priors for infinite networks, in *Bayesian Learning for Neural Networks* (Springer New York, New York, NY, 1996) pp. 29–53.
[13] C. Williams, Computing with infinite networks, in *Advances in Neural Information Processing Systems*, Vol. 9, edited by M. Mozer, M. Jordan, and T. Petsche (MIT Press, 1996).
[14] A. G. de G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani, Gaussian process behaviour in wide deep neural networks, in *International Conference on Learning Representations* (2018).
[15] J. Lee, J. Sohl-dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri, Deep neural networks as gaussian processes, in *International Conference on Learning Representations* (2018).
[16] A. Garriga-Alonso, C. E. Rasmussen, and L. Aitchison, Deep convolutional networks as shallow gaussian processes, in *International Conference on Learning Representations* (2019).
[17] R. Novak, L. Xiao, Y. Bahri, J. Lee, G. Yang, D. A. Abolafia, J. Pennington, and J. Sohl-dickstein, Bayesian deep convolutional networks with many channels are gaussian processes, in *International Conference on Learning Representations* (2019).
[18] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Advances in Neural Information Processing Systems*, Vol. 31, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018).
[19] L. Chizat, E. Oyallon, and F. Bach, On lazy training in differentiable programming, in *Advances in Neural Information Processing Systems*, Vol. 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019).
[20] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, Wide neural networks of any depth evolve as linear models under gradient descent, in *Advances in Neural Information Processing Systems*, Vol. 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019).
[21] C. Cortes and V. Vapnik, Support-vector networks, Machine Learning **20**, 273 (1995).
[22] B. Bordelon, A. Canatar, and C. Pehlevan, Spectrum dependent learning curves in kernel regression and wide

neural networks, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 119, edited by H. D. III and A. Singh (PMLR, 2020) pp. 1024–1034.
[23] R. Dietrich, M. Opper, and H. Sompolinsky, Statistical mechanics of support vector networks, Phys. Rev. Lett. **82**, 2975 (1999).
[24] M. Seleznova and G. Kutyniok, Neural tangent kernel beyond the infinite-width limit: Effects of depth and initialization, arXiv preprint arXiv:2206.00553 (2022).
[25] N. Vyas, Y. Bansal, and N. Preetum, Limitations of the ntk for understanding generalization in deep learning, arXiv preprint arXiv:2206.10012 (2022).
[26] J. M. Antognini, Finite size corrections for neural network gaussian processes (2019), arXiv:1908.10030 [cs.LG].
[27] S. Yaida, Non-Gaussian processes and neural networks at finite widths, in *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, Proceedings of Machine Learning Research, Vol. 107, edited by J. Lu and R. Ward (PMLR, 2020) pp. 165–192.
[28] B. Hanin, Random fully connected neural networks as perturbatively solvable hierarchies (2023), arXiv:2204.01058 [math.PR].
[29] J. Zavatone-Veth and C. Pehlevan, Exact marginal prior distributions of finite bayesian neural networks, in *Advances in Neural Information Processing Systems*, Vol. 34, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021) pp. 3364–3375.
[30] Y. Bengio and O. Delalleau, On the expressive power of deep architectures, in *International conference on algorithmic learning theory* (Springer, 2011) pp. 18–36.
[31] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks, The Journal of Machine Learning Research **20**, 2285 (2019).
[32] P. Rotondo, M. C. Lagomarsino, and M. Gherardi, Counting the learnable functions of geometrically structured data, Phys. Rev. Research **2**, 023169 (2020).
[33] P. Rotondo, M. Pastore, and M. Gherardi, Beyond the storage capacity: Data-driven satisfiability transition, Phys. Rev. Lett. **125**, 120601 (2020).
[34] M. Pastore, P. Rotondo, V. Erba, and M. Gherardi, Statistical learning theory of structured data, Phys. Rev. E **102**, 032119 (2020).
[35] M. Gherardi, Solvable model for the linear separability of structured data, Entropy **23**, 10.3390/e23030305 (2021).
[36] M. Pastore, Critical properties of the SAT/UNSAT transitions in the classification problem of structured data, Journal of Statistical Mechanics: Theory and Experiment **2021**, 113301 (2021).
[37] F. Aguirre-López, M. Pastore, and S. Franz, Satisfiability transition in asymmetric neural networks, Journal of Physics A: Mathematical and Theoretical **55**, 305001 (2022).
[38] A. M. Saxe, J. L. McClelland, and S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, edited by Y. Bengio and Y. LeCun (2014).
[39] A. M. Saxe, J. L. McClelland, and S. Ganguli, A mathematical theory of semantic development in deep neural networks, Proceedings of the National Academy of Sciences **116**, 11537 (2019), https://www.pnas.org/doi/pdf/10.1073/pnas.1820226116.
[40] J. Zavatone-Veth, A. Canatar, B. Ruben, and C. Pehlevan, Asymptotics of representation learning in finite bayesian neural networks, in *Advances in Neural Information Processing Systems*, Vol. 34, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021) pp. 24765–24777.
[41] G. Naveh and Z. Ringel, A self consistent theory of gaussian processes captures feature learning effects in finite cnns, in *Advances in Neural Information Processing Systems*, Vol. 34, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021) pp. 21352–21364.
[42] J. A. Zavatone-Veth, W. L. Tong, and C. Pehlevan, Contrasting random and learned features in deep bayesian linear regression, Phys. Rev. E **105**, 064118 (2022).
[43] J.-M. Bardet and D. Surgailis, Moment bounds and central limit theorems for gaussian subordinated arrays, Journal of Multivariate Analysis **114**, 457 (2013).
[44] I. Nourdin, G. Peccati, and M. Podolskij, Quantitative Breuer-Major theorems (2010).
[45] P. Breuer and P. Major, Central limit theorems for nonlinear functionals of gaussian fields, Journal of Multivariate Analysis **13**, 425 (1983).
[46] F. Gerace, B. Loureiro, F. Krzakala, M. Mezard, and L. Zdeborova, Generalisation error in learning with random features and the hidden manifold model, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 119, edited by H. D. III and A. Singh (PMLR, 2020) pp. 3452–3462.
[47] B. Loureiro, C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mezard, and L. Zdeborová, Learning curves of generic features maps for realistic datasets with a teacher-student model, Advances in Neural Information Processing Systems **34** (2021).
[48] S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mézard, and L. Zdeborová, The gaussian equivalence of generative models for learning with shallow neural networks, arXiv preprint arXiv:2006.14709 (2020).
[49] E. Dobriban and S. Wager, High-dimensional asymptotics of prediction: ridge regression and classification, The Annals of Statistics **46**, 247 (2018).
[50] S. Mei and A. Montanari, The generalization error of random features regression: Precise asymptotics and the double descent curve, Communications on Pure and Applied Mathematics (2019).
[51] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, Linearized two-layers neural networks in high dimension, The Annals of Statistics **49**, 1029 (2021).
[52] S. Ariosto, R. Pacelli, F. Ginelli, M. Gherardi, and P. Rotondo, Universal mean-field upper bound for the generalization gap of deep neural networks, Phys. Rev. E **105**, 064309 (2022).
[53] A. Shah, A. Wilson, and Z. Ghahramani, Student-t Processes as Alternatives to Gaussian Processes, in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 33, edited by S. Kaski and J. Corander (PMLR, Reykjavik, Iceland, 2014) pp. 877–885.

