

Statistical Physics of Deep Learning: Optimal Learning of a Multilayer Perceptron near Interpolation

Jean Barbier^{1,*}, Francesco Camilli², Minh-Toan Nguyen¹, Mauro Pastore¹, and Rudy Skerk^{3,†}

¹*The Abdus Salam International Centre for Theoretical Physics Strada Costiera 11, 34151 Trieste, Italy*

²*Alma Mater Studiorum-Università di Bologna,*

Dipartimento di Matematica Piazza di Porta S. Donato 5, 40126 Bologna, Italy

³*International School for Advanced Studies, Via Bonomea 265, 34136 Trieste, Italy*



(Received 12 December 2025; accepted 30 April 2026; published 22 July 2026)

For four decades, statistical physics has been providing a framework to analyze neural networks. A long-standing question remained on its capacity to tackle deep-learning models capturing rich feature-learning effects, thus going beyond the narrow networks or kernel methods analyzed until now. We positively answer through the study of the supervised learning of a multilayer perceptron. Importantly, (i) its width scales as the input dimension, making it more prone to feature learning than ultrawide networks and more expressive than narrow ones or ones with fixed embedding layers, and (ii) we focus on the challenging interpolation regime where the number of trainable parameters and data are comparable, which forces the model to adapt to the task. We consider the matched teacher-student setting. Therefore, we provide the fundamental limits of learning random deep neural-network targets and identify the sufficient statistics describing what is learned by an optimally trained network as the data budget increases. A rich phenomenology emerges with various learning transitions. With enough data, optimal performance is attained through the model’s “specialization” toward the target, but it can be hard to reach for training algorithms, which get attracted by suboptimal solutions predicted by the theory. Specialization occurs inhomogeneously across layers, propagating from shallow toward deep ones but also across neurons in each layer. Furthermore, deeper targets are harder to learn. Despite its simplicity, the Bayes-optimal setting provides insights into how the depth, nonlinearity, and finite (proportional) width influence neural networks in the feature-learning regime that are potentially relevant in much more general settings.

DOI: [10.1103/56sb-pdh6](https://doi.org/10.1103/56sb-pdh6)

Subject Areas: Complex Systems, Statistical Physics

I. INTRODUCTION

Neural networks (NNs) are the powerhouse of modern machine learning, with applications in all fields of science and technology. Their use is now widespread in society much beyond the scientific realm. Understanding their expressive power and generalization capabilities is therefore not only a stimulating intellectual activity, producing surprising results that seem to defy established common sense in statistics and optimization [1], but it is also of major practical and economic importance.

One issue is that even the models dating back to the inception of deep learning [2] are not theoretically

well understood when operating in the “feature-learning regime” (a task-dependent term that will be clear later). The simplest deep-learning model is the multilayer fully connected feedforward neural network, also called multilayer perceptron (MLP). It corresponds to a function $\mathcal{F}_{\theta}(\mathbf{x}) = \mathbf{v}^T \sigma(\mathbf{W}^{(L)} \sigma(\mathbf{W}^{(L-1)} \dots \sigma(\mathbf{W}^{(1)} \mathbf{x}) \dots)$ going from $\mathbb{R}^d$ to $\mathbb{R}$, parametrized by $L + 1$ matrices $\theta = [\mathbf{v} \in \mathbb{R}^{k_L \times 1}, (\mathbf{W}^{(l)} \in \mathbb{R}^{k_l \times k_{l-1}})_{l \leq L}]$, with k_l denoting the width of the l th hidden layer (with $k_0 = d$), and an activation function $\sigma(\cdot)$ applied entrywise to vectors.

Until now, the quantitative theories for such NNs predicting which relevant features they can extract, how much data n they need to do so, and how well they generalize beyond their training data, relied on oversimplified architectural and/or data-abundance assumptions. This prevented the precise capture of the combined role of the depth and nonlinearity of NNs when they were trained from sufficiently many data for them to fully express their representation power. This paper offers answers to these questions in a richer scenario than what current statistical approaches could tackle.

*Contact author: jbarbier@ictp.it

†Contact author: rskerk@sissa.it

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

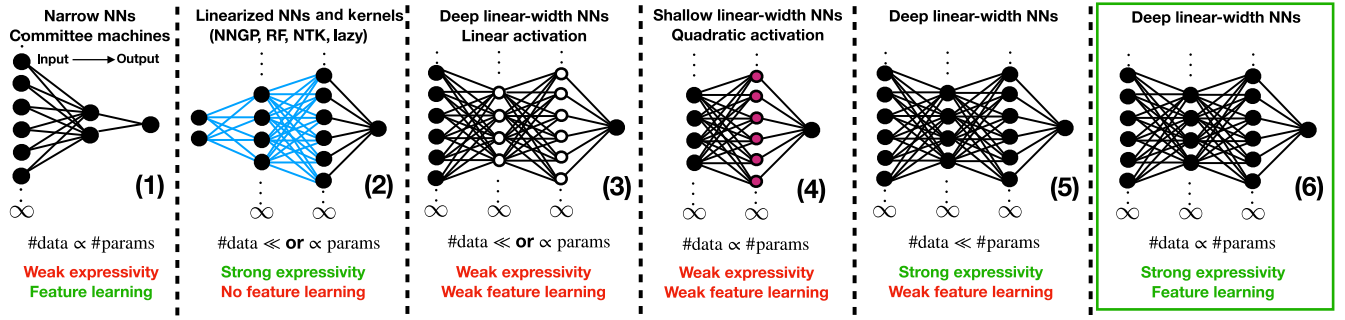


FIG. 1. Classification of models of fully connected feed-forward neural networks analyzed in the theoretical literature (see the main text for references). Class (1) models are very narrow, i.e., with a width independent of the large input dimension d . This class includes the perceptron and committee machines studied in the statistical mechanics literature (the latter are linked to the so-called multi-index models). They are analyzed near interpolation, where the number of data and model’s parameters are proportional. In this regime, feature learning emerges through phase transitions, but these models suffer from their limited expressivity. Class (2) encompasses all “kernel-like models” whose inner weights are frozen to random values (which is represented by the blue color) either by construction as in the random feature model or as a consequence of their overwhelming width or overparametrization (as in neural-network Gaussian processes, or in gradient-based dynamics in the lazy regime, where the weights effectively remain at initialization and the networks behave as neural tangent kernels). These models are expressive but do not learn task-relevant features due to their effectively frozen embedding layers. However, with readout weights (black last layer) to be $O(1/d)$ rather than the standard scaling $O(1/\sqrt{d})$, feature learning emerges despite the width being infinite. Another tractable simplification is class (3) deep linear networks, where the weights are learnable but the activation functions are linear (white circled nodes), thus reducing expressivity to that of a linear model, which allows for limited feature learning. A recent simplification is the linear-width shallow network with quadratic activation (pink nodes), class (4). Even if trained near interpolation, it can learn only a quadratic approximation of the target, which limits its expressivity. Moreover, we will see that in the teacher-student setting we consider, it cannot recover the target weights, thus preventing strong feature-learning in this sense. Class (5) models are the same as studied here, but set in the strongly overparametrized “proportional regime” with a sample size scaling only as the width. Yet, weak forms of feature learning can emerge: In the teacher-student setting as used in this paper, by “weak feature learning” we mean regimes where feature learning does occur but the model’s inner weights have zero overlap with the teacher’s one. The present paper considers fully trainable proportional-width nonlinear NNs trained near interpolation, class (6). This is the most challenging regime, where the model expressivity can fully manifest via strong task adaptation (i.e., recovery of the target).

A. A pit in the neural-network landscape

Given the difficulty of the theoretical analysis of NNs, a zoology of tractable simplifications reported here has emerged, each coming with pros and cons.

- (1) *Narrow networks:* Triggered by pioneering works employing spin-glass techniques to study NNs [3–5], the interest of the statistical physics community in the equilibrium properties of the narrow committee machines [$L = 1$ with $k_1 = \Theta(1)$ while $n = \Theta(d) \rightarrow \infty$, (1) in Fig. 1] rose quickly in the 1990s [6–20]. This line of classical works is at the inception of the discovery of learning phase transitions (found concurrently also in single-layer architectures with constrained weights or peculiar activations [21–26]). The main issue with narrow NNs is their restricted expressivity. Nevertheless, their analysis yielded important insights into NN learning mechanisms, some of which are also occurring in more expressive models. One that is of particular importance is the so-called specialization transition [27,28], where hidden neurons start learning different features. However, as we will see, in more expressive models a richer phenomenology emerges. This field has since then remained

very much alive, with the goal of treating more complex architectures as in the present paper; see Refs. [29,30] for reviews. Narrow NNs are multi-index functions, i.e., functions projecting their argument on a low-dimensional subspace (see Ref. [31] for a review). Their study allows us to understand which properties of nonlinearities make learning hard for gradient-based or message-passing algorithms [32–36].

- (2) *Kernel limit:* Ultrawide and linearized NNs, and the mean-field regime. On the other hand, in the ultrawide limit (L fixed, $k_l \gg n$) fully connected Bayesian NNs behave as kernel machines (the so-called neural-network Gaussian processes) [37–41], and hence suffer from these models’ limitations. Indeed, kernel machines infer the decision rule by first embedding the data in a fixed *a priori* feature space, the renowned kernel trick, then operating linear regression or classification over the features. In this respect, they do not learn task-relevant features and therefore need larger and larger feature spaces and training sets to fit their higher-order statistics [42–47]. The same conclusions hold for NNs trained with gradient-based methods but linearized around

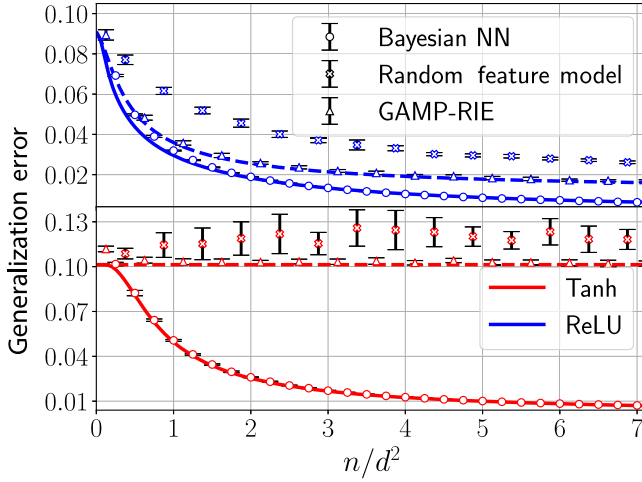


FIG. 2. Bayes-optimal mean-square generalization error achievable by a two-layer NN $\mathcal{F}_\theta(\mathbf{x}) = \mathbf{v}^\top \sigma(\mathbf{W}\mathbf{x})$ as a function of the amount of training data n over the squared input dimension d^2 when d , n , and the NN width k all diverge with n/d^2 and $k/d \rightarrow 0.5$ fixed (solid curves), with activation $\sigma(x) = \text{ReLU}(x)$ and $\tanh(2x)$ (same setting as the right panel of Fig. 5). These theoretical curves follow from the results in Sec. II. The task is regression with standard Gaussian inputs $(\mathbf{x}_\mu)_{\mu \leq n}$ and noisy responses $(y_\mu)_{\mu \leq n}$ generated by a target two-layer NN $\mathcal{F}_\theta(\mathbf{x})$ with Gaussian random weights and same activation. In the experiments, $d = 150$ and $k = 75$. Empty circles are obtained by training the Bayesian NN using Hamiltonian Monte Carlo initialized close to the target (yielding the best achievable error) and then computing its generalization error on 10^5 test data (error bars are the standard deviation over ten instances of the training set and target). Crosses show the generalization error empirically achievable by the random feature model $\mathcal{F}_\mathbf{a}^{\text{RF}}(\mathbf{x}) = \mathbf{a}^\top \sigma(\mathbf{W}^{\text{RF}}\mathbf{x})$ trained by exact empirical risk minimization $\mathbf{a}^* = \text{argmin}\{\sum_{\mu \leq n} [y_\mu - \mathcal{F}_\mathbf{a}^{\text{RF}}(\mathbf{x}_\mu)]^2 + t\|\mathbf{a}\|^2\}$ with optimized $\mathbf{a} \in \mathbb{R}^r$ and L_2 -regularization strength t picked by cross-validation. The fixed Gaussian features matrix $\mathbf{W}^{\text{RF}} \in \mathbb{R}^{r \times d}$ has width $r = 3kd$ roughly 3 times larger than the total number $k(d+1)$ of parameters of the NN. Triangles are the error of GAMP-RIE [96] extended to generic activations (see Appendix B 4), which reaches the performance of an optimally regularized kernel.

initialization [48], i.e., with frozen weights represented in blue in Fig. 1. These cases include the random feature model (RF) with fixed random inner layer [49] (which is a finite-size approximation of kernel machines), or the closely related neural tangent kernel [50] and lazy regimes [51]; see model (2) in Fig. 1. Such models are thus “effectively linearized” because only the readouts are learned. Figure 2 illustrates the importance of feature learning: Despite having a larger number of parameters, the best RF model or kernel is outperformed by an optimally trained NN; see also Refs. [52,53].

One way to probe feature-learning effects is through perturbative expansions around the

ultrawide limit, where $k \gg n$ but $O(1/k)$ corrections are kept [54–61]. This connects analysis to expansions around free fields in quantum field theory, where diagrammatic rules are used to manage complex combinatorial sums; see Refs. [62,63] for introductions. Another way to force feature learning in infinitely wide models is the mean-field scaling obtained by taking the readout weights $\mathbf{v}$ vanishingly small [$O(1/k)$ rather than the standard $O(1/\sqrt{k})$ scaling we consider]. Originally proposed as a means to escape the lazy regime of gradient-based dynamics [51] via a specific weights initialization [64–70], it was later extended to the Bayesian framework [71–75]. NNs in this scaling converge to kernel machines with data-dependent kernels (rather than fixed *a priori*, as in the lazy regime). We also mention Refs. [76,77], which alternatively rescale the Bayesian likelihood to induce feature learning at infinite width.

- (3) *Deep linear networks*: Another way to linearize networks, thus making them tractable, is by allowing fully trainable weights while placing linear activations in the inner layers. Linear networks are a major theoretical playground from the dynamical perspective [78] but also at the equilibrium (Bayesian) level [72,79–82]. In the same vein, theoreticians considered linear diagonal networks of the form $\mathcal{F}_{\mathbf{w},\mathbf{v}}^{\text{diag}}(\mathbf{x}) = \mathbf{v}^\top \text{diag}(\mathbf{w})\mathbf{x}$ [83,84], which exhibit an implicit bias in gradient descent learning [85–88]. A main issue with linear networks, however, is their intrinsically small expressivity, so only weak notions of feature learning can manifest.
- (4) *Shallow quadratic networks*: Various works have recently exploited the fact that a shallow NN with quadratic activation $\sigma(x) = x^2$ simplifies drastically [89–99]. However, we will see that this prevents strong feature learning to emerge. The closest settings to ours are Refs. [96–98]. There, the analysis based on results for the GLM [100] and matrix denoising [101–104] follows thanks to a specific mapping to a linear matrix-sensing problem, where the goal is to infer a Wishart-like matrix given its projection along random rank-one matrices.
- (5) *Proportional data regime*: This overparametrized regime considers a sample size much smaller than the number of model parameters [L fixed, d large, $k_l, n = \Theta(d)$]. Recent works show how a limited amount of feature learning makes the network equivalent to optimally regularized kernels [73,79,105–107] (see Ref. [108] for a large-depth analysis). MLPs thus reduce to linear networks (GLMs) in the sense conjectured in Ref. [109] and proven in Refs. [110,111] and thus suffer from their limitations. This could be a consequence of the fully connected architecture, as, e.g., convolutional

networks can learn more informative features in this regime [71,72,112,113]. In a similar data regime, Yoshino and co-workers have developed a replica theory for overparametrized deep NNs with a nonstandard architecture [114–116].

- (6) *A timely challenge: deep nonlinear networks of linear width trained near interpolation:* Despite the wealth of methods developed to study the aforementioned models, none is able to tackle NNs enjoying all the following realistic properties: (P_1) a width proportional to the input dimension, (P_2) with broad classes of nonlinear activations, (P_3) with possibly multiple hidden layers, (P_4) learning in the interpolation regime.

The property (P_1) combined with (P_2), (P_3) allows us to capture finite-width effects in NNs that are highly expressive, while still allowing us to take the large system limit needed to obtain sharp theoretical predictions. Even if it is not entirely clear whether a finite width improves the performance of Bayesian MLPs compared to their kernel limit [117,118], it is certainly one of the most natural ways to allow for the emergence of representation learning, which really is the crux of deep learning [2].

The interpolation regime (P_4) means L fixed, d large with $k_l = \Theta(d)$ [from (P_1)], and $n = \Theta(d^2)$, i.e., a sample size comparable to the number of trainable parameters. This regime is difficult to analyze for expressive models but also very interesting, because it forces them to adapt to the data in order to perform well. Hence, task-dependent feature learning emerges, thus escaping the reduction to linear models discussed above. Analyzing MLPs in the interpolation regime has been an open problem for decades and is widely recognized as one of the major theoretical challenges in the physics of learning [96,109]. That statistical mechanics is up to the task is an encouraging signal for physicists working on deep learning [119,120].

This setting is relevant and timely also from a practical perspective. Indeed, the latest NN architectures such as generative diffusion and large language models (LLMs) do operate near interpolation: The compute-optimal training scales parameters and tokens in equal proportion [121,122], with typical sizes ranging 10^{10} – 10^{12} . These models of utmost interest are highly expressive and also exhibit signs of feature learning [123,124]. From that perspective, it places them in a similar regime as considered in the present paper. LLMs are far more intricate than MLPs. Yet, they have things in common: In addition to the fact that one of the basic building blocks of LLMs is actually the MLP (together with the attention head), both correspond to deep nonlinear architectures. We thus consider it essential to tackle the interpolation regime of MLPs, with the hope that some insights brought forward by our theoretically tractable idealized setting

remain qualitatively relevant for the NN architectures deployed in applications.

B. Main contributions and setting

We address questions pertaining to the foundations of learning theory for NN models possessing all four properties (P_1)–(P_4). The first one is information-theoretic:

- Q1: Assuming the training data are generated by a target MLP, how much is needed to achieve a certain generalization performance using an MLP with same architecture and Bayes-optimally trained in a supervised manner?

The answer, provided analytically by Result 2 in Sec. II yields the Bayes-optimal limits of learning an MLP target function, thus bounding the performance of any model trained on the same data. The setting where the data-generating process is itself an MLP may look artificial at first, but given the high expressivity and universal approximation property of neural networks, studying their learning provides insights applicable to very general classes of functions. For this reason, and the analytical tractability of the teacher-student scenario explained below, this question has always been a starting point in the statistical physics literature on NNs [5].

Second, using statistical physics, we will answer another important question concerning interpretability:

- Q2: Given a certain data budget, which target features can the MLP learn?

This is key in order to understand the evolution of the best learning strategy for an MLP as a function of the amount of available data. Consequently, we will precisely explain what a perfectly trained MLP does to beat the random feature model or an optimally regularized kernel (see Fig. 2). In a few words, the reason is that given enough data, strong feature learning emerges within the NN, in the sense of recovery of the target weights. This happens through a specialization phase transition; in the deep case, there will be one transition per layer. This mechanism is not possible with the random feature model, which explains the gap in performance. These insights will follow from the detailed analysis of the sufficient statistics (i.e., the order parameters, OPs) of the model as the data increase, obtained from the large deviation perspective provided in Results 1, 3, and 4. The OPs carry more information than merely computing the achievable generalization performance (see Q1).

Finally, based on experiments, we provide in Sec. III algorithmic insights for MLPs with $L \leq 2$ layers:

- Q3: Given a reasonable compute and data budget, can practical training algorithms reach optimal performance or are they blocked by statistical-computational gaps?

The short answer is that it depends on the target, in particular, on whether its readout weights are discrete.

Answering these questions will provide a phase diagram depicting the optimal performance, the features that are learned to attain it, and the limitations faced by algorithms, as a function of the data budget; see Sec. III. Before presenting the setting needed to do so, let us emphasize once more that the theoretical component of this paper will be concerned only with static aspects, i.e., generalization capabilities of trained networks (after a manageable or unconstrained compute time). We do not provide any theoretical claims on how learning occurs during training.

1. Teacher-student setup

We consider the supervised learning of an MLP with L hidden layers when the data-generating model, i.e., the target function (or “teacher”), is also an MLP of the same form with unknown weights. These weights are the readouts $\mathbf{v}^0 \in \mathbb{R}^k$ and inner weights $\mathbf{W}^{(\ell)0} \in \mathbb{R}^{k_\ell \times k_{\ell-1}}$ for $\ell \leq L$ (with $k_0 = d$) drawn entrywise independent identically distributed (i.i.d.) from P_v^0 and P_W^0 , respectively (the latter distribution being the same law for all $\ell \leq L$). We assume P_W^0 to be centered while P_v^0 has mean $\bar{v}$, and both priors have unit second moment. We denote the set of unknown parameters of the target as $\boldsymbol{\theta}^0 = [\mathbf{v}^0, (\mathbf{W}^{(\ell)0})_{\ell \leq L}]$.

For a given input vector $\mathbf{x}_\mu \in \mathbb{R}^d$, for $\mu \leq n$, the response or label y_μ is drawn from a kernel P_{out}^0 :

$$y_\mu \sim P_{\text{out}}^0(\cdot | \lambda_\mu^0) \quad \text{with} \quad \lambda_\mu^0 := \mathcal{F}_{\boldsymbol{\theta}^0}^{(L)}(\mathbf{x}_\mu), \quad (1)$$

where the MLP target function is defined as

$$\mathcal{F}_{\boldsymbol{\theta}^0}^{(L)}(\mathbf{x}) := \frac{\mathbf{v}^{0\top}}{\sqrt{k^L}} \sigma \left(\frac{\mathbf{W}^{(L)0}}{\sqrt{k_{L-1}}} \sigma \left(\frac{\mathbf{W}^{(L-1)0}}{\sqrt{k_{L-2}}} \cdots \right. \right. \\ \left. \left. \times \sigma \left(\frac{\mathbf{W}^{(1)0}}{\sqrt{k_0}} \mathbf{x} \right) \cdots \right) \right).$$

We will analyze the case of an arbitrary number of layers L (that remains d independent). However, we will give special attention to the shallow, one-hidden-layer MLP (we drop layer indices in this case)

$$\mathcal{F}_{\boldsymbol{\theta}^0}^{(1)}(\mathbf{x}) = \frac{1}{\sqrt{k}} \mathbf{v}^{0\top} \sigma \left(\frac{1}{\sqrt{d}} \mathbf{W}^0 \mathbf{x} \right), \quad (2)$$

as well as the MLP with two hidden layers

$$\mathcal{F}_{\boldsymbol{\theta}^0}^{(2)}(\mathbf{x}) = \frac{1}{\sqrt{k_2}} \mathbf{v}^{0\top} \sigma \left(\frac{1}{\sqrt{k_1}} \mathbf{W}^{(2)0} \sigma \left(\frac{1}{\sqrt{d}} \mathbf{W}^{(1)0} \mathbf{x} \right) \right). \quad (3)$$

The kernel can be stochastic or model a deterministic rule if $P_{\text{out}}^0(y|\lambda) = \delta(y - \mathbf{f}^0(\lambda))$ for some function $\mathbf{f}^0$. Our main example $P_{\text{out}}^0(y|\lambda) = \exp\{-[1/(2\Delta)](y - \lambda)^2\} / \sqrt{2\pi\Delta}$ is the linear readout with Gaussian label noise.

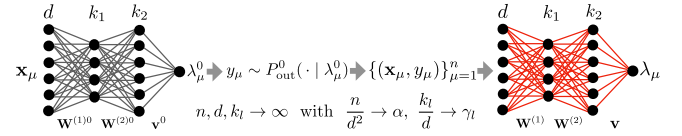


FIG. 3. The teacher-student scenario for the case of two hidden layers. The teacher NN is used to produce the responses given the inputs. A student NN with matched architecture (but who is not aware of the parameters of the teacher) is then trained in a Bayesian manner given the training data. We also display the scaling limit considered given by Eq. (4).

We will first consider i.i.d. standard Gaussian vectors as inputs $\mathbf{x}_\mu$. In that case, the whole data structure is dictated by the input-output relation only, allowing us to focus solely on the influence of the target function on the learning. In Sec. III, we generalize the results to include structured data: Gaussian with a covariance and real data (MNIST). The input and output pairs $\mathcal{D} = \{(\mathbf{x}_\mu, y_\mu)\}_{\mu \leq n}$ form the training set for a student network with matching architecture.

The Bayesian student learns via the posterior distribution of the weights matrices $\boldsymbol{\theta} = [\mathbf{v}, (\mathbf{W}^{(\ell)})_{\ell \leq L}]$ (of the same respective sizes as the teacher's) given the training data

$$dP(\boldsymbol{\theta}|\mathcal{D}) := \mathcal{Z}(\mathcal{D})^{-1} dP_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{\mu \leq n} P_{\text{out}}(y_\mu | \lambda_\mu(\boldsymbol{\theta})),$$

where $dP_{\boldsymbol{\theta}}(\boldsymbol{\theta}) := dP_v(\mathbf{v}) \prod_{\ell \leq L} dP_W(\mathbf{W}^{(\ell)})$ [with the notation $dP(\mathbf{M}) := \prod_{i,j} dP(M_{ij})$], with postactivations

$$\lambda_\mu(\boldsymbol{\theta}) := \mathcal{F}_{\boldsymbol{\theta}}^{(L)}(\mathbf{x}_\mu), \quad \mu \leq n.$$

The posterior normalization $\mathcal{Z}(\mathcal{D}) = \mathcal{Z}^{(L)}(\mathcal{D})$ for the model with L hidden layers is the partition function, and P_W, P_v are the priors assumed by the student. We focus on the Bayes-optimal setting $P_W = P_W^0, P_v = P_v^0$, and $P_{\text{out}} = P_{\text{out}}^0$, but the approach can be extended to account for a mismatch.

As stated above, we study the linear-width regime with quadratically many samples, which places the model in the interpolation regime, i.e., a large size limit

$$d, k_l, n \rightarrow +\infty \quad \text{with} \quad \frac{k_l}{d} \rightarrow \gamma_l \quad \text{for } l \leq L, \quad \frac{n}{d^2} \rightarrow \alpha. \quad (4)$$

Given the cost of training deep Bayesian MLPs and specific difficulties discussed below associated with an increasing number of layers, we distinguish the cases of one, two, and more than two hidden layers for what concerns the hypotheses we impose on the activation σ .

(H_1): For shallow NNs with $L = 1$ hidden layer, our results are valid for an arbitrary activation function as long as it admits an expansion in Hermite polynomials with coefficients $(\mu_\ell)_{\ell \geq 0}$ (see Appendix A 2)

$$\sigma(x) = \sum_{\ell \geq 0} \frac{\mu_\ell}{\ell!} \text{He}_\ell(x). \quad (5)$$

We also assume it has vanishing zeroth Hermite coefficient $\mu_0 = 0$, i.e., that it is centered $\mathbb{E}_{z \sim \mathcal{N}(0,1)} \sigma(z) = 0$. In Appendix B 1 g we relax this assumption. We will mainly consider tanh, ReLU, and Hermite polynomial activations.

Through Hermite expansion, the MLP function can be decomposed as

$$\mathcal{F}_\theta^{(1)}(\mathbf{x}) = \frac{\mu_1}{\sqrt{d}} \frac{\mathbf{v}^\top \mathbf{W}}{\sqrt{k}} \mathbf{x} + \frac{\mu_2}{2d} \text{Tr} \left[\frac{\mathbf{W}^\top \text{diag}(\mathbf{v}) \mathbf{W}}{\sqrt{k}} (\mathbf{x}^{\otimes 2} - I_d) \right] + \dots,$$

where $\dots$ contains terms made of tensors of all orders constructed from θ , contracted with input rank-one tensors $(\mathbf{x}^{\otimes \ell})_\ell$. In each such term, at least one tensor is of order $\ell \geq 3$. Therefore, an equivalent interpretation of the learning problem of an MLP target is that of a “tensor-sensing problem” where the tensors $\mathbf{v}^{\otimes \ell} \mathbf{W}^0 \in \mathbb{R}^d$, $\mathbf{W}^{0\top} \text{diag}(\mathbf{v}^0) \mathbf{W}^0 \in \mathbb{R}^d$ entering the observed responses are all constructed from the same fundamental parameters θ^0 (see, e.g., Ref. [125]). The first term in the above expansion is called the “linear term or component.” The one of the target is perfectly learnable in the quadratic data regime we consider. The second term is the “quadratic term or component.” Both will play a special role because, as we will see, the terms $\dots$ effectively behave as (Gaussian) noise when $n = \Theta(d^2)$, unless θ^0 is partially recovered. Learning through recovery of θ^0 is called specialization. In contrast, the linear and quadratic terms are learnable without specialization. This is reassuring given that we will argue that for many targets, it takes a time growing as $\exp(cd)$ for the network to specialize, for some positive c -dependent constant $c < 1$. This separation in algorithmic learnability of the first and second components versus all the others is at the root of the emergence of different learning strategies employed by the network and the crux of the generalization of a learning algorithm in Appendix B 4 coined GAMP-RIE (generalized approximate message passing with rotational invariant estimator) [96].

(H_2): For $L = 2$, we require $\mu_0 = \mu_2 = 0$, which is, e.g., the case for odd activations. Our main example is tanh. In the tensor inference problem appearing when expanding all activations, $\mu_2 = 0$ means that no quadratic term is present. However, a “product term” $\mathbf{W}^{(2)} \mathbf{W}^{(1)}$ appears (in addition to $\mathbf{v}^\top \mathbf{W}^{(2)} \mathbf{W}^{(1)}$). We will see in Sec. III that skipping the quadratic term implies that learning terms beyond the linear ones will be possible only through specialization. However, the presence of the product term will have interesting consequences on the learning curves. Importantly, $\mathbf{W}^{(2)} \mathbf{W}^{(1)}$ is a matrix learnable partly

independently of its factors and consequently requires its own OP in the analysis.

(H_3): For $L \geq 3$, we require $\mu_0 = \mu_1 = \mu_2 = 0$. This does not include standard activations, and we consider the hyperbolic tangent after setting $\mu_1 = 0$ in its Hermite decomposition. $\mu_2 = 0$ again entails that learning beyond-linear terms requires the network to specialize, and $\mu_1 = 0$ prevents the multiplication of OPs by avoiding the presence of many product terms.

Related to this last comment, we wish to emphasize that these hypotheses are not due to restrictions of the techniques we develop. The issue is purely practical: Relaxing them while increasing the number of layers yields a combinatorial explosion (in L) of the number of OPs to track in the theory as well as cumbersome formulas. We have therefore decided to leave for future work the analysis of the most general case and focus here on these special ones which already yield an extremely rich picture while remaining interpretable.

C. Replica method and HCIZ combined

A key component of our approach is the way we blend tools from spin glasses (the replica method [126]) and matrix models, in particular, the so-called Harish-Chandra–Itzykson–Zuber (HCIZ) “spherical” integral [127–129]. Here, we review the growing corpus of works utilizing it jointly with the replica method. Let us first define this matrix integral

$$\mathcal{Z}_{\text{HCIZ}}^{(\beta)}(\mathbf{A}, \mathbf{B}) := \int d\mu^{(\beta)}(\mathbf{O}) \exp\left(\frac{\beta N}{2} \text{Tr}[\mathbf{O} \mathbf{A} \mathbf{O}^\top \mathbf{B}]\right), \quad (6)$$

where $\beta = 2$ if $\mathbf{A}, \mathbf{B}$ are $N \times N$ Hermitian matrices, $\beta = 1$ if real symmetric. Respectively, the integral is over the unitary $\mathcal{U}(N)$ or orthogonal group $\mathcal{O}(N)$, with respect to the corresponding uniform Haar measure $\mu^{(\beta)}$. For $\beta = 2$, it admits a closed form for any N [127] and a known large- N limit for $\beta = 1, 2$ [128–131]. It is crucial to analyze matrix models in physics and random geometry [132–134]. In spite of having an “explicit” limit, it can be tackled only in a few cases [102,135]. However, if one matrix, say $\mathbf{A}$, has small rank compared to $N \gg 1$, the corresponding low-rank spherical integral is simple [136,137].

Spherical integrals were used in the replica method for spin glasses with correlated disorder in the seminal paper [138]. It triggered a long series of works in spin glasses [139–143], in analyzing simple NNs [144,145], and in inference and message-passing algorithms [146–162]. In these papers, the degrees of freedom are a few vectors (e.g., the replicas of the system forming a low-rank matrix $\mathbf{A}$) interacting with a quenched rotationally invariant matrix $\mathbf{B}$ of rank $\Theta(N)$. Rotational invariance, a crucial property for employing spherical integrals, means distributional invariance under orthogonal transformations, i.e., $\mathbb{P}(\mathbf{B}) = \mathbb{P}(\mathbf{O} \mathbf{B} \mathbf{O}^\top) \forall \mathbf{O} \in \mathcal{O}(N)$ if $\mathbf{B}$ is symmetric. Consequently,

only the low-rank spherical integral intervenes when integrating $\mathbf{B}$'s eigenvectors.

An active research line tries to include models where the degrees of freedom themselves are linear-rank matrices in addition to the quenched disorder. This presents a whole new challenge. Seminal papers in the context of matrix denoising are Refs. [163,164], which provided a spectral denoising algorithm (on which the GAMP-RIE [96] relies), also analyzed in Refs. [103,104]. Extensions to nonsymmetric matrix denoising exist [165,166]. An early attempt at combining linear-rank spherical integration [where both $\mathbf{A}$, $\mathbf{B}$ in Eq. (6) have rank $\Theta(N)$] with the replica method is Ref. [167], which tried to improve on the replica approach for matrix denoising in Refs. [168,169] that was missing important correlations among variables. It was followed by two concurrent papers yielding intractable [101] or perturbative [102] results for non-Gaussian signals.

Remark 1. No method in the aforementioned papers is satisfactory beyond the realm of denoising problems involving strictly rotational invariant signal matrices (e.g., Gaussian, Wishart). For example, the HCIZ and replica combination in the latest works [96,170] requires it, because after using the replica trick to integrate the quenched disorder, the HCIZ is used directly to integrate the annealed matrix degrees of freedom (representing the replicas of the signal matrix), which is possible by rotational invariance.

Recently, matrix denoising without rotational invariance was analyzed in Ref. [171] by assuming that the model behaves as a pure matrix model (due to an “effective rotational invariance”) in a first phase, and then as a “standard” planted mean-field spin glass in a second. The phases were thus treated separately via different formalisms—HCIZ in one phase, a cavity method under mean-field decoupling assumptions in the other—and then joined using a criterion to locate the transition. This approach yielded a good match with numerics. However, we now understand that this treatment can be improved, because the “matrix nature” of the model and associated correlations discarded by mean-field methods do play a role also in the second phase. Thus, a major conceptual (and technical) issue remained: whether there exists a theory based on a unified formalism able to describe the whole phase diagram of inference and learning problems involving linear-rank matrices which lack rotational invariance. Ideally, it should be able to handle the correlations induced by the matrix nature of the problem while still capturing the phase transitions and symmetry-breaking effects connected to its mean-field component. The present paper provides this theory in the context of NNs, through a replica and HCIZ combination of a different nature from previous works; see Sec. IV.

Related to this last point, we emphasize that previous works on extensive-width shallow NNs [$k = \Theta(d^c)$ for $0 < c \leq 1$] considered either purely quadratic activation

[93,94,96–99] or, on the contrary, with $\mu_{\leq 2} = 0$ [172]. Both settings enjoy intrinsic simplifications. On one hand, the quadratic NN reduces to a matrix-sensing problem [96,98]. It is therefore a “pure matrix model” with rotational invariance when considering Gaussian weights: The target (and model) depends only on them via $\mathbf{W}^{0\top} \mathbf{W}^0$. Therefore, by rotational invariance (from the left and right) of the Gaussian matrix $\mathbf{W}^0$, it cannot be recovered, so no specialization transitions can occur. The advantage is that a large toolbox from random matrix theory is then available: The HCIZ integral to study static aspects [96–98] or Oja’s flow and matrix Riccati differential equations for the dynamics [94,99,173]. On the other hand, Ref. [172] considers $\mu_\ell \neq 0$ for $\ell \geq 3$ only. In this case, the model is “purely mean field”: Strong decoupling phenomena take place which allow a treatment in terms of an effective one-body equivalent system as in mean-field spin systems.

In contrast, the techniques we develop in the present paper can deal with truly hybrid models where the two types of characteristics manifest and are taken into account using a single formalism: the correlations among the entries of the matrix degrees of freedom entering the problem, and specialization phase transitions induced by mean-field terms. The emerging phase diagram will consequently be extremely rich. In particular, we are able to treat the shallow MLP with generic activation function (Result 1), or the two-layer MLP with $\mu_1 \neq 0$ (Result 3). The case of $L \geq 3$ requires hypothesis (H_3) on σ , which, in turn, makes the model purely mean field.

D. Organization of the paper

Section II first discusses the main hypothesis underlying the theory and the meaning of the functions entering it. We then present the theoretical results: replica-symmetric formulas for the free entropy and OPs for shallow NNs (Result 1), with two hidden layers (Result 3), or arbitrary L (Result 4). These provide an answer to Q2. The Bayes generalization error is deduced automatically from Result 2 in all cases, thus answering Q1.

Section III is the core experimental part. It validates the theory through the numerical exploration of the rich learning phase diagram. Our main message concerning Q2 is as follows: As α increases two phases appear.

- (i) *Universal phase:* Before a critical sample rate α_{sp} , the NN makes predictions by exploiting specific nonlinear combinations of the teacher’s features without disentangling them; effectively, the student learns the best “quadratic network approximation” of the target. In this phase, performance is (asymptotically) independent of the detailed law of the target hidden weights (hence, the term “universal”). Yet, the (effectively quadratic) NN outperforms kernel ridge regression (and thus the random feature model, too; see Fig. 2); see Refs. [96,98].

- (ii) *Specialization phase*: Increasing the data beyond α_{sp} triggers specialisation transitions: Individual hidden units start aligning with target units. Which features specialize first is governed by the readout strengths of the target: Stronger features (larger readout amplitudes) emerge earlier. For heterogeneous readouts, this yields a sequence of specialization events; for homogeneous readouts, a collective transition occurs. If $L \geq 2$, possible heterogeneity both in the rows and columns of individual weight matrices induces nontrivial specialization profiles in each layer. In turn, different layers can experience different phases and do not necessarily specialize concurrently. We will also show that learning propagates from inner to outer hidden layers, because deeper layers require more data to be recovered through specialization. Consequently, deeper target functions appear harder to learn than shallow ones.

In summary, despite the model's matrix nature at the source of the universal phase, additional mean-field-like terms in the free entropy (in information theory parlance, Gaussian scalar channels) imply the existence of specialization events. These terms depend explicitly on the weights prior and interact with the matrix degrees of freedom, and ultimately break the numerous effective symmetries holding before the transition.

The theoretical phase diagram will be extensively tested against various training algorithms: two Monte-Carlo-based Bayesian samplers, a first-order optimization procedure (ADAM), and a mixed spectral and approximate message-passing algorithm generalizing the GAMP-RIE of [96] to accommodate general activation functions σ when $L = 1$. The performance of these algorithms belonging to different classes, even when suboptimal, can be exactly (or, for ADAM, at least accurately) predicted by nonequilibrium solutions of the theoretical equations.

Focusing on $L \leq 2$ for what pertains to algorithmic hardness of learning, Q3, we will show empirically that specialization is potentially hard to reach for some target functions, in particular, when the readouts are discrete. The tested algorithms fail to find it and instead get trapped by suboptimal nonspecialized solutions, probably due to statistical-computational gaps.

We will also generalize the theory to structured data, i.e., Gaussian with a covariance. It will capture the model's performance when trained from non-Gaussian inputs, too. Tests with real (MNIST images) and synthetic data generated by one layer of an NN will confirm it.

Section IV contains the main steps of our replica theory, with an emphasis on its novel ingredients. Along the derivation, the mixed matrix model and mean-field planted spin-glass nature of the problem will become apparent. Finally, Sec. V summarizes our contributions and discusses the numerous perspectives this work opens.

The appendixes are found after the references. Appendix A gathers some important prerequisites: Appendix A 1 summarizes all notations used in the paper (we advise the reader to give it a look before reading the main results). The definition of the Hermite polynomials and Mehler's formula are found in Appendix A 3, the Nishimori identities in Bayes-optimal inference in Appendix A 2, the link between free entropy and mutual information in Appendix A 4, and a simplification of the expression for the optimal mean-square generalization error in Appendix A 5.

Appendix B groups all subappendixes related to the shallow MLP. Appendix B 1 details all the steps of the replica calculation. Appendix B 2 proposes alternative routes to take care of the entropy of the order parameters associated with the matrix degrees of freedom in the model. Appendix B 3 analyzes the large sampling rate limit of the theoretical free entropy. Appendix B 4 provides the generalization of the GAMP-RIE algorithm needed to deal with general σ . Appendix B 5 is an empirical analysis of the hardness of learning shallow targets. Appendix B 6 is a partial proof for a special case of activation function. Finally, Appendix B 7 provides additional experimental validations of the fact that the readout weights of the model being learnable or fixed has no effect on its optimal performance.