[54] J. A. Zavatone-Veth, W. L. Tong, and C. Pehlevan, Contrasting random and learned features in deep bayesian linear regression, Phys. Rev. E **105**, 064118 (2022).

[55] J. Zavatone-Veth, A. Canatar, B. Ruben, and C. Pehlevan, Asymptotics of representation learning in finite bayesian neural networks, in *Advances in Neural Information Processing Systems*, Vol. 34, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021) pp. 24765–24777.

[56] J. A. Zavatone-Veth, A. Canatar, B. S. Ruben, and C. Pehlevan, Asymptotics of representation learning in finite bayesian neural networks*, Journal of Statistical Mechanics: Theory and Experiment **2022**, 114008 (2022).

[57] S. Ariosto, R. Pacelli, M. Pastore, F. Ginelli, M. Gherardi, and P. Rotondo, *Supplemental Material for "A statistical mechanics framework for Bayesian deep neural networks beyond the infinite-wdith limit"* (2023).

[58] B. Hanin and A. Zlokapa, Bayesian interpolation with deep linear networks, Proceedings of the National Academy of Sciences **120**, e2301345120 (2023).

[59] A. C. C. Coolen, M. Sheikh, A. Mozeika, F. Aguirre-López, and F. Antenucci, Replica analysis of overfitting in generalized linear regression models, Journal of Physics A: Mathematical and Theoretical **53**, 365001 (2020).

[60] A. Mozeika, M. Sheikh, F. Aguirre-López, F. Antenucci, and A. C. C. Coolen, Exact results on high-dimensional linear regression via statistical physics, Phys. Rev. E **103**, 042142 (2021).

[61] Y. Uchiyama, H. Oka, and A. Nono, Student's t-process regression on the space of probability density functions, Proceedings of the ISCIE International Symposium on Stochastic Systems Theory and its Applications **2021**, 1 (2021).

[62] H. Lee, E. Yun, H. Yang, and J. Lee, Scale mixtures of neural network gaussian processes, in *International Conference on Learning Representations* (2022).

[63] L. Aitchison, Why bigger is not always better: on finite and infinite neural networks, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 119, edited by H. D. III and A. Singh (PMLR, 2020) pp. 156–164.

[64] J. A. Zavatone-Veth and C. Pehlevan, Depth induces scale-averaging in overparameterized linear bayesian neural networks, in *2021 55th Asilomar Conference on Signals, Systems, and Computers* (2021) pp. 600–607.

[65] A. X. Yang, M. Robeyns, E. Milsom, N. Schoots, and L. Aitchison, A theory of representation learning in deep neural networks gives a deep generalisation of kernel methods (2023), arXiv:2108.13097 [stat.ML].

[66] Y. Cho and L. Saul, Kernel methods for deep learning, in *Advances in Neural Information Processing Systems*, Vol. 22, edited by Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Curran Associates, Inc., 2009).

[67] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, in *Advances in Neural Information Processing Systems*, Vol. 29, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016).

[68] G. Yang and S. Schoenholz, Mean field residual networks: On the edge of chaos, in *Advances in Neural Information Processing Systems*, Vol. 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017).

[69] B. D. Tracey and D. Wolpert, Upgrading from gaussian processes to student'st processes, in *2018 AIAA Non-Deterministic Approaches Conference* (2018) p. 1659.

[70] D. A. Roberts, S. Yaida, and B. Hanin, *The Principles of Deep Learning Theory* (Cambridge University Press, 2022) `https://deeplearningtheory.com`, arXiv:2106.10165 [cs.LG].

[71] F. Gerace, F. Krzakala, B. Loureiro, L. Stephan, and L. Zdeborová, Gaussian universality of linear classifiers with random labels in high-dimension, arXiv preprint arXiv:2205.13303 (2022).

[72] H. Cui, F. Krzakala, and L. Zdeborov'a, Optimal learning of deep random networks of extensive-width, in *International Conference on Machine Learning* (2023).

[73] J. Lee, S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein, Finite versus infinite neural networks: an empirical study, in *Advances in Neural Information Processing Systems*, Vol. 33, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., 2020) pp. 15156–15172.

[74] G. Pang, L. Yang, and G. E. Karniadakis, Neural-net-induced gaussian process regression for function approximation and pde solution, Journal of Computational Physics **384**, 270 (2019).

[75] R. Pacelli, rpacelli/FC_deep_bayesian_networks: FC_deep_bayesian_networks (2023).

[76] V. A. Marčenko and L. A. Pastur, Distribution of eigenvalues for some sets of random matrices, Mathematics of the USSR-Sbornik **1**, 457 (1967).

[77] P. Forrester, *Log-Gases and Random Matrices (LMS-34)*, London Mathematical Society Monographs (Princeton University Press, 2010).

ACKNOWLEDGEMENTS

M. P. has been supported by a grant from the Simons Foundation (grant No. 454941, S. Franz). P. R. acknowledges funding from the Fellini program under the H2020-MSCA-COFUND action, Grant Agreement No. 754496, INFN (IT). The authors would like to thank Silvio Franz, Luca Molinari, Fabian Aguirre-López, Raffaella Burioni, Alessandro Vezzani, Riccardo Aiudi, Federico Bassetti, Bruno Bassetti and the Computing Sciences group at Bocconi University in Milan for discussions and suggestions.

AUTHOR CONTRIBUTION

P.R., S.A. and M.P. performed the analytical calculations, supported by F.G, M.G. and R.P. Numerical experiments, data analysis and data visualization were carried out by R.P. All the authors contributed to discussing and interpreting the results and to writing and editing the manuscript. S.A and R.P. contributed equally to the work.

Supplemental material

I. DERIVATION OF AN EFFECTIVE ACTION FOR 1HL NEURAL NETWORKS WITH MULTIPLE OUTPUTS

In this section, we sketch the calculation to derive an effective action for 1HL neural networks with multiple outputs $\kappa > 1$. We stress that κ is finite and the case where the number of outputs scales with the width of the hidden layer N_1 has been the subject of investigations in [63–65]. We consider the following loss function:

$$\mathcal{L} = \frac{1}{2\kappa}\sum_{\mu=1}^{P}\sum_{a=1}^{\kappa}\left[y_a^\mu - (f_{\rm DNN}(\mathbf{x}^\mu))_a\right]^2 + \mathcal{L}_{\rm reg}\,, \tag{66}$$

$$\mathcal{L}_{\rm reg} = \frac{\lambda_1}{2\beta}\|v\|^2 + \frac{\lambda_0}{2\beta}\|W\|^2\,. \tag{67}$$

The partition function is defined as:

$$Z = \int\prod_{a,i_1}^{\kappa,N_1}\mathrm{d}v_{a,i_1}\prod_{i_1,i_0}^{N_1,N_0}\mathrm{d}w_{i_1i_0}\,\exp\left\{-\frac{\lambda_1}{2}\|v\|^2 - \frac{\lambda_0}{2}\|w\|^2 - \frac{\beta}{2\kappa}\sum_{\mu}^{P}\sum_{a}^{\kappa}\left[y_a^\mu - \frac{1}{\sqrt{N_1}}\sum_{i_1}^{N_1}v_{a,i_1}\sigma\left(\sum_{i_0}^{N_0}\frac{w_{i_1,i_0}x_{i_0}^\mu}{\sqrt{N_0}}\right)\right]^2\right\}. \tag{68}$$

We can decouple the layers in the loss through the addition of Dirac deltas, noticing that there will be one additional index a for the outputs.

$$\begin{aligned} Z = &\int\prod_{\mu,a}^{P,\kappa}\frac{\mathrm{d}s_a^\mu\,\mathrm{d}\bar{s}_a^\mu}{(2\pi)}\;\exp\left\{-\frac{\beta}{2\kappa}\sum_{\mu,a}^{P,\kappa}(y_a^\mu - s_a^\mu)^2 + i\sum_{\mu,a}^{P,\kappa}s_a^\mu\bar{s}_a^\mu\right\}\int\prod_{\mu,i_1}^{P,N_1}\frac{\mathrm{d}h_{i_1}^\mu\,\mathrm{d}\bar{h}_{i_1}^\mu}{(2\pi)}\exp\left\{i\sum_{\mu,i_1}^{P,N_1}h_{i_i}^\mu\bar{h}_{i_1}^\mu\right\}\\ &\int\prod_{a,i_1}^{\kappa,N_1}\mathrm{d}v_{a,i_1}\exp\left\{-\frac{\lambda_1}{2}\|v\|^2 - i\sum_{a,\mu}\bar{s}_a^\mu\sum_{i_1}\frac{v_{a,i_1}h_{i_1}^\mu}{\sqrt{N_1}}\right\}\int\prod_{i_1,i_0}^{N_1,N_0}\mathrm{d}w_{i_1,i_0}\exp\left\{-\frac{\lambda_0}{2}\|w\|^2 - i\sum_{i_1,\mu}\bar{h}_{i_1}^\mu\sum_{i_0}\frac{w_{i_1,i_0}x_{i_0}^\mu}{\sqrt{N_1}}\right\}. \end{aligned} \tag{69}$$

The integrals over the weights w_{i_1,i_0} and v_{a,i_1} are Gaussian and can be performed. As in the single-output case we can factorize the integrals in $h_{i_1}^\mu$ and $\bar{h}_{i_1}^\mu$ over the index i_1:

$$\begin{aligned} Z = &\int\prod_{\mu,a}\frac{\mathrm{d}s_a^\mu\,\mathrm{d}\bar{s}_a^\mu}{2\pi}e^{-\frac{\beta}{2\kappa}\sum_{\mu,a}(y_a^\mu - s_a^\mu)^2 + i\sum_{\mu,a}s_a^\mu\bar{s}_a^\mu}\\ &\left\{\int\prod_{\mu}^{P}\frac{\mathrm{d}h^\mu\,\mathrm{d}\bar{h}^\mu}{2\pi}e^{i\sum_{\mu}^{P}h^\mu\bar{h}^\mu - \frac{1}{2\lambda_1N_1}\sum_{a}^{\kappa}\left(\sum_\mu\bar{s}_a^\mu\sigma(h^\mu)\right)^2 - \frac{1}{2\lambda_0N_0}\sum_{i_0}^{N_0}\left(\sum_{\mu}^{P}\bar{h}^\mu x_{i_0}^\mu\right)^2}\right\}^{N_1}. \end{aligned} \tag{70}$$