Appendix C concerns only the deep MLP. Appendix C 1 is the replica calculation. Appendix C 2 shows the consistency of the formulas provided for structured inputs with $L = 1$ in the main text, and the ones for a special case of non-Gaussian data obtainable from the theory for two hidden layers when freezing the first one (which induces a structure for the inputs of the second, learnable layer).

Appendix D provides all information needed to reproduce the simulations with the provided codes [174].

II. MAIN RESULTS: THEORY OF THE MLP

We aim at evaluating the expected optimal generalization error in the teacher-student setting of Fig. 3. Let $[\mathbf{x}_{\text{test}}, y_{\text{test}} \sim P_{\text{out}}(\cdot | \lambda_{\text{test}}^0)]$ be a test sample independent of $\mathcal{D}$ drawn using the teacher, where λ_{test}^0 is defined as in Eq. (1) with $\mathbf{x}_\mu$ replaced by $\mathbf{x}_{\text{test}}$ [and similarly for $\lambda_{\text{test}}(\boldsymbol{\theta})$]. Given a prediction function $\mathbf{f}$, the Bayes estimator for the test response is $\hat{y}^{\mathbf{f}}(\mathbf{x}_{\text{test}}, \mathcal{D}) := \langle \mathbf{f}(\lambda_{\text{test}}(\boldsymbol{\theta})) \rangle$, where $\langle \cdot \rangle := \mathbb{E}[\cdot | \mathcal{D}]$. Then, for a performance measure $\mathcal{C}: \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}_{\geq 0}$ the Bayes generalization error is

$$\varepsilon^{\mathcal{C}, \mathbf{f}} := \mathbb{E}_{\boldsymbol{\theta}^0, \mathcal{D}, \mathbf{x}_{\text{test}}, y_{\text{test}}} \mathcal{C}(y_{\text{test}}, \langle \mathbf{f}(\lambda_{\text{test}}(\boldsymbol{\theta})) \rangle). \quad (7)$$

The case of square loss $\mathcal{C}(y, \hat{y}) = (y - \hat{y})^2$ with the choice $\mathbf{f}(\lambda) = \int dy y P_{\text{out}}(y | \lambda) =: \mathbb{E}[y | \lambda]$ yields the Bayes-optimal mean-square generalization error:

$$\varepsilon^{\text{opt}} := \mathbb{E}_{\theta^0, \mathcal{D}, \mathbf{x}_{\text{test}}, y_{\text{test}}} (y_{\text{test}} - \langle \mathbb{E}[y | \lambda_{\text{test}}(\theta)] \rangle)^2. \quad (8)$$

In order to access $\varepsilon^{\text{C}, \text{f}}$, ε^{opt} , and other relevant observables, one can tackle the computation of the average log-partition function, or free entropy in statistical physics:

$$f_n := \frac{1}{n} \mathbb{E}_{\theta^0, \mathcal{D}} \ln \mathcal{Z}(\mathcal{D}). \quad (9)$$

The mutual information $I(\theta^0; \mathcal{D})$ between the target and data is related to the free entropy f_n ; see Appendix A 4.

Before presenting the results, we will first detail the main hypothesis for their derivation and explain the physical meaning of the quantities entering them. This will ease their interpretation. We postpone the core of the theoretical derivations to Sec. IV.

A. Main hypothesis

Let $s \in \mathbb{N}_{>0}$ independent of d and define a Gaussian vector $(\lambda^a)_{a=0}^s := (\lambda^0, \lambda^1, \dots, \lambda^s)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{K}^*)$ with covariance (for $a, b = 0, \dots, s$)

$$(\mathbf{K}^*)_{ab} := \mathbb{E} \lambda^a \lambda^b = K^* + (K_d - K^*) \delta_{ab}. \quad (10)$$

Let $(\theta^a)_{a=1}^s$ be i.i.d. from the posterior $dP(\cdot | \mathcal{D})$, and θ^0 are the random target weights. Our main assumption is that there exists a nonrandom $\mathbf{K}^*$ such that, under the randomness of a common test input $\mathbf{x}_{\text{test}} \notin \mathcal{D}$ and $(\theta^a)_{a=0}^s$, the postactivations $[\lambda_{\text{test}}(\theta^a)]_{a=0}^s$ (called “replicas”) converge in law toward $(\lambda^a)_{a=0}^s$ in the limit (4):

$$\text{Hypothesis: } \exists \mathbf{K}^* | [\lambda_{\text{test}}(\theta^a)]_{a=0}^s \xrightarrow{\text{Law}} (\lambda^a)_{a=0}^s. \quad (11)$$

The goal of the replica method will be to derive $\mathbf{K}^*$ in terms of fundamental low-dimensional OPs capturing the statistical dependences among $(\theta^a)_{a=0}^s$. The above convergence can be equivalently assumed conditionally on $(\theta^a)_{a=0}^s$, if well sampled, by concentration of the OPs. $\mathbb{E}_{(\lambda^a)}[\cdot]$ must therefore be interpreted as the asymptotic equivalent of the expectation with respect to the “quenched Gibbs measure” (i.e., the whole randomness): Given a function $f: \mathbb{R}^{s+1} \mapsto \mathbb{R}^s$ of $s+1$ replicas of the postactivation (with $s, \bar{s}$ independent of d),

$$\mathbb{E}_{\theta^0, \mathcal{D}, \mathbf{x}_{\text{test}}, y_{\text{test}}} \langle f([\lambda_{\text{test}}(\theta^a)]_a) \rangle \rightarrow \mathbb{E}_{(\lambda^a)} f((\lambda^a)_a).$$

Notice that the covariance for $a \neq b$ does not depend on whether one of the indices is the teacher’s index 0. That the teacher is statistically indistinguishable from the other replicas $a \geq 1$ is a consequence of the Bayes-optimal setting and the Nishimori identities; see Appendix A 3. In the non-Bayes-optimal setting, which is treatable with a

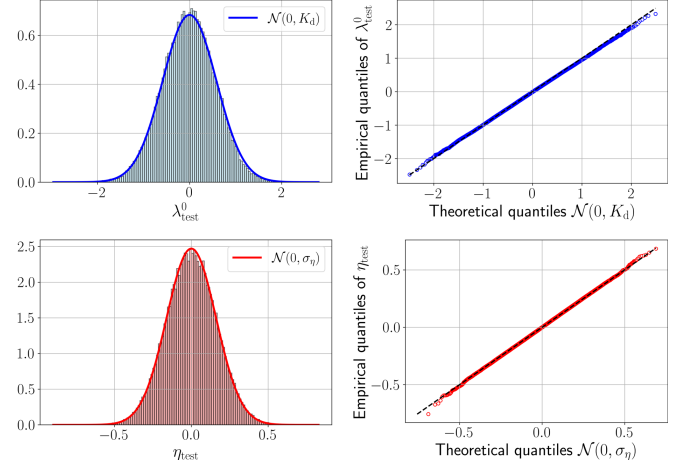


FIG. 4. Experimental evidence for the Gaussian hypothesis. In all experiments, $d = 300, \gamma = 0.5, \alpha = 3.0, \Delta = 0.1, \sigma(x) = \text{ReLU}(x) - 1/\sqrt{2\pi}$, both readout and inner weights have standard Gaussian prior. Empirical evaluations are based on a test set of size 5×10^4 . The results have been averaged over ten instances of the training set and teacher. Top left: histogram of the teacher postactivations λ_{test}^0 evaluated on $\mathbf{x}_{\text{test}}$ compared with the theoretically predicted Gaussian density $\mathcal{N}(0, K_d)$ [see Eq. (13) for the definition of K_d]. Top right: quantile-quantile plot comparing the theoretical quantiles of $\mathcal{N}(0, K_d)$ with the empirical ones of λ_{test}^0 . Bottom left: histogram of the student’s projection along the orthogonal direction to the teacher $\eta_{\text{test}} = \lambda_{\text{test}} - [\mathbb{E}_{\mathbf{x}_{\text{test}}} \lambda_{\text{test}}^0 \lambda_{\text{test}} / \mathbb{E}_{\mathbf{x}_{\text{test}}} (\lambda_{\text{test}}^0)^2] \lambda_{\text{test}}^0 \approx \lambda_{\text{test}} - (K^*/K_d) \lambda_{\text{test}}^0$, where $\lambda_{\text{test}}(\mathbf{v}, \mathbf{W})$ is the student postactivation with both $\mathbf{v}$ and $\mathbf{W}$ sampled from the posterior via Hamiltonian Monte Carlo, evaluated on the same test set, and compared with the theoretical density $\mathcal{N}(0, \sigma_\eta)$, where $\sigma_\eta = K_d - K^{*2}/K_d$ [see Eq. (16) for K^*]. Bottom right: quantile-quantile plot comparing the theoretical quantiles of $\mathcal{N}(0, \sigma_\eta)$ with the empirical quantiles of η_{test} .

similar approach, the covariance would be more complicated, with the teacher playing a special role.

The Gaussian hypothesis (11) will be justified *a posteriori* in Sec. III, by the excellent match between our predictions for the learning curves and OPs and the experimental ones. It can also be tested directly: Figure 4 displays the histogram of the teacher postactivation for multiple test inputs (blue) and of the projection of the student postactivation along the orthogonal direction to the teacher, when trained by Hamiltonian Monte Carlo and evaluated on the same test data (red). Our hypothesis implies that they should both be Gaussian distributed and indeed they are, with laws correctly predicted by the theory (see next two sections). We also compare the empirical moment generating function of $(\lambda_{\text{test}}^0, \lambda_{\text{test}})$ and its theoretical prediction based on the Gaussian hypothesis, respectively, given by $M_{\text{emp}}(t_0, t_1) = \mathbb{E}_{\mathbf{x}_{\text{test}}} \exp(t_0 \lambda_{\text{test}}^0 + t_1 \lambda_{\text{test}})$, where $\mathbb{E}_{\mathbf{x}_{\text{test}}}$ is an average over a test set of size 10^5 and $M_{\text{th}}(t_0, t_1) = \exp[K_d(t_0^2 + t_1^2)/2 + K^* t_0 t_1]$. Their relative error $|1 - M_{\text{emp}}(t_0, t_1)/M_{\text{th}}(t_0, t_1)|$ computed over a 21×21 regular grid in $[0, 1]^2$ has mean 0.015 and

standard deviation 0.016. This confirms that the theoretical Gaussian laws provide a remarkably accurate fit of the observed ones.

Remark 2. Gaussian assumptions on the postactivations are at the core of a fruitful series of works mapping the generalization capabilities of random feature models [175–180] and overparametrized NNs [109–111] to the ones of equivalent Gaussian covariate models. In these settings, formal proofs support the hypothesis, but the covariance of the postactivations matches the one of a statistically equivalent model which is linear in the input data. For this reason, many results in these settings go under the name of “Gaussian equivalence principle” (GEP) or “theorem.” The failure of the approaches based on the GEP to capture nonlinear effects in learning around interpolation has been attributed to non-Gaussian corrections becoming relevant in this regime [109]. Instead, we show here that the Gaussian hypothesis (11), once nonlinear effects are taken into account in the form of $\mathbf{K}^*$ as in our Results section below, works remarkably well to make predictions in the interpolation asymptotics, to the point we conjecture it to be exact in some cases; see Remark 4. A recent rigorous work provides examples where GEPs break down and describes how they can be redeemed [181].

B. Auxiliary potentials and their interpretation

As usual with the replica formalism used in the context of inference [182,183], the derived formulas are expressed in terms of auxiliary potential functions that are related to the log-normalization constants of the posterior distributions of auxiliary inference problems. These potentials shall be denoted by ψ_{P_W} , $\phi_{P_{\text{out}}}$, ι , and $\tilde{\iota}$. We describe their meaning hereby.

- (i) Let $w^0, w \sim P_W$, and $\xi \sim \mathcal{N}(0, 1)$ be all independent. We define the potential

$$\psi_{P_W}(x) := \mathbb{E}_{w^0, \xi} \ln \mathbb{E}_w \exp \left(-\frac{1}{2} x w^2 + x w^0 w + \sqrt{x} \xi w \right).$$

This is the free entropy of a scalar Gaussian observation channel $y_G = \sqrt{x} w^0 + \xi$ with prior P_W on the signal w^0 . Parameter x plays the role of signal-to-noise ratio (SNR).

- (ii) Let $\xi, u, u^0 \sim \mathcal{N}(0, 1)$ be all independent. Define

$$\phi_{P_{\text{out}}}(x; r) := \int dy \mathbb{E}_{\xi, u^0} P_{\text{out}}(y | \sqrt{x} \xi + \sqrt{r - x} u^0) \times \ln \mathbb{E}_u P_{\text{out}}(y | \sqrt{x} \xi + \sqrt{r - x} u).$$

This is the free entropy associated with the scalar observation channel $y_{\text{out}} \sim P_{\text{out}}(\cdot | \sqrt{x} \xi + \sqrt{r - x} u^0)$ with Gaussian signal u^0 , given a quenched variable ξ .

- (iii) In contrast with the two previous free entropies, which are associated with scalar inference problems, $\iota(x)$ is the mutual information between signal and

data of a high-dimensional, yet tractable, problem: matrix denoising. In this inference problem, the goal is to recover a generalized Wishart matrix $\tilde{\mathbf{S}}^0 := \tilde{\mathbf{W}}^0 \text{diag}(\mathbf{v}^0) \tilde{\mathbf{W}}^0 / \sqrt{k d} \in \mathbb{R}^{d \times d}$ given the matrix observation $\mathbf{Y}(x) = \sqrt{x} \tilde{\mathbf{S}}^0 + \mathbf{Z}$. Here, $\tilde{\mathbf{W}}^0 \in \mathbb{R}^{k \times d}$ has i.i.d. standard Gaussian entries, the noise $\mathbf{Z}$ is a matrix from the Gaussian orthogonal ensemble (GOE, i.e. symmetric matrices with the upper triangular part made of entries i.i.d. from $\mathcal{N}(0, (1 + \delta_{ij})/d)$), and x is the SNR. The potential is then defined as $\iota(x) := \lim_{d \rightarrow \infty} (1/d^2) I(\mathbf{Y}(x), \tilde{\mathbf{S}}^0)$. It was conjectured [101,102] and proven [103] that this mutual information is linked to the HCIZ integral:

$$\iota(x) = \frac{x}{2} \int s^2 \rho_{\tilde{\mathbf{S}}^0}(s) ds - \lim_d \frac{1}{d^2} \ln \mathcal{Z}_{\text{HCIZ}}^{(1)}[\sqrt{x} \tilde{\mathbf{S}}^0, \mathbf{Y}(x)],$$

where $\rho_{\tilde{\mathbf{S}}^0}$ is the limiting spectral density of $\tilde{\mathbf{S}}^0$ as $d \rightarrow \infty$. The limit of the log-HCIZ integral is generally intractable in practice despite admitting a dimension-independent variational expression [128,129]. Luckily, the one needed in the present setting is explicit [102,103]. The most convenient expression for numerical evaluation is based on the so-called I-MMSE relation [103,184], which links the mutual information between the signal and the data to the minimum mean square error (MMSE) in matrix denoising.

Using the results of Ref. [103], the limiting MMSE for matrix denoising verifies

$$\begin{aligned} \text{MMSE}_S(x) &:= \lim_{d \rightarrow \infty} \frac{1}{d} \mathbb{E} \|\tilde{\mathbf{S}}^0 - \mathbb{E}[\tilde{\mathbf{S}}^0 | \mathbf{Y}(x)]\|^2 \\ &= \frac{1}{x} \left(1 - \frac{4\pi^2}{3} \int \rho_{\mathbf{Y}(x)}(y)^3 dy \right). \end{aligned}$$

Using this expression and the I-MMSE relation mentioned above, $\iota(x)$ admits a compact expression:

$$\iota(x) = \frac{1}{4} \int_0^x \text{MMSE}_S(t) dt.$$

- (iv) Consider now a rectangular matrix denoising problem with observations $\tilde{\mathbf{Y}}(x) = \sqrt{x/(pk)} \mathbf{U}^0 \mathbf{V}^0 + \mathbf{N}/\sqrt{p} \in \mathbb{R}^{p \times d}$, where $\mathbf{U}^0 \in \mathbb{R}^{p \times k}$, $\mathbf{V}^0 \in \mathbb{R}^{k \times d}$, and $\mathbf{N} \in \mathbb{R}^{p \times d}$ are all made of i.i.d. standard Gaussian entries. $\tilde{\iota}(x; \eta, \gamma)$ is then defined as the limit (when $d, k, p \rightarrow \infty$) of the mutual information $[1/(pd)] I[\tilde{\mathbf{Y}}(x); \mathbf{U}^0 \mathbf{V}^0]$ while fixing $p/d \rightarrow \eta$, and $k/d \rightarrow \gamma$. Similar to its symmetric version, the mutual information is computed by means of a “rectangular spherical integral” [185]; see Eq. (C28). The noise being Gaussian, we can again exploit the I-MMSE relation [184]. The MMSE function for this problem [165,166] is

$$\begin{aligned} \text{MMSE}(x; \eta, \gamma) &:= \lim_{d \rightarrow \infty} \frac{1}{pkd} \mathbb{E} \|\mathbf{U}^0 \mathbf{V}^0 - \mathbb{E}[\mathbf{U}^0 \mathbf{V}^0 | \tilde{\mathbf{Y}}(x)]\|^2 \\ &= \frac{1}{x} \left[1 - \int \left(\eta \left(\frac{1}{\eta} - 1 \right) \right)^2 y^{-2} \tilde{\rho}_{\tilde{\mathbf{Y}}(x)}(y) \right. \\ &\quad \left. - \frac{\pi^2 \eta}{3} \tilde{\rho}_{\tilde{\mathbf{Y}}(x)}(y)^3 \right] dy. \end{aligned}$$

Here, $\tilde{\rho}_{\tilde{\mathbf{Y}}(x)}$ is the limiting singular value density of $\tilde{\mathbf{Y}}(x)$, which is the so-called rectangular free convolution [165,185] between the asymptotic singular value density of $\sqrt{x/(pk)} \mathbf{U}^0 \mathbf{V}^0$ and a Marchenko-Pastur distribution of parameter η . The potential $\tilde{t}(x; \eta, \gamma)$ is then given by

$$\tilde{t}(x; \eta, \gamma) = \frac{1}{2} \int_0^x \text{MMSE}(t; \eta, \gamma) dt.$$

C. Shallow MLP

Starting with $L = 1$, our first result is a formula for the free entropy based on the Gaussianity assumption (11). The strategy to evaluate it, based on the replica method, relies on identifying the sufficient statistics (order parameters), the free entropy being related to their large deviations rate function. This formula will therefore also give us access to their equilibrium values. We note that before the present work, no existing method could tackle linear-width NNs in the interpolation regime with a generic activation even for this shallow case.

1. Order parameters

In the definitions below, a superscript $*$ emphasizes that the student is sampled at equilibrium $\boldsymbol{\theta} \sim dP(\cdot|\mathcal{D})$, and that the thermodynamic limit (4) is taken (even if not explicit). All the OPs involve the target $\boldsymbol{\theta}^0$ and student $\boldsymbol{\theta}$ weights. By Bayes optimality and the Nishimori identities (Appendix A 3), the target weights can be equivalently replaced by $\boldsymbol{\theta}' \sim dP(\cdot|\mathcal{D})$ coming from another independent student.

- (i) $R_2^* \propto \text{Tr}[\mathbf{W}^{0\top} \text{diag}(\mathbf{v}^0) \mathbf{W}^0 \mathbf{W}^\top \text{diag}(\mathbf{v}) \mathbf{W}]$ measures the alignment between the teacher's and student's quadratic terms, which is nontrivial with $n = \Theta(d^2)$ data even when the student is not able to reconstruct $\mathbf{W}^0$ itself (i.e., to specialize).
- (ii) $\mathcal{Q}^*(\mathbf{v}) \propto \sum_{\{i|v_i^0=\mathbf{v}\}} (\mathbf{W}^0 \mathbf{W}^\top)_{ii}$ measures the overlap between the teacher and student's inner weights that are connected to readouts with the same amplitude $\mathbf{v}$: $\mathcal{Q}^*(\mathbf{v}) \neq 0$ signals that the student learns part of $\mathbf{W}^0$. Thus, the specialization transition for the neurons connected to readouts with amplitude $\mathbf{v}$ is defined as

$$\alpha_{\text{sp},\mathbf{v}}(\gamma) := \sup\{\alpha | \mathcal{Q}^*(\mathbf{v}) = 0\}. \quad (12)$$

For nonhomogeneous readouts, the specialization transition is defined as

$$\alpha_{\text{sp}}(\gamma) := \min_{\mathbf{v}} \alpha_{\text{sp},\mathbf{v}}(\gamma) = \min_{\mathbf{v}} \sup\{\alpha | \mathcal{Q}^*(\mathbf{v}) = 0\}.$$

Associated with these OPs, the “hat variables” $\hat{R}_2^*, \hat{\mathcal{Q}}^*(\mathbf{v})$ in Result 1 are conjugate OPs. Their meaning is that of effective fields (called “cavity fields” in spin glasses), which self-consistently determine the OPs through the replica-symmetric saddle-point equations given in the result.

To state our first result, we need additional definitions. Let $\mathcal{Q}(\mathbf{v}), \hat{\mathcal{Q}}(\mathbf{v}) \in \mathbb{R}$ for $\mathbf{v} \in \text{Supp}(P_v)$, $\mathcal{Q} := \{\mathcal{Q}(\mathbf{v}) | \mathbf{v} \in \text{Supp}(P_v)\}$ and similarly for $\hat{\mathcal{Q}}$. Let also [see Eq. (B6)] for a more explicit expression of g

$$g(x) := \sum_{\ell \geq 3} x^\ell \mu_\ell^2 / \ell!,$$

$$K(x, \mathcal{Q}) := \mu_1^2 + \mu_2^2 x / 2 + \mathbb{E}_{v \sim P_v} v^2 g(\mathcal{Q}(v)),$$

$$K_d := \mu_1^2 + \mu_2^2 (1 + \gamma \bar{v}^2) / 2 + g(1). \quad (13)$$

The physical meaning of $K(\cdot, \cdot)$, when evaluated at the equilibrium $R_2^*, \mathcal{Q}^*$, is that of the covariance K^* appearing in Eq. (10) [i.e., the large d limiting covariance between two postactivations $\lambda_{\text{test}}(\boldsymbol{\theta}^a), \lambda_{\text{test}}(\boldsymbol{\theta}^b)$ evaluated from the same test input $\mathbf{x}_{\text{test}}$ but with weights $\boldsymbol{\theta}^a, \boldsymbol{\theta}^b$ i.i.d. from the posterior]; K_d is instead their variance, which matches that of the target by Bayes optimality.

2. Replica-symmetric formulas

We are ready to state the replica-symmetric (RS) formula giving access to the equilibrium order parameters. From now on, we denote the joint $d, k, n \rightarrow \infty$ limit with rates (4) simply by “lim”.

Result 1. (Replica-symmetric free entropy for the MLP with $L = 1$) Assume that $\mu_0 = 0$ in the Hermite decomposition (5). Let the functional

$$\tau(\mathcal{Q}) := \text{MMSE}_S^{-1}[1 - \mathbb{E}_{v \sim P_v} v^2 \mathcal{Q}(v)^2]. \quad (14)$$

The replica-symmetric formula for the limiting free entropy $\lim f_n$ is $f_{\text{RS}}^{(1)}(R_2^*, \hat{R}_2^*, \mathcal{Q}^*, \hat{\mathcal{Q}}^*)$ with RS potential $f_{\text{RS}}^{(1)} = f_{\text{RS}}^{(1)}(R_2, \hat{R}_2, \mathcal{Q}, \hat{\mathcal{Q}})$ which, given (α, γ) , reads

$$\begin{aligned} f_{\text{RS}}^{(1)} &:= \phi_{P_{\text{out}}}(K(R_2, \mathcal{Q}); K_d) + \frac{1}{4\alpha} (1 + \gamma \bar{v}^2 - R_2) \hat{R}_2 \\ &\quad + \frac{\gamma}{\alpha} \mathbb{E}_{v \sim P_v} \left[\psi_{P_w}(\hat{\mathcal{Q}}(v)) - \frac{1}{2} \mathcal{Q}(v) \hat{\mathcal{Q}}(v) \right] \\ &\quad + \frac{1}{\alpha} [\iota(\tau(\mathcal{Q})) - \iota(\hat{R}_2 + \tau(\mathcal{Q}))]. \end{aligned} \quad (15)$$

The order parameter equilibrium values $(R_2^*, \hat{R}_2^*, Q^*, \hat{Q}^*)$ are obtained from the RS saddle-point equations (B41) derived from the extremization condition $\nabla f_{\text{RS}}^{(1)} = \mathbf{0}$ as a solution (there may be more than one) maximizing $f_{\text{RS}}^{(1)}$.

Result 1 provides R_2^*, Q^* through the solution of a tractable variational problem. Consequently, under our joint-Gaussianity hypothesis (11) on $[\lambda_{\text{test}}(\theta^a)]_{a=0}^s$ with i.i.d. $\theta^a \sim dP(\cdot|\mathcal{D})$ for $a = 1, \dots, s$, we can also access their asymptotic covariance (and thus their law) given by

$$(\mathbf{K}^*)_{ab} = K^* + (K_d - K^*)\delta_{ab}, \quad K^* = K(R_2^*, Q^*). \quad (16)$$

The Bayes error can then be computed as in Appendix A 5.

Result 2. (Bayes generalization error) Let $(\lambda^a)_{a \geq 0} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}^*)$ with covariance (16), $y_{\text{test}}|\lambda^0 \sim P_{\text{out}}(\cdot|\lambda^0)$. Assume $\mathcal{C}$ has series expansion $\mathcal{C}(y, \hat{y}) = \sum_{i \geq 0} c_i(y)\hat{y}^i$. The RS formula for the $\lim \varepsilon^{\mathcal{C},f}$ of the Bayes error (7) is

$$\mathbb{E}_{(\lambda^a)} \mathbb{E}_{y_{\text{test}}|\lambda^0} \sum_{i \geq 0} c_i(y_{\text{test}}) \prod_{a=1}^i f(\lambda^a). \quad (17)$$

Letting $\mathbb{E}[\cdot|\lambda] = \int dy(\cdot) P_{\text{out}}(y|\lambda)$, the RS formula for the $\lim \varepsilon^{\text{opt}}$ of the Bayes-optimal mean-square generalization error (8) is

$$\mathbb{E}_{(\lambda^0, \lambda^1)} (\mathbb{E}[y^2|\lambda^0] - \mathbb{E}[y|\lambda^0]\mathbb{E}[y|\lambda^1]). \quad (18)$$

This result assumed $\mu_0 = 0$; see Appendix B 1 g if $\mu_0 \neq 0$. We remind the reader that Sec. III provides a generalization of the theory able to tackle structured input data (valid also for the deep case $L \geq 2$).

The presence of the HCIZ matrix integral in our replica formulas suggests that the usual asymptotic decoupling of the finite marginals of the posterior in terms of products of the single-variable marginals does not occur here, in contrast with standard Bayes-optimal inference problems [186]. In the related context of matrix denoising, this may explain why the approximate message-passing algorithms proposed in Refs. [169, 187, 188] are, as stated by the authors, not properly converging or matching their corresponding theoretical predictions based on the cavity method, as it relies on such decouplings.

Remark 3. No OP related to the readout weights appears in our results. The reason is the following. The $kd = \Theta(d^2)$ inner weights $\mathbf{W}^0$ and $n = \Theta(d^2)$ data are overwhelmingly many compared to the k unknowns $\mathbf{v}^0$, which thus contribute trivially to the leading order of the thermodynamic equilibrium quantities we aim for. Let us prove that the mutual information stays the same at leading order if the readouts are fixed to $\mathbf{v}^0$ rather than learnable or unknown. By the chain rule for mutual information $I[(\mathbf{W}^0, \mathbf{v}^0); \mathcal{D}] = I(\mathbf{W}^0; \mathcal{D}|\mathbf{v}^0) + I(\mathbf{v}^0; \mathcal{D})$. Moreover $I(\mathbf{v}^0; \mathcal{D}) = H(\mathbf{v}^0) - H(\mathbf{v}^0|\mathcal{D})$. For a discrete-valued $\mathbf{v}^0$, both of these

Shannon entropies are non-negative. Additionally, $H(\mathbf{v}^0) = O(k)$. Because $H(\mathbf{v}^0) \geq H(\mathbf{v}^0|\mathcal{D})$, then $H(\mathbf{v}^0|\mathcal{D}) = O(k)$, too. Therefore, in the limit (4),

$$\frac{1}{n} I[(\mathbf{W}^0, \mathbf{v}^0); \mathcal{D}] = \frac{1}{n} I(\mathbf{W}^0; \mathcal{D}|\mathbf{v}^0) + O(1/d),$$

and similarly for the free entropy. The same also holds for the generalization error given its link with the mutual information; see Ref. [100]. The argument can be extended to continuous-valued readouts and $L \geq 2$.

Another way to understand this is through the symmetry of the NNs under permutation of their k hidden neurons. It implies that only the law of $\mathbf{v}^0$ matters. Consequently, if one draws $\mathbf{v}'$ from the correct P_v and fixes it in the student (thus learning only $\mathbf{W}$), it will have the same law as $\mathbf{v}^0$ (up to small fluctuations) and is therefore equally good as $\mathbf{v}^0$ when $d \gg 1$. This implies that in the Bayes-optimal setting, knowing P_v is equivalent to $\mathbf{v}^0$ for large d . For additional illustration, in the paper we test our theory with numerical experiments both with fixed and learnable readouts, as explicit from the caption of each figure. Moreover, Fig. 30 of Appendix B 5 (obtained with learnable readouts), to be compared directly with Figs. 5 (right) and 7 (bottom), obtained with fixed readouts, is showing that equilibrated Bayesian NNs achieve the same generalization performance independent of whether the readouts are trainable or fixed to the truth. The same holds for $L = 2$ (see Fig. 31). Figure 4 is another confirmation because the theory there, which describes well the empirical distribution of postactivations over test samples, is done fixing directly the readouts, while they are trained in the experiments for this figure. However, the readouts being fixed or learnable does influence the learning dynamics (see, e.g., the difference for ADAM in Figs. 9 and 27), but its theoretical analysis is beyond the scope of this paper.

3. Main steps of the replica approach

Before moving to the results for deep networks, let us provide a summary of the main steps of the derivation of Result 1, highlighting the conceptual novelties (the main ideas are similar for the deeper architectures we consider later on). More details are given in Sec. IV and Appendix B. We adopt the replica method but encounter and resolve a major technical obstacle that is absent in standard applications.

- (i) *Standard replica setup:* Having quenched the readouts, which is possible by the last remark, the replica trick $\lim_{n \rightarrow \infty} (1/n) \mathbb{E} \ln Z = \lim_{s \rightarrow 0^+} \lim_{n \rightarrow \infty} [1/(ns)] \ln \mathbb{E} Z^s$ [126] introduces s independent posterior samples of the inner weights (replicas) $\{\mathbf{W}^a\}_{a=1}^s$ alongside the teacher's $\mathbf{W}^0$. Under our main Gaussian hypothesis (11), the key quantity governing learning is the covariance of the replicated postactivations $\lambda^a = (1/\sqrt{k}) \mathbf{v}^{0\top} \sigma[(1/\sqrt{d}) \mathbf{W}^a \mathbf{x}]$ under the shared input

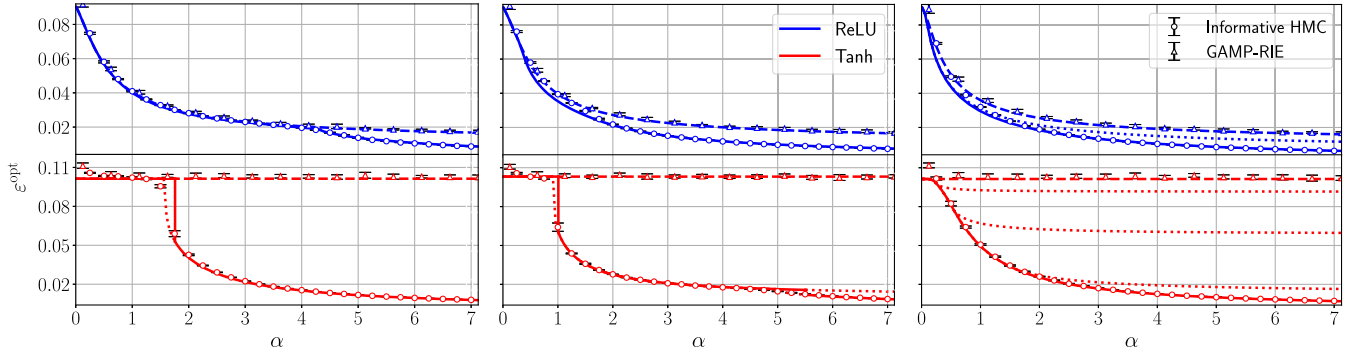


FIG. 5. Theoretical prediction (solid curves) of the Bayes-optimal mean-square generalization error (with the irreducible error Δ removed) for $L = 1$ with Gaussian inner weights, $\text{ReLU}(x)$ (blue curves), and $\tanh(2x)$ activation (red curves), $d = 200$, $\gamma = 0.5$, $\Delta = 0.1$, and different P_v laws. Dashed and dotted lines denote, respectively, the universal and multiple specialization branches where they are metastable (i.e., a solution of the RS equations not corresponding to the equilibrium). The readouts are fixed to the teacher's during sampling. Left: homogeneous readouts. Center: four-point readouts. Right: Gaussian readouts. For these two latter cases, the specialization transitions correspond to partial specialization (of just some neurons). The numerical points correspond to the half Gibbs error obtained with HMC with informative initialization on the target. Triangles are the error of GAMP-RIE [96] extended to generic activation; see Appendix B 4. Each point has been averaged over 12 instances of the training set (including the teacher). Error bars are the standard deviation over instances. The empirical test error is computed empirically from 10^5 i.i.d. test samples.

$\mathbf{x}$: $K^{ab} = \mathbb{E}_{\mathbf{x}}[\lambda^a \lambda^b]$. Using the Hermite expansion of σ and Mehler's formula (A9), this covariance decomposes as an infinite series of overlap order parameters $R_\ell^{ab} := (1/k) \sum_{i,j} v_i^0 v_j^0 (\Omega_{ij}^{ab})^\ell$, where $\Omega_{ij}^{ab} := \mathbf{W}_i^{a\top} \mathbf{W}_j^b / d$.

- (ii) *Reduction to fundamental order parameters*: A first nontrivial step is reducing this infinite family of overlaps (R_ℓ^{ab}) to a small manageable set. The first ingredient is to assume that specialization of neurons can happen only in a one-to-one manner by Bayes optimality (i.e., a neuron of the student cannot align with more than one teacher neuron). Consequently, all but one entry per row or column of the overlap matrix $(\Omega_{ij}^{ab})_{i,j \leq k}$ are $O(1/\sqrt{d})$. The second ingredient is that the permutation symmetry among neurons can be broken in a way that the k entries of the overlap matrix that are potentially $O(1)$ are put on the diagonal $(\Omega_{ii}^{ab})_{i \leq k}$, which amounts to select the “natural” permutation of student neurons with respect to the teacher (and similarly in any given layer in the $L \geq 2$ -hidden-layers case). The permutation symmetry can then be discarded, which is justified by the fact that the number of such permutations grows much slower than $O(\exp(n))$, and therefore plays no role at leading exponential order in the large system limit. The third ingredient is that the “Wishart-like” overlap matrix—with $O(1)$ diagonal and $O(1/\sqrt{d})$ off diagonal—taken at the ℓ th Hadamard power $(\Omega_{ij}^{ab})^\ell$ can be considered diagonal in the thermodynamic limit (4) for $\ell \geq 3$: The off diagonal becomes too weak to be relevant. This procedure leaves only three types of OPs: the linear overlap $R_1^{ab} \rightarrow 1$ [i.e., the linear

component of the target is perfectly learnable at $n = \Theta(d^2)$ and requires no further tracking], the quadratic overlap R_2^{ab} associated with the second Hermite component, and the overlaps $\Omega_{ii}^{ab} = Q^{ab}(\mathbf{v})$ for any hidden unit $i \in [k]$ connected to a readout with amplitude $\mathbf{v}$ (by symmetry among these neurons). This step yields the reduction of OPs via $R_\ell^{ab} \approx (1/k) \sum_i (v_i^0)^2 (\Omega_{ii}^{ab})^\ell \approx \mathbb{E}_{v \sim P_v} v^2 Q^{ab}(v)^\ell$ for $\ell \geq 3$. The covariance then takes the compact form $K^{ab} = K[R_2^{ab}, Q^{ab}(\cdot)]$ given in Eq. (13).

- (iii) *The entropic obstacle for matrix degrees of freedom*: In a standard narrow-network replica calculation, the weights are a few vectors. Consequently, the overlap order parameters are few neuron-neuron overlaps and their entropic degeneracy factorizes over components, yielding a scalar free-entropy term of the type ψ_{P_w} . Here the situation is fundamentally different: The second-order overlap $R_2^{ab} = (1/d^2) \text{Tr} \mathbf{S}_2^a \mathbf{S}_2^b$ depends on the matrix $\mathbf{S}_2^a := \mathbf{W}^{a\top} \text{diag}(\mathbf{v}^0) \mathbf{W}^a / \sqrt{k} \in \mathbb{R}^{d \times d}$. Fixing R_2^{ab} therefore imposes a quartic constraint on the weight entries (W_{ia}^a), rather than the quadratic constraints arising in standard problems. A direct computation of the entropy of (R_2^{ab}) conditional on $\mathcal{Q} := \{Q^{ab}(\mathbf{v}) | \mathbf{v} \in \mathbf{V}, a \leq b\}$, which is of the form

$$\ln \int dP((\mathbf{S}_2^a)_a | \mathcal{Q}) \prod_{a \leq b}^{0,s} \delta(d^2 R_2^{ab} - \text{Tr} \mathbf{S}_2^a \mathbf{S}_2^b)$$

for some complicated measure $dP[(\mathbf{S}_2^a)_a | \mathcal{Q}]$ is therefore out of reach via standard methods. In

contrast, the entropy of $\mathcal{Q}$ is easy to compute using Gaussian tricks.

- (iv) *Measure relaxation by moment matching*: The resolution is to shift perspective: Rather than integrating over the weight entries W_{ia}^a subject to the constraints on R_2^{ab} , we work directly with the matrices $\mathbf{S}_2^a$ as the fundamental degrees of freedom. Their true conditional law $dP[(\mathbf{S}_2^a)_a | \mathcal{Q}]$ is too complex; it is a coupled, non-Gaussian measure over generalized Wishart-like matrices. We replace it by a simpler effective measure $d\tilde{P}[(\mathbf{S}_2^a)_a | \mathcal{Q}] = d\tilde{P}[(\mathbf{S}_2^a)_a | \tau(\mathcal{Q})]$. The effective measure enforces what we claim to be the only relevant characteristics in the large system limit of the original measure $dP[(\mathbf{S}_2^a)_a | \mathcal{Q}]$, but it has the advantage of being tractable. It depends on a tilt parameter $\tau(\mathcal{Q})$ determined by enforcing exact moment matching: The second moment $\mathbb{E}_{\tilde{P}} \text{Tr} \mathbf{S}_2^a \mathbf{S}_2^b$ must reproduce the value computed under the true measure, which is accessible. The measure simplification we introduce differs crucially from the earlier approach [168] (developed for matrix factorization), which considers a factorized Gaussian effective measure, thereby discarding the Wishart-type correlations among entries of $\mathbf{S}_2^a$. Our Wishart-based relaxation retains the essential matrix correlations while remaining tractable and leads to improved agreement with experiments.
- (v) *Spherical integration*: With the effective measure in place, the entropic contribution of $R_2^{ab} = R_2$ reduces—after standard replica algebra, Hubbard-Stratonovich linearization of inter-replica couplings, and the $s \rightarrow 0^+$ limit under the replica-symmetric Ansatz [126]—to the free entropy of a rotationally invariant matrix denoising problem. This free entropy is computable via the HCIZ spherical integral yielding the two $\iota(\cdot)$ terms in $f_{\text{RS}}^{(1)}$ (15). The function $\text{MMSE}_S(\cdot)$ is the MMSE of this auxiliary denoising problem, and the moment-matching equation relating true to effective measure is Eq. (14). Via this equation, the matrix denoising problem interacts with the specialization order parameter $\mathcal{Q}^{ab}(\cdot) = \mathcal{Q}(\cdot)$. The entropy of $\mathcal{Q}$ yields instead the free entropy of a standard scalar Gaussian denoising problem (one per $\mathbf{v}$), i.e., the terms ψ_{P_w} .

Summary of the structure of the replica symmetric free entropy. The resulting free entropy $f_{\text{RS}}^{(1)}$ admits a transparent decomposition. The term $\phi_{P_{\text{out}}}$ is the standard output channel free entropy identical in form to what appears in the GLM calculation [100]. The terms ψ_{P_w} are the scalar Gaussian channel free entropies encoding the prior on the weights and governing the specialization transition(s). The two ι terms encode the matrix nature of the problem through the auxiliary matrix denoising problem and are the only places where the HCIZ integral enters. The

interplay between the mean-field specialization mechanism (encoded in ψ_{P_w}) and the matrix model (encoded in ι) is what generates the rich phase diagram of the model and constitutes the principal conceptual novelty of the approach with respect to prior works that treated these two aspects separately or only partially.