Once the integrals over the variables $\bar{h}^\mu$ are performed we obtain that the h^μ are Gaussian-distributed with zero mean and covariance matrix C, in analogy with the single-output case. The critical step is to consider the joint probability distribution of the following random variables:

$$q_a = \frac{1}{\sqrt{\lambda_1 N_1}}\sum_{\mu}^{P}\bar{s}_a^\mu\sigma(h^\mu). \tag{71}$$

As in the single-output case, in the asymptotic proportional limit $P/N_1 \sim O(1)$ we can conjecture a Gaussian equivalence, based on the reasonable assumption that the BM theorem can be generalised to the multivariate case. We therefore have that $P(\{q_a\}) \to \mathcal{N}(0, Q)$ where now Q is the covariance matrix given by:

$$Q(\bar{s}, C)_{a,b} = \frac{1}{\lambda_1 N_1}\sum_{\mu,\nu}^{P}\bar{s}_a^\mu\left[\int\mathrm{d}^P h\,P_1(\{h^\rho\})\sigma(h^\mu)\sigma(h^\nu)\right]\bar{s}_b^\nu = \frac{1}{\lambda_1 N_1}\sum_{\mu,\nu}^{P}\bar{s}_a^\mu K_{\mu\nu}(C)\bar{s}_b^\nu \tag{72}$$

and K/λ_1 is the NNGP kernel, as in the single-output case. We now integrate over the set of variables $\{q_a\}$:

$$\int \prod_a^\kappa \mathrm{d}q_a \frac{1}{\sqrt{\det(Q)}} e^{-\frac{1}{2}\sum_a^\kappa (q_a)^2 - \frac{1}{2}\sum_{a,b}^\kappa q_a Q^{-1}_{a,b} q_b} = \det\left(\mathbb{1}_\kappa + \mathbf{Q}\right)^{-\frac{1}{2}} . \tag{73}$$

Differently from the single-output case, we need to introduce a $\kappa \times \kappa$ matrix order parameter $Q_{a,b}$ as:

$$1 = \int \prod_{a,b} \mathrm{d}Q_{a,b}\, \delta\left[Q_{a,b} - \frac{1}{\lambda_1 N_1} \sum_{\mu,\nu}^P \bar{s}^\mu_a K_{\mu\nu}(C) \bar{s}^\nu_b \right] \tag{74}$$

and its dual $\bar{Q}_{a,b}$ via the Fourier representation of the deltas:

$$\begin{aligned} Z = &\int \mathrm{d}\mathbf{Q}\bar{\mathbf{Q}}\ \det\left[\mathbb{1}_\kappa + \mathbf{Q}\right]^{-\frac{N_1}{2}} e^{i\sum_{a,b} Q_{a,b}\bar{Q}_{a,b}} \int \prod_{a,\mu} \mathrm{d}\bar{s}^\mu_a \exp\left\{ \frac{i}{\lambda_1 N_1} \sum_{a,b} \bar{Q}_{a,b} \sum_{\mu,\nu} \bar{s}^\mu_a K_{\mu\nu} \bar{s}^\nu_b \right\} \\ &\int \prod_{a,\mu} \mathrm{d}s^\mu_a e^{-\frac{\beta}{2\kappa}\sum_{a,\mu}(y^\mu_a - s^\mu_a)^2 + i\sum_{a,\mu} s_{a,\mu}\bar{s}_{a,\mu}} . \end{aligned} \tag{75}$$

In conclusion, the integrals in s and $\bar{s}$ can be solved and we land with the following effective action S:

$$S(\mathbf{Q},\bar{\mathbf{Q}}) = -\operatorname{Tr}[\mathbf{Q}\bar{\mathbf{Q}}^\top] + \operatorname{Tr}\log(\mathbb{1}_\kappa + \mathbf{Q}) + \frac{\alpha_1}{P}\operatorname{Tr}\log\frac{\beta}{\kappa}\left[\frac{\kappa}{\beta}\,\mathbb{1}_\kappa \otimes \mathbb{1}_P + \frac{\bar{\mathbf{Q}}\otimes K}{\lambda_1}\right] + \frac{\alpha_1}{P} y^\top \left[\frac{\kappa}{\beta}\,\mathbb{1}_\kappa \otimes \mathbb{1}_P + \frac{\bar{\mathbf{Q}}\otimes K}{\lambda_1}\right]^{-1} y. \tag{76}$$

II. 1HL EFFECTIVE ACTION AND SINGLE OUTPUT: SPECIAL CASES

In this section we report cases of activation functions for which we are able to evaluate analytically the probability distribution of the variable q, defined in the main text as

$$q = \frac{1}{\sqrt{\lambda N_1}} \sum_{\mu=1}^P \bar{s}^\mu \sigma(h^\mu) , \tag{77}$$

at fixed instance of the vector $\bar{s}$, clarifying the conditions to impose on the data for q to be Gaussian. Its characteristic function is defined as

$$\psi(t) = \mathbb{E}_q\left\{\exp(iqt)\right\} = \mathbb{E}_h\left\{ \exp\left[\frac{it}{\sqrt{\lambda N_1}} \sum_\mu \bar{s}^\mu \sigma(h^\mu) \right] \right\} . \tag{78}$$

If q is Gaussian, then $\psi = \phi$, where

$$\phi(t) = \exp\left(-\frac{t^2 Q}{2}\right) = \exp\left(-\frac{t^2}{2\lambda N_1} \sum_{\mu,\nu} \bar{s}^\mu K_{\mu\nu} \bar{s}^\nu \right) \tag{79}$$

is the characteristic function of a Gaussian variable with variance given by

$$Q = \frac{1}{\lambda N_1} \sum_{\mu,\nu} \bar{s}^\mu K_{\mu\nu} \bar{s}^\nu , \qquad K_{\mu\nu} = \mathbb{E}_h[\sigma(h^\mu)\sigma(h^\nu)] . \tag{80}$$

A. Linear activation function: q is Gaussian at finite P, N_0, N_1

The case of $\sigma = \mathrm{id}$ has been already worked out in the literature, see [5, 58]. We report it here for reference, and to stress that our theory reduces to known cases as it should. Indeed, when the activation function is linear the average over h in Eq. (78) can be computed exactly at finite P, N_1, and gives

$$\psi_{\mathrm{lin}}(t) = \exp\left(-\frac{t^2}{2\lambda N_1} \sum_{\mu,\nu} \bar{s}^\mu C_{\mu\nu} \bar{s}^\nu \right) . \tag{81}$$

Note that C is the value of the kernel for $\sigma = \mathrm{id}$. This is strictly true as long as C has no zero eigenvalue, so at least for $N_0 > P$; however, a small regularization proportional to the identity matrix can be added to C to avoid this problem.

This result is simply due to the fact that the sum of jointly Gaussian variables is Gaussian, which is true for generic Gram matrices C and any value of P, N_1, even far from the asymptotic limit $P \sim N_1$ large. In order to evaluate the remaining integrals over the order parameters at the saddle-point of the effective action, this limit is still required, as performed indeed in [5], to which our theory reduces; otherwise, for P, N_1 finite one can express the partition function exactly in terms of Meijer G-functions, see [58].

B. Quadratic activation function

Let us take now $C_{\mu\mu} = 1$ (normalized data) and quadratic (zero-mean) activation function:

$$\sigma(x) = x + a(x^2 - 1)\,. \tag{82}$$

The kernel is given by

$$K_{\mu\nu} = \mathbb{E}_h[\sigma(h^\mu)\sigma(h^\nu)] = C_{\mu\nu} + 2a^2(C_{\mu\nu})^2\,. \tag{83}$$

Also in this case the characteristic function in (78) can be evaluated exactly:

$$\psi_{\mathrm{quad}}(t) = \frac{\exp\left\{-\frac{t^2}{2\lambda N_1}\bar{s}^\top C\left[\mathbb{1}_P - \frac{2iat}{\sqrt{\lambda N_1}}\,\mathrm{diag}(\bar{s})C\right]^{-1}\bar{s}\right\}}{\det[(\mathbb{1}_P - \frac{2iat}{\sqrt{\lambda N_1}}\,\mathrm{diag}(\bar{s})C)]^{1/2}}\exp\left(-\frac{iat}{\sqrt{\lambda N_1}}\sum_\mu \bar{s}^\mu\right)\,. \tag{84}$$

We can express the non-trivial matrices appearing in this expression as Neumann series:

$$\left[\mathbb{1}_P - \frac{2iat}{\sqrt{\lambda N_1}}\,\mathrm{diag}(\bar{s})C\right]^{-1} = \sum_{n=0}^{+\infty}\left(\frac{2iat}{\sqrt{\lambda N_1}}\right)^n[\mathrm{diag}(\bar{s})C]^n\,, \tag{85}$$

$$-\frac{1}{2}\,\mathrm{Tr}\log\left[\mathbb{1}_P - \frac{2iat}{\sqrt{\lambda N_1}}\,\mathrm{diag}(\bar{s})C\right] = -\sum_{n=1}^{+\infty}\frac{1}{n}\left(\frac{2ia}{\sqrt{\lambda N_1}}\right)^n\mathrm{Tr}\{[\mathrm{diag}(\bar{s})C]^n\}\,. \tag{86}$$

To prove Gaussianity, we need to require the following asymptotic behaviors:

$$\frac{1}{N_1^{1+n/2}}\bar{s}^\top C[\mathrm{diag}(\bar{s})C]^n\bar{s} = O(P/N_1^{1+n/2})\,, \tag{87}$$

$$\frac{1}{N_1^{n/2}}\,\mathrm{Tr}\{[\mathrm{diag}(\bar{s})C]^n\} = O(P/N_1^{n/2})\,, \tag{88}$$

so that in the regime where $\alpha_1 = P/N_1$ is finite only the $n = 0$ term counts for (85) and the $n = 1, 2$ terms for (86). Using

$$-\frac{1}{2}\,\mathrm{Tr}\log\left[\mathbb{1}_P - \frac{2iat}{\sqrt{\lambda N_1}}\,\mathrm{diag}(\bar{s})C\right] \approx \frac{iat}{\sqrt{\lambda N_1}}\sum_\mu \bar{s}^\mu - \frac{a^2t^2}{\lambda N_1}\sum_{\mu,\nu}\bar{s}_\mu(C_{\mu\nu})^2\bar{s}_\nu\,, \tag{89}$$

we get

$$\psi_{\mathrm{quad}}(t) \sim \exp\left[-\frac{t^2}{2\lambda N_1}\sum_{\mu,\nu}\bar{s}^\mu K_{\mu\nu}\bar{s}^\nu\right]\,. \tag{90}$$

The conditions (87), (88) should be interpreted as hypothesis on the Gram matrix of the data C and on the realization of the vector $\bar{s}$ in order for the property of Gaussianity to hold. Let us see the simplest case of i.i.d. standard normal input data and $\bar{s}^\top = (1, \cdots, 1)$. Then, C is a Wishart matrix with a finite spectrum in the regime $P \sim N_0$ [76], and

$$\frac{1}{P}\,\mathrm{Tr}(C^n) = O(1)\,, \qquad \frac{1}{P}\sum_{\mu,\nu}(C^n)_{\mu\nu} = O(1)\,. \tag{91}$$

The first behaviour follows from the fact that the eigenvalues are $O(1)$, while the second can be proven using

$$C = O(1)\mathbb{1}_P + O(1/\sqrt{N_0})H\,, \tag{92}$$

where H is a symmetric random matrix with elements ± 1, or, more formally, exploiting the fact that the eigenvectors of a Wishart matrix are random and uniformly distributed on the sphere [77], so that

$$\frac{1}{P}\sum_{\mu,\nu}(C^n)_{\mu\nu} = \frac{1}{P}\sum_{\rho}\lambda_\rho^n\sum_{\mu}U_{\mu\rho}\sum_{\nu}U^{-1}_{\rho\nu} = O(1)\,, \tag{93}$$

where λ_ρ is the ρ-th eigenvalue of C and U the matrix whose ρ-th column is the corresponding eigenvector. Given that, properties (87), (88) follow and q is Gaussian.