D. Two-hidden-layer MLP

For $L = 2$, we consider activations without zeroth and second Hermite components; see (H_2) . The results are obtained by an expansion of the nested activations in the Hermite basis. This produces different terms that can be interpreted as equivalent subnetworks with “effective” readouts and inner weights built as combinations of the original ones, as detailed in Sec. IV C. When the linear component of the last activation is involved, the readouts $\mathbf{v}$ combine with the second layer inner weights and give rise to “effective readouts” $\mathbf{v}^{(2)} := \mathbf{W}^{(2)\top} \mathbf{v} / \sqrt{k_2}$ that act on the nonlinear first layer. By binning through finite discretization the distribution of the components of this vector, we denote the admitted amplitudes as $\mathbf{v}^{(2)}$. Similarly, when the linear component of the first layer is considered, the two sets of inner weights combine in an effective layer with weights $\mathbf{W}^{(2;1)} := \mathbf{W}^{(2)} \mathbf{W}^{(1)} / \sqrt{k_1}$, which can be reconstructed partly independently of its factors $\mathbf{W}^{(2)}$ and $\mathbf{W}^{(1)}$, and thus comes with an OP.

1. Order parameters

Already with two hidden layers, the OPs detailed below describe a much richer phase diagram than in the shallow case. Until now, it was unclear which OPs should be tracked.

- (i) $\mathcal{Q}_1^*(\mathbf{v}^{(2)}) \propto \sum_{\{i|v_i^{(2)0}=\mathbf{v}^{(2)}\}} (\mathbf{W}^{(1)0} \mathbf{W}^{(1)\top})_{ii}$ is the overlap between the teacher and student’s first-layer weights connected to the effective readouts $\mathbf{v}^{(2)}$ with amplitude $\mathbf{v}^{(2)}$. As in Remark 3, the vector $\mathbf{v}^{(2)}$ can be treated as quenched on the teacher’s one. In virtue of this, from its definition, $\mathbf{v}^{(2)}$ has Gaussian distributed entries by the central limit theorem.
- (ii) $\mathcal{Q}_2^*(\mathbf{v}, \mathbf{v}^{(2)}) \propto \sum_{\{i,j|v_i^0=\mathbf{v}, v_j^{(2)0}=\mathbf{v}^{(2)}\}} W_{ij}^{(2)0} W_{ij}^{(2)}$ is the overlap for the second layer. It is labeled by two values. The first, $\mathbf{v}$, for the shallow case, is the value of a readout. It takes into account the learning inhomogeneity along the output dimension ($i \leq k_2$) of the second-layer weight matrix induced by the readouts $\mathbf{v}$. $\mathbf{v}^{(2)}$ is instead the same variable labeling $\mathcal{Q}_1^*(\mathbf{v}^{(2)})$. It captures the inhomogeneity along the input dimension ($j \leq k_1$) of the second layer induced by the inhomogeneity of the first-layer output, itself induced by (and therefore labeled according to) the effective readouts $\mathbf{v}^{(2)}$. Notice that this implies a nontrivial feedback loop of interactions: Inhomogeneities of $\mathbf{W}^{(2)}$ influence $\mathbf{W}^{(1)}$ via $\mathbf{v}^{(2)}$, and at the same time the inhomogeneities in

$\mathbf{W}^{(1)}$'s rows influence the columns of $\mathbf{W}^{(2)}$ directly.

We wish to emphasize a conceptually important point. $\mathcal{Q}_2^*(\mathbf{v}, \mathbf{v}^{(2)})$ being a matrix may lead us to believe that it does not help in reducing the dimensionality of the problem, because the “microscopic degrees of freedom” are weight matrices, too. However, $\mathbf{v}$ and $\mathbf{v}^{(2)}$ are indexing intensive, d -independent dimensions. Indeed, the binning of $\mathcal{K}_{12} := \{1, \dots, k_1\} \times \{1, \dots, k_2\}$ in terms of the non-overlapping sets $\{i, j | v_i^0 = \mathbf{v}, v_j^{(2)0} = \mathbf{v}^{(2)}\}$ entering $\mathcal{Q}_2^*$'s definition (i.e., the mapping from $\mathcal{K}_{12}$ to $\{\mathbf{v}\} \times \{\mathbf{v}^{(2)}\}$) is done as follows: First, $\mathcal{K}_{12}$ is partitioned into finitely many “macroscopic” sets. Second, the thermodynamic limit $d \rightarrow +\infty$ is taken; see Eq. (4). Finally, only after this limit is the number of bins allowed to diverge. This implies that each set always includes a number of terms growing to infinity as d^2 times a small constant. Consequently, $\mathcal{Q}_2^*(\mathbf{v}, \mathbf{v}^{(2)})$ (or any OP function with continuous argument) is a proper macroscopic (or intensive) OP summarizing the behavior of a large assembly of degrees of freedom for each pair of arguments. Dimensionality reduction therefore takes place and justifies the use of saddle-point integration with respect to the OPs when evaluating the (replicated) log-partition function in Sec. IV.

- (iii) Lastly, $\mathcal{Q}_{2:1}^*(\mathbf{v}) \propto \sum_{\{i | v_i^0 = \mathbf{v}\}} (\mathbf{W}^{(2:1)0} \mathbf{W}^{(2:1)\top})_{ii}$ is the teacher-student overlap for specific rows of $\mathbf{W}^{(2:1)}$. This OP arises from the linear term in the Hermite expansion of the inner activation. It is needed because the product $\mathbf{W}^{(2:1)}$ between the first- and second-layer weights can in principle be learned partly independently of $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$. Observe that $\mathbf{W}^{(2:1)}$ “connects” the input directly to the output, which is why it is labeled only by the readout values $\mathbf{v}$.

The reader eager to already gain intuition on the behavior of these functional “vector” and “matrix order parameters” can look at Figs. 15 and 16. We remark that extrapolating to the linear-width setting, the replica techniques successful for narrow NNs also yield overlap matrix order parameters, but with a prohibitive dimension. However, if in the shallow case one takes $k \rightarrow \infty$ after $d \rightarrow \infty$, simple parametrizations of the $k \times k$ overlaps allow us to solve it [28]. This double limit is, however, not an extensive-width limit and, indeed, the resulting formulas are similar to those for GLMs [189–191].

For $L = 2$, specialization transitions can happen layer-wise: We defined the specialization transitions as

$$\alpha_{\text{sp},l} := \sup\{\alpha | \mathcal{Q}_l^* \equiv 0\} \quad \text{for } l = 1, 2,$$

where $\mathcal{Q}_l^* \equiv 0$ means the constant null function. Keep in mind that a nonvanishing overlap $\mathcal{Q}_{2:1}^*$ entails another kind of learning mechanism other than specialization.

Our result for the two-hidden-layer MLP requires the following function: Letting $v^{(2)} \sim \mathcal{N}(0, 1)$, $v \sim P_v$,

$$\begin{aligned} K^{(2)}(\bar{\mathcal{Q}}) &:= \mu_1^4 + \mu_1^2 \mathbb{E}_{v^{(2)}} (v^{(2)})^2 g(\mathcal{Q}_1(v^{(2)})) \\ &\quad + \mathbb{E}_v v^2 g\left(\mu_1^2 \mathcal{Q}_{2:1}(v)\right) \\ &\quad + \mathbb{E}_{v^{(2)}} \mathcal{Q}_2(v, v^{(2)}) g(\mathcal{Q}_1(v^{(2)})) \Big), \end{aligned}$$

with $\bar{\mathcal{Q}} := \{\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_{2:1}\}$, which are functions of $\mathbf{v}, \mathbf{v}^{(2)}$. Analogous notations hold for the conjugate OPs $\hat{\mathcal{Q}}_1, \hat{\mathcal{Q}}_2, \hat{\mathcal{Q}}_{2:1}$. The meaning of $K^{(2)}(\bar{\mathcal{Q}}^*)$ evaluated at equilibrium is, as in the shallow case, that of asymptotic covariance between different replicas of the postactivation with the same test input as Eq. (10):

$$(\mathbf{K}^*)_{ab} = K^* + (1 - K^*)\delta_{ab}, \quad K^* = K^{(2)}(\bar{\mathcal{Q}}^*). \quad (19)$$

That the variance is 1 is a consequence of our convention $\mathbb{E}_{z \sim \mathcal{N}(0,1)} \sigma(z)^2 = 1$, which greatly simplifies notations in the deep case; see Appendix C 1 for an explanation.

2. Replica-symmetric formula

Recall the definitions of $\text{MMSE}(x; \eta, \gamma), \tilde{l}(x; \eta, \gamma)$ in Sec. II. The equilibrium OPs are determined by the following RS formula:

Result 3. (Replica-symmetric free entropy for the MLP with $L = 2$) Consider an activation with $\mu_0 = \mu_2 = 0$ in Eq. (5) and $\mathbb{E}_{z \sim \mathcal{N}(0,1)} \sigma(z)^2 = 1$. Let $v^{(2)} \sim \mathcal{N}(0, 1)$, $v \sim P_v$. Define $\text{MMSE}_v := \text{MMSE}[\cdot; \gamma_2 P_v(\mathbf{v}), \gamma_1]$, $\tilde{l}_v := \tilde{l}[\cdot; \gamma_2 P_v(\mathbf{v}), \gamma_1]$ and $\tau_v = \tau_v(\mathcal{Q}_1, \mathcal{Q}_2)$ solves

$$\text{MMSE}_v(\tau_v) := 1 - \mathbb{E}_{v^{(2)}} \mathcal{Q}_2(\mathbf{v}, v^{(2)}) \mathcal{Q}_1(v^{(2)}).$$

The RS formula for the limiting free entropy $\lim f_n$ for the MLP with $L = 2$ hidden layers is given by $f_{\text{RS}}^{(2)}(\mathcal{Q}_1^*, \hat{\mathcal{Q}}_1^*, \mathcal{Q}_2^*, \hat{\mathcal{Q}}_2^*, \mathcal{Q}_{2:1}^*, \hat{\mathcal{Q}}_{2:1}^*)$ with RS potential

$$\begin{aligned} f_{\text{RS}}^{(2)} &:= \phi_{P_{\text{out}}} (K^{(2)}(\bar{\mathcal{Q}}); 1) \\ &\quad + \frac{\gamma_1}{\alpha} \mathbb{E} \left[\psi_{P_{w_1}}(\hat{\mathcal{Q}}_1(v^{(2)})) - \frac{1}{2} \mathcal{Q}_1(v^{(2)}) \hat{\mathcal{Q}}_1(v^{(2)}) \right] \\ &\quad + \frac{\gamma_1 \gamma_2}{\alpha} \mathbb{E} \left[\psi_{P_{w_2}}(\hat{\mathcal{Q}}_2(v, v^{(2)})) - \frac{1}{2} \mathcal{Q}_2(v, v^{(2)}) \hat{\mathcal{Q}}_2(v, v^{(2)}) \right] \\ &\quad + \frac{\gamma_2}{\alpha} \mathbb{E} \left[\frac{\hat{\mathcal{Q}}_{2:1}(v)}{2} (1 - \mathcal{Q}_{2:1}(v)) \right. \\ &\quad \left. - \tilde{l}_v(\tau_v + \hat{\mathcal{Q}}_{2:1}(v)) + \tilde{l}_v(\tau_v) \right]. \end{aligned}$$

The order parameter equilibrium values $(Q_1^*, \hat{Q}_1^*, Q_2^*, \dots)$ are obtained from the RS saddle-point equations (C31) derived from the extremization condition $\nabla f_{\text{RS}}^{(2)} = \mathbf{0}$ as a solution (there may be more than one) maximizing $f_{\text{RS}}^{(2)}$.

Deducing the Bayes error is done as in the shallow case: From Result 3, we get $\bar{Q}^*$ and thus the covariance $\mathbf{K}^*$ given by Eq. (19), which simply replaces Eq. (16) in Result 2.

E. Three-or-more hidden-layer MLP

1. Order parameters

For $L \geq 3$, we consider activations verifying (H_3) . In this setting, our theory predicts specialization of all layers as the only nontrivial learning mechanism. Accordingly, the OPs are as follows:

- (i) $Q_l^* \propto \text{Tr}(\mathbf{W}^{(l)0} \mathbf{W}^{(l)\top})$ for $l \leq L-1$ are the teacher-student layerwise overlaps. They are simple scalars rather than functions: Indeed, the neurons in all layers but the last enter the theory in a symmetric way, such that we can freely sum over their indices.
- (ii) $Q_L^*(\mathbf{v}) \propto \sum_{\{i|v_i^0=\mathbf{v}\}} (\mathbf{W}^{(L)0} \mathbf{W}^{(L)\top})_{ii}$ is the overlap between teacher and student's L th layer weights connected to the readouts $\mathbf{v}^0$ with amplitude $\mathbf{v}$. As before, weights connected to larger readouts are learned from fewer data.

For the considered class of activation functions, a single specialization transition occurs jointly for all layers at

$$\alpha_{\text{sp}} := \sup\{\alpha | Q_1^* = \dots = Q_{L-1}^* = 0 \text{ and } Q_L^* \equiv 0\}.$$

Redefining $\bar{Q} := \{(Q_l)_{l \leq L-1}, Q_L\}$, and letting $v \sim P_v$, we introduce

$$K^{(L)}(\bar{Q}) := \mathbb{E}_v v^2 g \left(Q_L(v) g(Q_{L-1} g(\dots Q_2 g(Q_1) \dots)) \right).$$

The asymptotic covariance $\mathbf{K}^*$ between replicas of the postactivation with same test input is of the form (19) with $K^* = K^{(L)}(\bar{Q}^*)$.

2. Replica-symmetric formula

The equilibrium order parameters and Bayes error are derived from the replica-symmetric formula below, which applies to MLPs with any number of layers as long as σ verifies (H_3) and has normalized variance.

Result 4. (Replica-symmetric free entropy for the MLP with arbitrary L) Consider an activation σ with $\mu_0 = \mu_1 = \mu_2 = 0$ in Eq. (5) such that $\mathbb{E}_{z \sim \mathcal{N}(0,1)} \sigma(z)^2 = 1$. The replica-symmetric formula for the limiting free entropy $\lim f_n$ for the MLP with L hidden layers is given by

$f_{\text{RS}}^{(L)}(Q_1^*, \hat{Q}_1^*, \dots, Q_{L-1}^*, \hat{Q}_{L-1}^*, Q_L^*, \hat{Q}_L^*)$ with RS potential

$$\begin{aligned} f_{\text{RS}}^{(L)} &:= \phi_{P_{\text{out}}}(K^{(L)}(\bar{Q}); 1) \\ &+ \frac{\gamma_{L-1} \gamma_L}{\alpha} \mathbb{E}_{v \sim P_v} \left[\psi_{P_{W_L}}(\hat{Q}_L(v)) - \frac{1}{2} Q_L(v) \hat{Q}_L(v) \right] \\ &+ \sum_{l=1}^{L-1} \frac{\gamma_{l-1} \gamma_l}{\alpha} \left[\psi_{P_{W_l}}(\hat{Q}_l) - \frac{1}{2} Q_l \hat{Q}_l \right], \end{aligned}$$

where $\gamma_0 := 1$. The order parameters equilibrium values $\dots^*$ are obtained from the RS saddle-point equations derived from the extremization condition $\nabla f_{\text{RS}}^{(L)} = \mathbf{0}$, as a solution (there may be more than one) maximizing $f_{\text{RS}}^{(L)}$.

The Bayes error follows by plugging $\mathbf{K}^*$ into Result 2. Our results provide a precise quantitative theory for the sufficient statistics and generalization capabilities of shallow and deep Bayesian MLPs with data generated by a random MLP target with matched architecture, for broad classes of activations and weight distributions.

Remark 4. For $L = 1$, we conjecture that our theory is exact for activations σ with $\mu_2 = 0$. This is strengthened by a partial proof in Appendix B 6. The case $\mu_2 \neq 0$ is special, as it involves the HCIZ integral with possibly approximative steps; see the discussion in Appendix B 2 b. When the theory does not rely on matrix integrals, the assumptions we make and which we believe are exact are mostly (and in order) (i) the Gaussian hypothesis (11) on preactivations which can be accurately tested, (ii) that entries of Wishart-like overlap matrices can be considered small with respect to their diagonal when taken at a large enough power, and (iii) the identification and indexing of the OPs as well as their concentration in the thermodynamic limit (4) (i.e., replica symmetry), which is justified in the Bayes-optimal setting we consider [186,192].

For NNs with arbitrary L , we are confident that Result 4 is exact as, again, matrix integrals do not appear. See Fig. 19 that confirms its high accuracy. Another nice property is its simplicity. Even more so if $P_v = \delta_1$ when all OPs are scalars. Yet, it takes into account all key aspects of the model: its depth, the linear width of the layers, and the interpolation regime. Consequently, despite not capturing all intricacies emerging when considering a more general σ , it has a high pedagogical value.

Beyond these cases, when matrix integrals appear in the formulas, due to the unconventional nature of their derivation we cannot confidently assess or discard their exactness despite their excellent match with numerics. One reason is that it is numerically difficult to test our theory against the rigorous result [97] for the special case $L = 1$, $\sigma(x) = x^2$, $P_W = \mathcal{N}(0, 1)$ that they cover. When numerically solving the extremization of Eq. (15), the saddle-point equations seem to predict a maximizer at $Q(\mathbf{v}) > 0$ when $\gamma \lesssim 1$. The equations of Ref. [97] instead match the universal branch of the theory, i.e., $Q(\mathbf{v}) = 0 \forall \mathbf{v}$, for

any (α, γ) . Yet, we cannot confidently discard the exactness of the theory because the difference between the correct free entropy and the predicted one never exceeds approximately 1%: Our RS potential is very flat in $\mathcal{Q}$. It could be that the true maximizer is at $\mathcal{Q}(\mathbf{v}) = 0$ even when $\gamma \lesssim 1$, and that we observe otherwise due to numerical errors. Indeed, evaluating the spherical integrals $\iota(\cdot)$ in $f_{\text{RS}}^{(1)}$ is challenging, in particular when γ is small. Actually, for $\gamma \gtrsim 1$ we correctly get that $\mathcal{Q}(\mathbf{v}) = 0$ is the maximizer.

III. TESTING THE THEORY, AND ALGORITHMIC INSIGHTS

A. Experimental setting

In this section, we compare our theory with simulations. For all experiments but the ones in the dedicated paragraph on structured data, we use standard Gaussian input vectors $\mathbf{x}_\mu \sim \mathcal{N}(\mathbf{0}, I_d)$. We tested both the case of frozen and learnable readouts. For the equilibrium values obtained through sampling algorithms, it makes no difference, as explained in Remark 3 and Appendix B 7, and further tested in Figs. 4 and 30. For the ADAM optimizer we tested, this can change its dynamics but the overall conclusions remain the same.

We consider three different priors for the readouts: the standard Gaussian prior $P_v = \mathcal{N}(0, 1)$, homogeneous readouts $P_v = \delta_1$, and the four-point prior $P_v = \frac{1}{4}(\delta_{-3/\sqrt{5}} + \delta_{-1/\sqrt{5}} + \delta_{1/\sqrt{5}} + \delta_{3/\sqrt{5}})$ (which is centered and has unit variance). To reduce finite-size sampling fluctuations we fix the empirical frequencies of the entries in each readout vector, rather than sampling them. For the four-point prior, this means enforcing 1/4 frequencies to each symbol. For the Gaussian prior, we use an almost-deterministic Gaussian readout: The k entries are set to the population quantiles of a standard normal. The case of random Gaussian readouts is presented in Appendix B 7. For the activation functions, we remind the reader about the hypotheses (H_1) , (H_2) , (H_3) depending on the NN depth L . We consider polynomial activations made of sums of Hermite polynomials used in conjunction with Rademacher inner weights $P_W = \frac{1}{2}(\delta_{-1} + \delta_1)$ in Fig. 6. For the other figures on standard Gaussian inner weights for all hidden layers, we take $\sigma(x) = \text{ReLU}(x)$ as an example of activation with both $\mu_1 \neq 0$ and $\mu_2 \neq 0$, and $\sigma(x) = \tanh(2x)$ with $\mu_1 \neq 0$ but $\mu_2 = 0$. We also consider its normalized version which is analytically convenient (but not necessary) when there is more than one hidden layer: $\sigma(x) = \tanh(2x)/\sigma_{\tanh}$ with $\sigma_{\tanh}$ enforcing $\mathbb{E}_{z \sim \mathcal{N}(0,1)} \sigma(z)^2 = 1$.

In all experiments, we consider the regression task with linear readout and Gaussian label noise of variance Δ . We thus focus on the mean-square generalization error. We always remove from it the irreducible error present in definition (18) for the linear readout $\epsilon^{\text{opt}} \rightarrow \epsilon^{\text{opt}} - \Delta$, still denoting it ϵ^{opt} by a slight abuse of notation.

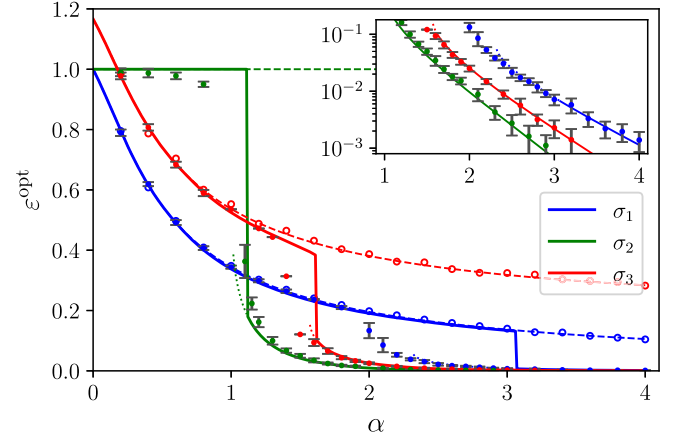


FIG. 6. Theoretical prediction (solid curves) of the Bayes-optimal mean-square generalization error for $L = 1$ with binary inner weights and polynomial activations: $\sigma_1 = \text{He}_2/\sqrt{2}$, $\sigma_2 = \text{He}_3/\sqrt{6}$, $\sigma_3 = \text{He}_2/\sqrt{2} + \text{He}_3/6$, with $\gamma = 0.5, d = 150, \Delta = 1.25$, and quenched homogeneous readouts $\mathbf{v} = \mathbf{1}$. Dots are the half Gibbs error computed using the Metropolis-Hastings algorithm initialized informatively. Circles are the error of GAMP-RIE [96] extended to generic activation. Points are averaged over 16 data and teacher instances. Error bars for the MCMC are the standard deviation over instances (omitted for GAMP-RIE, but of the same order). Dashed and dotted lines denote, respectively, the universal and specialization branches where they are metastable.

B. Probing the solutions of the RS saddle-point equations

The various theoretical errors (with Δ removed) we will analyze are all obtained from the same formula:

$$\epsilon^\square = \mathbb{E}_{(\lambda^0, \lambda)} [(\lambda^0)^2 - \lambda^0 \lambda] = K_d - K^\square, \quad (20)$$

where $(\lambda^0, \lambda) \sim \mathcal{N}(\mathbf{0}, \mathbf{K}^\square)$, with $\square \in \{*, \text{uni}, \text{sp}\}$ and $K^\square, K_d$ are, respectively, the covariance off diagonal and diagonal (the latter being 1 for $L \geq 2$). For $L = 1$, $\mathbf{K}^\square$ has the form (16), but where the equilibrium solution $*$ of the RS saddle-point equations (simply called “RS equations” from now on) can also be replaced by the universal solution (or branch) yielding ϵ^{uni} or by the specialization solution yielding ϵ^{sp} . The latter probes the performance of a Bayesian student initialized in the vicinity of the target rather than completely randomly. The equilibrium solution corresponds to the Bayes-optimal error: $\epsilon^* = \epsilon^{\text{opt}}$. In the same way, for $L \geq 2$, $\mathbf{K}^\square$ generalizes Eq. (19) using $K^{(2)}, K^{(L)}$, and we get the errors similarly.

We now explain how to concretely find the solutions of the RS equations. For any L , the universal solution is obtained using the fully uninformative initialization, i.e., setting all physical order parameters to 0 in the RS equations and then solving them by fixed-point iterations (the conjugate OPs never require an initialization). In

contrast, the specialization solution is obtained from the fully informative initialization where all physical OPs start from a strictly positive value (generally close to 1 to speed up convergence). When the universal solution is the equilibrium one (i.e., maximizes the RS potential among all fixed points), it defines the universal phase. Similarly, when the specialization solution is the equilibrium one, it defines the specialization phase.

For $L = 1$, inhomogeneous readouts imply multiple specialization transitions associated with different solutions of the RS equations, in addition to the one discussed above. Each one is associated with a different “state” where only some of the (macroscopic) subpopulations of neurons connected to the same readout value have specialized; see Eq. (12). These “partially specialized solutions” are accessed by initializing the RS equations with $\mathcal{Q}(\mathbf{v}) = c\mathbb{1}(\mathbf{v} \geq \bar{\mathbf{v}})$ for some $\bar{\mathbf{v}}$ and constant c close to 1.

For $L = 2$, each layer can live in a different phase (i.e., specialize at different sampling rates), which are defined similarly as for $L = 1$ but layerwise using the layer-indexed overlaps. Additionally, in a given layer, partial specialization as described above is also possible, making the overall picture extremely rich: For deep NNs, specialization transitions can happen inhomogeneously across layers, but also across neurons in a given layer. The theory predicts these two types of learning inhomogeneities observed in simulations.

In order to get the corresponding solutions of the RS equations, we proceed as before by playing with the initialization of the OPs. We focus on three representative specialization scenarios across layers, the equilibrium solution always corresponding to one of them: (i) $\mathcal{Q}_{2:1} > 0$ (i.e., positive for any argument value) and $\mathcal{Q}_1, \mathcal{Q}_2 \equiv 0$, because the product matrix $\mathbf{W}^{(2)0}\mathbf{W}^{(1)0}$ can be learned without specialization, (ii) $\mathcal{Q}_1 > 0$ and $\mathcal{Q}_{2:1}, \mathcal{Q}_2 \equiv 0$ to probe partial specialization of the first layer only, and recall (iii) the fully informative initialization $\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_{2:1} > 0$ for complete specialization. Other initializations converge either to the universal solution or match the solution reached from (iii). Having access to this rich family of solutions, the corresponding errors are again obtained by plugging them in $\mathbf{K}^\square$ in Eq. (20).

For $L \geq 3$, under the (H_3) hypothesis, specialization of all layers occurs concurrently, so we consider only the universal and specialization solutions.

Notice that Eq. (20) relies on the simplification of the optimal mean-square generalization error in Appendix A 5, which is a direct consequence of the Nishimori identity at equilibrium (Appendix A 3). By construction of the theory, any solution of the RS equations verifies the Nishimori identities, which justifies using Eq. (20) beyond the equilibrium solution $*$. This reflects the property that metastable states “behave as the equilibrium” for what concerns the validity of the Nishimori identities and concentration properties; see Remark 5 for a discussion and Figs. 8, 18, and 21 for numerical confirmations.

C. Tested algorithms

The theory is tested against four algorithms: The first two are based on Monte Carlo, the third is a spectral algorithm combined with approximate message passing, and the last is a popular first-order optimizer: ADAM. We thus cover different classes of algorithms, and we will see that our theory is linked to all.

(Algo₁) Hamiltonian Monte Carlo (HMC) initialized uninformatively, i.e., from a random initialization, will be used to sample the posterior when the inner weights have a Gaussian prior. We will also use HMC to sample it but starting from an informative (i.e., on the teacher) initialization. These initializations may lead to different results and are used to probe the two solutions of the theory: universal and specialization.

(Algo₂) Another algorithm used for sampling the posterior but with binary-valued weight matrices is the standard Metropolis-Hastings algorithm. It will also be tested from the two kinds of initializations.

Remark 5. The optimal way to construct a predictor for a test sample using these Monte Carlo sampling algorithms is Bayesian, i.e., through an empirical average of the network output over sampled configurations: $\langle \lambda_{\text{test}}(\boldsymbol{\theta}) \rangle_{\text{Monte Carlo}}$. This is costly, as we would need to do that for many instances of the problem and hyperparameters. A computationally more efficient alternative, but in general suboptimal, is a one-shot estimator $\lambda_{\text{test}}(\boldsymbol{\theta})$: a student constructed from one sample $\boldsymbol{\theta}$ of the parameters. The average mean-square generalization error of the latter estimator is called the Gibbs error: $\varepsilon^{\text{Gibbs}} := \mathbb{E}_{\boldsymbol{\theta}^0, \mathcal{D}, \mathbf{x}_{\text{test}}} [\langle \lambda_{\text{test}}(\boldsymbol{\theta}) - \lambda_{\text{test}}^0 \rangle^2]$. At equilibrium, $\varepsilon^{\text{Gibbs}}/2 = \varepsilon^{\text{opt}}$; see Remark 8 or Ref. [100] for a justification based on the Nishimori identities.

For the experiments, we use this formula also during sampling. In practice, we compute the half Gibbs error as $\mathbb{E}_{\mathbf{x}_{\text{test}}} [\lambda_{\text{test}}(\boldsymbol{\theta}_t) - \lambda_{\text{test}}^0]^2/2$ based on a single sample $\boldsymbol{\theta}_t$ at time t (per α value and dataset), where $\mathbb{E}_{\mathbf{x}_{\text{test}}}$ is an empirical average over many test inputs (10^4 – 10^5). When the chains have mixed and the samples are correctly drawn according to the posterior, which is guaranteed for long enough times, this replacement is justified if also assuming the concentration of the square error with respect to $\boldsymbol{\theta}_t$ onto the Gibbs error, i.e., $\mathbb{E}_{\mathbf{x}_{\text{test}}} [\lambda_{\text{test}}(\boldsymbol{\theta}_t) - \lambda_{\text{test}}^0]^2 = \mathbb{E}_{\mathbf{x}_{\text{test}}} \langle [\lambda_{\text{test}}(\boldsymbol{\theta}) - \lambda_{\text{test}}^0]^2 \rangle + o_d(1)$. This concentration is numerically verified to hold for t sufficiently large. Consequently, our way to evaluate the Bayes error holds at equilibrium for large d .

However, out of equilibrium these guarantees are lost. This is *a priori* an issue given that we will need to evaluate errors in “metastable states” which are empirically the only reachable ones in polynomial time. Yet, we claim that when probing the error at a metastable state, the relation $\mathbb{E}_{\boldsymbol{\theta}^0, \mathcal{D}, \mathbf{x}_{\text{test}}} \langle [\lambda_{\text{test}}(\boldsymbol{\theta}) - \lambda_{\text{test}}^0]^2 \rangle_{\text{meta}}/2 = \mathbb{E}_{\boldsymbol{\theta}^0, \mathcal{D}, \mathbf{x}_{\text{test}}} [\langle \lambda_{\text{test}}(\boldsymbol{\theta}) \rangle_{\text{meta}} - \lambda_{\text{test}}^0]^2$, where $\langle \cdot \rangle_{\text{meta}}$ means “sampling at the metastable state,” remains valid, but not while dynamically reaching it (where half the error of the one-shot estimator is merely a

proxy for the Bayesian one). In other words, a hypothesis we make for the rest of the discussion is that metastable states (if present) “behave” as the equilibrium for what concerns Nishimori identities and concentration properties, which are the only ones we need to use in this relation. This is verified in other Bayes-optimal inference problems [100], and it will be justified *a posteriori* by the match of the theoretical predictions (which always agree with the Nishimori identities) associated with a metastable state and the numerics.

One can also build s -shot estimators from HMC samples of (meta)stable states, namely, $\lambda_{\text{test}}[(\theta^{(p)})_{p \leq s}] := (1/s) \sum_{p=1}^s \lambda_{\text{test}}(\theta^{(p)})$ where $\theta^{(p)}$ are HMC samples for long enough times. Assuming, as above, that Nishimori identities hold also at metastable states, and that the generalization error of such an estimator concentrates in $(\theta^{(p)})_{p \leq s}$, i.e., $\mathbb{E}_{\mathbf{x}_{\text{test}}} \{\lambda_{\text{test}}[(\theta^{(p)})_{p \leq s}] - \lambda_{\text{test}}^0\}^2 = \mathbb{E}_{\theta^0, \mathcal{D}, \mathbf{x}_{\text{test}}} \langle \{\lambda_{\text{test}}[(\theta^{(p)})_{p \leq s}] - \lambda_{\text{test}}^0\}^2 \rangle_{\text{meta}} + o_d(1)$, then we can predict the generalization error of an s -shot estimator. It suffices to expand the square in the $\langle \cdot \rangle_{\text{meta}}$ in the last expression to find $\mathbb{E}_{\theta^0, \mathcal{D}, \mathbf{x}_{\text{test}}} \langle \{\lambda_{\text{test}}[(\theta^{(p)})_{p \leq s}] - \lambda_{\text{test}}^0\}^2 \rangle_{\text{meta}} = [(s+1)/s] \mathbb{E}_{\theta^0, \mathcal{D}, \mathbf{x}_{\text{test}}} [\langle \lambda_{\text{test}}(\theta) \rangle_{\text{meta}} - \lambda_{\text{test}}^0]^2$. The latter expression, for $s = 1$, indeed yields the correct relation for one-shot estimators. It is also clear that when s becomes larger, and HMC is sampling the equilibrium state, the above is approaching ε^{opt} .

As a numerical test of these claims, Figs. 8, 10, and 18 show that the theory (which is “Nishimori compliant”) captures the HMC error, both for one- and s -shot estimators, not only at long times, but also during the earlier plateau it experiences when sampling a metastable state. On top of this evidence, in Appendix A 3 and Fig. 21 we also plot the evolution of a deviation from the Nishimori identities during posterior sampling by HMC. We see that on the states that HMC finds for long enough times, the Nishimori identities are verified, regardless of them being stable or only metastable.

(Algo₃) We extended the GAMP-RIE of Ref. [96], publicly available [193], to obtain a polynomial-time predictor for test data in the shallow networks case. Extending this algorithm initially proposed for quadratic activation to a generic one is possible thanks to the identification of an effective GLM onto which the learning problem can be mapped; see Appendix B 4 [the mapping being exact only when $\sigma(x) = x^2$ [96]]. The key observation is that our effective GLM representation holds not only from a theoretical perspective when describing the universal phase, but also algorithmically. The GAMP-RIE is P_W independent, as it exploits only the asymptotic spectral law of $\mathbf{W}^0 \text{diag}(\mathbf{v}^0) \mathbf{W}^0$, which is the same for Gaussian or binary weight matrices by spectral universality [194]. It is therefore in general suboptimal. In order to evaluate the generalization error of GAMP-RIE in the experiments, we plug the estimator (B75) into Eq. (8).

(Algo₄) We also test the standard PYTHON implementation of the ADAM optimizer [195] initialized uninformatively for Gaussian teacher weights. The generalization error for ADAM for a given training set is evaluated as $\mathbb{E}_{\mathbf{x}_{\text{test}}} [\lambda_{\text{test}}(\theta_t) - \lambda_{\text{test}}^0]^2$ using parameters θ_t obtained by training a student network through empirical risk minimization with nonregularized cost function $C(\theta) = (1/n) \sum_{\mu \leq n} (\lambda_{\mu}(\theta) - y_{\mu})^2$. Adding weight decay does not change the global picture. Notice that in contrast with the Monte Carlo algorithms where the Gibbs error (divided by 2) is a computationally simpler way to access their mean-square generalization error, the error of ADAM is not divided by 2 because it provides a one-shot estimator and is used as such for predictions.

The codes needed to reproduce our experiments are accessible online [174].

D. Shallow MLP

1. Generalization error and specialization transition

Starting with the shallow case, in Figs. 5 and 6 we report the theoretical generalization errors from Result 2 for both the universal and specialization solutions.

Figure 5 considers networks with Gaussian inner weights sampled with informatively initialized HMC in order to focus on the specialization solution. Tests with uninformative initialization are discussed later on. Experiments and theory show that HMC initialized close to the target precisely follows the theoretical specialization solution ε^{sp} (which is not always the equilibrium). In contrast, GAMP-RIE’s generalization error follows the universal branch of the theory ε^{uni} . It can actually be shown analytically that it is the case when $d \rightarrow \infty$. An interesting observation is that nonhomogeneous readouts trigger the appearance of the specialization transition earlier and shrink the region where the equilibrium solution coexists with a metastable one (dotted line). In the case of a four-point prior (middle panel), we see two partial specialization transitions as defined in Eq. (12). The first corresponds to the specialization of the neurons connected to readouts with the largest amplitude only, and thus yields a greater improvement in the error than the second.

With continuous readouts, there are infinitely many specialization transitions; Fig. 5 right shows different representative branches. The equilibrium is obtained iteratively by initializing $\mathcal{Q}(\mathbf{v}) = c \mathbb{1}(\mathbf{v} \geq \bar{\mathbf{v}})$ with $c \lesssim 1$ for several values of $\bar{\mathbf{v}}$, selecting the fixed point with the largest free entropy. This selected equilibrium may coincide with the specialized solution (reached from a fully informative initialization), as in the case of ReLU, or it may slightly differ from it, as for tanh. This distinction leads to very small differences in the generalization curves. Furthermore, experimental findings in Appendix B 5 suggest that, while for discrete readouts (left and center panel), the specialized solution is hard to reach; continuous readouts ease the

learning task, with no evidence of hardness for standard algorithms (see paragraph below).

Figure 6 concerns networks with Rademacher inner weights. The numerical points are of two kinds: the dots obtained from Metropolis-Hastings sampling and the circles for the GAMP-RIE. We report analogous simulations for ReLU and ELU activations in Fig. 24, Appendix B 4. Notice an important thing here: Although our theoretical framework presented in Sec. IV uses the HCIZ integral, which relies on the strict rotational invariance of the matrices involved, it is able to accommodate any prior on the weights. It is thus able to deal with nonrotationally invariant matrices, as in the case of Rademacher weights.

In the two considered setups (Gaussian P_W of Fig. 5 and Rademacher of Fig. 6), when data are scarce, $\alpha < \alpha_{\text{sp}}$, the student cannot break the numerous symmetries of the problem, resulting in an effective rotational invariance at the source of the prior universality of the free entropy and OPs, with posterior samples having a vanishing overlap with $\mathbf{W}^0$. In this universal phase, feature learning occurs because the student tunes its weights to match a quadratic approximation of the teacher, rather than aligning to those weights themselves. This phase is universal in the law of the i.i.d. teacher inner weights (centered, with unit variance): Our numerics obtained both with binary and Gaussian inner weights match the theory well. This phase is superseded at α_{sp} by a specialisation phase where the prior P_W matters. There, a finite fraction of the student weights aligns with the teacher's, which lowers the generalization error.

The phenomenology depends on the activation function due to the following reason. Recall the interpretation in terms of a tensor inference problem discussed in Sec. IB, in particular, that before the specialization, components in the Hermite expansion of the target beyond the first two play the role of effective noise when learning. Only with $\alpha > \alpha_{\text{sp}}$ can the student realize that they are informative and exploit them. Consequently, for odd activation ($\tanh$ in Fig. 5, σ_2 in Fig. 6), where $\mu_2 = 0$, we observe that the generalization error is constant for $\alpha < \alpha_{\text{sp}}$, whereas at the phase transition it suddenly drops. This is because the learning of the second component is skipped entirely, and the only way to perform better is to learn all terms jointly through specialization.

We emphasize that our theory is consistent with Ref. [109], which considers the simpler regime of strong overparametrization $n = \Theta(d)$ rather than the interpolation one $n = \Theta(d^2)$: Our generalization curves at $\alpha \rightarrow 0$ match theirs at $\alpha_1 := n/d \rightarrow +\infty$, which is when the student learns perfectly the linear component $\mathbf{v}^{0\top} \mathbf{W}^0$ of the target but nothing more. This is also the best a network can do in the quadratic data regime when $\mu_2 = 0$ if it does not specialize.

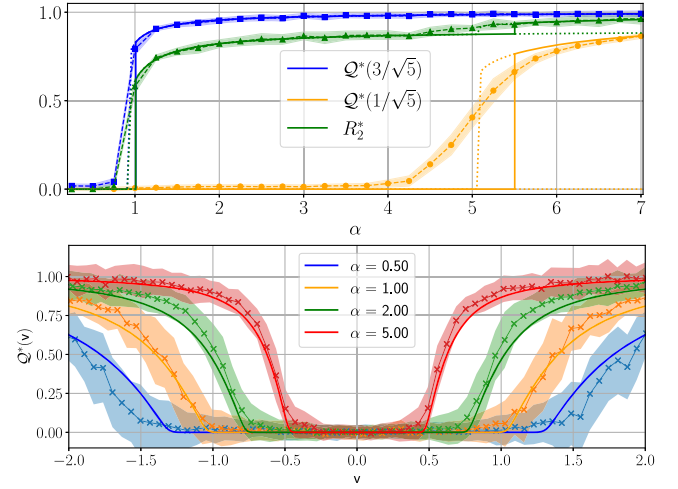


FIG. 7. Theoretical prediction (solid curves) for the equilibrium overlaps as a function of the sampling ratio α for $L = 1$ with Gaussian inner weights $d = 200, \gamma = 0.5, \Delta = 0.1$. The empirical crossed curves were obtained from informed HMC using a single posterior sample $\mathbf{W}$ (per α and data instance), and shaded regions around them correspond to 1 standard deviation with respect to data instances. Top: $\sigma(x) = \tanh(2x)$ and four-point readouts, and average over 12 instances of the data. Bottom: $\sigma(x) = \text{ReLU}(x)$ and Gaussian readouts. $Q(v)$ is evaluated numerically by dividing the interval $[-2, 2]$ into bins and then computing the value of the overlap associated with the readout value in that bin. We averaged over 100 data instances. Readouts are fixed to $\mathbf{v}^0$.

2. Order parameters and learning mechanisms

Figure 7 reveals sequences of phase transitions in α . The top panel shows the evolution of the two relevant overlaps $Q^*(v)$ in the case of readouts with discrete values: As α increases, the student weights start aligning with the target weights with highest readout amplitude, marking the first phase transition. At the same time, R_2^* jumps, indicating that learning of the quadratic term of the target occurs concurrently. As these alignments strengthen, the last transition occurs when the weights corresponding to the next largest readout amplitude are learned. We see the (relatively small) effect of this latter transition also at the level of the generalization error; see red middle curve in Fig. 5 at $\alpha \approx 5$. Through the same mechanism, continuous readouts produce an infinite sequence of learning transitions in the limit (4), as supported by the lower part of Fig. 7 for Gaussian readouts. From these observations, we conclude that the readout amplitudes $|v_j|$ controlling the strength with which the responses (y_μ) depend on feature (neuron) $\mathbf{W}_j^0$ play the role of an SNR.

3. Algorithmic hardness of specializing, and ADAM as an approximate Bayesian sampler

Even when dominating the posterior measure, the specialization solution can be algorithmically hard to reach.

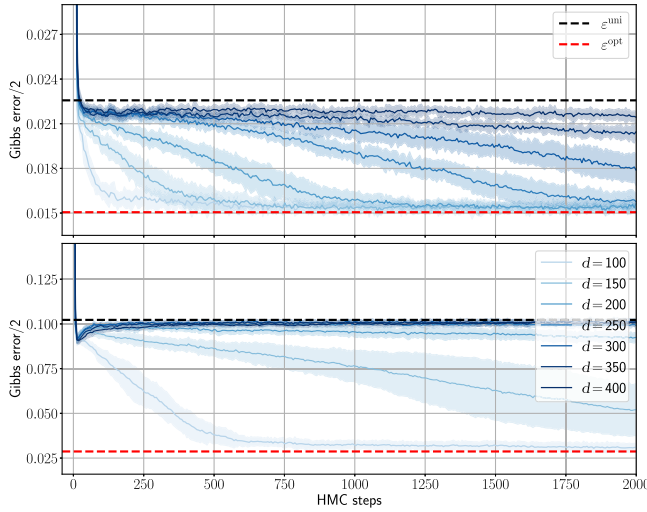


FIG. 8. Half Gibbs error of HMC from random initialization as a function of the number of updates for various d with $L = 1$ and Gaussian inner weights. The errors are averaged over ten data instances, and shaded regions represent 1 standard deviation. The black dashed line corresponds to the error associated with the universal solution, while the red corresponds to the specialized solution. Top: $\sigma(x) = \text{ReLU}(x)$, $\alpha = 3.0$, $\gamma = 0.5$, $\Delta = 0.1$, and four-point quenched readouts. Bottom: $\sigma(x) = \tanh(2x)$, $\alpha = 2.5$, $\gamma = 0.5$, $\Delta = 0.1$, and homogeneous quenched readouts. For both activations, larger d lead to an important slowing-down of the convergence toward the specialized solution happening precisely when crossing the error predicted by the universal solution ϵ^{uni} of the theory.

With discrete readouts, simulations for binary inner weights exhibit specialization only when sampling with informative initialization. Moreover, even in cases where algorithms (such as ADAM or HMC for Gaussian inner weights) are able to find the specialization solution, they manage to do so only after a training time increasing exponentially with d . For the continuous distribution $P_v = \mathcal{N}(0, 1)$, our tests are inconclusive on hardness and deserve numerical investigation at a larger scale. We refer to Appendix B 5 for a detailed discussion and systematic tests. As an illustration of the conclusions reached in this appendix, Figs. 8 and 9 display the evolution of the generalization error reached with HMC and ADAM, respectively (recall that for HMC we plot the half-Gibbs-error proxy).

Figure 8 shows that HMC, for a discrete readout prior, converges fast to the universal solution where it is abruptly stopped (for d large), before very slowly approaching the specialization solution. The time it takes to escape the plateau scales exponentially with the dimension, meaning that improving upon ϵ^{uni} is hard given quadratically many data. The behavior is the same for both $\text{ReLU}(x)$ and $\tanh(2x)$ activations. We observe the same phenomenology when the teacher’s inner weights are drawn from a binary distribution, while HMC sampling wrongly assumes a

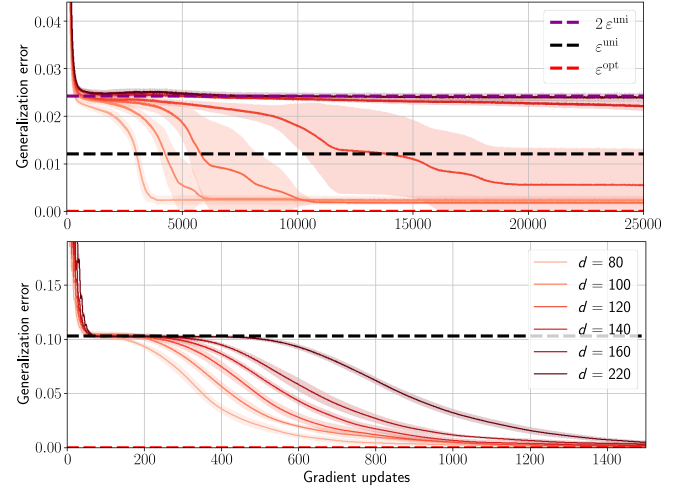


FIG. 9. Generalization error of ADAM from random initialization as a function of the gradient updates for various d with $L = 1$ and Gaussian inner weights. The initial learning rate is 0.01 and batch size $\lfloor n/4 \rfloor$. The error is averaged over ten data instances, and shaded regions represent 1 standard deviation and computed empirically from 10^4 i.i.d. test samples. In both plots $\alpha = 5.0$, $\gamma = 0.5$, $\Delta = 10^{-4}$, the target readouts are homogeneous, while the student has learnable readouts. Top: $\sigma(x) = \text{ReLU}(x)$. The error plateaus at the purple dashed line corresponding to twice the error associated with the universal solution of the theory ϵ^{uni} . Bottom: $\sigma(x) = \tanh(2x)$. The error plateaus at the black dashed line corresponding to the universal solution. The number of gradient updates necessary to improve upon the universal solution (or twice its value) grows exponentially with d ; see Appendix B 5.

Gaussian prior, indicating prior universality of the metastable state.

Concerning ADAM, in Fig. 9, the picture remains globally the same: An initial fast convergence followed by a long plateau before a slow descent toward a close-to-specialization solution (the precise analysis of where ADAM lands in this case and its associated generalization error is beyond the scope of this paper). In the case of $\sigma = \text{ReLU}$ (top panel) and homogeneous readouts the plateau is at $2\epsilon^{\text{uni}}$. It takes again exponentially many updates to escape it. ADAM thus reaches precisely the same error as the Gibbs error of HMC, i.e., when also using HMC as a one-shot estimator. This suggests that ADAM is essentially sampling the posterior the best it can given only a polynomial-in- d number of updates, and ends up in a metastable state similar to the one produced by HMC. The same observation was made for pure gradient descent in the special case of one hidden layer with quadratic activation [96]. Our observations contribute, in a quantitatively precise manner within a rather general NN model, to the recent line of works on “stochastic gradient descent behaves as a Bayes sampler” [196–199] but for ADAM.

There is however one major difference compared to HMC in the case of $\sigma(x) = \tanh(2x)$ (bottom panel):

ADAM plateaus close to ϵ^{uni} instead of $2\epsilon^{\text{uni}}$. By calling $\theta_{\text{algorithm}}$ a sample from that algorithm, this means that the ADAM one-shot estimator $\lambda_{\text{test}}(\theta_{\text{ADAM}})$ performs almost as a Bayesian, ensemble-averaged estimator $\langle \lambda_{\text{test}}(\theta_{\text{HMC}}) \rangle_{\text{meta}}$ sampling the metastable state. The performance of the latter estimator is what we conjecture to be the best achievable in polynomial time. Notice instead that HMC, when also used as a one-shot estimator, does not perform as well (see again Fig. 8, bottom panel). This is at odds with the ReLU case, where ADAM and one-shot HMC were comparable, and both worse than the Bayesian estimator.

We believe that these different behaviors are a consequence of the fact that $\mu_2 \neq 0$ for ReLU, while it is vanishing for tanh, a crucial element also in our replica analysis of Sec. IV. Indeed, with no second component in the activation, the theory predicts the linear term as the only one learnable without specializing. Thus, the nonspecialized NN effectively behaves as a linear model trying to fit a noisy linear target. In this respect, the picture is similar to what happens in the proportional regime $n = \Theta(d)$, where the mapping to a GLM is known to hold in the whole phase diagram [109–111]. In Ref. [109], it is shown that optimizers, such as optimally regularized ridge regression, achieve the performance of the Bayes estimator. Here we observe that the picture changes with $n = \Theta(d^2)$ when also the quadratic term is present, as in the top panel of Fig. 9. A gap between the performance of optimizers and Bayesian estimators has been recently shown in Ref. [98] limited to purely quadratic activation and requires future investigation in our more general setting.

In Fig. 10, we show the impact of overparametrization with respect to the target function in ADAM. This figure refers to a setting where $L = 1$, and the number of hidden units for the target is fixed to 100. K in the left panel is instead the number of hidden neurons of the (possibly mismatched) student. In the right panel, we show the performance that an HMC sampler would attain with an s -shot estimator. Both plots display generalization errors for the ReLU activation. The dashed lines correspond to the theoretical predictions for s -shot estimators as discussed in Remark 5. As we can see, overparametrized students are able to reach, with ADAM, the same performance predicted with an s -shot estimator produced via HMC. There is an intuitive reason for this: When the student has, say, $K = 500$ hidden units, which means 5 times more than the target, then ADAM is somehow picking five sets of weights which, when combined via the readouts, yield the performance of a five-shot estimator. This suggests that ADAM is effectively “sampling” $s = 5$ i.i.d. configurations from the metastable state; this aligns with the results of Refs. [196–199].

This experiment shows that introducing overparametrization can help to reduce the generalization error to ϵ^{uni} in polynomial time with ADAM, even when matched students would get stuck at $2\epsilon^{\text{uni}}$. However, the specialization value

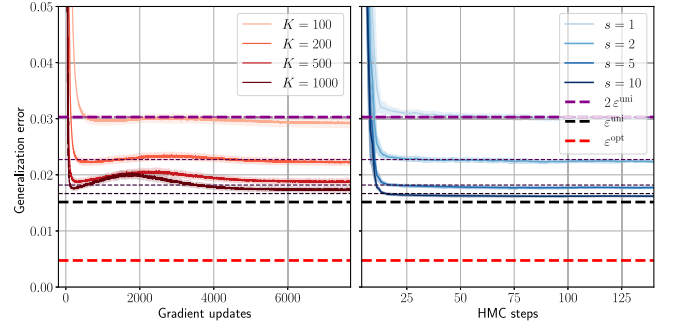


FIG. 10. Generalization error of different estimators, initialized randomly, as a function of the number of gradient updates or HMC steps. The errors are averaged over ten data instances, and shaded regions represent 1 standard deviation. The black and purple dashed lines correspond, respectively, to the error associated with the universal solution and twice the universal solution, while the red dashed line corresponds to the specialized solution. The dashed lines between the black and purple lines indicate the universal performance of s -shot estimators, which is simply $[(s+1)/s]\epsilon^{\text{uni}}$; see Remark 5. In both panels, $\sigma(x) = \text{ReLU}(x)$, $\alpha = 4.0$, $\gamma = 0.5$, $\Delta = 0.03$, $d = 200$, and readouts are homogeneous. The number of hidden units of the teacher is $k = 100$. Left: generalization error of an overparametrized student trained with ADAM as a function of the gradient updates; the readouts are learnable during training. K represents the width of the possibly mismatched student. Right: generalization error of an s -shot estimator obtained averaging the output of s posterior HMC samples, as a function of the HMC steps; the readouts are fixed during sampling.

remains out of reach in polynomial time according to this picture.

In summary, these experiments show a clear timescale separation between the universal solution (or twice the universal solution) reachable in polynomial time and the specialization solution requiring exponential time (when it corresponds to the equilibrium). The suboptimal solution is due to the presence of an attractive metastable state experienced by both algorithms. Moreover, ADAM behaves similarly to a Bayesian sampler like HMC.

4. Structured data

Let us consider structured data where the input distribution is different from the standard Gaussian. The most basic example of this is Gaussian inputs with covariance $\mathbf{C}$. In this case, the model can be transformed into one with standard Gaussian inputs at the cost of losing independence among entries within the same row $\mathbf{W}$. Indeed, by writing $\mathbf{x}_\mu = \mathbf{C}^{1/2}\tilde{\mathbf{x}}_\mu$, where $\tilde{\mathbf{x}}_\mu \sim \mathcal{N}(\mathbf{0}, I_d)$, the model can be viewed as having $\tilde{\mathbf{x}}_\mu$ as input and $\tilde{\mathbf{W}} = \mathbf{W}\mathbf{C}^{1/2}$ as first-layer weights. The weight matrix $\tilde{\mathbf{W}}$ has independent rows but dependent entries within the same row.

With this relaxed condition, where $\tilde{\mathbf{W}}$'s rows are independent and follow a law $P_{\mathbf{w}}$ in $\mathbb{R}^d$, for activations that have zero second Hermite coefficient, Result 1 still holds, with

the only modification being the replacement of the scalar free-entropy function $\psi_{P_{\mathbf{w}}}$ by its vector version:

$$\lim_{d \rightarrow \infty} \frac{1}{d} \mathbb{E}_{\mathbf{w}^0, \boldsymbol{\xi}} \ln \mathbb{E}_{\mathbf{w}} e^{-\frac{1}{2} \|\mathbf{w}\|^2 + x \mathbf{w}^T \mathbf{w}^0 + \sqrt{x} \boldsymbol{\xi}^T \mathbf{w}}, \quad (21)$$

where $\mathbf{w}, \mathbf{w}^0$ are i.i.d. from $P_{\mathbf{w}}$ in $\mathbb{R}^d$ and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I_d)$. This is evident from the replica computation, where the i.i.d. assumption on the weights is required in Eq. (B23) to factorize the integral over the weights, yielding the log-integral term that later becomes the scalar free entropy. If dependences within the same row are allowed, the factorization can be performed only over the rows, yielding the free entropy (21). For activations with a nonzero second Hermite coefficient, analyzing the model with structured data is an open problem. Even for $\mathcal{N}(\mathbf{0}, \mathbf{C})$ input, the task requires solving a denoising problem for the matrix $\mathbf{C}^{1/2} \mathbf{W} \text{diag}(\mathbf{v}) \mathbf{W} \mathbf{C}^{1/2}$, which is not analytically tractable due to the lack of rotational invariance.

In general, there is no analytical simplification for Eq. (21) except when $\mathbf{w}$ has i.i.d. entries or when $\mathbf{w}$ is a Gaussian vector. In the latter case, suppose $P_{\mathbf{w}} = \mathcal{N}(\mathbf{0}, \mathbf{C})$, then $\psi_{P_{\mathbf{w}}}(x) = \lim_{d \rightarrow \infty} [1/(2d)] [x \text{Tr}(\mathbf{C}) - \ln \det(I_d + x \mathbf{C})]$. If $\mathbf{C}$ admits a limiting spectral density $\rho_{\mathbf{C}}$,

$$\psi_{P_{\mathbf{w}}}(x) = \frac{1}{2} \int [xs - \ln(1 + xs)] \rho_{\mathbf{C}}(s) ds. \quad (22)$$

In Appendix C 2, we show that by quenching the first-hidden-layer's weights and taking the first activation to be linear via a procedure similar to that followed to derive Result 3, it is possible to describe such structured data within our replica formalism.

Figure 11 shows that structure in the input helps reduce the optimal generalization error. Here, the covariance matrices are of the Wishart form $\mathbf{W}_0 \mathbf{W}_0^T / d_0$, where $\mathbf{W}_0 \in \mathbb{R}^{d \times d_0}$ is a Gaussian matrix, with varying ratios d_0/d . It is known that as d_0/d increases, the spectrum of the Wishart matrix approaches that of the identity matrix, meaning the inputs become less structured. The figure demonstrates that as d_0/d grows, the curves quickly approach that of the standard Gaussian inputs.

For a broad class of non-Gaussian inputs, our experiments show that the Bayes-optimal error is unchanged if the inputs entering the random MLP target are replaced by Gaussian vectors with the same covariance. This behavior, implied in our theory by the Gaussian hypothesis (11), was verified using inputs generated by feeding standard Gaussian vectors into one NN layer with Gaussian weights as well as real data from MNIST; see Fig. 12. The discrepancy between theoretical and experimental results at low α for MNIST data can be attributed to the few large outlier eigenvalues, while our theory is dependent only on the spectral density of the data covariance, which the outliers do not influence.

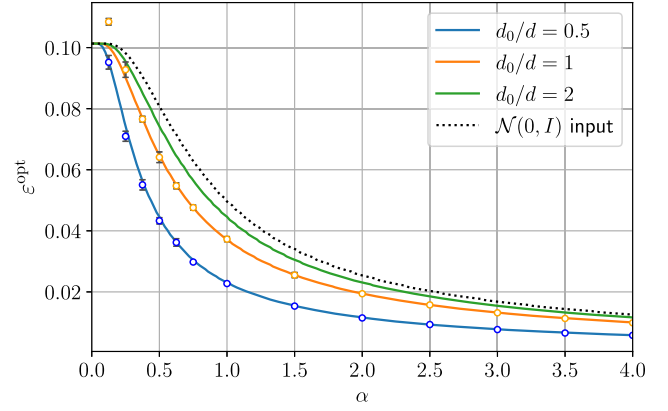


FIG. 11. Theoretical prediction for the Bayes-optimal mean-square generalization error for $L = 1$ with fixed Gaussian readouts, Gaussian inner weights, structured inputs drawn from $\mathcal{N}(\mathbf{0}, \mathbf{C})$, where $\mathbf{C} = \mathbf{W}_0 \mathbf{W}_0^T / d_0$, where $\mathbf{W}_0 \in \mathbb{R}^{d \times d_0}$ is a Gaussian matrix, $\tanh(2x)$ activation, $d = 150$, $\gamma = 0.5$, $\Delta = 0.1$. The dotted line shows the theoretical result for standard Gaussian inputs, with the other settings unchanged. Experimental points are obtained with informative HMC by averaging over nine data instances, with error bars representing the standard deviation. As the ratio d_0/d grows, the data become less structured, and the theoretical curve rapidly approaches that of standard Gaussian input.

E. Two-hidden-layer MLP

In this section, we present experiments for $L = 2$ hidden layers, similar to the ones conducted for the shallow case. Adding one layer already makes the picture richer, in particular, in terms of the learning phase transitions taking place and the information flow across layers.

1. Learning propagates from inner toward outer layers

Figure 13 displays the mean-square generalization error for a network with hyperbolic tangent activation. The general picture is similar to the shallow case: At small α , a single prior-independent solution exists; with more data, the specialization solution branches out continuously. A clear transition occurs in the specialization solution. The mechanism behind it is explained by Fig. 14 showing the evolution of the average over neurons of the overlap profile in each layer.

We first discuss the case of Gaussian $\mathbf{v}^0$, bottom of Fig. 14 and right of Fig. 13. The ordering of the specializations along layers is clear: (i) The network starts specializing from the inner layer $\mathbf{W}^{(1)}$. The information then propagates outward. (ii) The deep layer $\mathbf{W}^{(2)}$ is then getting learned. This experimentally observed shallow-to-deep layer ordering of the learning is encoded in our formulas (including for $L \geq 3$): The RS equations imply that the overlap for the second layer can become nonzero only if the one for the first layer is itself nonvanishing.

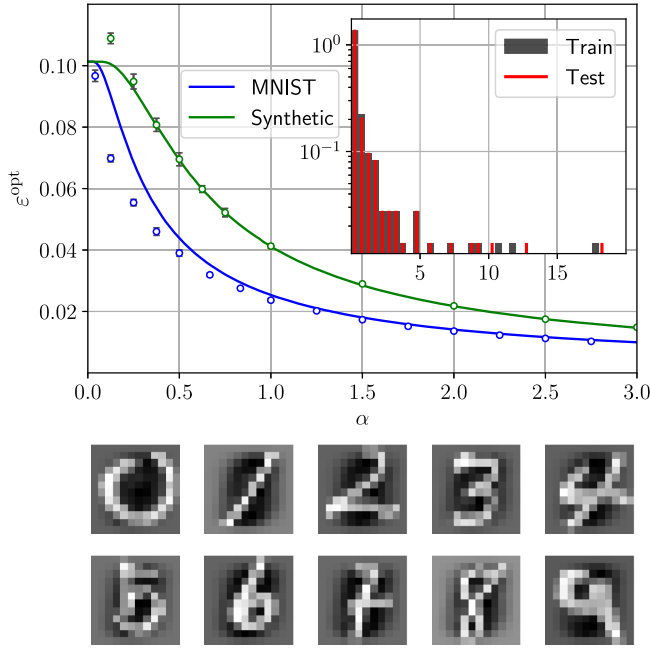


FIG. 12. Top: theoretical prediction of the optimal mean-square generalization error for non-Gaussian data. The inputs are taken from the MNIST image dataset or as the outputs of one layer of another NN fed with standard Gaussian vectors (synthetic data). More precisely, the synthetic data are generated as $\mathbf{x}_\mu = \sigma_0(\mathbf{W}^{(0)} \mathbf{x}_\mu^{(0)} / \sqrt{d_0})$ where $\mathbf{W}^{(0)} \in \mathbb{R}^{d \times d_0}$ is a Gaussian matrix with $d_0/d = 0.5$, $\mathbf{x}_\mu^{(0)} \sim \mathcal{N}(\mathbf{0}, I_{d_0})$, $\sigma_0(x) = [\text{ReLU}(x) - \mu_0]/c_0$ where μ_0 is the zeroth Hermite coefficient of $\text{ReLU}(x)$ and c_0 enforces $\mathbb{E}_{z \sim \mathcal{N}(0,1)} \sigma_0(z)^2 = 1$. $(\mathbf{x}_\mu)$ are then passed through the random MLP target with $L = 1$, $\sigma(x) = \tanh(2x)$, $\gamma = 0.5$, and Gaussian weights (inner and readouts) to generate the noisy responses (y_μ) with $\Delta = 0.1$. The trainable NN has fixed $\mathbf{v} = \mathbf{v}^0$. The MNIST dataset consists of 60 000 training samples and 10 000 test samples. Each one is a 28×28 pixel image representing a digit from 0 to 9. To make the dataset manageable for HMC, each image is downsampled to a 12×12 resolution. This is achieved by partitioning each side of the original image into blocks of sizes 4, 2, 2, ..., 2, 4, resulting in 12×12 regions, over which pixel values are averaged. The maximum value for the sampling rate is thus $60\,000/144^2 \approx 2.9$. Subsequently, the images are centered and normalized to have zero mean and a covariance matrix $\mathbf{C}$ satisfying $\text{Tr}(\mathbf{C}) = d = 144$. Importantly, the responses are still generated by a random target NN with the same architecture as the trained one: The purpose of this experiment is to test input data with realistic correlations. Inset: histogram of the eigenvalues of covariance matrices computed from the training and test datasets. Both exhibit a few eigenvalues that are significantly larger than the rest, which may explain the discrepancy between theoretical and experimental results at low α . Bottom: examples of MNIST images after being downsampled, centered, and normalized, showing that their integrity is preserved after the process.

Concerning the recovery of $\mathbf{W}^{(2)}\mathbf{W}^{(1)}$, it should be understood that products of weight matrices from different layers are learnable partly independently of their factors. This is similar to the shallow case, where the quadratic term $\mathbf{W}^{0\top} \text{diag}(\mathbf{v}^0) \mathbf{W}^0$ in the target can be partially recovered without learning $\mathbf{W}^0$, and thus comes with its own OP which can be nonzero even if the one for $\mathbf{W}^0$ vanishes (in the universal phase). The equations consistently encompass this possibility.

The learning transitions occur more abruptly with homogeneous readouts (top of Fig. 14 and left of Fig. 13): The learning of the product matrix and the deep one occur jointly and sharply. The homogeneity of the readouts is the source of the discontinuity of the learning transition which makes learning harder. Interestingly, continuous rather than discrete readouts induce smoother transitions. Nevertheless, notice the richer behavior of the generalization error and overlaps right after the first transition for homogeneous readouts, and also that in that region of α , the ordering of overlap values is different between the two readout distributions.

We mention that Refs. [114–116] also predicted learning inhomogeneities across layers in a teacher-student setting but in a strongly overparametrized data regime.

2. Inhomogeneous learning profile across neurons and matrix order parameters

When there are two or more hidden layers, nontrivial overlap profiles emerge in each. This effect is a joint consequence of the depth and linear width of the network but also of the complex interactions among the layers [recall the discussion when we introduced $\mathcal{Q}_2^*(\mathbf{v}, \mathbf{v}^{(2)})$ in Sec. II D].

For the first hidden layer, top panel of Fig. 15, the overlap inhomogeneity is related to the fluctuations in the effective readouts of the target $\mathbf{v}^{(2)0} := \mathbf{W}^{(2)0\top} \mathbf{v}^0 / \sqrt{k_2}$: Its components are Gaussian random variables. It implies that neurons are not all “measured equally well” (in particular, through the linear term in the Hermite expansion of σ in the first layer). The profile therefore manifests itself along the output (row) dimension.

For the second layer, with homogeneous readouts, the only inhomogeneity is along its input (column) dimension and is induced by the output profile of the first-layer weights (second panel). For completeness, we checked that, consistently with the theory, the profile of the second-layer overlap along its output dimension is indeed constant (third panel). Because of the homogeneity of readouts, a similar constant overlap profile along the output dimension of the product matrix $\mathbf{W}^{(2:1)}$ appears (last panel).

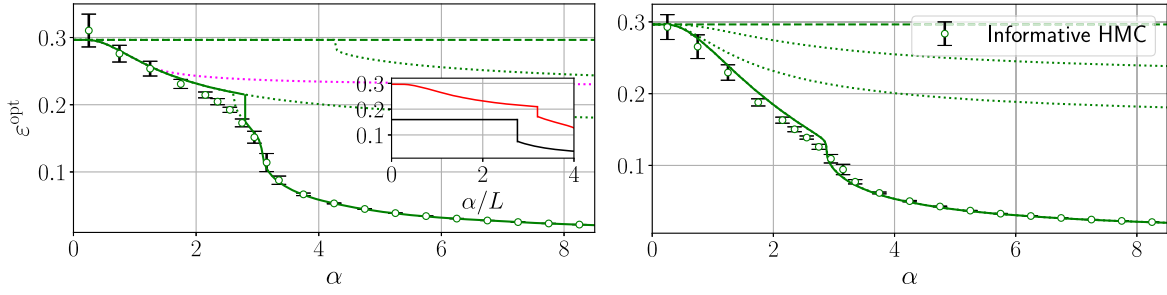


FIG. 13. Theoretical prediction (green solid curve) of the Bayes-optimal mean-square generalization error for $L = 2$ with Gaussian inner weights, $\sigma(x) = \tanh(2x)/\sigma_{\tanh}$, $d = 200$, $\gamma_1 = \gamma_2 = 0.5$, $\Delta = 0.2$, and different P_v laws. The dashed line represents the universal branch. Dotted lines denote metastable specialization branches of the RS saddle-point equations reached from different initializations for the overlaps. From the top, the first green dotted line represents the solution reached by initializing $Q_{2:1} > 0$ (i.e., positive for any argument value) and $Q_1, Q_2 \equiv 0$ (0 for all arguments), the second to $Q_1, Q_{2:1}, Q_2 > 0$ (it yields in the left panel the small metastable solution just before the transition around $\alpha = 2.8$; in the right panel, this solution collapses on the equilibrium curve), the third is $Q_1 > 0, Q_{2:1}, Q_2 \equiv 0$. The magenta dotted curve corresponds to initialization $Q_1(v^{(2)}) > 0$ only for sufficiently large $v^{(2)}$, while $Q_{2:1}, Q_2 \equiv 0$: This solution is an inhomogeneous specialization across layers (only the first specializes), and across the neurons in that layer (only some neurons specialize). Points are obtained with Hamiltonian Monte Carlo with informative initialization. Each point has been averaged over 20 instances of the data, with error bars representing 1 standard deviation. The generalization error is computed empirically from 10^4 i.i.d. test samples. The readouts are fixed to the teacher's during sampling. Left: homogeneous readouts. Inset: optimal generalization error of a shallow MLP (black line, $L = 1$) and deep MLP (red line, $L = 2$). The activation for both curves is $\sigma(x) = \tanh(2x)/\sigma_{\tanh}$, $\gamma, \gamma_1, \gamma_2 = 1$, and $\Delta = 0.2$, while α is divided by L for comparison. Right: Gaussian readouts.

These experiments probing all possible overlap inhomogeneities in a three-layer MLP vindicate our definitions and indexing of OPs in the theory.

As an illustration of what is learned in the deep NN when increasing the data, we plot in Fig. 16 the three functional overlap OPs using heat maps. Two carry a

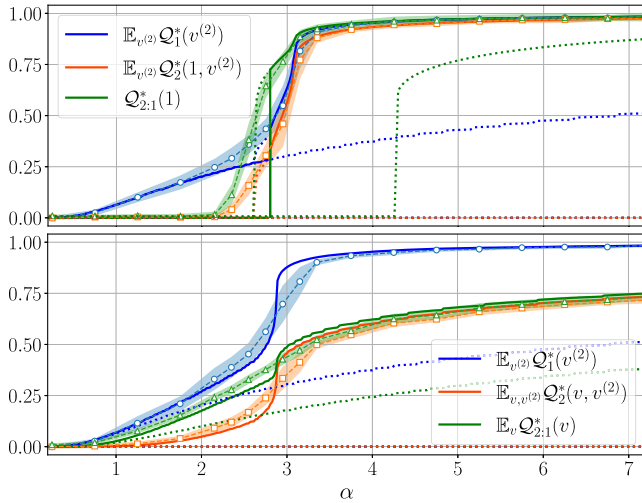


FIG. 14. Solid and dotted curves represent, respectively, the mean of different overlaps at equilibrium and in metastable specialized states, as a function of the sampling ratio α for $L = 2$ with Gaussian inner weights, $\sigma(x) = \tanh(2x)/\sigma_{\tanh}$, $d = 200$, $\gamma_1 = \gamma_2 = 0.5$, $\Delta = 0.2$. The shaded curves were obtained from informed HMC. Each point has been averaged over 20 instances of the training set, with 1 standard deviation depicted. The readouts are fixed to the teacher's during sampling. Top: homogeneous readouts. Bottom: Gaussian readouts.

single argument and the one for the second inner layer possesses two, as it captures at a macroscopic level the learning inhomogeneities along both the rows and columns of $\mathbf{W}^{(2)}$ (we restricted the domains of the OPs for visualization despite the fact that the true ones are in principle infinite). As α exceeds the smooth transition happening around $\alpha = 2.8$ in the same setting of the bottom panel of Fig. 14, we see that specialization nucleates in $\mathbf{W}^{(2)}$ starting from its neurons (i.e., rows) indexed by the largest readout amplitudes $|v|$, and concurrently from its “dual neurons” (i.e., its columns) connected to the largest effective readout amplitudes $|v^{(2)}|$. Specialization then propagates toward lower values as α increases. The figures emphasize how the learning of the other matrices (the first-layer weights and the product matrix, which both display learning inhomogeneities along one dimension only) interact with the deep one and yield such an intricate behavior.

Finally, in Fig. 17, we display the result of the numerical experiment for $Q_2(\mathbf{v}, \mathbf{v}^{(2)})$. This figure was realized by averaging over different instances of a data-student pair, with the overlap values ordered as those of Fig. 16 for each pair, and then by performing a “local average” of neighboring indices on the grid $(\mathbf{v}, \mathbf{v}^{(2)})$ in order to suppress the “microscopic fluctuations.” The latter average should be interpreted in the thermodynamic limit as being over a relatively small patch, which still contains $\Theta(d^2)$ weights. Remarkably, this figure could have also been generated from a single instance of the student, as the local average alone is sufficient to reproduce the patterns of Fig. 16.

Notice that in contrast with the shallow case where $Q(\mathbf{v})$ vanishes for $\mathbf{v}$ small even at large sampling ratios (see

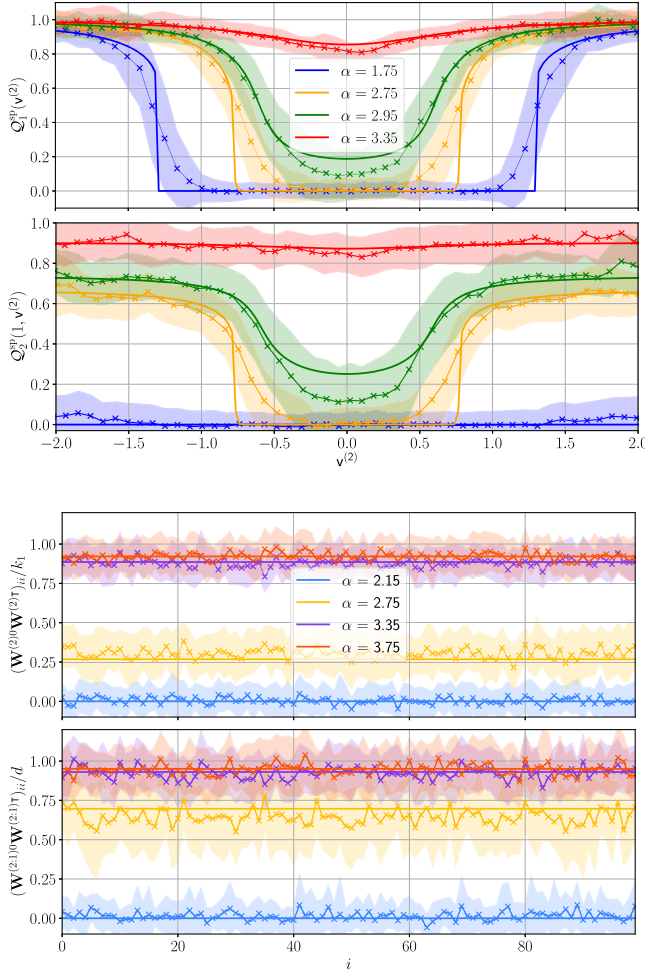


FIG. 15. Theoretical predictions (solid curves) for the overlaps obtained from informative initialization as functions of $\mathbf{v}^{(2)0}$ or $i = 1, \dots, k_2$ for $L = 2$ with activation $\tanh(2x)/\sigma_{\tanh}$, $d = 200$, $\gamma_1 = \gamma_2 = 0.5$, $\Delta = 0.2$, Gaussian inner weights, homogeneous quenched readouts, and different α values. The shaded curves were obtained from informed HMC. Using singles posterior samples, the overlaps have been evaluated numerically by dividing the interval $[-2, 2]$ into bins and by computing their value in each bin. Each point has been averaged over 20 instances of the data, and shaded regions around them correspond to 1 standard deviation. First (top): first-layer overlap $Q_1^*(\mathbf{v}^{(2)})$ profile ordered according to the amplitude of the effective readouts $\mathbf{v}^{(2)0}$. Second: the input- (or column-) indexed overlap for the second layer $Q_2^*(1, \mathbf{v}^{(2)})$ also ordered according to the effective readouts. Third: the neuron (i.e., output- or row-) indexed overlap profile for the second layer. Last: the output-indexed overlap profile for the product matrix $\mathbf{W}^{(2:1)}$.

Fig. 7), the top panel of Fig. 15 and Figs. 16, 17 show that, for sufficiently large α , overlaps indexed by $\mathbf{v}^{(2)}$ become nonzero for any value of the index. This occurs because the effective readouts enter the covariance $K^{(2)}(\bar{\mathbf{Q}})$ from Result 3 in a different way compared to the actual readouts $\mathbf{v}$.

3. Algorithmic hardness and partial specialization

We repeated experiments probing the behavior of HMC and ADAM similar to the shallow case; see Fig. 18. Starting with HMC, two noticeable differences appear compared to $L = 1$. First, reaching the specialized equilibrium state from uninformative initialization seems much harder and costly than in the shallow case; we do observe the descent toward it only for a rather small size ($d = 50$).

The most striking difference, however, is the nature of the state that HMC experiences. For $L = 1$, HMC was getting attracted by the metastable state associated with the universal solution (top panel of Fig. 8), with the single inner layer not specializing. With one more hidden layer a richer picture emerges. The top panel of Fig. 18 shows that in polynomial-in- d time, before complete specialization ultimately occurs when the chain equilibrates, HMC is now stuck in a partially specialized metastable state. There, the first layer has specialized (i.e., has been partly recovered) while the second has not. We tracked the experimental overlaps, not depicted here, and they confirm this picture. The theory correctly predicts this state: It is the third green dotted curve from the top in Fig. 13 left panel, which at $\alpha = 4$ corresponds to the dashed blue curve $e^{\text{meta}} = 0.2$ in Fig. 18 (the two plots are done in the same setting). This is very interesting: Somehow, depth helps from that perspective, since this mechanism cannot be observed in the shallow case. The reason is the presence of the effective readouts in the target $\mathbf{v}^{(2)0} = \mathbf{W}^{(2)0T} \mathbf{v}^0 / \sqrt{k_2}$ when $L = 2$, that inhomogeneously “measure” the inner layer.

Apart from this stable metastable state, there are other solutions of the RS equations. The magenta curve in Fig. 13 depicts a case where only the first layer specializes and only for a subset of its neurons. The same effect is observed in shallow MLPs for their single hidden layer (see Fig. 5, right panel). Although these solutions exist, their stability is not systematically analyzed here, as this poses a significant challenge both for the theory and even more so in the numerical experiments due to strong finite-size effects and the difficulty in training large deep Bayesian NNs. We leave an investigation of their stability for future work.

When learning with ADAM (bottom panel of Fig. 18), we instead observe a similar scenario as with $L = 1$ with tanh activation (lower panel of Fig. 9): As d increases, ADAM gets stopped by the metastable state associated with the universal solution and does not reach the better partially specialized one. This suggests that HMC is able to “see” metastable states that outperform ϵ^{uni} , while ADAM is not in the tested cases. It would be interesting to understand why.

F. Three or more hidden layers

1. Deeper is harder

Figure 19 displays the theoretical and experimental Bayes error and the theoretical overlaps corresponding to informative initialization. A number of observations can be

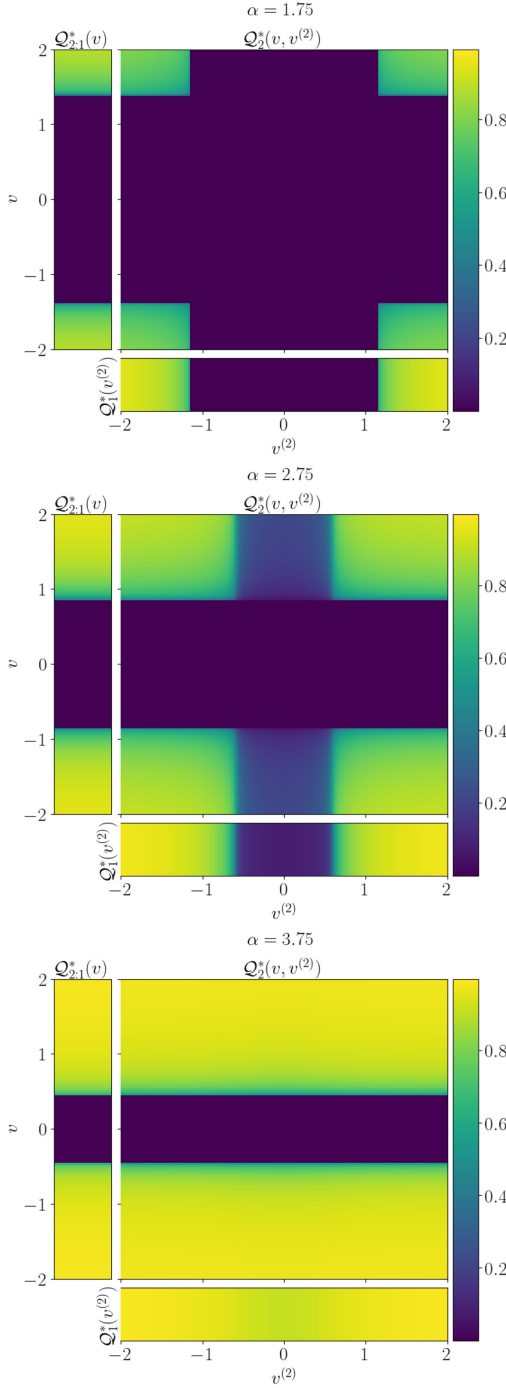


FIG. 16. Heat maps of all theoretical equilibrium overlaps as function of the sampling rate $\alpha \in \{1.75, 2.75, 3.75\}$ (increasing from top to bottom) for $L = 2$ with Gaussian inner and readout weights, $\sigma(x) = \tanh(2x)/\sigma_{\tanh}$, $\gamma_1 = \gamma_2 = 0.5$, $\Delta = 0.2$, which is the same setting as the right panel of Fig. 13 and bottom panel of Fig. 14. Left column: product matrix overlap $Q_{2,1}^*(v)$. Bottom row: first-layer overlap $Q_1^*(v^{(2)})$. Central square: second-layer overlap $Q_2^*(v, v^{(2)})$. The overlap arguments in the theory are amplitudes $v, v^{(2)} > 0$. However, we plot them here as a function of the actual signed readouts and effective readout values $v, v^{(2)}$ for better visualization of what is going on in the network. These figures fully capture the features learned along the layers and across the rows and columns of each layer weight matrix, in a three-layer neural network.