In principle, Gaussianity can be also proven via diagrammatic techniques. Take for example the quartic moment of the variable q in (77). One can see, via Wick's theorem, that

$$\begin{aligned}
&\mathbb{E}_h[\sigma(h^{\mu_1})\sigma(h^{\mu_2})\sigma(h^{\mu_3})\sigma(h^{\mu_4})] - (K_{\mu_1\mu_2}K_{\mu_3\mu_4} + K_{\mu_1\mu_3}K_{\mu_2\mu_4} + K_{\mu_1\mu_4}K_{\mu_2\mu_3}) = \\
&\quad 16a^4(C_{\mu_1\mu_2}C_{\mu_1\mu_3}C_{\mu_2\mu_4}C_{\mu_3\mu_4} + C_{\mu_1\mu_2}C_{\mu_1\mu_4}C_{\mu_2\mu_3}C_{\mu_3\mu_4} + C_{\mu_1\mu_3}C_{\mu_1\mu_4}C_{\mu_2\mu_3}C_{\mu_2\mu_4}) \\
&\quad + 4a^2(C_{\mu_1\mu_2}C_{\mu_1\mu_3}C_{\mu_2\mu_4} + C_{\mu_1\mu_2}C_{\mu_1\mu_3}C_{\mu_3\mu_4} + C_{\mu_1\mu_2}C_{\mu_1\mu_4}C_{\mu_2\mu_3} + \\
&\qquad C_{\mu_1\mu_2}C_{\mu_1\mu_4}C_{\mu_3\mu_4} + C_{\mu_1\mu_2}C_{\mu_2\mu_3}C_{\mu_3\mu_4} + C_{\mu_1\mu_2}C_{\mu_2\mu_4}C_{\mu_3\mu_4} + \\
&\qquad C_{\mu_1\mu_3}C_{\mu_1\mu_4}C_{\mu_2\mu_3} + C_{\mu_1\mu_3}C_{\mu_1\mu_4}C_{\mu_2\mu_4} + C_{\mu_1\mu_3}C_{\mu_2\mu_3}C_{\mu_2\mu_4} + \\
&\qquad C_{\mu_1\mu_3}C_{\mu_2\mu_4}C_{\mu_3\mu_4} + C_{\mu_1\mu_4}C_{\mu_2\mu_3}C_{\mu_2\mu_4} + C_{\mu_1\mu_4}C_{\mu_2\mu_3}C_{\mu_3\mu_4})\,,
\end{aligned} \tag{94}$$

while the quartic term from (79) involves only the diagrams

$$\begin{aligned}
&K_{\mu_1\mu_2}K_{\mu_3\mu_4} + K_{\mu_1\mu_3}K_{\mu_2\mu_4} + K_{\mu_1\mu_4}K_{\mu_2\mu_3} = \\
&\quad 4a^4(C^2_{\mu_1\mu_2}C^2_{\mu_3\mu_4} + C^2_{\mu_1\mu_3}C^2_{\mu_2\mu_4} + C^2_{\mu_1\mu_4}C^2_{\mu_2\mu_3}) \\
&\quad + 2a^2(C^2_{\mu_1\mu_2}C_{\mu_3\mu_4} + C_{\mu_1\mu_2}C^2_{\mu_3\mu_4} + C^2_{\mu_1\mu_3}C_{\mu_2\mu_4} + C_{\mu_1\mu_3}C^2_{\mu_2\mu_4} + C^2_{\mu_1\mu_4}C_{\mu_2\mu_3} + C_{\mu_1\mu_4}C^2_{\mu_2\mu_3}) \\
&\quad + C_{\mu_1\mu_2}C_{\mu_3\mu_4} + C_{\mu_1\mu_3}C_{\mu_2\mu_4} + C_{\mu_1\mu_4}C_{\mu_2\mu_3}\,.
\end{aligned} \tag{95}$$

This is not surprising: the variables $\sigma(h^\mu)$ are not Gaussian due to the non-linearity. However, when summed over all the indices, the diagrams in Eq. (94) are of the form $\operatorname{Tr} C^4$ or $\sum_{\mu,\nu}(C^3)_{\mu\nu}$, both $O(P)$ under the hypothesis stated above, while the diagrams in (95) are of the form $(\sum_{\mu,\nu}(C^2)_{\mu\nu})^2$, $(\sum_{\mu,\nu}C_{\mu\nu})^2$ or $(\sum_{\mu,\nu}(C^2)_{\mu\nu})(\sum_{\mu,\nu}C_{\mu\nu})$, which are all $O(P^2)$ and leading over the first ones.

As long as $\bar{s}^\mu \sim O(1)$ for all μ, we do not expect the previous derivation to change. On the other hand, we point out that there exist special configurations $\bar{s}$, such as $\bar{s}^\top = (1, 0, \cdots, 0)$, for which this reasoning breaks down. As such, we are assuming that the contribution of these special configurations to the effective action is negligible in the thermodynamic limit.

III. GENERALISATION TO DEEP NEURAL NETWORKS WITH L HIDDEN LAYERS: DERIVATION OF THE SADDLE-POINT EQUATIONS IN SPECIAL CASES

In the next sections we consider two cases where simplifications arise. These special cases correspond to kernels K such that $K(\alpha C) = \alpha^s K(C)$, where α is any positive scalar and $s \geq 0$ is an integer. It turns out that $s = 0$ is realized by the sign activation function, whereas $s = 1$ holds for piece-wise linear activations such as ReLU or Leaky-ReLU.