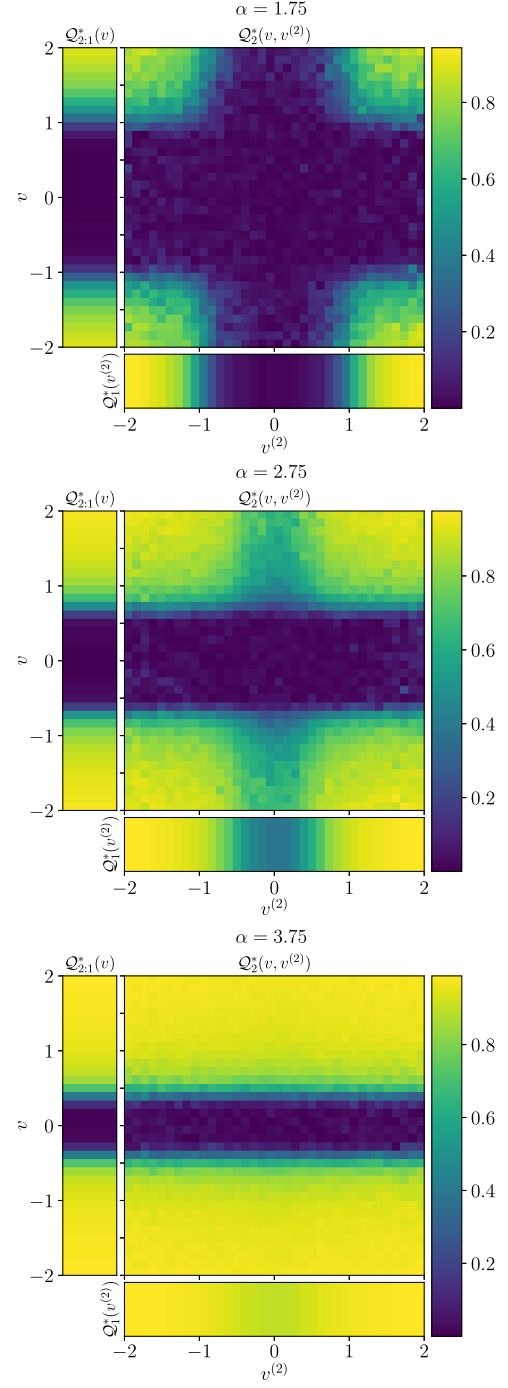


FIG. 17. Heat map of all the empirical overlaps on grids of (effective) readouts $(v, v^{(2)})$. Besides the local average performed by binning the distribution of the weights according to the grid of readouts, we average over 100 instances of data-student pairs. Here, $d = 200$, $\alpha = 1.75$ (top), $\alpha = 2.75$ (middle), and $\alpha = 3.75$ (bottom), while the rest is as in Fig. 16. For the second-layer overlap, the values are rearranged putting those associated with highest readout values on the corners of the image, as in Fig. 16. Up to finite-size fluctuations, the same patterns as the theoretically predicted ones, top and bottom panels of Fig. 16, clearly appear.

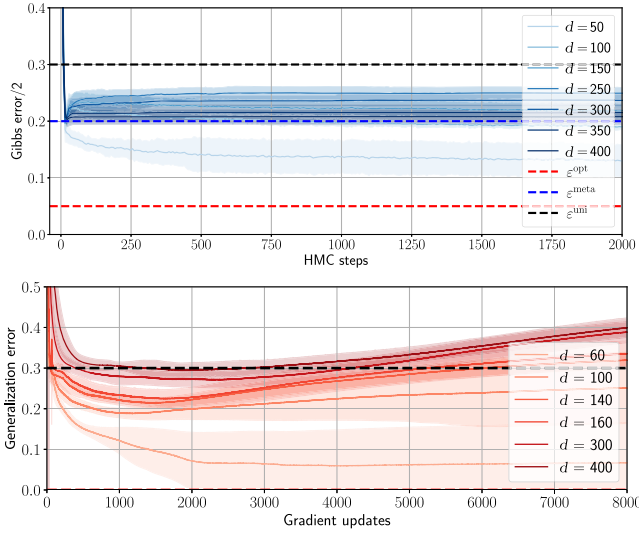


FIG. 18. Generalization errors computed empirically from 10^4 i.i.d. test samples of HMC (ADAM) as a function of the number of steps (gradient updates). Errors are averaged over ten instances; shaded areas indicate 1 standard deviation. Here, $L = 2$, $\sigma(x) = \tanh(2x)/\sigma_{\tanh}$, $\gamma_1 = \gamma_2 = 0.5$, $\alpha = 4.0$ with Gaussian inner weights and homogeneous target readouts for both plots. Dashed lines represent the theoretical errors associated with equilibrium and metastable solutions. Top: Half Gibbs error of HMC from random initialization as a function of the number of updates for various d with $\Delta = 0.2$. The readouts are quenched during sampling. Bottom: Generalization error of ADAM from random initialization as a function of the gradient updates for various d with $\Delta = 10^{-4}$. The initial learning rate is 0.01 and batch size $\lfloor n/4 \rfloor$. The student has a learnable readout layer.

made. Let us start with the left panel. Note that the sampling rate α in the abscissa is rescaled by the number of layers to make a fair comparison between targets with different depths. We conclude based on the location of the specialization transition [common to all layers under (H_3)] and the ordering of the generalization error curves with L that *the deeper the target, the more data per layer it requires to be recovered, and the higher is the ϵ^{opt} at a given α/L* . This vindicates information theoretically the intuitive picture that the more “nonlinear” a task or target is, in the present case through more layers, the harder it should be to learn it. Another confirmation of this fact is provided in the inset of Fig. 13: Also in the case of normalized tanh activation, the amount of data per layer remains greater for the deep case, in spite of the fact that $\mu_1 \neq 0$ implies the presence of effective Gaussian readouts $\mathbf{v}^{(2)0} := \mathbf{W}^{(2)0} \mathbf{r} \mathbf{v}^0 / \sqrt{k_2}$ “measuring” the inner layer $\mathbf{W}^{(1)0}$, which allow it to specialize for small sampling rate.

Another way to see from the results that depth is linked to hardness is through the right panel depicting the overlaps in each layer: For a given $L \geq 2$ and at fixed sampling rate (nonrescaled by L this time), overlaps Q_l^* are monotonically decreasing with the layer index l . In other words, deeper features are harder to learn than the shallow ones, which confirms for generic L the shallow-to-deep ordering of the learning observed for $L = 2$ in the previous section. This is again rather intuitive and matches what is observed when deploying neural networks on real tasks [200]; see also Ref. [201] for the role of depth in NNs when learning in a hierarchical task.

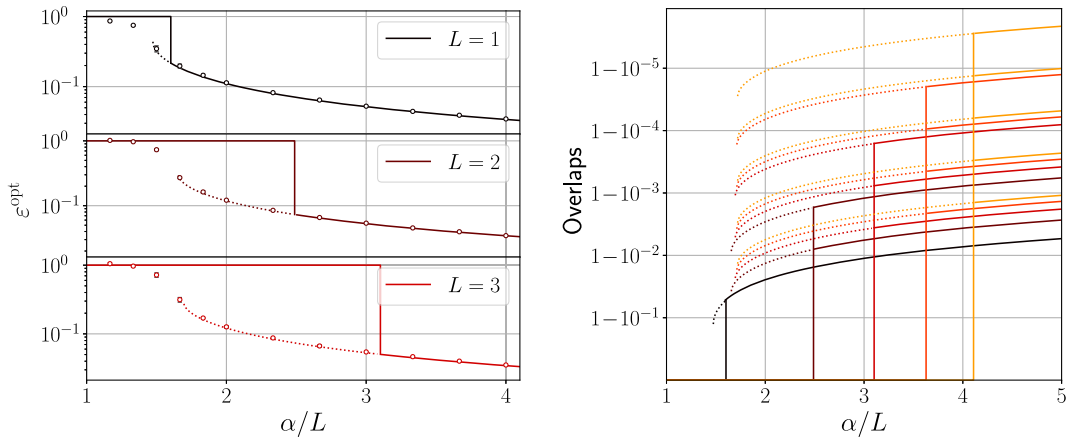


FIG. 19. Theoretical predictions for deep NNs with homogeneous readout $v_i = 1$ and the activation on every layer given by $\sigma(x) = [\tanh(2x) - \mu_1 x]/c$, where μ_1 is the first Hermite coefficient of $\tanh(2x)$, and the constant c is such that $\mathbb{E}_{z \sim \mathcal{N}(0,1)} \sigma(z)^2 = 1$; $\gamma_L = \dots = \gamma_1 = 1$, and the number of hidden layers $L \in \{1, \dots, 5\}$. Left: Bayes-optimal error (solid curves) as function of α/L for $L \leq 3$, while dotted lines are metastable solutions of the RS equations. The points and error bars represent the mean and standard deviation of half the Gibbs errors evaluated on nine data instances at $d = 50$ for all points except the second and third points (from left to right) for $L = 3$, which are computed at $d = 30$ since HMC remains stuck at initialization for higher d . Right: overlap $Q_l^{(L)*}$ of the l th layer weights in the L -hidden layers NN for each pair $(l \leq L, L)$. For each L , all phase transitions for different $Q_l^{(L)*}$ occur concurrently, with the overlaps decreasing with the layer index: $Q_1^{(L)*} > \dots > Q_L^{(L)*}$ after the transition.

In addition to the ordering of overlaps across layers within the same NN, another ordering is also evident from Fig. 19. Letting $Q_l^{(L+1)*}$ be the l th layer equilibrium overlap for an NN with L hidden layers, we have

$$Q_{l+1}^{(L+1)*} \leq Q_l^{(L+1)*} \quad \text{for } 1 \leq l \leq L.$$

It follows from $Q_{l+1}^{(L+1)*} \leq Q_{l+1}^{(L+1)*}|_{\mathbf{w}^{(1)}=\mathbf{w}^{(1)0}} = Q_l^{(L)*}$, where the equality is a consequence of the fact that, since σ has $\mu_0 = \mu_1 = \mu_2 = 0$ and $\mathbb{E}_{z \sim \mathcal{N}(0,1)} \sigma(z)^2 = 1$, the data after the first layer have a covariance indistinguishable from I_d from the perspective of the NN. Thus, the $(L+1)$ -hidden-layer NN with a quenched first layer is equivalent to one with L hidden layers and standard Gaussian inputs.

IV. REPLICAS PLUS HCIZ, REVAMPED

The goal is to compute the asymptotic free entropy by the replica method [126], a powerful approach from spin-glass theory also used in machine learning [29], combined with the HCIZ integral. We focus first on the derivation of the results for the shallow case $L = 1$ which comes with its own set of difficulties due to the presence of $\mu_2 \neq 0$ in σ . We will later move on to the deep case where, even when considering $\mu_2 = 0$, a different kind of difficulty will appear due to the multilayer structure.

Our derivation is based on three key ingredients.

- (i) The first ingredient is a Gaussian *Ansatz* on the replicated postactivations, which generalizes Conjecture 3.1 of Ref. [109], now proved in Ref. [111], where it is specialized to the case of linearly many data [$n = \Theta(d)$]. To obtain this generalization, we will write the kernel arising from the covariance of the postactivations as an infinite series of scalar OPs derived from the expansion of the activation function in the Hermite basis, following an approach recently devised in Ref. [202] in the context of the random feature model (see also Refs. [48,203]).
- (ii) The second ingredient exposed in Sec. IV B 2 amounts to a drastic reduction of the number of OPs entering the covariance of the postactivations through the realization that infinitely many of them are expressible in terms of a few, more fundamental (functional) OPs.
- (iii) The last ingredient is a generalization of an *Ansatz* used in the replica method by Ref. [168] for dictionary learning, which will allow us to capture important correlations discarded by these earlier approaches [168,169]. Our *Ansatz*, explained in Sec. IV B 3, is the crux for capturing the lack of rotational invariance and matrix nature of the problem when σ possesses $\mu_2 \neq 0$ when $L = 1$ or $\mu_1 \neq 0$ when $L = 2$. We will see that, surprisingly, the HCIZ integral remains central despite the absence

of rotational symmetry. Appendix B 2 provides a comparison with the approach of Ref. [168].

For the sake of presentation, we discuss in the main text only the nonstandard steps corresponding to these ingredients. The complete derivations are presented in Appendix B 1 for the shallow case and Appendix C 1 for the deep.

A. Fixing the readouts

We use the fact that, having as a goal the computation of the leading order of the free entropy, the readouts $\mathbf{v}$ of the learner can be fixed from the beginning to those of the target $\mathbf{v}^0$. The proof has been given in Remark 3 for the mutual information (and further discussed and tested in Appendix B 7) and implies directly at the level of free entropy that

$$\frac{1}{n} \mathbb{E} \ln \mathcal{Z}_{\mathbf{v}=\mathbf{v}^0} = \frac{1}{n} \mathbb{E} \ln \mathcal{Z}_{\mathbf{v}^{\text{learnable}}} + O(1/d).$$

Consequently, for the rest of the derivations, we set without loss of generality $\mathbf{v} = \mathbf{v}^0$, thus leaving as learnable parameters the (many more) inner weights. If keeping learnable readouts, the (equivalent) replica calculation would be more cumbersome and ultimately yield that the overlap between $\mathbf{v}$ and $\mathbf{v}^0$ is irrelevant for all the other OPs in the universal phase, and once the student specializes, the readouts corresponding to specialized neurons are concurrently exactly recovered.

B. Shallow MLP

We start with the shallow case $L = 1$. Having directly fixed the readouts, the partition function is redefined as

$$\mathcal{Z}(\mathcal{D}) := \mathcal{Z}_{\mathbf{v}=\mathbf{v}^0} = \int dP_W(\mathbf{W}) \prod_{\mu \leq n} P_{\text{out}}(y_\mu | \lambda_\mu(\boldsymbol{\theta}))$$

with $\lambda_\mu(\boldsymbol{\theta}) := \mathcal{F}_\theta^{(1)}(\mathbf{x}_\mu)$ and $\boldsymbol{\theta} = (\mathbf{W}, \mathbf{v}^0)$. The quenched variables, averaged by the symbol $\mathbb{E} = \mathbb{E}_{\mathcal{D}}$, are the data $\mathcal{D}$ which depend on inputs and the teacher. Equivalently, $\mathbb{E}[\cdot] = \mathbb{E}_{(\mathbf{x}_\mu)} \mathbb{E}_{\boldsymbol{\theta}^0} \mathbb{E}_{(y_\mu | (\mathbf{x}_\mu), \boldsymbol{\theta}^0)}[\cdot]$.

1. Replicated system and order parameters

The starting point to tackle the data average is the usual replica trick:

$$\begin{aligned} \lim_{n \rightarrow 0^+} \frac{1}{n} \mathbb{E} \ln \mathcal{Z}(\mathcal{D}) &= \lim_{s \rightarrow 0^+} \lim_{ns} \frac{1}{ns} \ln \mathbb{E} \mathcal{Z}^s \\ &= \lim_{s \rightarrow 0^+} \lim_{ns} \frac{1}{ns} \ln \mathbb{E} \mathcal{Z}^s \end{aligned}$$

assuming the limits commute. Consider first $s \in \mathbb{N}^+$. Let $\boldsymbol{\theta}^a = (\mathbf{W}^a, \mathbf{v}^0)$, $(\mathbf{x}, y) = (\mathbf{x}_1, y_1)$, and the replicas of the

postactivation, including the teacher's $a = 0$:

$$\left\{ \lambda^a(\theta^a) := \frac{1}{\sqrt{k}} \mathbf{v}^{0\top} \sigma \left(\frac{1}{\sqrt{d}} \mathbf{W}^a \mathbf{x} \right) \right\}_{a=0}^s.$$

We directly get

$$\mathbb{E} \mathcal{Z}^s = \mathbb{E}_{\mathbf{v}^0} \int \prod_a^{0,s} dP_W(\mathbf{W}^a) \left[\mathbb{E}_{\mathbf{x}} \int dy \prod_a^{0,s} P_{\text{out}}(y | \lambda^a(\theta^a)) \right]^n. \quad (23)$$

The key is to identify the law of the replicas $\{\lambda^a\}_{a=0}^s$, which are dependent random variables due to the common random Gaussian input $\mathbf{x}$, conditionally on (θ^a) . As explained and checked numerically in Sec. II, our main hypothesis is that $\{\lambda^a\}$ are jointly Gaussian for the (θ^a) that dominate the partition function (i.e., posterior samples), an *Ansatz* we cannot prove but that we validate *a posteriori* thanks to the excellent match between the theory and the empirical generalization curves; see Eq. (11), Remark 2, and Sec. III.

Given two replica indices $a, b \in \{0, \dots, s\}$ we define the neuron-neuron overlap matrix

$$\Omega_{ij}^{ab} := \frac{1}{d} \mathbf{W}_i^{a\top} \mathbf{W}_j^b, \quad i, j \in [k].$$

Recalling σ 's Hermite expansion, Mehler's formula (see Appendix A 2) implies that the postactivations covariance is

$$K^{ab} := \mathbb{E}[\lambda^a \lambda^b | \theta^a, \theta^b] = \sum_{\ell \geq 1} \frac{\mu_\ell^2}{\ell!} R_\ell^{ab}, \quad (24)$$

with the infinitely many overlap OPs

$$R_\ell^{ab} := \frac{1}{k} \sum_{i,j \leq k} v_i^0 v_j^0 (\Omega_{ij}^{ab})^\ell, \quad \ell \geq 1. \quad (25)$$

This covariance $\mathbf{K}$ is complicated but, as we argue hereby, simplifications occur as $d \rightarrow \infty$. In particular, the first two overlaps R_1^{ab}, R_2^{ab} are special. We claim that higher-order overlaps $(R_\ell^{ab})_{\ell \geq 3}$ can be simplified as functions of simpler OPs.

2. Simplifying the order parameters

In this section, we show how to drastically reduce the number of OPs (25) to track. To build some intuition, it is convenient to define the symmetric tensors $\mathbf{S}_\ell^a$ with

$$S_{\ell; \alpha_1, \dots, \alpha_\ell}^a := \frac{1}{\sqrt{k}} \sum_{i \leq k} v_i^0 W_{i\alpha_1}^a \cdots W_{i\alpha_\ell}^a. \quad (26)$$

Indeed, the generic ℓ th overlap (25) can be written as $R_\ell^{ab} = (\mathbf{S}_\ell^a \cdot \mathbf{S}_\ell^b) / d^\ell$ (where “ $\cdot$ ” is the inner product among tensors obtained by contracting all the indices), e.g., $R_2^{ab} = \text{Tr} \mathbf{S}_2^a \mathbf{S}_2^b / d^2$. The following assumptions consider how these tensors behave for $\ell = 1, 2$ and $\ell \geq 3$. Let us start from the latter case.

First, we assume that for Hadamard powers $\ell \geq 3$, the off-diagonal of the overlap $(\Omega^{ab})^{\circ \ell}$ obtained from i.i.d. weight matrices sampled from the posterior, is small enough to be discarded:

$$(\Omega_{ij}^{ab})^\ell \approx \delta_{ij} (\Omega_{ii}^{ab})^\ell \quad \text{if } \ell \geq 3. \quad (27)$$

Approximate equality is up to a matrix with $o_d(1)$ operator norm. In other words, the vector $\mathbf{W}_i^a$ of a student (replica) is assumed to possibly align, for each i , only with a single $\mathbf{W}_j^b$ of the teacher (or, by Bayes optimality, of another replica) indexed by $j = \pi_i$, with π a permutation. The model is symmetric under permutations of hidden neurons with the same readout value. We thus take π to be the identity without loss of generality. The same “concentration on the diagonal” happens, e.g., for a standard Wishart matrix, which is the extreme case for Ω^{ab} if $\mathbf{W}^a = \mathbf{W}^b$ and $P_W = \mathcal{N}(0, 1)$. Its eigenvectors and those of its Hadamard square are delocalized, while higher Hadamard powers $\ell \geq 3$ have strongly localized eigenvectors [136] (consequently, R_2^{ab} will require a separate treatment).

Moreover, assume that the readout prior has discrete support $\text{Supp}(P_v) =: \mathbf{V} = \{\mathbf{v}\}$; this can be relaxed by binning to a continuous support, as mentioned in Sec. II. By exchangeability among neurons with the same readout value, we further assume that all diagonal elements $\{\Omega_{ii}^{ab} | i \in \mathcal{I}_v\}$ concentrate onto the constant $\mathcal{Q}^{ab}(\mathbf{v})$, where $\mathcal{I}_v := \{i \leq k | v_i^0 = \mathbf{v}\}$:

$$(\Omega_{ij}^{ab})^\ell \approx \delta_{ij} \mathcal{Q}^{ab}(\mathbf{v})^\ell \quad \text{if } \ell \geq 3 \text{ and } i \text{ or } j \in \mathcal{I}_v. \quad (28)$$

Equivalently, under the neuron exchangeability assumption, by summing over the indices $i \in \mathcal{I}_v$ and dividing by their number, the constant $\mathcal{Q}^{ab}(\mathbf{v})$ can be written as

$$\mathcal{Q}^{ab}(\mathbf{v}) := \frac{1}{|\mathcal{I}_v|d} \sum_{i \in \mathcal{I}_v} (\mathbf{W}^a \mathbf{W}^{b\top})_{ii}.$$

This definition is directly related to the way we measure overlaps in numerical experiments, as empirical averages are less affected by finite-size effects than specific choices of $i \in \mathcal{I}_v$; thus, we adopt this definition also in our theoretical analysis. The advantage in switching from Eq. (27) to Eq. (28), i.e., in labeling the neurons with their readout value, is an expression suitable for the asymptotic regime we are considering, where the neurons are infinitely many. Indeed, with these simplifications we can write

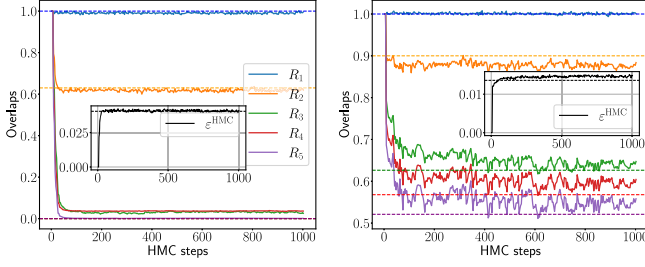


FIG. 20. Hamiltonian Monte Carlo dynamics of the overlaps $R_\ell = R_\ell^{01}$ between student and teacher weights for $\ell \in [5]$, with activation function for $L = 1$ with $\text{ReLU}(x)$ activation, $d = 200$, $\gamma = 0.5$, linear readout with $\Delta = 0.1$ and two choices of sample rates and readouts: $\alpha = 1.0$ with $P_v = \delta_1$ (left) and $\alpha = 3.0$ with $P_v = \mathcal{N}(0, 1)$ (right). The teacher weights $\mathbf{W}^0$ are Gaussian, and the readouts are fixed during sampling to the teacher ones. The dynamics is initialized informatively, i.e., on $\mathbf{W}^0$. The overlap R_1 always fluctuates around 1. Left: the overlaps R_ℓ for $\ell \geq 3$ at equilibrium converge to 0, while R_2 is well estimated by the theory (orange dashed line). Right: at higher sample rate α , also the R_ℓ 's for $\ell \geq 3$ nonzero and agree with their theoretical prediction (dashed lines). Insets show the mean-square generalization error attained by HMC (solid) and the theoretical prediction ϵ^{opt} (dashed).

$$R_\ell^{ab} = \mathbb{E}_{v \sim P_v} v^2 Q^{ab}(v)^\ell + o_d(1) \quad \text{for } \ell \geq 3. \quad (29)$$

This assumption has also a natural interpretation in terms of the tensors $\mathbf{S}_\ell^a$: In the absence of specialization, $Q^{ab}(\mathbf{v}) = 0$ for all $\mathbf{v}$, so $R_\ell^{ab} = (\mathbf{S}_\ell^a \cdot \mathbf{S}_\ell^b)/d^\ell = 0$ according to Eq. (29). Indeed, a nonspecialized model with $kd = \Theta(d^2)$ parameters and $n = \Theta(d^2)$ data cannot learn these tensors, as this would require the knowledge of $\Theta(d^\ell/\ell!)$ entries and a comparable amount of tunable parameters. On the contrary, once specialization occurs the model is able to factorize them using the rhs of Eq. (26).

Our assumption is verified numerically *a posteriori* as follows. Identity (29) is true [without $o_d(1)$] for the predicted theoretical values of the OPs by construction of our theory. Figure 7 verified the good agreement between theoretical and experimental overlap profiles $Q^{01}(\mathbf{v})$ for all $\mathbf{v} \in \mathbf{V}$ [which is statistically the same as $Q^{ab}(\mathbf{v})$ for any $a \neq b$ by the so-called Nishimori identity following from Bayes optimality; see Appendix. A 3], while Fig. 20 checks the agreement at the level of (R_ℓ^{ab}) . Consequently, Eq. (29) also holds for the experimental overlaps.

Having simplified the $\ell \geq 3$ terms in the series (24), let us pass to the $\ell = 1$ case. Given that the number of data $n = \Theta(d^2)$ and that the corresponding $(\mathbf{S}_1^a)$ are only d dimensional, they are reconstructed perfectly (the same argument was used in Remark 3 to argue that the readouts $\mathbf{v}$ can be quenched). We thus assume right away that at equilibrium the overlaps $R_1^{ab} = 1$ (or saturate to their maximum value; if tracked, the corresponding saddle-point

equations end up being trivial and do fix this). In other words, in the quadratic data regime, the μ_1 contribution in the Hermite decomposition of σ for the target is perfectly learnable, while higher-order ones play a nontrivial role. In contrast, Ref. [109] studies the regime $n = \Theta(d)$ where μ_1 is the only learnable term.

Then, the average replicated partition function reads

$$\mathbb{E} Z^s = \int d\mathbf{R}_2 d\mathbf{Q} \exp(F_S + nF_E),$$

where F_E, F_S depend on $\mathbf{R}_2 = (R_2^{ab})$ and $\mathbf{Q} := \{Q^{ab} | a \leq b\}$, where $Q^{ab} := \{Q^{ab}(\mathbf{v}) | \mathbf{v} \in \mathbf{V}\}$.

The “energetic potential” is defined as

$$e^{nF_E} := \left[\int dy d\lambda \frac{\exp(-\frac{1}{2} \lambda^\top \mathbf{K}^{-1} \lambda)}{[(2\pi)^{s+1} \det \mathbf{K}]^{1/2}} \prod_a^{0,s} P_{\text{out}}(y | \lambda^a) \right]^n.$$

It takes this form due to our Gaussian assumption on the replicated postactivations and is thus easily computed; see Appendix B 1 a.

The “entropic potential” F_S taking into account the degeneracy of the OPs entering the covariance of the postactivations is obtained by averaging delta functions fixing their definitions with respect to the microscopic degrees of freedom $(\mathbf{W}^a)$. It can be written compactly using the following conditional law over the tensors $(\mathbf{S}_2^a)_{a=0}^s$:

$$\begin{aligned} P((\mathbf{S}_2^a) | \mathbf{Q}) &:= V_W^{kd}(\mathbf{Q})^{-1} \int \prod_a^{0,s} dP_W(\mathbf{W}^a) \\ &\times \prod_{a \leq b}^{0,s} \prod_{\mathbf{v} \in \mathbf{V}} \delta(|\mathcal{I}_v| d Q^{ab}(\mathbf{v}) - \sum_{i \in \mathcal{I}_v} \mathbf{W}_i^{a\top} \mathbf{W}_i^b) \\ &\times \prod_a^{0,s} \delta(\mathbf{S}_2^a - \mathbf{W}^{a\top} \text{diag}(\mathbf{v}^0) \mathbf{W}^a / \sqrt{k}), \end{aligned} \quad (30)$$

with normalization $V_W^{kd} = V_W^{kd}(\mathbf{Q})$ given by

$$\begin{aligned} V_W^{kd} &= \int \prod_a^{0,s} dP_W(\mathbf{W}^a) \\ &\times \prod_{a \leq b}^{0,s} \prod_{\mathbf{v} \in \mathbf{V}} \delta(|\mathcal{I}_v| d Q^{ab}(\mathbf{v}) - \sum_{i \in \mathcal{I}_v} \mathbf{W}_i^{a\top} \mathbf{W}_i^b). \end{aligned}$$

The entropy of $(\mathbf{R}_2, \mathbf{Q})$, which is the challenging term to compute, then reads

$$e^{F_S} := V_W^{kd} \int dP((\mathbf{S}_2^a) | \mathbf{Q}) \prod_{a \leq b}^{0,s} \delta(d^2 R_2^{ab} - \text{Tr} \mathbf{S}_2^a \mathbf{S}_2^b).$$

3. Tackling the entropy: Measure simplification by moment matching

The delta functions above fixing R_2^{ab} in the entropy of $\mathbf{R}_2$ conditional on $\mathcal{Q}$ induce quartic constraints between the degrees of freedom (W_{ia}^a) instead of quadratic as usual. A direct computation thus seems out of reach. However, we will exploit the fact that the constraints are quadratic in the matrices ($\mathbf{S}_2^a$). Consequently, shifting our focus toward ($\mathbf{S}_2^a$) as the basic degrees of freedom to integrate rather than (W_{ia}^a) will allow us to move forward by simplifying their measure (30). Note that while $(W_{ia}^a)_{i,a}$ are i.i.d. under the prior P_W , $\mathbf{S}_2^a$ has dependent entries under its corresponding prior. This important fact is taken into account as follows.

Define P_S as the probability density of a generalized Wishart random matrix $\tilde{\mathbf{W}}^\top \text{diag}(\mathbf{v}^0) \tilde{\mathbf{W}} / \sqrt{k}$, where $\tilde{\mathbf{W}} \in \mathbb{R}^{k \times d}$ is made of i.i.d. standard *Gaussian* entries. The simplification we consider consists in replacing Eq. (30) by the effective measure

$$\begin{aligned} & \tilde{P}((\mathbf{S}_2^a) | \mathcal{Q}) \\ & := \tilde{V}_W^{kd}(\mathcal{Q})^{-1} \prod_a^{0,s} P_S(\mathbf{S}_2^a) \prod_{a < b}^{0,s} e^{\frac{1}{2} \tau(\mathcal{Q}^{ab}) \text{Tr} \mathbf{S}_2^a \mathbf{S}_2^b}, \end{aligned} \quad (31)$$

where $\tilde{V}_W^{kd} = \tilde{V}_W^{kd}(\mathcal{Q})$ is a normalization constant, and

$$\tau(\mathcal{Q}^{ab}) := \text{MMSE}_S^{-1} \{1 - \mathbb{E}_{v \sim P_v} [v^2 \mathcal{Q}^{ab}(v)^2]\}. \quad (32)$$

The rationale behind this choice goes as follows. The matrices ($\mathbf{S}_2^a$) are, under the measure (30), (i) similar to generalized Wishart matrices, but instead constructed from (ii) non-Gaussian factors ($\mathbf{W}^a$), which (iii) are coupled between different replicas, thus inducing a coupling among replicas ($\mathbf{S}^a$). The proposed simplified measure captures all three aspects while remaining tractable, as we explain now.

The first assumption is that in the measure (30) the details of the (centered, unit variance) prior P_W enter only through $\mathcal{Q}$ at leading order. Because of the conditioning, we can thus relax it to Gaussian (with the same two first moments) by universality, as is often the case in random matrix theory. P_W will instead explicitly enter in the entropy of $\mathcal{Q}$ related to V_W^{kd} . Point (ii) is thus taken care of by the conditioning. Then, the generalized Wishart prior P_S encodes (i) and, finally, the exponential tilt in $\tilde{P}$ induces the replica couplings of point (iii).

It now remains to capture the correct dependence of measure (30) on $\mathcal{Q}$. This is done by realizing that

$$\int dP((\mathbf{S}_2^a) | \mathcal{Q}) \frac{1}{d^2} \text{Tr} \mathbf{S}_2^a \mathbf{S}_2^b = \mathbb{E}_{v \sim P_v} [v^2 \mathcal{Q}^{ab}(v)^2] + \gamma \bar{v}^2.$$

This is shown in Appendix B 1 c. The Lagrange multiplier $\tau(\mathcal{Q}^{ab})$ to plug into $\tilde{P}$ enforcing this moment-matching condition between true and simplified measures as $s \rightarrow 0^+$

is Eq. (32); see Appendix B 1 e. For completeness, we provide in Appendix B 2 alternatives to the simplification (31), whose analysis is left for future work.

4. Final steps and spherical integration

Combining all our findings, the average replicated partition function is simplified as

$$\begin{aligned} \mathbb{E} \mathcal{Z}^s &= \int d\mathbf{R}_2 d\mathcal{Q} e^{n F_E + kd \ln V_W(\mathcal{Q}) - kd \ln \tilde{V}_W(\mathcal{Q})} \\ &\times \prod_a^{0,s} P_S(\mathbf{S}_2^a) \prod_{a < b}^{0,s} e^{\frac{1}{2} \tau(\mathcal{Q}^{ab}) \text{Tr} \mathbf{S}_2^a \mathbf{S}_2^b} \\ &\times \prod_{a < b}^{0,s} \delta(d^2 R_2^{ab} - \text{Tr} \mathbf{S}_2^a \mathbf{S}_2^b). \end{aligned}$$

The equality should be interpreted as holding at leading exponential order $\exp[\Theta(n)]$, assuming the validity of our previous measure simplification. All remaining steps but the last are standard:

- (i) Express the delta functions fixing $\mathcal{Q}, \mathbf{R}_2$ in exponential form using their Fourier representation; this introduces additional Fourier conjugate OPs $\hat{\mathcal{Q}}, \hat{\mathbf{R}}_2$ of same respective dimensions.
- (ii) Once this is done, the terms coupling different replicas of ($\mathbf{W}^a$) or of ($\mathbf{S}^a$) are all quadratic. Using the Hubbard-Stratonovich transformation [i.e., $\mathbb{E}_{\mathbf{Z}} \exp[(d/2) \text{Tr} \mathbf{M} \mathbf{Z}] = \exp[(d/4) \text{Tr} \mathbf{M}^2]$ for a $d \times d$ symmetric matrix $\mathbf{M}$ with $\mathbf{Z}$ a standard GOE matrix] therefore allows us to linearize all replica-replica coupling terms, at the price of introducing new Gaussian fields interacting with all replicas.
- (iii) After these manipulations, we identify at leading exponential order an effective action $\mathcal{S}$ depending on the OPs only, which allows a saddle-point integration with respect to them as $n \rightarrow \infty$:

$$\begin{aligned} \lim_{ns} \frac{1}{ns} \ln \mathbb{E} \mathcal{Z}^s &= \lim_{ns} \frac{1}{ns} \ln \int d\mathbf{R}_2 d\hat{\mathbf{R}}_2 d\mathcal{Q} d\hat{\mathcal{Q}} e^{n \mathcal{S}} \\ &= \frac{1}{s} \text{extr} \mathcal{S}. \end{aligned}$$

- (iv) Next, the replica limit $s \rightarrow 0^+$ of the previously obtained expression has to be considered. To do so, we make a replica-symmetric assumption; i.e., we consider that at the saddle point, all OPs entering the action $\mathcal{S}$, and thus K^{ab} , too, take a simple form of the type $R^{ab} = R_d \delta_{ab} + R(1 - \delta_{ab})$. Replica symmetry is rigorously known to be correct in Bayes-optimal learning and is thus justified here; see Refs. [186,204].
- (v) The resulting expression still includes two high-dimensional integrals related to the $\mathbf{S}_2$ matrices. They correspond to the free entropies associated

with the Bayes-optimal denoising of a generalized Wishart matrix, described above Result 1, for two signal-to-noise ratios. The last step deals with these matrix integrals over rotationally invariant matrices using the HCIZ integral, whose form is tractable in this case [102,103]. These free entropies yield the two last terms $\iota(\cdot)$ in $f_{\text{RS}}^{(1)}$, Eq. (15).

The complete derivation in Appendix B 1 gives Result 1. From the meaning of the OPs, this analysis also yields the postactivations covariance $\mathbf{K}$ and thus Result 2.

As a final remark, we emphasize again a key difference between our approach and earlier works on extensive-rank systems. If, instead of taking the generalized Wishart prior P_S as the base measure over the matrices $(\mathbf{S}_2^a)$ in the simplified $\tilde{P}$ with moment matching, one takes a factorized Gaussian measure, thus entirely forgetting the dependences among $\mathbf{S}_2^a$ entries; this mimics the Sakata-Kabashima replica method [168]. Our *Ansatz* thus captures important correlations neglected in Refs. [168,169,171,188] in the context of linear-rank matrix inference. For completeness, we show in Appendix B 2 that our *Ansatz* indeed improves the prediction compared to these earlier approaches.

C. Two-hidden-layer MLP

We now move to the deep MLP by first considering the $L = 2$ case. We highlight here only the crucial steps that make the derivation different with respect to the one sketched above, referring the reader to Appendix C 1 for more details. We assume this time that $\mu_2 = 0$ (in addition to $\mu_0 = 0$): In this way, our approach simplifies considerably, as the matrix degrees of freedom involved in the second Hermite components of the activation functions do not appear in the theory. We will see however that, due to the deep structure of the network, matrix degrees of freedom coming from combinations of the model's hidden weights are still entering the theory and will require the use of a rectangular spherical integral (see Appendix C 1). Popular activation functions (e.g., all the odd ones) comply with the requirement $\mu_2 = 0$.

1. Replicated system, order parameters, and their simplification

Replicas of the postactivations can now be written recursively as

$$\left\{ \lambda^a(\theta^a) := \frac{1}{\sqrt{k_2}} \mathbf{v}^{0\top} \sigma^{(2)}(\mathbf{h}^{(2)a}) \right\}_{a=0}^s, \quad \left\{ \mathbf{h}^{(l)a} := \frac{1}{\sqrt{k_{l-1}}} \mathbf{W}^{(l)a} \sigma^{(l-1)}(\mathbf{h}^{(l-1)a}) \right\}_{a=0, \dots, s; l=1,2},$$

where we allowed different activations at each layer, and we used the notation $\sigma^{(0)}(x) := x$, $k_0 = d$ and $\mathbf{h}^{(0)a} = \mathbf{x}$ for all a . For the sake of presentation, we further require the

normalization $\mathbb{E}_{z \sim \mathcal{N}(0,1)} \sigma^{(l)}(z)^2 = 1$ for all l , to avoid tracking this variance in the following (the case of generic variance can be derived from Appendix A 2). The expectation over the input $\mathbf{x}$ at given weights can be done as in Eq. (23) by assuming the same joint Gaussianity of the postactivations (λ^a) as in the shallow case. Moreover, to use recursively Mehler's formula, we also assume that the pair of preactivations $(h_i^{(2)a}, h_j^{(2)b})$ is jointly Gaussian for any choice of $a, b = 0, \dots, s$ and $i, j \leq k_2$. With these assumptions, the covariances $K^{ab} := \mathbb{E}_{\mathbf{x}} \lambda^a \lambda^b$ and $\Omega^{(l)ab} := \mathbb{E}_{\mathbf{x}} \mathbf{h}^{(l)a} \mathbf{h}^{(l)b\top}$ can be written as

$$\begin{aligned} K^{ab} &= \frac{(\mu_1^{(2)})^2}{k_2} \mathbf{v}^{0\top} \Omega^{(2)ab} \mathbf{v}^0 + \frac{1}{k_2} \mathbf{v}^{0\top} g^{(2)}(\Omega^{(2)ab}) \mathbf{v}^0, \\ \Omega^{(2)ab} &= \frac{(\mu_1^{(1)})^2}{k_1} \mathbf{W}^{(2)a} \Omega^{(1)ab} \mathbf{W}^{(2)b\top} \\ &\quad + \frac{1}{k_1} \mathbf{W}^{(2)a} g^{(1)}(\Omega^{(1)ab}) \mathbf{W}^{(2)b\top}, \\ \Omega^{(1)ab} &= \frac{1}{k_0} \mathbf{W}^{(1)a} \mathbf{W}^{(1)b\top}, \end{aligned}$$

where the functions

$$g^{(l)}(x) := \sum_{\ell=3}^{\infty} \frac{(\mu_{\ell}^{(l)})^2}{\ell!} x^{\ell}$$

are applied entrywise to matrices, and $\mu_{\ell}^{(l)}$ is the ℓ th Hermite coefficient of $\sigma^{(l)}$.

Unfolding the above recursion, the covariance K^{ab} can be written in terms of overlaps of “effective” hidden weights and readout vectors

$$\begin{aligned} \mathbf{W}^{(2:1)a} &:= \frac{\mathbf{W}^{(2)a} \mathbf{W}^{(1)a}}{\sqrt{k_1}}, \\ \mathbf{v}^{(1)a} &:= \frac{\mathbf{W}^{(1)a\top} \mathbf{W}^{(2)a\top} \mathbf{v}^0}{\sqrt{k_2 k_1}}, \quad \mathbf{v}^{(2)a} := \frac{\mathbf{W}^{(2)a\top} \mathbf{v}^0}{\sqrt{k_2}}, \end{aligned}$$

each of them arising from combinations of the activation's linear components. We also set $\mathbf{v}^{(3)a} := \mathbf{v}^0$. Moreover, simplifications can be taken along the three following lines:

- (i) $\mathbf{v}^{(1)a\top} \mathbf{v}^{(1)b} / k_0$ is an overlap of d -dimensional vectors: As explained above (see the discussion on R_1^{ab} in the shallow case), it can be directly taken to be 1 in the quadratic data regime.
- (ii) Wherever a function $g^{(l)}$ involving only Hadamard powers greater than 2 is applied to a matrix overlap, we assume the resulting matrix to be diagonal in the limit $g^{(l)}(\Omega^{(l)ab})_{ij} \approx \delta_{ij} g^{(l)}(\Omega^{(l)ab})_{ii}$, as we did in Eq. (27).

- (iii) The components $v_i^{(2)a}, v_i^{(2)b}$ of the effective readouts enter the above expressions only if $\Omega_{ii}^{(1)ab} \neq 0$, that is if some specialization has occurred in the previous layer. As these components are $\Theta(k_1)$ in number, they can be reconstructed exactly with $\Theta(d^2)$ data. We can thus take these vectors as given, $\mathbf{v}^{(2)a\top} = \mathbf{v}^{(2)0\top}$. From central limit theorem, the components of $\mathbf{v}^{(2)0\top}$ are standard Gaussian variables $v_i^{(2)0} \sim \mathcal{N}(0, 1)$.

From point (ii), and extending the approach we followed for the shallow case, we are naturally led to consider as OPs the diagonal profiles of the overlap matrices $(\Omega_{ii}^{(l)ab})_i$. Moreover, from point (iii) we can label internal neurons (say, the ones in layer l) with the value of the effective readout to which they are connected ($v_i^{(l+1)0}$) rather than with their index ($i \leq k_l$). By binning the distribution of the elements of $\mathbf{v}^{(l)0}$, we define the sets of indices $\mathcal{I}_{v^{(l)}} := \{i \leq k_{l-1} | v_i^{(l)0} = v^{(l)}\}$, while keeping $\mathcal{I}_v$ (with no layer label) as in Eq. (28). As before, in order to define the OPs, we further assume exchangeability among neurons with the same effective readout value [e.g., $(\mathbf{W}^{(1)a}\mathbf{W}^{(1)b\top})_{ii}/d =: Q_1^{ab}(\mathbf{v}^{(2)})$ for all $i \in \mathcal{I}_{v^{(2)}}$]. Equivalently, by summing over these indices and normalizing by their number, we obtain

$$\begin{aligned} Q_1^{ab}(\mathbf{v}^{(2)}) &:= \frac{1}{|\mathcal{I}_{v^{(2)}}|d} \sum_{i \in \mathcal{I}_{v^{(2)}}} (\mathbf{W}^{(1)a}\mathbf{W}^{(1)b\top})_{ii}, \\ Q_2^{ab}(\mathbf{v}, \mathbf{v}^{(2)}) &:= \frac{1}{|\mathcal{I}_{v^{(2)}}||\mathcal{I}_v|} \sum_{i \in \mathcal{I}_{v^{(2)}}, j \in \mathcal{I}_v} W_{ji}^{(2)a} W_{ji}^{(2)b}, \\ Q_{2:1}^{ab}(\mathbf{v}) &:= \frac{1}{|\mathcal{I}_v|d} \sum_{i \in \mathcal{I}_v} (\mathbf{W}^{(2:1)a}\mathbf{W}^{(2:1)b\top})_{ii}. \end{aligned}$$

The bold notations $\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{Q}_{2:1}$ are defined analogously to the shallow case. In terms of these, the covariance of the postactivations reads

$$\begin{aligned} K^{ab} &= (\mu_1^{(2)}\mu_1^{(1)})^2 + (\mu_1^{(2)})^2 \mathbb{E}_{v^{(2)} \sim \mathcal{N}(0,1)} (v^{(2)})^2 g^{(1)}(Q_1^{ab}(v^{(2)})) \\ &\quad + \mathbb{E}_{v \sim P_v} v^2 g^{(2)}[(\mu_1^{(1)})^2 Q_{2:1}^{ab}(v)] \\ &\quad + \mathbb{E}_{v^{(2)} \sim \mathcal{N}(0,1)} Q_2^{ab}(v, v^{(2)}) g^{(1)}(Q_1^{ab}(v^{(2)})). \end{aligned}$$

The structure of this covariance informs us on how the learning can or cannot take place. E.g., if the first layer does not specialize $Q_1^{ab}(v^{(2)}) = 0$, then the second layer cannot either because its associated overlap disappears.

From this point on, the energetic potential F_E follows: It is the very same as for $L = 1$ but using the above covariance.

2. Tackling the entropy using the rectangular spherical integral

The entropic potential accounting for the degeneracy of the OPs requires some care. We first define the conditional law over the matrices $(\mathbf{W}^{(2:1)a})$:

$$\begin{aligned} P((\mathbf{W}^{(2:1)a}) | \mathbf{Q}_1, \mathbf{Q}_2) &\propto \int \prod_{a=0}^s \prod_{l=1}^2 dP_{W_l}(\mathbf{W}^{(l)a}) \\ &\quad \times \delta(\mathbf{W}^{(2:1)a} - \mathbf{W}^{(2)a}\mathbf{W}^{(1)a}/\sqrt{k_1}) \\ &\quad \times \prod_{a \leq b}^{0,s} \prod_{v^{(2)} \in \mathbf{V}^{(2)}} \delta\left(|\mathcal{I}_{v^{(2)}}| d Q_1^{ab}(\mathbf{v}^{(2)}) \right. \\ &\quad \left. - \sum_{i \in \mathcal{I}_{v^{(2)}}} (\mathbf{W}^{(1)a}\mathbf{W}^{(1)b\top})_{ii}\right) \\ &\quad \times \prod_{v \in \mathbf{V}} \delta\left(|\mathcal{I}_{v^{(2)}}| |\mathcal{I}_v| Q_2^{ab}(\mathbf{v}, \mathbf{v}^{(2)}) \right. \\ &\quad \left. - \sum_{i \in \mathcal{I}_{v^{(2)}}, j \in \mathcal{I}_v} W_{ji}^{(2)a} W_{ji}^{(2)b}\right), \end{aligned}$$

where $\mathbf{v}^{(2)}$ is the binned support of $\mathcal{N}(0, 1)$. In this way, we can write the entropic contribution as

$$\begin{aligned} e^{F_s} &= V_{2:1}^{k_2 d} \int dP((\mathbf{W}^{(2:1)a}) | \mathbf{Q}_1, \mathbf{Q}_2) \prod_{a \leq b}^{0,s} \prod_{v \in \mathbf{V}} \\ &\quad \times \delta\left(|\mathcal{I}_v| d Q_{2:1}^{ab}(\mathbf{v}) - \sum_{i \in \mathcal{I}_v} (\mathbf{W}^{(2:1)a}\mathbf{W}^{(2:1)b\top})_{ii}\right), \end{aligned}$$

where $V_{2:1}^{k_2 d}$, depending implicitly on $\mathbf{Q}_1, \mathbf{Q}_2$, is the normalization factor of $P((\mathbf{W}^{(2:1)a}) | \mathbf{Q}_1, \mathbf{Q}_2)$.

The evaluation of the last integral involves coupled replicas of matrices with correlated entries. We deal with it by relaxing the measure $P((\mathbf{W}^{(2:1)a}) | \mathbf{Q}_1, \mathbf{Q}_2)$ to a tractable one, still able to capture the correlations between these degrees of freedom. To this aim, we first observe that, asymptotically, under the true conditional measure

$$\begin{aligned} \frac{1}{|\mathcal{I}_v|d} \sum_{i \in \mathcal{I}_v} \mathbb{E}[(\mathbf{W}^{(2:1)a}\mathbf{W}^{(2:1)b\top})_{ii} | \mathbf{Q}_1, \mathbf{Q}_2] \\ \approx \mathbb{E}_{v^{(2)} \sim \mathcal{N}(0,1)} Q_2^{ab}(\mathbf{v}, v^{(2)}) Q_1^{ab}(v^{(2)}). \end{aligned}$$

In order to match this moment in our relaxation, we take a tractable base measure with exponential tilts for each value $v \in \mathbf{V}$:

$$\begin{aligned} d\bar{P}((\mathbf{W}^{(2:1)a}) | \mathbf{Q}_1, \mathbf{Q}_2) \\ = \prod_{v \in \mathbf{V}} V(\tau_v)^{-1} \prod_{a=0}^s dP_{UV}(\mathbf{W}_v^{(2:1)a}) e^{\sum_{a < b, 0} \tau_v^{ab} \text{Tr} \mathbf{W}_v^{(2:1)a} \mathbf{W}_v^{(2:1)b\top}}, \end{aligned}$$

where $\mathbf{W}_v^{(2:1)a} = (\mathbf{W}_i^{(2:1)a})_{i \in \mathcal{I}_v}$ and dP_{UV} is the law of the product of two matrices with i.i.d. Gaussian entries ($\mathbf{U} \in \mathbb{R}^{|\mathcal{I}_v| \times k_1}$, $\mathbf{V} \in \mathbb{R}^{k_1 \times d}$), $\tau_v = (\tau_v^{ab})_{a,b}$ is a function of $\mathcal{Q}_1, \mathcal{Q}_2$ fixed by the previous moment matching, and $V(\tau_v)$ is a normalization factor. With this relaxation, the entropic contribution can be evaluated explicitly using the rectangular spherical integral, leading eventually to Result 3 (see Appendix C 1 for more details).

D. Three-or-more hidden-layer MLP

To tackle the $L \geq 3$ case, one could push forward the approach we presented in the previous section. However, even with the simplification of considering only activations with no second Hermite component, the analysis remains very challenging, as the postactivation covariance involves the “effective weights”

$$\mathbf{W}^{(l':l)a} := \frac{1}{\sqrt{k_{l'-1}k_{l-2} \dots k_l}} \mathbf{W}^{(l')a} \mathbf{W}^{(l'-1)a} \dots \mathbf{W}^{(l)a},$$

as shown in Appendix C 1. These terms appear through combinations of the linear components of the activation functions $\sigma^{(l'-1)}, \dots, \sigma^{(l)}$. To evaluate the entropic contributions of the OPs $\mathcal{Q}_{l':l}$ defined as overlaps between replicas of the above effective weights, one has to consider all the possible learning mechanisms the network could adopt: a totally unspecialized strategy, where no $\mathbf{W}^{(l')a}$ entering $\mathbf{W}^{(l':l)a}$ is learned by itself, but still $\mathbf{W}^{(l')a}$ is learned as a whole; a totally specialized strategy, where the model is able to learn separately all the $\mathbf{W}^{(l')a}$; mixed strategies where some subsets of layers are specialized while others are not. All these mechanisms give explicit contributions to the correlation between teacher and student and could correspond to different phases of the system. We leave the study of this rich phase diagram for future works.

To simplify the picture, we require activations such that $\mu_0^{(l)} = \mu_1^{(l)} = \mu_2^{(l)} = 0$ and set $\mathbb{E}_{z \sim \mathcal{N}(0,1)} \sigma^{(l)}(z)^2 = 1$ for all $l \leq L$, which this time implies $g^{(l)}(1) = 1$. In this case, the effective weights $\mathbf{W}^{(l':l)a}$ do not enter the postactivations covariance, simplifying considerably our analysis because the only possible learning strategy is specialization at all layers. Indeed,

$$\{\lambda^a(\theta^a) := \frac{1}{\sqrt{k_L}} \mathbf{v}^{0\top} \sigma^{(L)}(\mathbf{h}^{(L)a})\}_{a=0}^s, \text{ and } K^{ab} := \mathbb{E}_{\mathbf{x}} \lambda^a \lambda^b$$

can now be written recursively as

$$K^{ab} = \frac{1}{k_L} \mathbf{v}^{0\top} g^{(L)}(\mathbf{\Omega}^{(L)ab}) \mathbf{v}^0, \\ \mathbf{\Omega}^{(l)ab} = \frac{1}{k_{l-1}} \mathbf{W}^{(l)a} g^{(l-1)}(\mathbf{\Omega}^{(l-1)ab}) \mathbf{W}^{(l)b\top}, \quad (33)$$

with $\mathbf{\Omega}^{(0)ab} = I_d$ for all a, b , and $g^{(0)} = \text{id}$ (the identity map). We make the same concentration assumption as before, namely, $g^{(l)}(\mathbf{\Omega}^{(l)ab})_{ij} \approx \delta_{ij} g^{(l)}(\mathbf{\Omega}^{(l)ab})_{ii}$. Moreover, we notice that only the neurons in the L th layer can contribute with different importance to the output of the network, being connected to the (potentially) nonhomogeneous readout vector $\mathbf{v}^0$; nothing in the postactivation covariance distinguishes neurons in a layer $l < L$. For this reason, the OPs in this case are

$$\mathcal{Q}_l^{ab} := \frac{1}{k_l k_{l-1}} \text{Tr} \mathbf{W}^{(l)a} \mathbf{W}^{(l)b\top} \quad \text{for } l = 1, \dots, L-1, \\ \mathcal{Q}_L^{ab}(\mathbf{v}) := \frac{1}{|\mathcal{I}_v| k_{L-1}} \sum_{i \in \mathcal{I}_v} (\mathbf{W}^{(L)a} \mathbf{W}^{(L)b\top})_{ii}.$$

In terms of these, the postactivations covariance reads

$$K^{ab} = \mathbb{E}_{v \sim P_v} v^2 g^{(L)} \left(\mathcal{Q}_L^{ab}(v) g^{(L-1)}(\mathcal{Q}_{L-1}^{ab}) \right. \\ \left. \times g^{(L-2)}(\dots \mathcal{Q}_2^{ab} g^{(1)}(\mathcal{Q}_1^{ab}) \dots) \right). \quad (34)$$

The entropic contribution of the OPs is easily computable, as they are all independent of each other. The calculation reported in Appendix C 1 ultimately yields the free entropy reported in Result 4.

V. CONCLUSION AND PERSPECTIVES

In this paper, we have derived a quantitatively accurate statistical physics description of the optimal generalization capability of fully trained linear-width MLPs with an arbitrary number of layers, for broad classes of activation functions, in the challenging scaling regime where the number of parameters is comparable to that of the training data. Even for shallow MLPs, this feature-learning regime has resisted for a long time to mean-field approaches used, e.g., in the study of the narrow committee machines [7,13,15,27,205].

Our theory has been validated through extensive numerical experiments using Monte Carlo samplers, a generalization we proposed of the GAMP-RIE algorithm, and a popular algorithm used to train NNs, ADAM, showing in all cases a phenomenological picture consistent with our predictions.

Phase transitions in supervised learning are known in the statistical physics literature at least since Ref. [22], when the analysis was limited to linear models. In this sense, our theory contributes to enrich this landscape, unveiling numerous phase transitions in the learning of different layers' weights for MLPs. Such transitions can occur heterogeneously across layers and inside each layer. This rich behavior is captured by functional order parameters

with up to two arguments, which nevertheless allow a tractable dimensionality reduction of the problem.

Concerning limitations, for $L \geq 2$ we made some simplifying assumptions on the activation functions in order to reduce the number of order parameters and possible “learning strategies” accessible to the trained network. A direction we intend to pursue is to relax them. Specifically, we aim at a theory encompassing a broader class of activation functions for an arbitrary finite number of layers. An additional difficulty is the systematic stability analysis of the potentially many fixed points of the RS equations, each corresponding to a different learning strategy of the network. Our theory opens an avenue for these extensions, as all the order parameters needed are now identifiable, but the analytical treatment and numerical exploration of the problem require further effort. Moreover, as hinted at in Appendix B 6, for $L = 1$ and activations without the second Hermite coefficient, we foresee a path for a possible rigorous proof of our results.

A key novelty of our approach is the way we blend matrix models and spin-glass techniques in a unified formalism, which is able to handle matrix degrees of freedom which are not necessarily rotationally invariant. It applies to NNs, but also to matrix-sensing problems. We foresee that it will be useful beyond the realm of inference and learning problems. Another limitation of the approach is linked to the restricted class of solvable matrix models [132,134]. Indeed, as explained in Appendix B 2, possible improvements would require additional order parameters. Taking them into account yields matrix models when computing their entropy which, to the best of our knowledge, are not currently solvable. This is an exciting program at the crossroads of matrix models, inference, and learning of extensive-rank matrices.

Accounting for structured inputs is another challenging perspective. Here we took into consideration a rather simple data model, i.e., Gaussian data with a covariance in the vein of Refs. [207,208]. It would be desirable to study richer data models like mixture models [209,210], hidden manifolds [176], object manifolds and simplexes [211–214], and hierarchical data [201,215].

We have considered the idealized matched teacher-student setting as a first step, with the goal of tackling the methodological bottlenecks associated with the depth and linear width of the NN. Even in this simpler Bayes-optimal scenario, which in particular prevents replica symmetry breaking [126], the solution required the development of a nonstandard approach. A natural next step is to consider targets belonging to a different function class than the trainable MLP and focus on training by empirical risk minimization (zero temperature) rather than Bayesian (finite temperature). We believe this generalization can be carried out without major modifications of the theory, at least in the case where the target remains an MLP but with a different architecture. A complementary natural

continuation is to push our formalism further in order to tackle deep architectures beyond the MLP, such as convolutional networks, restricted Boltzmann machines, or transformers.

The identification of the relevant order parameters for the characterization of the equilibrium state carried out in our contribution paves the way for the study of the learning dynamics of first-order methods in similar settings. Indeed, there exist classical methods rooted in physics to study the learning dynamics of NNs [216–219]. Recently, Ref. [220] exploited these techniques to study the learning dynamics of a large NN trained on a GLM target, observing a separation of timescales between generalization and overfitting: It could be interesting to use the insights from our equilibrium analysis to extend their approach to more expressive targets. In the context of the dynamics of learning, it is also relevant to consider power-law-distributed readouts in the target, as many groups are currently doing in order to capture neural scaling laws [172,221–224] in extensive-width shallow NNs.

On a more speculative footing, we would like to point out that the specialization transition we uncovered in the interpolation asymptotics is only one of the known mechanisms for feature learning: kernel adaptation [63,113], grokking [77,225], etc. One of the ambitions of an asymptotic theory is to provide a good approximation for a large but realistic network, where, however, these different mechanisms can be at play together. We thus believe that understanding the interplay between them, as proposed in Refs. [73,226], and eventually unifying their taxonomy, is of practical relevance for the future.

ACKNOWLEDGMENTS

M. P. thanks Vittorio Erba and Pietro Rotondo for interesting discussions and suggestions. F. C. was affiliated with the Abdus Salam International Centre for Theoretical Physics during the conduction of this work. J. B., F. C., M.-T. N., and M. P. were funded by the European Union (ERC, CHORAL, Project No. 101039794).

All authors contributed equally to this work, and the names are ordered alphabetically. R. S. has carried out the numerical experiments with M.-T. N. in addition to theoretical work.

The views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

DATA AVAILABILITY

The data that support the findings of this article are openly available [174].

APPENDIX A: NOTATIONS, PREREQUISITES, AND AUXILIARY RESULTS

1. Notations

- (i) *Fonts*: Bold symbols are reserved for vectors and matrices. For the order parameters, a calligraphic symbol such as $\mathcal{Q}$ will emphasize that it is a function, while Q is a scalar.
- (ii) *Thermodynamic limit*: The limit $\lim$ without further specification will always correspond to the joint limit of the large input dimension, NN layer widths, and number of data all diverging, $d, k_l, n \rightarrow +\infty$, with scaling (4); it will be called the thermodynamic limit.
- (iii) *Hermite decomposition of the activation*: (μ_ℓ) are the Hermite coefficients of the activation σ when expressed in the orthogonal basis of Hermite polynomials $\text{He}_\ell(x)$; see Eq. (5). When different activations at each layer are considered, $\sigma^{(l)}$ denotes the activation at layer l , and $\mu_\ell^{(l)}$ is its ℓ th Hermite coefficient.
- (iv) *Vectors and matrices*: A vector $\mathbf{x}$ is always considered to be in column form, its transpose $\mathbf{x}^\top$ is a row, and the inner product is thus $\mathbf{u}^\top \mathbf{v} = \sum_i u_i v_i$. The norm $\|\mathbf{A}\| = (\sum_{ij} A_{ij}^2)^{1/2}$ is the Frobenius norm and the Euclidean norm for a vector. The trace operator for matrices is Tr . The ℓ th Hadamard (entrywise) power of a matrix is denoted with a superscript $\circ \ell$.
- (v) *Replicas*: The superscript 0 will always indicate a quantity associated with the target function, while superscript $a, b = 1, \dots, s$ will be used for replicas in the replica method or for i.i.d. samples from the posterior distribution.
- (vi) *Probability and expectations*: The symbol $\sim$ expresses that a random variable is drawn from a certain law. $\mathbb{P}(\cdot)$ is a probability, $P(\cdot)$ is a density function with respect to the Lebesgue measure, $dP(\cdot)$ the associated probability, $P(\cdot|Y)$ is the conditional density given Y . $\mathcal{N}(m, \Sigma)$ is the density of a Gaussian with mean m and variance Σ ; $\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$ is the multivariate version. The expectation operator with respect to a generic random variable X is denoted $\mathbb{E}_X$, the conditional expectation of X given Y is $\mathbb{E}[\cdot|Y]$, and the expectation with respect to all ensuing random variables entering an expression is simply $\mathbb{E}$. The bracket notation is reserved for an expectation with respect to an arbitrary (but d -independent) number of samples $(\theta^a)_a$ from the posterior given the training data $dP(\cdot|\mathcal{D})^{\otimes \infty}$: $\langle f((\theta^a)_a) \rangle := \mathbb{E}[f((\theta^a)_a)|\mathcal{D}]$.
- (vii) *Integrals and densities*: When unspecified, the integration domain of an integral $\int f(X) dX$ is $\mathbb{R}^{\dim(X)}$. For a sequence of symmetric real matrices $\mathbf{X} = \mathbf{X}_d$ indexed by the dimension d , the density $\rho_{\mathbf{X}}(s)$ with respect to the Lebesgue measure is the weak limit of

the empirical law of its (real) eigenvalues $\lim_{d \rightarrow \infty} (1/d) \sum_{i \leq d} \delta(s - \lambda_i(\mathbf{X}_d))$, where $\delta(\cdot)$ is the delta Dirac function; the Kronecker delta is denoted δ_{ij} .

- (viii) *Scalings and proportionalities*: The symbol $\propto$ means equality up to a multiplicative constant (which may be d dependent). We use the standard BigO and smallO notations $O(\cdot), o(\cdot)$. In particular, $o_d(1)$ means a sequence vanishing as $d \rightarrow \infty$. The notation $f_d = \Theta(g_d)$ means that the two sequences verify $f_d/g_d \rightarrow C$ for some constant $C \in (0, +\infty)$ as $d \rightarrow \infty$. $\approx$ means equality up to a correction $o_d(1)$.
- (ix) *Information-theoretic notions*: The mutual information between two random variables X, Y with joint law $P_{X,Y}$ and marginals P_X, P_Y is the Kullback-Leibler divergence $I(X; Y) = I(Y; X) := D_{\text{KL}}(P_{X,Y} \| P_X \otimes P_Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$, where $H(X)$ is the Shannon entropy if X is discrete, and it is instead the differential entropy if continuous valued. Similarly, $H(Y|X)$ is the conditional Shannon or differential entropy. We refer to Ref. [227] if these information-theoretic notions are not familiar.

2. Hermite basis and Mehler's formula

Recall the Hermite expansion of the activation:

$$\sigma(x) = \sum_{\ell=0}^{\infty} \frac{\mu_\ell}{\ell!} \text{He}_\ell(x). \quad (\text{A1})$$

We are expressing it on the basis of the probabilist's Hermite polynomials, generated through

$$\text{He}_\ell(z) = \frac{d^\ell}{dt^\ell} \exp(tz - t^2/2) \Big|_{t=0}. \quad (\text{A2})$$

The Hermite basis has the property of being orthogonal with respect to the standard Gaussian measure, which is the distribution of the input data. Specifically, if $z \sim \mathcal{N}(0, 1)$,

$$\mathbb{E} \text{He}_k(z) \text{He}_\ell(z) = \ell! \delta_{k\ell}. \quad (\text{A3})$$

By orthogonality, the coefficients of the expansions can be obtained as

$$\mu_\ell = \mathbb{E} \text{He}_\ell(z) \sigma(z). \quad (\text{A4})$$

Moreover,

$$\mathbb{E}[\sigma(z)^2] = \sum_{\ell=0}^{\infty} \frac{\mu_\ell^2}{\ell!}. \quad (\text{A5})$$

These coefficients for some popular choices of σ are reported in Table I for reference. The Hermite basis can

TABLE I. First Hermite coefficients of some activation functions reported in the figures. θ is the Heaviside step function.

$\sigma(z)$	μ_0	μ_1	μ_2	μ_3	μ_4	μ_5	$\cdots$	$\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma(z)^2]$
$\text{ReLU}(z) = z\theta(z)$	$1/\sqrt{2\pi}$	$1/2$	$1/\sqrt{2\pi}$	0	$-1/\sqrt{2\pi}$	0	$\cdots$	$1/2$
$\tanh(2z)$	0	$0.729\,48$	0	$-0.613\,98$	0	1.5632	$\cdots$	$0.635\,26$
$\tanh(2z)/\sigma_{\tanh}$	0	$0.915\,24$	0	$-0.770\,33$	0	1.9613	$\cdots$	1

be generalized to an orthogonal basis with respect to the Gaussian measure with generic variance. Let $z \sim \mathcal{N}(0, r)$, then

$$\text{He}_\ell^{[r]}(z) = \frac{d^\ell}{dt^\ell} \exp(tz - t^2 r/2) \Big|_{t=0}. \quad (\text{A6})$$

For this basis one has

$$\mathbb{E} \text{He}_k^{[r]}(z) \text{He}_\ell^{[r]}(z) = \ell! r^\ell \delta_{k\ell}. \quad (\text{A7})$$

Consider now a couple of jointly Gaussian random variables $\mathbf{x} = (u, v) \sim \mathcal{N}(0, \mathbf{C})$ with

$$\mathbf{C} = \begin{pmatrix} r & q \\ q & r \end{pmatrix}. \quad (\text{A8})$$

Then, by Mehler's formula

$$\begin{aligned} & \frac{1}{2\pi\sqrt{r^2 - q^2}} \exp\left[-\frac{1}{2} \mathbf{x}^\top \mathbf{C}^{-1} \mathbf{x}\right] \\ &= \frac{e^{-\frac{u^2}{2r}}}{\sqrt{2\pi r}} \frac{e^{-\frac{v^2}{2r}}}{\sqrt{2\pi r}} \sum_{\ell=0}^{+\infty} \frac{q^\ell}{\ell! r^{2\ell}} \text{He}_\ell^{[r]}(u) \text{He}_\ell^{[r]}(v), \end{aligned} \quad (\text{A9})$$

and by orthogonality of the Hermite basis, Eq. (24) readily follows by noticing that the variables $[h_i^a = (\mathbf{W}^a \mathbf{x})_i / \sqrt{d}]_{i,a}$ at given $(\mathbf{W}^a)$ are Gaussian with covariances $\Omega_{ij}^{ab} = \mathbf{W}_i^{a\top} \mathbf{W}_j^b / d$:

$$\begin{aligned} \mathbb{E} \sigma(h_i^a) \sigma(h_j^b) &= \sum_{\ell=0}^{\infty} \frac{(\mu_\ell^{[r]})^2}{\ell! r^{2\ell}} (\Omega_{ij}^{ab})^\ell, \\ \mu_\ell^{[r]} &= \mathbb{E}_{z \sim \mathcal{N}(0, r)} \text{He}_\ell^{[r]}(z) \sigma(z). \end{aligned} \quad (\text{A10})$$

Moreover, as Ω_{ii}^{aa} , to be identified with r above, converges to the variance of the prior of $\mathbf{W}^0$ for large d by Bayes optimality, whenever $\Omega_{ii}^{aa} \rightarrow 1$ we can specialize this formula to the simpler case $r = 1$ we reported in the main text.

3. Nishimori identities

The Nishimori identities are a set of symmetries arising in inference in the Bayes-optimal setting as a consequence

of Bayes's rule. To introduce them, consider a test function f of the teacher weights, collectively denoted by $\boldsymbol{\theta}^0$, of $s-1$ replicas of the student's weights $(\boldsymbol{\theta}^a)_{2 \leq a \leq s}$ drawn conditionally i.i.d. from the posterior, and possibly also of the training set $\mathcal{D}$: $f(\boldsymbol{\theta}^0, \boldsymbol{\theta}^2, \dots, \boldsymbol{\theta}^s; \mathcal{D})$. Then

$$\mathbb{E}_{\boldsymbol{\theta}^0, \mathcal{D}} \langle f(\boldsymbol{\theta}^0, \boldsymbol{\theta}^2, \dots, \boldsymbol{\theta}^s; \mathcal{D}) \rangle = \mathbb{E}_{\boldsymbol{\theta}^0, \mathcal{D}} \langle f(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \dots, \boldsymbol{\theta}^s; \mathcal{D}) \rangle. \quad (\text{A11})$$

The Nishimori identities thus allow us to replace the teacher's weights with another replica from the posterior measure. The proof follows from Bayes's theorem; see, e.g., Ref. [100].

The Nishimori identities have some consequences also on our replica symmetric *Ansatz* for the free entropy. In particular, they constrain the values of the asymptotic mean of some OPs. For instance, consider the shallow case

$$\begin{aligned} R_2^{a0} &= \lim \frac{1}{d^2} \mathbb{E}_{\mathcal{D}, \boldsymbol{\theta}^0} \langle \text{Tr}[\mathbf{S}_2^a \mathbf{S}_2^{0}] \rangle \\ &= \lim \frac{1}{d^2} \mathbb{E}_{\mathcal{D}} \langle \text{Tr}[\mathbf{S}_2^a \mathbf{S}_2^b] \rangle = R_2^{ab}, \quad \text{for } a \neq b. \end{aligned} \quad (\text{A12})$$

In addition, within the replica symmetry assumption, all the above overlaps are equal to one another.

Combined with the concentration of OPs, which can be proven in great generality in Bayes-optimal inference [186,192], the Nishimori identities fix the values of some of them. For instance, we have that with high probability

$$\begin{aligned} \frac{1}{d^2} \text{Tr}[(\mathbf{S}_2^a)^2] &\rightarrow R_d = \lim \frac{1}{d^2} \mathbb{E}_{\mathcal{D}} \langle \text{Tr}[(\mathbf{S}_2^a)^2] \rangle \\ &= \lim \frac{1}{d^2} \mathbb{E}_{\boldsymbol{\theta}^0} \text{Tr}[(\mathbf{S}_2^0)^2] \\ &= 1 + \gamma \bar{v}^2, \end{aligned} \quad (\text{A13})$$

with $\mathbb{E}v = \bar{v}$. When this happens, as for R_d , then the respective Fourier conjugates $\hat{R}_d$ vanish, since the desired constraints were already asymptotically enforced without the need of additional delta functions. This is because the configurations in which the OPs take those values exponentially (in n) dominate the posterior measure, so these constraints are automatically imposed by the measure. Another OP for which we have a similar consequence is

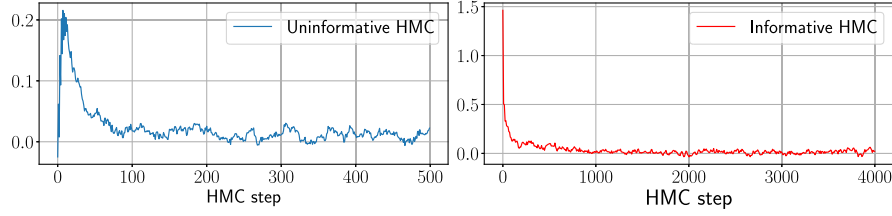


FIG. 21. Trajectories of $\mathbb{E}_{\mathbf{x}_{\text{test}}}(\lambda_t^1 - \lambda_t^2)^2 / \mathbb{E}_{\mathbf{x}_{\text{test}}}(\lambda_t^1 - \lambda^0)^2 - 1$, where $\lambda_t^a := \lambda_{\text{test}}(\theta_t^a)$ and $\lambda^0 := \lambda_{\text{test}}(\theta^0)$. θ_t^1 is the HMC sample at step t for the first chain independently initialized from the second; $\mathbb{E}_{\mathbf{x}_{\text{test}}}$ is an average over $5 \cdot 10^4$ test samples. Here, $L = 1, d = 150, \gamma = 0.5, \Delta = 0.1, \alpha = 5, \sigma = \text{ReLU}, \mathbf{W}$ is Gaussian, and $\mathbf{v}$ is homogenous. HMC runs are initialized uninformatively (left) and informatively (right) in order to probe the metastable and equilibrium states, respectively. This quantity approaches zero for long enough times, indicating the empirical validity of the Nishimori identity $\mathbb{E}(\lambda^1 \lambda^2) = \mathbb{E}(\lambda^1) \lambda^0$ both for the posterior average $\langle \cdot \rangle$ and the average over the metastable state $\langle \cdot \rangle_{\text{meta}}$ (see Remark 5 in the main text).

$$\begin{aligned} Q^{aa}(\mathbf{v}) &= \lim_{d|\mathcal{I}_v|} \frac{1}{d|\mathcal{I}_v|} \sum_{i \in \mathcal{I}_v} \mathbb{E}_{\mathcal{D}}(\mathbf{W}_i^a \cdot \mathbf{W}_i^a) \\ &= \lim_{d|\mathcal{I}_v|} \frac{1}{d|\mathcal{I}_v|} \sum_{i \in \mathcal{I}_v} \mathbb{E}_{\theta^0} \|\mathbf{W}_i^0\|^2 = 1, \end{aligned} \quad (\text{A14})$$

and consequently, $\hat{Q}^{aa}(\mathbf{v}) = 0$.

Let us now draw some more generic conclusions that we shall need in the following. Given a generic set of OPs labeled by replica indices, say, $\mathbf{Q} = (Q^{ab})_{a,b=0,\dots,s}$, a replica-symmetric *Ansatz* for it would enforce the following form:

$$\mathbf{Q} = \begin{pmatrix} \rho & m\mathbf{1}_s^T \\ m\mathbf{1}_s & (Q_d - Q)I_s + Q\mathbf{1}_s\mathbf{1}_s^T \end{pmatrix} \in \mathbb{R}^{s+1 \times s+1}, \quad (\text{A15})$$

where $\mathbf{1}_s = (1, 1, \dots, 1) \in \mathbb{R}^s$, and $I_s \in \mathbb{R}^{s \times s}$ is the identity matrix. Under rather general terms, the Nishimori identities are actually enforcing the constraints $\rho = Q_d$, $m = Q$, yielding

$$\mathbf{Q} = \begin{pmatrix} Q_d & Q\mathbf{1}_s^T \\ Q\mathbf{1}_s & (Q_d - Q)I_s + Q\mathbf{1}_s\mathbf{1}_s^T \end{pmatrix} \in \mathbb{R}^{s+1 \times s+1}. \quad (\text{A16})$$

As we explained in this section, the Nishimori identities are a property of posterior measures sampled at equilibrium in the Bayes-optimal setting. In the numerical part of this paper exploiting Monte Carlo sampling, we checked their validity not only at equilibrium, but also whenever the algorithm is stuck in a metastable state captured by suboptimal branches of our theory (see Remark 5 in the main text). We report an explicit check of this fact in Fig. 21.

4. Linking free entropy and mutual information

It is possible to relate the mutual information (MI) of the inference problem to the free entropy $f_n = \mathbb{E} \ln \mathcal{Z}$ introduced in the main text. Indeed, we can write the MI as

$$\frac{I(\theta^0; \mathcal{D})}{n} = \frac{H(\mathcal{D})}{n} - \frac{H(\mathcal{D}|\theta^0)}{n}, \quad (\text{A17})$$

where $H(Y|X)$ is the conditional Shannon entropy of Y given X . Using the chain rule for the entropy, and the definition (9), the free entropy can be recast as

$$-f_n = \frac{H(\{y_\mu\}_{\mu \leq n} | \{\mathbf{x}_\mu\}_{\mu \leq n})}{n} = \frac{H(\mathcal{D})}{n} - \frac{H(\{\mathbf{x}_\mu\}_{\mu \leq n})}{n}. \quad (\text{A18})$$

On the other hand, $H(\mathcal{D}|\theta^0) = H(\{y_\mu\}|\theta^0, \{\mathbf{x}_\mu\}) + H(\{\mathbf{x}_\mu\})$, i.e.,

$$\frac{H(\mathcal{D}|\theta^0)}{n} \approx -\mathbb{E}_\lambda \int dy P_{\text{out}}(y|\lambda) \ln P_{\text{out}}(y|\lambda) + \frac{H(\{\mathbf{x}_\mu\}_{\mu \leq n})}{n}, \quad (\text{A19})$$

where $\lambda \sim \mathcal{N}(0, K_d)$, with K_d given by Eq. (13) for $L = 1$, while for $L = 2$ and $L = 3$, given our normalization assumptions on the activation functions, $K_d = 1$ (assuming here that $\mu_0 = 0$; see Appendix B 1 g if the activation σ is noncentered). Equality holds asymptotically in the limit. This allows us to express the MI asymptotically as

$$\frac{I(\theta^0; \mathcal{D})}{n} = -f_n + \mathbb{E}_\lambda \int dy P_{\text{out}}(y|\lambda) \ln P_{\text{out}}(y|\lambda) + o_n(1). \quad (\text{A20})$$

Specializing the equation to the Gaussian channel, one obtains

$$\frac{I(\theta^0; \mathcal{D})}{n} = -f_n - \frac{1}{2} \ln(2\pi e \Delta). \quad (\text{A21})$$

We chose to normalize by n because, in our scaling, it is always proportional to the number of total parameters, that is, $\Theta(d^2)$. Hence, with this choice one can interpret the parameter α as an effective signal-to-noise ratio.

Remark 6. The arguments of Ref. [171] to show the existence of an upper bound on the mutual information per variable in the case of discrete variables, and the associated inevitable breaking of prior universality beyond a certain threshold in matrix denoising, apply to the present model, too. It implies, as in the aforementioned paper, that the mutual information per variable cannot go beyond $\ln 2$ for Rademacher inner weights. Our theory is consistent with this fact; this is a direct consequence of the analysis in Appendix B 3 carried out for the shallow case $L = 1$ [see, in particular, Eq. (B68)] specialized to binary prior over $\mathbf{W}$.

5. Alternative representation for ϵ^{opt} with $L = 1$

We recall that $\theta^0 = (\mathbf{v}^0, \mathbf{W}^0)$ and similarly for $\theta^1 = \theta, \theta^2, \dots$, which are replicas, i.e., conditionally i.i.d. samples from $dP(\mathbf{W}, \mathbf{v}|\mathcal{D})$ (the reasoning below applies whether $\mathbf{v}$ is learnable or quenched, so in general, we can consider a joint posterior over both). From its definition (8), the Bayes-optimal generalization error can be recast as

$$\epsilon^{\text{opt}} = \mathbb{E}_{\theta^0, \mathbf{x}_{\text{test}}} [\mathbb{E}[y_{\text{test}}^2 | \lambda^0] - 2\mathbb{E}_{\theta^0, \mathcal{D}, \mathbf{x}_{\text{test}}} [\mathbb{E}[y_{\text{test}} | \lambda^0] \langle \mathbb{E}[y | \lambda] \rangle] + \mathbb{E}_{\theta^0, \mathcal{D}, \mathbf{x}_{\text{test}}} [\langle \mathbb{E}[y | \lambda] \rangle^2], \quad (\text{A22})$$

where $\mathbb{E}[y | \lambda] = \int dy y P_{\text{out}}(y | \lambda)$, and $(\lambda^a)_{a=0, \dots, s}$ are the random variables (random due to the test input $\mathbf{x}_{\text{test}}$ drawn independently of the training data $\mathcal{D}$, and their respective weights θ^0, θ)

$$\lambda^a = \lambda_{\text{test}}(\theta^a) = \frac{\mathbf{v}^{a\top}}{\sqrt{k}} \sigma \left(\frac{\mathbf{W}^a \mathbf{x}_{\text{test}}}{\sqrt{d}} \right). \quad (\text{A23})$$

Recall that the bracket $\langle \cdot \rangle$ is the average with respect to the posterior and acts on $\theta^1 = \theta, \theta^2, \dots$. Notice that the last term on the rhs of Eq. (A22) can be rewritten as

$$\mathbb{E}_{\theta^0, \mathcal{D}, \mathbf{x}_{\text{test}}} [\langle \mathbb{E}[y | \lambda] \rangle^2] = \mathbb{E}_{\theta^0, \mathcal{D}, \mathbf{x}_{\text{test}}} [\langle \mathbb{E}[y | \lambda^1] \mathbb{E}[y | \lambda^2] \rangle],$$

with superscripts being replica indices.