A. Saddle-point equations for scale independent kernels of the form $K(\alpha C) = K(C)$ $(\alpha > 0)$

In the case of sign activation function, it is straightforward to show that the behavior under scalar multiplication of the kernel $K_L^{(R)}(C)$ follows from the property $\operatorname{sign}(\alpha x) = \operatorname{sign}(x)$. It turns out that in this special case the effective action for deep learning considerably simplifies, since the non-linear dependence of $K_L^{(R)}$ on the variables $\{\bar{Q}_\ell\}_{\ell \neq L}$ disappears. This allows to solve the saddle-point equations exactly. In particular:

$$Q^*_\ell = 0\,, \quad \bar{Q}^*_\ell = 1 \quad \forall\, \ell = 1, \ldots, L-1\,, \tag{96}$$

whereas the functional form of the solution for $\bar{Q}_L$ is the same as in the one-hidden layer case (in the zero temperature limit):

$$\bar{Q}_L^* = \frac{\sqrt{(\alpha_L - 1)^2 + 4\alpha_L \frac{1}{P} y^T K_L^{-1} y} - (\alpha_L - 1)}{2} \,. \tag{97}$$

In practice, such a solution shows that deep architectures with sign activation (that are problematic to employ in practice since it is challenging to backpropagate derivatives) essentially behave as one hidden layer neural networks in the proportional limit and the only marker of the depth L is retained in the infinite-width kernel K_L.

B. Saddle-point equations for piecewise linear kernels of the form $K(\alpha C) = \alpha K(C)$

The linear behavior of the kernel under scalar multiplication follows for ReLU and leaky ReLU activation function from the property $\mathrm{ReLU}(\alpha x) = \alpha\,\mathrm{ReLU}(x)$. It turns out that in this case the effective action reads:

$$\begin{aligned} S_{\mathrm{DNN}}(\{Q_\ell, \bar{Q}_\ell\}) = \sum_{\ell=1}^{L} \frac{\alpha_L}{\alpha_\ell} \left[-Q_\ell \bar{Q}_\ell + \log(1 + Q_\ell)\right] + \frac{\alpha_L}{P} \mathrm{Tr} \log \beta \left[\frac{\mathbb{1}}{\beta} + \left(\prod_{\ell=1}^{L} \bar{Q}_\ell\right) K_L(C)\right] \\ + \frac{\alpha_L}{P} y^T \left[\frac{\mathbb{1}}{\beta} + \left(\prod_{\ell=1}^{L} \bar{Q}_\ell\right) K_L(C)\right]^{-1} y \,. \end{aligned} \tag{98}$$

Exactly as for the one hidden layer case, the saddle-point equations simplify in the zero temperature limit and under the assumption that the L-hidden layers kernel K_L has only positive eigenvalues:

$$Q_\ell \bar{Q}_\ell - \alpha_\ell + \frac{\alpha_\ell}{\left(\prod_{\ell_1} \bar{Q}_{\ell_1}\right)} \frac{1}{P} y^T K_L^{-1} y = 0 \tag{99}$$

for all $\ell = 1, \ldots, L$.

Notice that if $\alpha_\ell = \alpha$ for all $\ell = 1, \ldots, L$, it is easy to show that the only solution must satisfy $Q_\ell^* = Q^*$ for all ℓ and we recover the heuristic mean field theory proposed in Ref. [5]. The reason for this equivalence is obvious: the authors of [5] found the heuristic mean field theory for ReLU activation by replacing the linear kernel with the corresponding NNGP kernel, noticing that the ReLU kernel transforms as the linear one under multiplication by a scalar. Our derivation shows that this replacement is not correct for general activation functions (see for instance the case of sign activation previously discussed), but it is possible in this particular case.

For completeness, we also show how to re-derive the self-consistent equations found by Li and Sompolinsky [5] in the linear case. The effective action for the linear case reads:

$$\begin{aligned} S_{\mathrm{DNN}}(\{Q_\ell, \bar{Q}_\ell\}) = \sum_{\ell=1}^{L} \frac{\alpha_L}{\alpha_\ell} \left[-Q_\ell \bar{Q}_\ell + \log(1 + Q_\ell)\right] + \frac{\alpha_L}{P} \mathrm{Tr} \log \beta \left[\frac{\mathbb{1}}{\beta} + \left(\prod_{\ell=1}^{L} \bar{Q}_\ell\right) C_L\right] \\ + \frac{\alpha_L}{P} y^T \left[\frac{\mathbb{1}}{\beta} + \left(\prod_{\ell=1}^{L} \bar{Q}_\ell\right) C_L\right]^{-1} y \,, \end{aligned} \tag{100}$$

where $C_L = C/(\prod_{\ell=1}^{L} \lambda_\ell)$ and $C_{\mu\nu} = \mathbf{x}^\mu \cdot \mathbf{x}^\nu / (\lambda_0 N_0)$. Let us consider the case of isotropic aspect ratios $\alpha_\ell = \alpha$, $\forall \ell = 1, \ldots, L$ and same Gaussian priors at each layer $\lambda_\ell = \lambda$, $\forall \ell = 0, \ldots, L$. The saddle-point equations for $\bar{Q}_\ell$ read:

$$1 - \bar{Q}_\ell = \alpha \left(1 - \frac{\lambda^L}{\left(\prod_{\ell_1} \bar{Q}_{\ell_1}\right)} \frac{1}{P} y^T C^{-1} y\right) \,. \tag{101}$$

It turns out that we recover the equation for the renormalization parameter u_0 in [5] by noticing that the only solution of this system of equations is of the form $\bar{Q}_\ell = \bar{Q}^*$ and by making the identification $u_0 = \bar{Q}^*/\lambda$.

IV. GENERALISING THE EFFECTIVE ACTION TO FINITE-MEAN ACTIVATION FUNCTIONS

In this section we show how the theory can be generalized in the case of finite-mean activation functions. In fact, up to this point, our derivation assumed that the integral of the activation function over a centered Gaussian is zero, i.e.

the activation function is zero-mean. The goal of this section is to show that removing such hypothesis modifies the effective action in the asymptotic limit. Since ReLU activation belongs to this more general case, the findings of this section imply that Li-Sompolinsky heuristic theory [5] should be modified as well. As for the rest of the manuscript, we start by considering one hidden layer architectures and we later extend the result to L hidden layers.