In order to show Result 2 for a generic P_{out} we assume the joint Gaussianity of the variables $(\lambda^0, \lambda^1, \lambda^2, \dots)$, with covariance given by K^{ab} with $a, b \in \{0, 1, 2, \dots\}$. Indeed, in the limit “lim”, the theory considers $(\lambda^a)_{a \geq 0}$ as jointly Gaussian under the randomness of a common input, here $\mathbf{x}_{\text{test}}$, conditionally on the weights (θ^a) . Their covariance depends on the weights (θ^a) through various overlap OPs introduced in the main text. In the large limit “lim”, these overlaps are assumed to concentrate under the quenched posterior average $\mathbb{E}_{\theta^0, \mathcal{D}}[\cdot]$ toward nonrandom asymptotic values corresponding to the extremizer globally maximizing the RS potential in Result 1, with the overlaps entering K^{ab} through Eq. (B10). Using the replica-symmetric *Ansatz* [see Eq. (B12)] and the Nishimori identities, the covariance evaluated on these overlaps shall be denoted $\mathbf{K}^*$ as in the main text, and its elements are written as

$K^{*ab} = K^* + \delta_{ab}(K_d - K^*)$. This hypothesis is then confirmed by the excellent agreement between our theoretical predictions based on this assumption and the experimental results. This implies directly Eq. (17) in Result 2 from definition (7). For the special case of optimal mean-square generalization, error it yields

$$\lim \epsilon^{\text{opt}} = \mathbb{E}_{\lambda^0} [\mathbb{E}[y_{\text{test}}^2 | \lambda^0] - 2\mathbb{E}_{\lambda^0, \lambda^1} [\mathbb{E}[y_{\text{test}} | \lambda^0] \mathbb{E}[y | \lambda^1]] + \mathbb{E}_{\lambda^1, \lambda^2} [\mathbb{E}[y | \lambda^1] \mathbb{E}[y | \lambda^2]], \quad (\text{A24})$$

where in the replica-symmetric *Ansatz*,

$$\begin{aligned} \mathbb{E}[(\lambda^0)^2] &= K_d, & \mathbb{E}[\lambda^0 \lambda^1] &= \mathbb{E}[\lambda^0 \lambda^2] = K^*, \\ \mathbb{E}[\lambda^1 \lambda^2] &= K^*, & \mathbb{E}[(\lambda^1)^2] &= \mathbb{E}[(\lambda^2)^2] = K_d. \end{aligned} \quad (\text{A25})$$

We thus have

$$\mathbb{E}_{\lambda^0, \lambda^1} [\mathbb{E}[y_{\text{test}} | \lambda^0] \mathbb{E}[y | \lambda^1]] = \mathbb{E}_{\lambda^1, \lambda^2} [\mathbb{E}[y | \lambda^1] \mathbb{E}[y | \lambda^2]]. \quad (\text{A26})$$

Plugging the above in Eq. (A24) yields Eq. (18).

Let us now prove a formula for the optimal mean-square generalization error written in terms of the overlaps which holds for the special case of linear readout with Gaussian label noise $P_{\text{out}}(y | \lambda) = \exp\{-[1/(2\Delta)](y - \lambda)^2\}/\sqrt{2\pi\Delta}$. The following derivation is exact and does not require any Gaussianity assumption on the random variables (λ^a) . For the linear Gaussian channel, the means verify $\mathbb{E}[y | \lambda] = \lambda$ and $\mathbb{E}[y^2 | \lambda] = \lambda^2 + \Delta$. Plugged into Eq. (A22), this yields

$$\epsilon^{\text{opt}} - \Delta = \mathbb{E}_{\theta^0, \mathbf{x}_{\text{test}}} [(\lambda^0)^2] - 2\mathbb{E}_{\theta^0, \mathcal{D}, \mathbf{x}_{\text{test}}} [\lambda^0 \langle \lambda \rangle] + \mathbb{E}_{\theta^0, \mathcal{D}, \mathbf{x}_{\text{test}}} [\langle \lambda^1 \lambda^2 \rangle], \quad (\text{A27})$$

whence we clearly see that the generalization error depends only on the covariance of $\lambda^0, \lambda^1, \lambda^2$ under the randomness of the shared input $\mathbf{x}_{\text{test}}$ at fixed weights, regardless of the validity of the Gaussian hypothesis on the postactivations (11) that we assume in the replica computation. This covariance was already computed in Eq. (24); we recall it here for the reader's convenience

$$\begin{aligned} K(\theta^a, \theta^b) &:= \mathbb{E}_{\mathbf{x}_{\text{test}}} [\lambda^a \lambda^b] = \sum_{\ell=1}^{\infty} \frac{\mu_{\ell}^2}{\ell!} \frac{1}{k} \sum_{i,j=1}^k v_i^a (\Omega_{ij}^{ab})^{\ell} v_j^b \\ &= \sum_{\ell=1}^{\infty} \frac{\mu_{\ell}^2}{\ell!} R_{\ell}^{ab}, \end{aligned} \quad (\text{A28})$$

where $\Omega_{ij}^{ab} := \mathbf{W}_i^{a\top} \mathbf{W}_j^b / d$, and R_{ℓ}^{ab} as introduced in Eq. (24) for $a, b = 0, 1, 2$. We stress that $K(\theta^a, \theta^b)$ is not the limiting covariance K^{ab} whose elements are in Eq. (B12), but rather the finite-size one. $K(\theta^a, \theta^b)$ provides us with an efficient way to compute the generalization error numerically used in Appendix B 4, Fig. 24, that is through the formula

$$\begin{aligned}
\varepsilon^{\text{opt}} - \Delta &= \mathbb{E}_{\theta^0} K(\theta^0, \theta^0) - 2\mathbb{E}_{\theta^0, \mathcal{D}} \langle K(\theta^0, \theta^1) \rangle \\
&\quad + \mathbb{E}_{\theta^0, \mathcal{D}} \langle K(\theta^1, \theta^2) \rangle \\
&= \sum_{\ell=1}^{\infty} \frac{\mu_{\ell}^2}{\ell!} \mathbb{E}_{\theta^0, \mathcal{D}} \langle R_{\ell}^{00} - 2R_{\ell}^{01} + R_{\ell}^{12} \rangle. \quad (\text{A29})
\end{aligned}$$

In the above, the posterior measure $\langle \cdot \rangle$ is taken care of by Monte Carlo sampling (when it equilibrates). In addition, as in the main text, we assume that in the large system limit the (numerically confirmed) identity (29) holds. Putting all ingredients together, we get

$$\begin{aligned}
\varepsilon^{\text{opt}} - \Delta &= \mathbb{E}_{\theta^0, \mathcal{D}} \left\langle \mu_1^2 (R_1^{00} - 2R_1^{01} + R_1^{12}) \right. \\
&\quad + \frac{\mu_2^2}{2} (R_2^{00} - 2R_2^{01} + R_2^{12}) \\
&\quad + \mathbb{E}_{v \sim P_v} v^2 [g(\mathcal{Q}^{00}(v)) - 2g(\mathcal{Q}^{01}(v)) \\
&\quad \left. + g(\mathcal{Q}^{12}(v))] \right\rangle. \quad (\text{A30})
\end{aligned}$$

In the Bayes-optimal setting, one can use again the Nishimori identities that imply $\mathbb{E}_{\theta^0, \mathcal{D}} \langle R_1^{12} \rangle = \mathbb{E}_{\theta^0, \mathcal{D}} \langle R_1^{01} \rangle$, and analogously $\mathbb{E}_{\theta^0, \mathcal{D}} \langle R_2^{12} \rangle = \mathbb{E}_{\theta^0, \mathcal{D}} \langle R_2^{01} \rangle$ and $\mathbb{E}_{\theta^0, \mathcal{D}} \langle g(\mathcal{Q}_W^{12}(v)) \rangle = \mathbb{E}_{\theta^0, \mathcal{D}} \langle g(\mathcal{Q}_W^{01}(v)) \rangle$. Inserting these identities into Eq. (A30), one gets

$$\begin{aligned}
\varepsilon^{\text{opt}} - \Delta &= \mathbb{E}_{\theta^0, \mathcal{D}} \langle \mu_1^2 (R_1^{00} - R_1^{01}) + \frac{\mu_2^2}{2} (R_2^{00} - R_2^{01}) \rangle \\
&\quad + \mathbb{E}_{v \sim P_v} v^2 [g(\mathcal{Q}^{00}(v)) - g(\mathcal{Q}^{01}(v))]. \quad (\text{A31})
\end{aligned}$$

This formula relies only on the validity of Eq. (29), and makes no assumption on the law of the λ 's. That it depends only on their covariance is simply a consequence of the quadratic nature of the mean-square generalization error.

Remark 7. Note that the derivation up to Eq. (A29) did not assume Bayes optimality [while Eq. (A31) did]. Therefore, one can consider it in cases where the true posterior average $\langle \cdot \rangle$ is replaced by one which does not verify the Nishimori identities. Equation (A29) is the formula we use to compute the generalization error of Monte-Carlo-based estimators in the inset of Fig. 24. There, MCMC cannot equilibrate and experiences a glassy regime. This regime, at variance with the metastable states described in the main text (see Remark 5), does not correspond to any suboptimal branch of our theory: We verified numerically that indeed the Nishimori identities do not hold there.

Remark 8. Using the Nishimori identity of Appendix A 3 and again that, for the linear readout with Gaussian label noise $\mathbb{E}[y|\lambda] = \lambda$ and $\mathbb{E}[y^2|\lambda] = \lambda^2 + \Delta$, it is easy to check that the so-called Gibbs error

$$\varepsilon^{\text{Gibbs}} := \mathbb{E}_{\theta^0, \mathcal{D}, \mathbf{x}_{\text{test}}, y_{\text{test}}} \{ \{ y_{\text{test}} - \mathbb{E}[y|\lambda_{\text{test}}(\theta)] \}^2 \} \quad (\text{A32})$$

is related for this channel to the Bayes-optimal mean-square generalization error through the identity

$$\varepsilon^{\text{Gibbs}} - \Delta = 2(\varepsilon^{\text{opt}} - \Delta). \quad (\text{A33})$$

We exploited this relationship together with the concentration of the Gibbs error with respect to the quenched posterior measure $\mathbb{E}_{\theta^0, \mathcal{D}} \langle \cdot \rangle$ when evaluating the numerical generalization error of the Monte Carlo algorithms reported in the main text.

APPENDIX B: SHALLOW MLP

1. Details of the replica calculation

In this section, we report all the details needed to derive our results with the replica method when $L = 1$. The starting point is the assumption of joint Gaussianity of the postactivations

$$\left\{ \lambda^a(\theta^a) := \frac{1}{\sqrt{k}} \mathbf{v}^{0\top} \sigma \left(\frac{1}{\sqrt{d}} \mathbf{W}^a \mathbf{x} \right) \right\}_{a=0}^s$$

under the randomness of $\mathbf{x}$, for typical (θ^a) . The covariance of this Gaussian family is

$$\begin{aligned}
K^{ab} &:= \mathbb{E}_{\mathbf{x}} \lambda^a(\theta^a) \lambda^b(\theta^b) \\
&= \frac{1}{k} \sum_{i,j=1}^k v_i^0 v_j^0 \mathbb{E}_{\mathbf{x}} \sigma \left(\frac{\mathbf{W}_i^a \cdot \mathbf{x}}{\sqrt{d}} \right) \sigma \left(\frac{\mathbf{W}_j^b \cdot \mathbf{x}}{\sqrt{d}} \right). \quad (\text{B1})
\end{aligned}$$

To compute it, we use Mehler's formula (A9) plus the assumption that $\|\mathbf{W}^a\|^2/d$ concentrates toward 1, which is verified by Nishimori identities:

$$K^{ab} = \sum_{\ell=1}^{\infty} \frac{\mu_{\ell}^2}{\ell! k} \sum_{i,j=1}^k v_i^0 v_j^0 (\Omega_{ij}^{ab})^{\ell}, \quad \Omega_{ij}^{ab} := \frac{\mathbf{W}_i^a \cdot \mathbf{W}_j^b}{d}. \quad (\text{B2})$$

Because of norm concentrations of the $\mathbf{W}_i^a$'s, one can show that for a fixed i only for a small number of indices, j $\mathbf{W}_j^b$ can have an $O(d)$ projection onto $\mathbf{W}_i^a$. As confirmed by our numerics, we assume this projection is large only for one index j , which by permutation invariance, we can assume to be i itself. In other words, we are assuming that

$$\Omega_{ii}^{ab} = O(1), \quad \Omega_{ij}^{ab} = O\left(\frac{1}{\sqrt{d}}\right) \quad \text{for } i \neq j. \quad (\text{B3})$$

Note that this is rigorously provable for the diagonal part of the covariance $a = b$. Then our assumptions imply

$$\frac{1}{k} \sum_{i \neq j} v_i^0 v_j^0 (\Omega_{ij}^{ab})^{\ell} = O(k/d^{\ell/2}) = O(d^{1-\ell/2}), \quad (\text{B4})$$

which vanishes for any $\ell \geq 3$. Hence, the covariance simplifies to

$$K^{ab} = \mu_1^2 R_1^{ab} + \frac{\mu_2^2}{2} R_2^{ab} + \frac{1}{k} \sum_{i=1}^k (v_i^0)^2 g(\Omega_{ii}^{ab}) + O(d^{-1/2}), \quad (\text{B5})$$

where

$$\begin{aligned} g(x) &= \sum_{\ell=3}^{\infty} \frac{\mu_{\ell}^2}{\ell!} x^{\ell} \\ &= \mathbb{E}_{(y,z)|x} [\sigma(y)\sigma(z)] - \mu_0^2 - \mu_1^2 x - \frac{\mu_2^2}{2} x^2, \\ (y, z) &\sim \mathcal{N}\left[(0, 0), \begin{pmatrix} 1 & x \\ x & 1 \end{pmatrix}\right]. \end{aligned} \quad (\text{B6})$$

Here, again by permutation symmetry, we can assume that all overlaps Ω_{ii}^{ab} for $i \in \mathcal{I}_v$ concentrate onto the same value $\mathcal{Q}^{ab}(v)$ labeled by v , thus leading to

$$\begin{aligned} K^{ab} &\approx \mu_1^2 R_1^{ab} + \frac{\mu_2^2}{2} R_2^{ab} + \sum_{v \in \mathcal{V}} v^2 \frac{|\mathcal{I}_v|}{k} \frac{1}{|\mathcal{I}_v|} \sum_{i \in \mathcal{I}_v} g(\Omega_{ii}^{ab}) \\ &\approx \mu_1^2 + \frac{\mu_2^2}{2} R_2^{ab} + \sum_{v \in \mathcal{V}} P_v(v) v^2 g(\mathcal{Q}^{ab}(v)). \end{aligned} \quad (\text{B7})$$

In the above, we have used the aforementioned permutation symmetry and concentration, and the fact that $(|\mathcal{I}_v|/k) \rightarrow P_v(v)$. Furthermore, we assumed that $R_1^{ab} \rightarrow 1$, which is justified by the fact that R_1^{ab} is a scalar overlap between two vectors $\mathbf{S}_1^a, \mathbf{S}_1^b$. In fact, at the present scaling of n, k, d , the vector $\mathbf{S}_1^0$ can be retrieved exactly.

K^{ab} is what governs the “energy” in our model, and the overlaps therein appearing thus play the role of OPs. Recall the form of the replicated partition function

$$\begin{aligned} \mathbb{E} \mathcal{Z}^s &= \mathbb{E}_{\mathbf{v}^0} \int \prod_a^{0,s} dP_W(\mathbf{W}^a) \\ &\quad \times \left[\mathbb{E}_{\mathbf{x}} \int dy \prod_a^{0,s} P_{\text{out}}(y | \lambda^a(\boldsymbol{\theta}^a)) \right]^n, \end{aligned}$$

which, after the above simplifications, reads

$$\mathbb{E} \mathcal{Z}^s = \int d\mathbf{R}_2 d\mathbf{Q} \exp[F_S(\mathbf{R}_2, \mathbf{Q}) + nF_E(\mathbf{R}_2, \mathbf{Q})], \quad (\text{B8})$$

where $\mathbf{R}_2 = (R_2^{ab})$ and $\mathbf{Q} := \{\mathcal{Q}^{ab} | a \leq b\}$, $\mathcal{Q}^{ab} := \{\mathcal{Q}^{ab}(v) | v \in \mathcal{V}\}$. We split the discussion in the evaluation of the energetic potential F_E , the entropic potential F_S , and finally the derivation of the saddle-point equations for the OPs. The calculation is performed assuming replica symmetry, as explained below.

a. Energetic potential

The replicated energetic term under our Gaussian assumption on the joint law of the postactivation replicas is reported here for the reader’s convenience:

$$F_E = \ln \int dy \int d\lambda \frac{e^{-\frac{1}{2}\lambda^T \mathbf{K}^{-1} \lambda}}{\sqrt{(2\pi)^{s+1} \det \mathbf{K}}} \prod_{a=0}^s P_{\text{out}}(y | \lambda^a), \quad (\text{B9})$$

with

$$K^{ab} = \mu_1^2 + \frac{\mu_2^2}{2} R_2^{ab} + \mathbb{E}_{v \sim P_v} v^2 g(\mathcal{Q}^{ab}(v)). \quad (\text{B10})$$

The energetic term F_E is already expressed as a low-dimensional integral, but it simplifies considerably under the RS *Ansatz* and after using the Nishimori identities. Let us denote $\mathbf{Q}(v) = [\mathcal{Q}^{ab}(v)]_{a,b=0}^s$, then, using Eqs. (A15) and (A16)

$$\begin{aligned} \mathbf{Q}(v) &= \begin{pmatrix} 1 & \mathbf{Q}(v)\mathbf{1}_s^T \\ \mathbf{Q}(v)\mathbf{1}_s & [1 - \mathbf{Q}(v)]I_s + \mathbf{Q}(v)\mathbf{1}_s\mathbf{1}_s^T \end{pmatrix} \\ \Leftrightarrow \hat{\mathbf{Q}}(v) &= \begin{pmatrix} 0 & -\hat{\mathbf{Q}}(v)\mathbf{1}_s^T \\ -\hat{\mathbf{Q}}(v)\mathbf{1}_s & \hat{\mathbf{Q}}(v)I_s - \hat{\mathbf{Q}}(v)\mathbf{1}_s\mathbf{1}_s^T \end{pmatrix}, \end{aligned}$$

and similarly,

$$\begin{aligned} \mathbf{R}_2 &= \begin{pmatrix} R_d & R_2\mathbf{1}_s^T \\ R_2\mathbf{1}_s & (R_d - R_2)I_s + R_2\mathbf{1}_s\mathbf{1}_s^T \end{pmatrix} \\ \Leftrightarrow \hat{\mathbf{R}}_2 &= \begin{pmatrix} 0 & -\hat{R}_2\mathbf{1}_s^T \\ -\hat{R}_2\mathbf{1}_s & \hat{R}_2I_s - \hat{R}_2\mathbf{1}_s\mathbf{1}_s^T \end{pmatrix}, \end{aligned}$$

where we reported the *Ansatz* also for the Fourier conjugates for future convenience, though not needed for the energetic potential. We are going to use repeatedly the Fourier representation of the delta function, namely, $\delta(x) = [1/(2\pi)] \int d\hat{x} \exp(i\hat{x}x)$. Because the integrals we will end up with will always be at some point evaluated by the saddle point, implying a deformation of the integration contour in the complex plane, tracking the imaginary unit i in the delta functions will be irrelevant. Similarly, the normalization $1/2\pi$ will always contribute to subleading terms in the integrals at hand. Therefore, we will allow ourselves to formally write $\delta(x) = \int d\hat{x} \exp(r\hat{x}x)$ for a convenient constant r , keeping in mind these considerations (again, as we evaluate the final integrals by the saddle point, the choice of r ends up being irrelevant).

The RS *Ansatz*, which is equivalent to an assumption of concentration of the OPs in the high-dimensional limit, is known to be exact when analyzing Bayes-optimal inference and learning, as in the present paper; see Refs. [182,186,192]. Under the RS *Ansatz*, $\mathbf{K}$ acquires a similar form:

$$\mathbf{K} = \begin{pmatrix} K_d & K\mathbf{1}_s^\top \\ K\mathbf{1}_s & (K_d - K)I_s + K\mathbf{1}_s\mathbf{1}_s^\top \end{pmatrix} \quad (\text{B11})$$

with

$$K \equiv K(R_2, \mathcal{Q}) = \mu_1^2 + \frac{\mu_2^2}{2}R_2 + \mathbb{E}_{v \sim P_v} v^2 g(\mathcal{Q}(v)),$$

$$K_d = \mu_1^2 + \frac{\mu_2^2}{2}R_d + g(1). \quad (\text{B12})$$

In the RS *Ansatz*, it is thus possible to give a convenient low-dimensional representation of the multivariate Gaussian integral of F_E in terms of white Gaussian random variables:

$$\lambda^a = \xi\sqrt{K} + u^a\sqrt{K_d - K} \quad \text{for } a = 0, 1, \dots, s, \quad (\text{B13})$$

where $\xi, (u^a)_{a=0}^s$ are i.i.d. standard Gaussian variables. Then

$$F_E = \ln \int dy \mathbb{E}_{\xi, u^0} P_{\text{out}}(y|\xi\sqrt{K} + u^0\sqrt{K_d - K})$$

$$\times \prod_{a=1}^s \mathbb{E}_{u^a} P_{\text{out}}(y|\xi\sqrt{K} + u^a\sqrt{K_d - K}). \quad (\text{B14})$$

The last product over the replica index a contains identical factors thanks to the RS *Ansatz*. Therefore, by expanding in $s \rightarrow 0^+$, we get

$$F_E = s \int dy \mathbb{E}_{\xi, u^0} P_{\text{out}}(y|\xi\sqrt{K} + u^0\sqrt{K_d - K})$$

$$\times \ln \mathbb{E}_u P_{\text{out}}(y|\xi\sqrt{K} + u\sqrt{K_d - K}) + O(s^2) \quad (\text{B15})$$

$$=: s\phi_{P_{\text{out}}}(K(R_2, \mathcal{Q}); K_d) + O(s^2). \quad (\text{B16})$$

Notice that the energetic contribution to the free entropy has the same form as in the generalized linear model [100]. For our running example of linear readout with Gaussian noise, the function $\phi_{P_{\text{out}}}$ reduces to

$$\phi_{P_{\text{out}}}(K(R_2, \mathcal{Q}); K_d) = -\frac{1}{2} \ln[2\pi e(\Delta + K_d - K)]. \quad (\text{B17})$$

b. Entropic potential

The entropic potential is obtained by counting the degeneracy of configurations yielding the same values of OPs appearing in $\mathbf{K}$:

$$e^{F_s} = \int \prod_{a=0}^s d\mathbf{S}_2^a \int \prod_{a=0}^s dP_W(\mathbf{W}^a) \delta\left(\mathbf{S}_2^a - \frac{\mathbf{W}^{a\top} \text{diag}(\mathbf{v}^0) \mathbf{W}^a}{\sqrt{k}}\right)$$

$$\times \prod_{a \leq b}^{0,s} \prod_{\mathbf{v} \in \mathcal{V}} \delta(d|\mathcal{I}_{\mathbf{v}}| \mathcal{Q}^{ab}(\mathbf{v}) - \sum_{i \in \mathcal{I}_{\mathbf{v}}} \mathbf{W}_i^{a\top} \mathbf{W}_i^b) \prod_{a \leq b}^{0,s} \delta(d^2 R_2^{ab} - \text{Tr} \mathbf{S}_2^a \mathbf{S}_2^b), \quad (\text{B18})$$

where we have introduced the integral over the Hermitian matrices $d\mathbf{S}_2^a = \prod_{\alpha_1 \leq \alpha_2} dS_{2;\alpha_1\alpha_2}^a$. Defining

$$V_W^{kd}(\mathcal{Q}) := \int \prod_{a=0}^s dP_W(\mathbf{W}^a) \prod_{a \leq b}^{0,s} \prod_{\mathbf{v} \in \mathcal{V}} \delta(d|\mathcal{I}_{\mathbf{v}}| \mathcal{Q}^{ab}(\mathbf{v}) - \sum_{i \in \mathcal{I}_{\mathbf{v}}} \mathbf{W}_i^{a\top} \mathbf{W}_i^b), \quad (\text{B19})$$

the entropic potential can be conveniently recast in terms of the following conditional measure:

$$P((\mathbf{S}_2^a) | \mathcal{Q}) = V_W^{kd}(\mathcal{Q})^{-1} \int \prod_a^{0,s} dP_W(\mathbf{W}^a) \delta(\mathbf{S}_2^a - \mathbf{W}^{a\top} \text{diag}(\mathbf{v}) \mathbf{W}^a / \sqrt{k}) \prod_{a \leq b}^{0,s} \prod_{\mathbf{v} \in \mathcal{V}} \delta(d|\mathcal{I}_{\mathbf{v}}| \mathcal{Q}^{ab}(\mathbf{v}) - \sum_{i \in \mathcal{I}_{\mathbf{v}}} \mathbf{W}_i^{a\top} \mathbf{W}_i^b) \quad (\text{B20})$$

as

$$e^{F_s} := V_W^{kd}(\mathcal{Q}) \int dP((\mathbf{S}_2^a) | \mathcal{Q}) \prod_{a \leq b}^{0,s} \delta(d^2 R_2^{ab} - \text{Tr} \mathbf{S}_2^a \mathbf{S}_2^b). \quad (\text{B21})$$

Recall that $\mathcal{V}$ is the support of P_v (assumed discrete for the moment). Recall also that we have quenched the readout weights to the ground truth. This is a measure over different replicas of the random matrices $\mathbf{S}_2^a$, defined in terms of the distribution of the matrices $\mathbf{W}^a$ by the first delta function in Eq. (B20) coupled through the term $\mathcal{Q}^{ab}$ in the second delta function. This coupling between replicas marks a difference with the computation of Ref. [96]: To proceed, we need to relax

this measure to something more manageable. We thus first evaluate the exact asymptotic of its trace second moment to eventually write a relaxation in a moment-matching scheme.

c. Exact second moment of $P((S_2^a)|\mathcal{Q})$

In this measure, one can compute the asymptotic of its second moment

$$\begin{aligned} \int dP((S_2^a)|\mathcal{Q}) \frac{1}{d^2} \text{Tr} S_2^a S_2^b &= V_W^{kd}(\mathcal{Q})^{-1} \int \prod_a^{0,s} dP_W(\mathbf{W}^a) \frac{1}{kd^2} \text{Tr} [\mathbf{W}^{a\top} \text{diag}(\mathbf{v}) \mathbf{W}^a \mathbf{W}^{b\top} \text{diag}(\mathbf{v}) \mathbf{W}^b] \\ &\times \prod_{a \leq b}^{0,s} \prod_{\mathbf{v} \in \mathbf{V}} \delta(d|\mathcal{I}_{\mathbf{v}}| \mathcal{Q}^{ab}(\mathbf{v}) - \sum_{i \in \mathcal{I}_{\mathbf{v}}} \mathbf{W}_i^{a\top} \mathbf{W}_i^b). \end{aligned} \quad (\text{B22})$$

The measure is coupled only through the last δ 's. We can decouple the measure at the cost of introducing Fourier conjugates whose values will then be fixed by a saddle-point computation. The second moment computed will not affect the saddle point; hence, it is sufficient to determine the value of the Fourier conjugates through the computation of $V_W^{kd}(\mathcal{Q})$, which rewrites as

$$\begin{aligned} V_W^{kd}(\mathcal{Q}) &= \int \prod_a^{0,s} dP_W(\mathbf{W}^a) \prod_{a \leq b}^{0,s} \prod_{\mathbf{v} \in \mathbf{V}} d\hat{B}^{ab}(\mathbf{v}) \exp \left[-\hat{B}^{ab}(\mathbf{v}) \left(d|\mathcal{I}_{\mathbf{v}}| \mathcal{Q}^{ab}(\mathbf{v}) - \sum_{i \in \mathcal{I}_{\mathbf{v}}} \mathbf{W}_i^{a\top} \mathbf{W}_i^b \right) \right] \\ &\approx \prod_{\mathbf{v} \in \mathbf{V}} \exp \left(d|\mathcal{I}_{\mathbf{v}}| \text{extr}_{[\hat{B}^{ab}(\mathbf{v})]} \left[-\sum_{a \leq b, 0}^s \hat{B}^{ab}(\mathbf{v}) \mathcal{Q}^{ab}(\mathbf{v}) + \ln \int \prod_{a=0}^s dP_W(w_a) e^{\sum_{a \leq b, 0}^s \hat{B}^{ab}(\mathbf{v}) w_a w_b} \right] \right). \end{aligned} \quad (\text{B23})$$

In the last line, we have used saddle-point integration over $\hat{B}^{ab}(\mathbf{v})$, and the approximate equality is up to a multiplicative $\exp[o(n)]$ constant. From the above, it is clear that the stationary $\hat{B}^{ab}(\mathbf{v})$ are such that

$$\mathcal{Q}^{ab}(\mathbf{v}) = \frac{\int \prod_{r=0}^s dP_W(w_r) w_a w_b \prod_{r \leq t, 0}^s e^{\hat{B}^{rt}(\mathbf{v}) w_r w_t}}{\int \prod_{r=0}^s dP_W(w_r) \prod_{r \leq t, 0}^s e^{\hat{B}^{rt}(\mathbf{v}) w_r w_t}} =: \langle w_a w_b \rangle_{\hat{\mathbf{B}}(\mathbf{v})}. \quad (\text{B24})$$

Using these notations, the asymptotic trace moment of the S_2 's at leading order becomes

$$\begin{aligned} \int dP((S_2^a)|\mathcal{Q}) \frac{1}{d^2} \text{Tr} S_2^a S_2^b &= \frac{1}{kd^2} \sum_{i,l=1}^k \sum_{j,p=1}^d \langle W_{ij}^a v_i^0 W_{ip}^a W_{lj}^b v_l^0 W_{lp}^b \rangle_{\{\hat{\mathbf{B}}(\mathbf{v})\}_{\mathbf{v} \in \mathbf{V}}} \\ &= \frac{1}{k} \sum_{\mathbf{v} \in \mathbf{V}} v^2 \sum_{i \in \mathcal{I}_{\mathbf{v}}} \left\langle \left(\frac{1}{d} \sum_{j=1}^d W_{ij}^a W_{ij}^b \right)^2 \right\rangle_{\hat{\mathbf{B}}(\mathbf{v})} + \frac{1}{k} \sum_{j=1}^d \left\langle \sum_{i=1}^k \frac{v_i^0 (W_{ij}^a)^2}{d} \sum_{l \neq i, 1}^k \frac{v_l^0 (W_{lj}^b)^2}{d} \right\rangle_{\{\hat{\mathbf{B}}(\mathbf{v})\}_{\mathbf{v} \in \mathbf{V}}}. \end{aligned} \quad (\text{B25})$$

We have used the fact that $\langle \cdot \rangle_{\hat{\mathbf{B}}(\mathbf{v})}$ is symmetric if the prior P_W is, thus forcing us to match j with p if $i \neq l$. Considering that by the Nishimori identities $\mathcal{Q}^{aa}(\mathbf{v}) = 1$, it implies $\hat{B}^{aa}(\mathbf{v}) = 0$ for any $a = 0, 1, \dots, s$ and $\mathbf{v} \in \mathbf{V}$. Furthermore, the measure $\langle \cdot \rangle_{\hat{\mathbf{B}}(\mathbf{v})}$ is completely factorized over neuron and input indices. Hence, every normalized sum can be assumed to concentrate to its expectation by the law of large numbers. Specifically, we can write that with high probability as $d, k \rightarrow \infty$,

$$\frac{1}{d} \sum_{i \in \mathcal{I}_{\mathbf{v}}} \sum_{j=1}^d W_{ij}^a W_{ij}^b \longrightarrow |\mathcal{I}_{\mathbf{v}}| \mathcal{Q}^{ab}(\mathbf{v}), \quad \frac{1}{k} \sum_{\mathbf{v}, \mathbf{v}' \in \mathbf{V}} \mathbf{w}' \sum_{j=1}^d \sum_{i \in \mathcal{I}_{\mathbf{v}}} \frac{(W_{ij}^a)^2}{d} \sum_{l \in \mathcal{I}_{\mathbf{v}'}, l \neq i} \frac{(W_{lj}^b)^2}{d} \approx \gamma \sum_{\mathbf{v}, \mathbf{v}' \in \mathbf{V}} \frac{|\mathcal{I}_{\mathbf{v}}| |\mathcal{I}_{\mathbf{v}}'|}{k^2} \mathbf{v} \mathbf{v}' \rightarrow \gamma \bar{v}^2, \quad (\text{B26})$$

where we used $|\mathcal{I}_{\mathbf{v}}|/k \rightarrow P_v(\mathbf{v})$ as k diverges. Consequently, the second moment at leading order appears as claimed:

$$\int dP((S_2^a)|\mathcal{Q}) \frac{1}{d^2} \text{Tr} S_2^a S_2^b = \sum_{\mathbf{v} \in \mathbf{V}} P_v(\mathbf{v}) v^2 \mathcal{Q}^{ab}(\mathbf{v})^2 + \gamma \bar{v}^2 = \mathbb{E}_{v \sim P_v} v^2 \mathcal{Q}^{ab}(v)^2 + \gamma \bar{v}^2. \quad (\text{B27})$$

d. Relaxation of $P((\mathbf{S}_2^a)|\mathcal{Q})$ via maximum entropy with moment matching

We now show how to obtain the relaxation $\tilde{P}((\mathbf{S}_2^a)|\mathcal{Q})$ in Eq. (31), which we report here for the reader's convenience:

$$\tilde{P}((\mathbf{S}_2^a)|\mathcal{Q}) := \tilde{V}_W^{-kd}(\mathcal{Q}) \prod_a^{0,s} P_S(\mathbf{S}_2^a) \prod_{a<b}^{0,s} e^{\frac{1}{2}\tau(\mathcal{Q}^{ab})\text{Tr}\mathbf{S}_2^a\mathbf{S}_2^b}, \quad (\text{B28})$$

where P_S is the probability density of a generalized Wishart random matrix, i.e., of $\tilde{\mathbf{W}}^\top \text{diag}(\mathbf{v}) \tilde{\mathbf{W}} / \sqrt{k}$ with $\tilde{\mathbf{W}} \in \mathbb{R}^{k \times d}$ made of i.i.d. standard *Gaussian* entries, $\tilde{V}_W^{kd}(\mathcal{Q})$ is the proper normalization constant, and $\tau(\mathcal{Q}^{ab})$ is such that

$$\int d\tilde{P}((\mathbf{S}_2^a)|\mathcal{Q}) \frac{1}{d^2} \text{Tr}\mathbf{S}_2^a\mathbf{S}_2^b = \mathbb{E}_{v \sim P_v} v^2 \mathcal{Q}^{ab} (v)^2 + \gamma \bar{v}^2. \quad (\text{B29})$$

We shall see that the above equation converts into a convenient relation involving the inverse function MMSE_S^{-1} when taking the replica-symmetric *Ansatz*. The effective law $\tilde{P}((\mathbf{S}_2^a)|\mathcal{Q})$ is the least restrictive choice among the Wishart-type distributions with a trace moment fixed precisely to the one above. In more specific terms, it is the solution of the following maximum entropy problem:

$$\inf_{P, \tau} \left\{ D_{\text{KL}}(P \| P_S^{\otimes s+1}) + \sum_{a \leq b, 0}^s \tau^{ab} \left(\mathbb{E}_P \frac{1}{d^2} \text{Tr}\mathbf{S}_2^a\mathbf{S}_2^b - \gamma \bar{v}^2 - \mathbb{E}_{v \sim P_v} v^2 \mathcal{Q}^{ab} (v)^2 \right) \right\}, \quad (\text{B30})$$

where P_S is a generalized Wishart distribution [as defined above Eq. (31)], and P is in the space of joint probability distributions over $s+1$ symmetric matrices of dimension $d \times d$. The rationale behind the choice of P_S as a base measure is that, in the absence of any other information, a statistician can always use a generalized Wishart measure

for the $\mathbf{S}_2$'s if they assume universality in the law of the inner weights. This *Ansatz* would still yield a nontrivial performance achieved by our adaptation of GAMP-RIE in Appendix B 4 for generic activations.

Note that if $a = b$ then, by the Nishimori identities, the second moment above matches precisely $R_d = 1 + \gamma \bar{v}^2$. This entails directly $\tau^{aa} = 0$, as the generalized Wishart prior P_S already imposes this constraint.

e. Entropic potential with the relaxed measure

We now use the results from the previous paragraphs to compute the entropic contribution F_S to the free entropy, Eq. (B21). Indeed, let us proceed with the relaxation of the measure $P((\mathbf{S}_2^a)|\mathcal{Q})$ by replacing it with $\tilde{P}((\mathbf{S}_2^a)|\mathcal{Q})$ derived above:

$$e^{F_S} = V_W^{kd}(\mathcal{Q}) \int d\hat{\mathbf{R}}_2 \exp\left(-\frac{d^2}{2} \sum_{a \leq b, 0}^s \hat{R}_2^{ab} R_2^{ab}\right) \frac{1}{\tilde{V}_W^{kd}(\mathcal{Q})} \times \int \prod_{a=0}^s dP_S(\mathbf{S}_2^a) \exp\left(\sum_{a \leq b, 0}^s \frac{\tau_{ab} + \hat{R}_2^{ab}}{2} \text{Tr}\mathbf{S}_2^a\mathbf{S}_2^b\right), \quad (\text{B31})$$

where we have introduced another set of Fourier conjugates $\hat{\mathbf{R}}_2$ for $\mathbf{R}_2$. The factor $V_W^{kd}(\mathcal{Q})$ was already treated in Eq. (B23). However, here it will contribute as a tilt of the overall entropic contribution, and the Fourier conjugates $\hat{\mathcal{Q}}^{ab}(\mathbf{v})$ will appear in the final variational principle.

As usual, the Nishimori identities impose $R_2^{aa} = R_d = 1 + \gamma \bar{v}^2$ without the need of any Fourier conjugate. Hence, similarly to τ^{aa} , $\hat{R}_2^{aa} = 0$, too. Furthermore, in the hypothesis of replica symmetry, we set $\tau^{ab} = \tau$ and $\hat{R}_2^{ab} = \hat{R}_2$ for all $0 \leq a < b \leq s$. Then, when the number of replicas s tends to 0^+ , we can recognize the free entropy of a matrix denoising problem. More specifically, using the Hubbard-Stratonovich transformation [i.e., $\mathbb{E}_{\mathbf{Z}} \exp[(d/2)\text{Tr}\mathbf{M}\mathbf{Z}] = \exp[(d/4)\text{Tr}\mathbf{M}^2]$ for a $d \times d$ symmetric matrix $\mathbf{M}$ with $\mathbf{Z}$ a standard GOE matrix], we get

$$J_n(\tau, \hat{R}_2) := \lim_{s \rightarrow 0^+} \frac{1}{ns} \ln \int \prod_{a=0}^s dP_S(\mathbf{S}_2^a) \exp\left(\frac{\tau + \hat{R}_2}{2} \sum_{a < b, 0}^s \text{Tr}\mathbf{S}_2^a\mathbf{S}_2^b\right) = \frac{1}{n} \mathbb{E} \ln \int dP_{\tilde{\mathbf{S}}}(\tilde{\mathbf{S}}_2) \exp \frac{d}{2} \text{Tr} \left[\sqrt{\tau + \hat{R}_2} \mathbf{Y} \tilde{\mathbf{S}}_2 - (\tau + \hat{R}_2) \frac{\tilde{\mathbf{S}}_2^2}{2} \right], \quad (\text{B32})$$

where $\mathbf{Y} = \mathbf{Y}(\tau + \hat{R}_2) = \sqrt{\tau + \hat{R}_2} \tilde{\mathbf{S}}_2^0 + \boldsymbol{\xi}$ with $\boldsymbol{\xi}$ a standard GOE matrix, $\tilde{\mathbf{S}}_2 = \mathbf{S}_2/\sqrt{d}$ and analogously for the ground truth matrix, and the outer expectation is with respect to $\mathbf{Y}$ (or $\tilde{\mathbf{S}}_2^0, \boldsymbol{\xi}$). Thanks to the fact that the base measure $P_{\tilde{\mathbf{S}}}$ is rotationally invariant, the above can be solved exactly in the limit $n \rightarrow \infty, n/d^2 \rightarrow \alpha$ (see, e.g., Ref. [103]):

$$J(\tau, \hat{R}_2) = \lim J_n(\tau, \hat{R}_2) = \frac{1}{\alpha} \left(\frac{(\tau + \hat{R}_2)R_d}{4} - \iota(\tau + \hat{R}_2) \right), \quad \text{with} \quad \iota(x) := \frac{1}{8} + \frac{1}{2} \Sigma(\rho_{\mathbf{Y}(x)}). \quad (\text{B33})$$

Here, $\iota(x) = \lim I(\mathbf{Y}(x); \tilde{\mathbf{S}}_2^0)/d^2$ is the limiting mutual information between data $\mathbf{Y}(x)$ and signal $\tilde{\mathbf{S}}_2^0$ for the channel $\mathbf{Y}(x) = \sqrt{x}\tilde{\mathbf{S}}_2^0 + \boldsymbol{\xi}$, the measure $\rho_{\mathbf{Y}(x)}$ is the asymptotic spectral law of the observation matrix $\mathbf{Y}(x)$, and $\Sigma(\mu) := \int \ln |x - y| d\mu(x) d\mu(y)$. Using free probability, the law $\rho_{\mathbf{Y}(x)}$ can be obtained as the free convolution of a generalized Marchenko-Pastur distribution [the asymptotic spectral law of $\tilde{\mathbf{S}}_2^0 = \tilde{\mathbf{W}}^{0\top} \text{diag}(\mathbf{v}^0) \tilde{\mathbf{W}}^0 / \sqrt{k d}$, which is a generalized Wishart random matrix] and the semicircular distribution (the asymptotic spectral law of $\boldsymbol{\xi}$); see Ref. [136]. We provide the code to obtain this distribution numerically in the attached repository. The function $\text{MMSE}_S(x)$ is obtained through a derivative of ι , using the so-called I-MMSE relation [103,184]:

$$4 \frac{d}{dx} \iota(x) = \text{MMSE}_S(x) = \frac{1}{x} \left(1 - \frac{4\pi^2}{3} \int \mu_{\mathbf{Y}(x)}^3(y) dy \right). \quad (\text{B34})$$

The normalization $[1/(ns)] \ln \tilde{V}_W^{kd}(\mathcal{Q})$ in the limit $n \rightarrow \infty, s \rightarrow 0^+$ can be simply computed as $J(\tau, 0)$.