The crucial difference wrt to the case studied in section II A is that if the activation function is not zero-mean, also the random variable

$$q = \frac{1}{\sqrt{N_1 \lambda_1}} \sum_{\mu=1}^{P} \bar{s}^\mu \sigma(h^\mu) \tag{102}$$

has now a finite mean. In particular:

$$\langle q \rangle_{P(q)} = \frac{1}{\sqrt{N_1 \lambda_1}} \sum_{\mu=1}^{P} \bar{s}^\mu m^\mu \,, \qquad m^\mu = \int \frac{dt}{\sqrt{2\pi C_{\mu\mu}}} e^{-t^2/(2C_{\mu\mu})} \sigma(t) \,. \tag{103}$$

A straightforward calculation shows that the result for finite-mean activation is found by performing the replacement:

$$\frac{\bar{Q}}{\lambda_1} K \to K^{(R)}(Q, \bar{Q}) = \frac{\bar{Q}}{\lambda_1} K - \frac{\left(\bar{Q} - \frac{1}{1+Q}\right)}{\lambda_1} K^{(1)} \,, \qquad K^{(1)}_{\mu\nu} = m^\mu m^\nu \tag{104}$$

in the effective action in Eq. (7) of the main text. As such, the one-hidden layer action for finite-mean activation functions reads:

$$S_{\rm 1HL} = \, - Q\bar{Q} + \log(1+Q) + + \frac{\alpha_1}{P} \, {\rm Tr} \log \beta \left[\frac{\mathbb{1}}{\beta} + K^{(R)}(Q, \bar{Q}) \right] + \frac{\alpha_1}{P} y^\top \left[\frac{\mathbb{1}}{\beta} + K^{(R)}(Q, \bar{Q}) \right]^{-1} y \tag{105}$$

It is worth noticing that while in the zero mean case there was a simple relations between Q and $\bar{Q}$ at any temperature, we now lose that property and the saddle-point equations are not exactly solvable anymore, not even in the zero temperature limit. On the contrary, one can check that in the infinite-width limit we recover the previous result $\bar{Q} = 1$, $Q = 0$ and the rank one matrix $K^{(1)}$ does not contribute to the generalization error, since it does always appear in combination with the scalar $\bar{Q} - 1/(1+Q)$ that vanishes in the infinite-width limit.

Let us move to the derivation of an effective action for L hidden layers. As for the derivation with zero-mean activation function, the key step is to understand how the joint probability of the pre-activations at layer ℓ is linked to the one at layer $\ell - 1$. While in the zero-mean activation case, the key observation was that P_2 is related to P_1 by the replacement $C \to \bar{Q}_1 K(C)/\lambda_1$ (see Eq. 60), here we find that the correct replacement is $C \to \bar{Q}_1 K(C)/\lambda_1 - (\bar{Q}_1 - 1/(1+Q_1))K^{(1)}/\lambda_1$. Differently from the zero-mean activation case, where the kernel at layer L was only depending on the variables $\{\bar{Q}_\ell\}$, here we find that the recurrence is given in terms of the $\{Q_\ell\}$ as well. In conclusion, this produces a more unpleasant action where all the $\{Q_\ell, \bar{Q}_\ell\}$ are coupled via the nested non-linear expression of the kernel. The explicit recurrence relation for finite-mean activation functions is given by:

$$K^{(R)}_\ell = \frac{\bar{Q}_\ell}{\lambda_\ell} K \circ \left[K^{(R)}_{\ell-1} \right] - \frac{\left(\bar{Q}_\ell - \frac{1}{1+Q_\ell}\right)}{\lambda_\ell} K^{(1)} \circ \left[K^{(R)}_{\ell-1} \right] \,, \qquad K^{(R)}_0 = C \tag{106}$$

and the effective saddle-point action reads:

$$S_{\rm DNN} = \sum_{\ell=1}^{L} \frac{\alpha_L}{\alpha_\ell} \left[-Q_\ell \bar{Q}_\ell + \log(1+Q_\ell) \right] + \frac{\alpha_L}{P} {\rm Tr} \log \beta \left(\frac{\mathbb{1}}{\beta} + K^{(R)}_L(\{\bar{Q}_\ell, Q_\ell\}) \right) + \frac{\alpha_L}{P} y^T \left(\frac{\mathbb{1}}{\beta} + K^{(R)}_L(\{\bar{Q}_\ell, Q_\ell\}) \right)^{-1} y \,. \tag{107}$$

In view of the above considerations, it should be now clear that the heuristic Li-Sompolinsky theory (re-derived in the previous section) amounts to disregard all the additional terms $K^{(1)}$ that arise from the approach presented in this section.

V. NUMERICAL ISSUES IN SAMPLING FROM THE BAYESIAN POSTERIOR

Obtaining a perfect agreement between theory and simulations when sampling from a Bayesian posterior (especially in the zero temperature limit) is prevented by a number of technical numerical issues presented in the following.

1. Finite-size effects certainly play a role in explaining the small mismatch between theory and experiment. To address this point, we are currently performing high-precision numerical simulations with fixed $\alpha = P/N_1$ and increasing values of N_1 and P.

2. The $T \to 0$ limit, which corresponds to perfect interpolation of the dataset and is the only case in which the saddle point equations can be solved analytically, was the most logic to address for starting, but turns out to be very hard to simulate. This is clear from some preliminary work we are doing, where we numerically solve the saddle point equations at generic T for the saddle point variables Q, $\bar{Q} = f(Q)$. We find that the function $Q(T)$ changes rapidly for small temperatures.

3. At $T = 0.001$, the autocorrelation time of the simulation is already very large, taking as little as $5 \cdot 10^6$ epochs to thermalize. As the temperature is decreased, the autocorrelation time increases, and we need hundreds of thousands of epochs to gain satisfactory statistics.

4. The effect of a finite learning rate η has to be taken into account as well. From our preliminary results, we empirically observe that finite-η effects are larger at higher temperature. The standard way to take into account finite-η effects is to perform the extrapolation to $\eta \to 0$ simulating different learning rates.

5. Computing the theory in the case of $L > 1$ networks requires to numerically minimize a complex nested saddle-point functional of the variables $\bar{Q}_\ell$. We are currently working on a numerical routine to efficiently perform this task.