For the other normalization, following the same steps as in the previous section, we can simplify $V_W^{kd}(\mathcal{Q})$ as follows:

$$\frac{1}{ns} \ln V_W^{kd}(\mathcal{Q}) \approx \frac{\gamma}{\alpha s} \sum_{\mathbf{v} \in \mathbf{V}} \frac{1}{k} \text{extr}[\mathcal{I}_{\mathbf{v}}] \left[- \sum_{a \leq b, 0}^s \hat{Q}_W^{ab}(\mathbf{v}) Q^{ab}(\mathbf{v}) + \ln \int \prod_{a=0}^s dP_W(w_a) e^{\sum_{a \leq b, 0}^s \hat{Q}_W^{ab}(\mathbf{v}) w_a w_b} \right], \quad (\text{B35})$$

as n grows, where extremization is with respect to the hatted variables only. Thanks to the Nishimori identities, we have that at the saddle point $\hat{Q}^{aa}(\mathbf{v}) = 0$ and $Q^{aa}(\mathbf{v}) = 1$. This, together with standard steps and the RS Ansatz, allows us to write the $d \rightarrow \infty, s \rightarrow 0^+$ limit of the above as

$$\lim_{s \rightarrow 0^+} \lim_{ns} \frac{1}{ns} \ln V_W^{kd}(\mathcal{Q}) = \frac{\gamma}{\alpha} \mathbb{E}_{v \sim P_v} \text{extr} \left[- \frac{\hat{Q}(v) Q(v)}{2} + \psi_{P_W}(\hat{Q}(v)) \right] \quad (\text{B36})$$

with $\psi_{P_W}(\cdot)$ as in the main text. Gathering all these results yields directly

$$\lim_{s \rightarrow 0^+} \lim_{ns} \frac{F_S}{ns} = \text{extr} \left\{ \frac{\hat{R}_2(R_d - R_2)}{4\alpha} - \frac{1}{\alpha} [\iota(\tau + \hat{R}_2) - \iota(\tau)] + \frac{\gamma}{\alpha} \mathbb{E}_{v \sim P_v} \left[\psi_{P_W}(\hat{Q}(v)) - \frac{\hat{Q}(v) Q(v)}{2} \right] \right\}. \quad (\text{B37})$$

Extremization is with respect to $\hat{R}_2, \hat{Q}$. τ has to be intended as a function of $\mathcal{Q} = \{Q(\mathbf{v}) | \mathbf{v} \in \mathbf{V}\}$ through the moment-matching condition

$$4\alpha \partial_\tau J(\tau, 0) = R_d - 4\iota'(\tau) = \mathbb{E}_{v \sim P_v} v^2 Q(v)^2 + \gamma \bar{v}^2, \quad (\text{B38})$$

which is the $s \rightarrow 0^+$ limit of the moment-matching condition between $P((\mathbf{S}_2^a) | \mathcal{Q})$ and $\tilde{P}((\mathbf{S}_2^a) | \mathcal{Q})$. Simplifying using the value of $R_d = 1 + \gamma \bar{v}^2$ according to the Nishimori identities, and using the I-MMSE relation between $\iota(\tau)$ and $\text{MMSE}_S(\tau)$, we get

$$\text{MMSE}_S(\tau) = 1 - \mathbb{E}_{v \sim P_v} v^2 Q(v)^2 \Leftrightarrow \tau = \text{MMSE}_S^{-1}[1 - \mathbb{E}_{v \sim P_v} v^2 Q(v)^2]. \quad (\text{B39})$$

Since MMSE_S is a monotonic decreasing function of its argument (and thus invertible), the above always has a solution, and it is unique for a given collection $\mathcal{Q}$.

f. RS free entropy and saddle-point equations

Putting the energetic Eq. (B16) and entropic Eq. (B37) contributions together, we obtain the variational replica-symmetric free-entropy potential:

$$f_{\text{RS}}^{(1)} := \phi_{P_{\text{out}}}(K(R_2, \mathcal{Q}); K_d) + \frac{1}{4\alpha}(1 + \gamma\bar{v}^2 - R_2)\hat{R}_2 + \frac{\gamma}{\alpha}\mathbb{E}_{v \sim P_v} \left[\psi_{P_W}(\hat{\mathcal{Q}}(v)) - \frac{1}{2}\mathcal{Q}(v)\hat{\mathcal{Q}}(v) \right] + \frac{1}{\alpha}[\iota(\tau(\mathcal{Q})) - \iota(\hat{R}_2 + \tau(\mathcal{Q}))], \quad (\text{B40})$$

which is then extremized with respect to $\{\hat{\mathcal{Q}}(\mathbf{v}), \mathcal{Q}(\mathbf{v}) | \mathbf{v} \in \mathbf{V}\}$, $\hat{R}_2$, R_2 , while τ is a function of $\mathcal{Q}$ through the moment-matching condition (B39). The saddle-point equations are then

$$\begin{cases} \mathcal{Q}(\mathbf{v}) = \mathbb{E}_{w^0, \xi}[w^0 \langle w \rangle_{\hat{\mathcal{Q}}(\mathbf{v})}], \\ P_v(\mathbf{v})\hat{\mathcal{Q}}(\mathbf{v}) = \frac{1}{2\gamma}[R_2 - \gamma\bar{v}^2 - \mathbb{E}_{v \sim P_v} v^2 \mathcal{Q}(v)^2] \partial_{\mathcal{Q}(\mathbf{v})} \tau(\mathcal{Q}) + 2\frac{\alpha}{\gamma} \partial_{\mathcal{Q}(\mathbf{v})} \phi_{P_{\text{out}}}(K(R_2, \mathcal{Q}); K_d), \\ R_2 = R_d - \frac{1}{\bar{R}_2 + \tau(\mathcal{Q})} \left[1 - \frac{4\pi^2}{3} \int \mu_{\mathbf{Y}[\hat{R}_2 + \tau(\mathcal{Q})]}^3(y) dy \right], \\ \hat{R}_2 = 4\alpha \partial_{R_2} \phi_{P_{\text{out}}}(K(R_2, \mathcal{Q}); K_d), \end{cases} \quad (\text{B41})$$

where $w^0 \sim P_W$, $\xi \sim \mathcal{N}(0, 1)$, and we define the measure

$$\langle \cdot \rangle_x = \langle \cdot \rangle_x(w^0, \xi) := \frac{\int dP_W(w) (\cdot) e^{(\sqrt{x}\xi + xw^0)w - \frac{1}{2}xw^2}}{\int dP_W(w) e^{(\sqrt{x}\xi + xw^0)w - \frac{1}{2}xw^2}}. \quad (\text{B42})$$

All the above formulas are easily specialized for the linear readout with Gaussian label noise using Eq. (B17). We report here the saddle-point equations in this case [recalling that g is defined in Eq. (B6)]:

$$\begin{cases} \mathcal{Q}(\mathbf{v}) = \mathbb{E}_{w^0, \xi}[w^0 \langle w \rangle_{\hat{\mathcal{Q}}(\mathbf{v})}], \\ \hat{\mathcal{Q}}(\mathbf{v}) = \frac{1}{2\gamma P_v(\mathbf{v})} [R_2 - \gamma\bar{v}^2 - \mathbb{E}_{v \sim P_v} v^2 \mathcal{Q}(v)^2] \partial_{\mathcal{Q}(\mathbf{v})} \tau(\mathcal{Q}) + \frac{\alpha}{\gamma} \frac{v^2 g'(\mathcal{Q}(\mathbf{v}))}{\Delta + \frac{1}{2}\mu_2^2(R_d - R_2) + g(1) - \mathbb{E}_{v \sim P_v} v^2 g(\mathcal{Q}(v))}, \\ R_2 = R_d - \frac{1}{\bar{R}_2 + \tau(\mathcal{Q})} \left[1 - \frac{4\pi^2}{3} \int \mu_{\mathbf{Y}[\hat{R}_2 + \tau(\mathcal{Q})]}^3(y) dy \right], \\ \hat{R}_2 = \frac{\alpha\mu_2^2}{\Delta + \frac{1}{2}\mu_2^2(R_d - R_2) + g(1) - \mathbb{E}_{v \sim P_v} v^2 g(\mathcal{Q}(v))}. \end{cases} \quad (\text{B43})$$

If one assumes that the overlaps appearing in Eq. (A31) are self-averaging around the values that solve the saddle-point equations (and maximize the RS potential), that is, $R_1^{00}, R_1^{01} \rightarrow 1$ (as assumed in this scaling), $R_2^{00} \rightarrow R_d, R_2^{01} \rightarrow R_2^*$, and $\mathcal{Q}^{00}(\mathbf{v}) \rightarrow 1, \mathcal{Q}^{01}(\mathbf{v}) \rightarrow \mathcal{Q}^*(\mathbf{v})$, then the limiting Bayes-optimal mean-square generalization error for the case of linear readout with Gaussian noise appears as

$$\varepsilon^{\text{opt}} - \Delta = K_d - K^* = \frac{\mu_2^2}{2}(R_d - R_2^*) + g(1) - \mathbb{E}_{v \sim P_v} v^2 g(\mathcal{Q}^*(v)). \quad (\text{B44})$$

g. Noncentered activations

Consider a noncentered activation function, i.e., $\mu_0 \neq 0$ in Eq. (A1). This reflects on the law of the postactivations, which will still be Gaussian, centered at

$$\mathbb{E}_{\mathbf{x}} \lambda^a = \frac{\mu_0}{\sqrt{k}} \sum_{i=1}^k v_i =: \mu_0 \Lambda, \quad (\text{B45})$$

and with the covariance given by Eq. (24) (we are assuming $\|\mathbf{W}_i^a\|^2/d \rightarrow 1$). In the above, we have introduced the new mean parameter Λ . Notice that, if the $\mathbf{v}$'s have a $\bar{v} = O(1)$ mean, then Λ scales as $\sqrt{k}$ due to our choice of normalization.

One can carry out the replica computation for a fixed Λ . This new parameter, being quenched, does not affect the entropic term. It will appear only in the energetic term as a shift to the means, yielding

$$F_E = F_E(\mathbf{K}, \Lambda) = \ln \int dy \int d\lambda \frac{e^{-\frac{1}{2}\lambda^T \mathbf{K}^{-1} \lambda}}{\sqrt{(2\pi)^{s+1} \det \mathbf{K}}} \prod_{a=0}^s P_{\text{out}}(y|\lambda^a + \mu_0 \Lambda). \quad (\text{B46})$$

Within the replica-symmetric *Ansatz*, the above turns into

$$e^{F_E} = \int dy \mathbb{E}_{\xi, u^0} P_{\text{out}} \left(y | \mu_0 \Lambda + \xi \sqrt{\frac{m_K^2}{K}} + u^0 \sqrt{\rho_K - \frac{m_K^2}{K}} \right) \prod_{a=1}^s \mathbb{E}_{u^a} P_{\text{out}}(y | \mu_0 \Lambda + \xi \sqrt{K} + u^a \sqrt{K_d - K}).$$

Therefore, the simplification of the potential F_E proceeds as in the centered activation case, yielding at leading order in the number s of replicas

$$\frac{F_E(K_d, K, \Lambda)}{s} = \int dy \mathbb{E}_{\xi, u^0} P_{\text{out}}(y | \mu_0 \Lambda + \xi \sqrt{K} + u^0 \sqrt{K_d - K}) \ln \mathbb{E}_u P_{\text{out}}(y | \mu_0 \Lambda + \xi \sqrt{K} + u \sqrt{K_d - K}) + O(s)$$

in the Bayes-optimal setting. In the case when $P_{\text{out}}(y|\lambda) = f(y - \lambda)$, then one can verify that the contributions due to the means containing μ_0 cancel each other. This is verified in our running example where P_{out} is the Gaussian channel:

$$\frac{F_E(K_d, K, \Lambda)}{s} = -\frac{1}{2} \ln[2\pi(\Delta + K_d - K)] - \frac{1}{2} - \frac{\mu_0^2}{2} \frac{(\Lambda - \Lambda)^2}{\Delta + K_d - K} + O(s) = -\frac{1}{2} \ln[2\pi(\Delta + K_d - K)] - \frac{1}{2} + O(s). \quad (\text{B47})$$

2. Alternative simplifications of $P((\mathbf{S}_2^a) | \mathcal{Q})$ through moment matching

A crucial step that allowed us to obtain a closed-form expression for the model's free entropy is the relaxation $\tilde{P}((\mathbf{S}_2^a) | \mathcal{Q})$ (31) of the true measure $P((\mathbf{S}_2^a) | \mathcal{Q})$ (30) entering the replicated partition function. The specific form we chose (tilted Wishart distribution with a matching second moment) has the advantage of capturing crucial features of the true measure, such as the fact that the matrices $\mathbf{S}_2^a$ are generalized Wishart matrices with coupled replicas, while keeping the problem solvable with techniques derived from random matrix theory of rotationally invariant ensembles. In this appendix, we report some alternative routes one can take to simplify, or potentially improve the theory.

a. A factorized simplified distribution

In the specialization phase, one can assume that the only crucial feature to keep track in relaxing $P((\mathbf{S}_2^a) | \mathcal{Q})$ (30) is the coupling between different replicas becoming more and more relevant as α increases. In this case, inspired by Refs. [168,169], in order to relax Eq. (30) we can propose the Gaussian *Ansatz*

$$d\tilde{P}((\mathbf{S}_2^a) | \mathcal{Q}) = \prod_{a=0}^s d\mathbf{S}_2^a \prod_{\alpha=1}^d \delta(S_{2;\alpha\alpha}^a - \sqrt{k}\bar{v}) \times \prod_{\alpha_1 < \alpha_2}^d \frac{e^{-\frac{1}{2} \sum_{a,b=0}^s S_{2;\alpha_1\alpha_2}^a \bar{\tau}^{ab}(\mathcal{Q}) S_{2;\alpha_1\alpha_2}^b}}{\sqrt{(2\pi)^{s+1} \det[\bar{\tau}(\mathcal{Q})^{-1}]}}}, \quad (\text{B48})$$

where $\bar{v}$ is the mean of the readout prior P_v , and $\bar{\tau}(\mathcal{Q}) := [\bar{\tau}^{ab}(\mathcal{Q})]_{a,b}$ is fixed by

$$[\bar{\tau}(\mathcal{Q})^{-1}]_{ab} = \mathbb{E}_{v \sim P_v} v^2 Q^{ab}(v)^2.$$

In words, first, the diagonal elements of $\mathbf{S}_2^a$ are d random variables whose $O(1)$ fluctuations cannot affect the free entropy in the asymptotic regime we are considering, being too few compared to $n = \Theta(d^2)$. Hence, we assume they concentrate to their mean. Concerning the $d(d-1)/2$ off-diagonal elements of the matrices $(\mathbf{S}_2^a)_a$, they are zero-mean variables whose distribution at given $\mathcal{Q}$ is assumed to be factorized over the input indices. The definition of $\bar{\tau}(\mathcal{Q})$ ensures matching with the true second moment (B27).

Equation (B48) is considerably simpler than Eq. (31): Following this *Ansatz*, the entropic contribution to the free entropy gives

$$\begin{aligned}
 e^{\bar{F}_s} &:= \int \prod_{a \leq b, 0}^s d\hat{R}_2^{ab} e^{kd \ln V_W(\mathcal{Q}) + \frac{d^2}{4} \text{Tr} \hat{\mathbf{R}}_2^T \mathbf{R}_2} \left[\int \prod_{a=0}^s dS_2^a \frac{e^{-\frac{1}{2} \sum_{a,b=0}^s S_2^a [\bar{\tau}^{ab}(\mathcal{Q}) + \hat{R}_2^{ab}] S_2^b}}{\sqrt{(2\pi)^{s+1} \det[\bar{\tau}(\mathcal{Q})^{-1}]}} \right]^{d(d-1)/2} \\
 &\times \int \prod_{a=0}^s \prod_{\alpha=1}^d dS_{2;\alpha\alpha}^a \delta(S_{2;\alpha\alpha}^a - \sqrt{k\bar{v}}) e^{-\frac{1}{4} \sum_{a,b=0}^s \hat{R}_2^{ab} \sum_{\alpha=1}^d S_{2;\alpha\alpha}^a S_{2;\alpha\alpha}^b},
 \end{aligned} \tag{B49}$$

instead of Eq. (B31). Integration over the diagonal elements $(S_{2;\alpha\alpha}^a)_\alpha$ can be done straightforwardly, yielding

$$e^{\bar{F}_s} = \int \prod_{a \leq b, 0}^s d\hat{R}_2^{ab} e^{kd \ln V_W(\mathcal{Q}) + \frac{d^2}{4} \text{Tr} \hat{\mathbf{R}}_2^T (\mathbf{R}_2 - \gamma \mathbf{1} \mathbf{1}^T \bar{v}^2)} \left[\int \prod_{a=0}^s dS_2^a \frac{e^{-\frac{1}{2} \sum_{a,b=0}^s S_2^a [\bar{\tau}^{ab}(\mathcal{Q}) + \hat{R}_2^{ab}] S_2^b}}{\sqrt{(2\pi)^{s+1} \det[\bar{\tau}(\mathcal{Q})^{-1}]}} \right]^{d(d-1)/2}. \tag{B50}$$

The remaining Gaussian integral over the off-diagonal elements of $\mathbf{S}_2$ can be performed exactly, leading to

$$e^{\bar{F}_s} = \int \prod_{a \leq b, 0}^s d\hat{R}_2^{ab} e^{kd \ln V_W(\mathcal{Q}) + \frac{d^2}{4} \text{Tr} \hat{\mathbf{R}}_2^T (\mathbf{R}_2 - \gamma \mathbf{1} \mathbf{1}^T \bar{v}^2) - \frac{d(d-1)}{4} \ln \det[I_{s+1} + \hat{\mathbf{R}}_2 \bar{\tau}(\mathcal{Q})^{-1}]}. \tag{B51}$$

In order to proceed and perform the $s \rightarrow 0^+$ limit, we use the RS *Ansatz* for the overlap matrices combined with the Nishimori identities, as explained above. The only difference with respect to the approach detailed in Appendix B 1 is the determinant in the exponent of the integrand of Eq. (B51), which reads

$$\ln \det[I_{s+1} + \hat{\mathbf{R}}_2 \bar{\tau}(\mathcal{Q})^{-1}] = s \ln[1 + \hat{R}_2(1 - \mathbb{E}_{v \sim P_v} v^2 \mathcal{Q}(v)^2)] - s \hat{R}_2 + O(s^2). \tag{B52}$$

After taking the replica and high-dimensional limits, the resulting free entropy is

$$\begin{aligned}
 f_{\text{sp}}^{(1)} &= \phi_{P_{\text{out}}}(K(R_2, \mathcal{Q}); K_d) + \frac{(1 + \gamma \bar{v}^2 - R_2) \hat{R}_2}{4\alpha} + \frac{\gamma}{\alpha} \mathbb{E}_{v \sim P_v} \left[\psi_{P_W}(\hat{\mathcal{Q}}(v)) - \frac{1}{2} \mathcal{Q}(v) \hat{\mathcal{Q}}(v) \right] \\
 &- \frac{1}{4\alpha} \ln[1 + \hat{R}_2(1 - \mathbb{E}_{v \sim P_v} v^2 \mathcal{Q}(v)^2)],
 \end{aligned} \tag{B53}$$

to be extremized with respect to $R_2, \hat{R}_2, \{\mathcal{Q}(v), \hat{\mathcal{Q}}(v)\}$. The main advantage of this expression over Eq. (B40) is its simplicity: The moment-matching condition fixing $\bar{\tau}(\mathcal{Q})$ is straightforward (and has been solved explicitly in the final formula), and the result does not depend on the nontrivial (and difficult to numerically evaluate) function $\iota(x)$, which is the mutual information of the associated matrix denoising problem (which has been effectively replaced by the much simpler denoising problem of independent Gaussian variables under Gaussian noise). Moreover, one can show, in the same fashion as done in Appendix B 3, that the generalization error predicted from this expression has the same large- α behavior than the one obtained from Eq. (B40). However, not surprisingly, being derived from an *Ansatz* ignoring the Wishart-like nature of the matrices $\mathbf{S}_2^a$, this expression does not reproduce the expected behavior of the model in the universal phase, i.e., for $\alpha < \alpha_{\text{sp}}(\gamma)$.

To fix this issue, one can compare the predictions of the theory derived from this *Ansatz*, with the ones obtained by plugging $\mathcal{Q}(v) = 0 \ \forall v$ (denoted $\mathcal{Q} \equiv 0$) into the theory devised in the main text [Eq. (15)]

$$\begin{aligned}
 f_{\text{uni}}^{(1)} &:= \phi_{P_{\text{out}}}(K(R_2, \mathcal{Q} \equiv 0); K_d) \\
 &+ \frac{1}{4\alpha} (1 + \gamma \bar{v}^2 - R_2) \hat{R}_2 - \frac{1}{\alpha} \iota(\hat{R}_2)
 \end{aligned} \tag{B54}$$

to be extremized now only with respect to the scalar parameters $R_2, \hat{R}_2$ [one can easily verify that, for $\mathcal{Q} \equiv 0$, $\tau(\mathcal{Q}) = 0$ and the extremization with respect to $\hat{\mathcal{Q}}$ in Eq. (15) gives $\hat{\mathcal{Q}} \equiv 0$]. Notice that $f_{\text{uni}}^{(1)}$ is not depending on the prior over the inner weights, which is the reason why we are calling it universal. For consistency, the two free entropies $f_{\text{sp}}^{(1)}, f_{\text{uni}}^{(1)}$ should be compared through a discrete variational principle; that is, the free entropy of the model is predicted to be

$$\bar{f}_{\text{RS}}^{(1)} := \max\{\text{extr} f_{\text{uni}}^{(1)}, \text{extr} f_{\text{sp}}^{(1)}\}, \tag{B55}$$

instead of the unified variational form (15). Quite generally, $\text{extr} f_{\text{uni}}^{(1)} > \text{extr} f_{\text{sp}}^{(1)}$ for low values of α , so that the behavior of the model in the universal phase is correctly predicted. The curves cross at a critical value

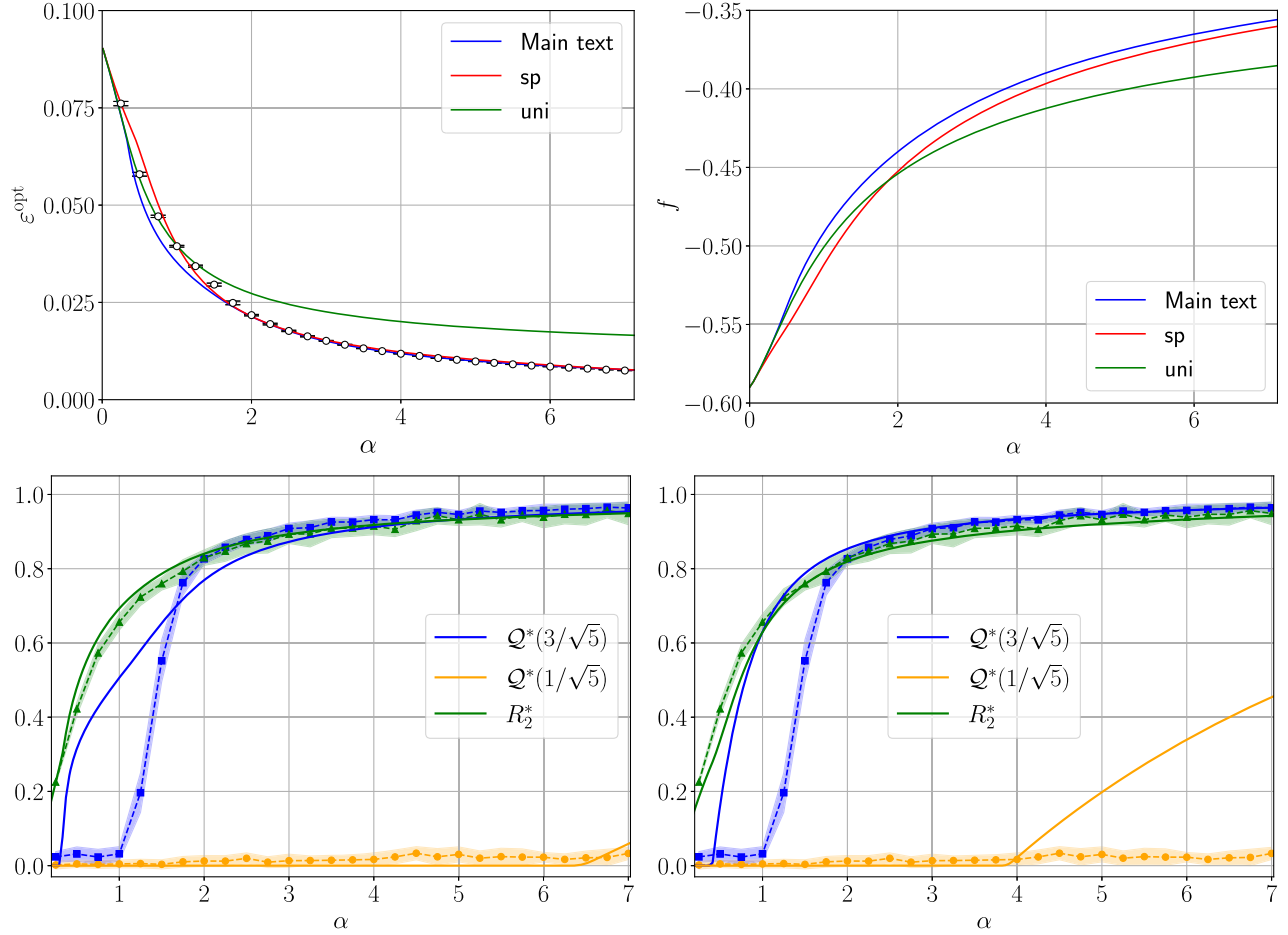


FIG. 22. Different theoretical curves and numerical results for $\text{ReLU}(x)$ activation, $P_v = \frac{1}{4}(\delta_{-3/\sqrt{5}} + \delta_{-1/\sqrt{5}} + \delta_{1/\sqrt{5}} + \delta_{3/\sqrt{5}})$, $d = 200$, $\gamma = 0.5$, with linear readout with Gaussian noise of variance $\Delta = 0.1$. Top left: optimal mean-square generalization error predicted by the theory reported in the main text (solid blue) versus the branch obtained from the simplified *Ansatz* (B48) (solid red). The green solid line shows the universal branch corresponding to $\mathcal{Q} \equiv 0$, and empty circles are HMC results with informative initialization and homogeneous quenched readouts. Top right: theoretical free-entropy curves (colors and line styles as top left). Bottom: predictions for the overlaps $\mathcal{Q}(v)$ and R_2 from the theory devised in the main text (left) and in Appendix B 2 a (right).

$$\bar{\alpha}_{\text{sp}}(\gamma) = \sup\{\alpha | \text{extr}f_{\text{uni}}^{(1)} > \text{extr}f_{\text{sp}}^{(1)}\}, \quad (\text{B56})$$

instead of the value $\alpha_{\text{sp}}(\gamma)$ reported in the main text. This approach has been profitably adopted in Ref. [171] in the context of matrix denoising, a problem sharing some of the challenges presented in this paper. In this respect, it provides a heuristic solution that quantitatively predicts the behavior of the model in most of its phase diagram. Moreover, for any activation σ with a second Hermite coefficient $\mu_2 = 0$ (e.g., all odd activations) the *Ansatz* (B48) yields the same theory as the one devised in the main text, as in this case, $K(R_2, \mathcal{Q})$ entering the energetic part of the free entropy does not depend on R_2 , so that the extremization selects $R_2 = \hat{R}_2 = 0$, and the remaining parts of Eq. (B53) match the ones of Eq. (15). Finally, Eq. (B48) is consistent with the observation that specialization never arises in the case of quadratic activation and Gaussian prior over the inner weights: In this case, one can check that the

universal branch $\text{extr}f_{\text{uni}}^{(1)}$ is always higher than $\text{extr}f_{\text{sp}}^{(1)}$, and thus never selected by Eq. (B55). For a convincing check on the validity of this approach, and a comparison with the theory devised in the main text and numerical results, see Fig. 22, top left panel.

However, despite its merits listed above, this appendix's approach presents some issues, both from the theoretical and practical points of view:

- (i) The final free entropy of the model is obtained by comparing curves derived from completely different *Ansätze* for the distribution $P((\mathbf{S}_2^d) | \mathcal{Q})$ (Gaussian with coupled replicas, leading to f_{sp} , vs pure generalized Wishart with independent replicas, leading to f_{uni}), rather than within a unified theory as in the main text.
- (ii) The predicted critical value $\bar{\alpha}_{\text{sp}}(\gamma)$ seems to be systematically larger than the one observed in experiments (see Fig. 22, top right panel, and

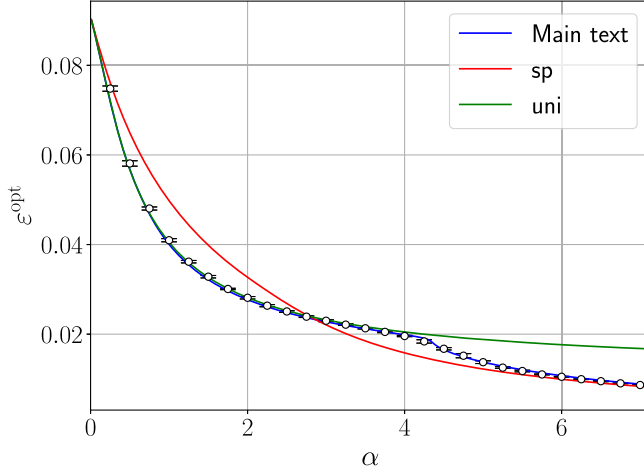


FIG. 23. Generalization error for ReLU activation and Rademacher readout prior P_v of the theory reported in the main text (solid blue) versus the branch obtained from the simplified *Ansatz* (B48) (solid red). The green solid line shows $\mathcal{Q} \equiv 0$ (universal branch), and empty circles are HMC results with informative initialization and homogeneous quenched readouts. All hyperparameters are the same as in Fig. 22.

compare the crossing point of the “sp” and “uni” free entropies with the actual transition where the numerical points depart from the universal branch in the top left panel).

- (iii) Predictions for the functional overlap $\mathcal{Q}^*$ from this approach are in much worse agreement with experimental data with respect to the ones from the theory presented in the main text (see Fig. 22, bottom panel, and compare with Fig. 7 in the main text).
- (iv) In the cases we tested, the prediction for the generalization error from the theory devised in the main text is in much better agreement with numerical simulations than the one from this appendix (see Fig. 23 for a comparison).

Therefore, the more elaborate theory presented in the main is not only more meaningful from the theoretical viewpoint, but also in overall better agreement with simulations.

b. Possible refined analyses with structured $\mathbf{S}_2$ matrices

In the main text, we kept track of the inhomogeneous profile of the readouts induced by the nontrivial distribution P_v , which is ultimately responsible for the sequence of specialization phase transitions occurring at increasing α , thanks to a functional OP $\mathcal{Q}(\mathbf{v})$ measuring how much the student’s hidden weights corresponding to all the readout elements equal to $\mathbf{v}$ have aligned with the teacher’s. However, when writing $\tilde{P}((\mathbf{S}_2^a)|\mathcal{Q})$ we treated the tensor $\mathbf{S}_2^a$ as a whole, without considering the possibility that its “components”

$$S_{2;\alpha_1\alpha_2}^a(\mathbf{v}) := \frac{\mathbf{v}}{\sqrt{|\mathcal{I}_v|}} \sum_{i \in \mathcal{I}_v} W_{i\alpha_1}^a W_{i\alpha_2}^a \quad (\text{B57})$$

could follow different laws for different $\mathbf{v} \in \mathbf{V}$. To do so, let us define

$$\begin{aligned} R_2^{ab} &= \frac{1}{k} \sum_{\mathbf{v}, \mathbf{v}'} \mathbf{v} \mathbf{v}' \sum_{i \in \mathcal{I}_v, j \in \mathcal{I}_{v'}} (\Omega_{ij}^{ab})^2 \\ &= \sum_{\mathbf{v}, \mathbf{v}'} \frac{\sqrt{|\mathcal{I}_v| |\mathcal{I}_{v'}|}}{k} \mathcal{Q}_2^{ab}(\mathbf{v}, \mathbf{v}'), \quad \text{where} \\ \mathcal{Q}_2^{ab}(\mathbf{v}, \mathbf{v}') &:= \frac{1}{d^2} \text{Tr} \mathbf{S}_2^a(\mathbf{v}) \mathbf{S}_2^b(\mathbf{v}')^\top. \end{aligned} \quad (\text{B58})$$

The generalization of Eq. (B27) then reads

$$\begin{aligned} \int dP((\mathbf{S}_2^a)|\mathcal{Q}) \frac{1}{d^2} \text{Tr} \mathbf{S}_2^a(\mathbf{v}) \mathbf{S}_2^b(\mathbf{v}')^\top \\ = \delta_{\mathbf{v} \mathbf{v}'} \mathcal{Q}^{ab}(\mathbf{v})^2 + \gamma \mathbf{v} \mathbf{v}' \sqrt{P_v(\mathbf{v}) P_v(\mathbf{v}')} \end{aligned} \quad (\text{B59})$$

with respect to the true distribution $P((\mathbf{S}_2^a)|\mathcal{Q})$ reported in Eq. (30). Despite the already good match of the theory in the main text with the numerics, taking into account this additional level of structure thanks to a refined simplified measure could potentially lead to further improvements. The simplified measure able to enforce this moment matching while taking into account the Wishart form (B57) of the matrices $[\mathbf{S}_2^a(\mathbf{v})]$ is

$$\begin{aligned} d\tilde{P}((\mathbf{S}_2^a)|\mathcal{Q}) &\propto \prod_{\mathbf{v} \in \mathbf{V}} \prod_a dP_S^{\mathbf{v}}(\mathbf{S}_2^a(\mathbf{v})) \\ &\times \prod_{\mathbf{v} \in \mathbf{V}} \prod_{a < b} e^{\frac{1}{2} \tilde{\tau}_v^{ab}(\mathcal{Q}) \text{Tr} \mathbf{S}_2^a(\mathbf{v}) \mathbf{S}_2^b(\mathbf{v})}, \end{aligned} \quad (\text{B60})$$

where $P_S^{\mathbf{v}}$ is the law of a random matrix $\mathbf{v} \bar{\mathbf{W}} \bar{\mathbf{W}}^\top |\mathcal{I}_v|^{-1/2}$ with $\bar{\mathbf{W}} \in \mathbb{R}^{d \times |\mathcal{I}_v|}$ having i.i.d. standard Gaussian entries. For properly chosen $(\tilde{\tau}_v^{ab})$, Eq. (B59) is verified for this simplified measure.

However, the OPs $[\mathcal{Q}_2^{ab}(\mathbf{v}, \mathbf{v}')]$ are difficult to deal with if keeping a general form, as they not only imply coupled replicas $[\mathbf{S}_2^a(\mathbf{v})]_a$ for a given $\mathbf{v}$ (a kind of coupling that is easily linearized with a single Hubbard-Stratonovich transformation, within the replica-symmetric treatment justified in Bayes-optimal learning), but also a coupling for different values of the variable $\mathbf{v}$. Linearizing it would yield a more complicated matrix model than the integral reported in Eq. (B32), because the resulting coupling field would break rotational invariance, and therefore the model does not have a form which is known to be solvable; see Ref. [132].

A first idea to simplify $P((\mathbf{S}_2^a)|\mathcal{Q})$ (30) while taking into account the additional structure induced by Eqs. (B58) and (B59) and keeping the model solvable, is to consider a generalization of the relaxation (B48). This entails dropping entirely the dependences among matrix entries, induced by their Wishart-like form (B57), for each $\mathbf{S}_2^a(\mathbf{v})$. In this case, the moment constraints (B59) can be exactly enforced by choosing the simplified measure

$$d\bar{P}((S_2^a)|\mathcal{Q}) = \prod_{v \in V} \prod_{a=0}^s dS_2^a(v) \prod_{\alpha=1}^d \delta(S_{2;\alpha\alpha}^a(v) - v\sqrt{|\mathcal{I}_v|}) \\ \times \prod_{v \in V} \prod_{\alpha_1 < \alpha_2}^d \frac{e^{-\frac{1}{2} \sum_{a,b=0}^s S_{2;\alpha_1\alpha_2}^a(v) \bar{\tau}_v^{ab}(\mathcal{Q}) S_{2;\alpha_1\alpha_2}^b(v)}}{\sqrt{(2\pi)^{s+1} \det[\bar{\tau}_v(\mathcal{Q})^{-1}]}}. \quad (\text{B61})$$

The parameters $[\bar{\tau}_v^{ab}(\mathcal{Q})]$ are then properly chosen to enforce Eq. (B59) for all $0 \leq a \leq b \leq s$ and $v, v' \in V$. Using this measure, the resulting entropic term, taking into account the degeneracy of the OPs $[Q_2^{ab}(v, v')]$ and $[Q^{ab}(v)]$, remains tractable through Gaussian integrals [the energetic term is obviously unchanged once we express (R_2^{ab}) entering it using these new OPs through the identity (B58), and keeping in mind that nothing changes for higher-order overlaps compared to the theory in the main text]. We leave for future work the analysis of this Gaussian relaxation and other possible simplifications of Eq. (B60) leading to solvable models.

3. Large sample rate limit of $f_{\text{RS}}^{(1)}$

In this section, we show that when the prior over the weights is discrete, the MI can never exceed the entropy of the prior itself. As for the main text, we consider the readouts quenched to the ground truth ones, since they cannot affect the MI between weights and data at this scaling. For this appendix, we restrict to $L = 1$, but the argument can be generalized to an arbitrary number of layers.

We first need to control the function MMSE when its argument is large. By a saddle-point argument, one can show that the leading term for $\text{MMSE}_S(\tau)$ when $\tau \rightarrow \infty$ is of the type $C(\gamma)/\tau$ for a proper constant C depending at most on γ .

We now notice that the equation for $\hat{Q}(v)$ in Eq. (B41) can be rewritten as

$$P_v(v) \hat{Q}(v) = \frac{1}{2\gamma} [\text{MMSE}_S(\tau) - \text{MMSE}_S(\tau + \hat{R}_2)] \partial_{Q(v)} \tau \\ + 2 \frac{\alpha}{\gamma} \partial_{Q(v)} \phi_{P_{\text{out}}}(K(R_2, \mathcal{Q}); K_d). \quad (\text{B62})$$

For $\alpha \rightarrow \infty$, we make the self-consistent *Ansatz* $Q(v) = 1 - o_\alpha(1)$. As a consequence, using the aforementioned scaling of the $\text{MMSE}(\tau)$, $1/\tau$ has to vanish by the moment-matching condition (B39) as $o_\alpha(1)$, too. Using the very same equation, we are also able to evaluate $\partial_{Q(v)} \tau$ as follows:

$$\partial_{Q(v)} \tau = \frac{-2P_v(v) v^2 Q(v)}{\text{MMSE}'(\tau)} \approx \frac{2P_v(v) v^2 Q(v) \tau^2}{C(\gamma)} \quad (\text{B63})$$

as $\alpha \rightarrow \infty$, where we have used $\text{MMSE}_S(\tau) \approx C(\gamma)/\tau$ to estimate the derivative. We use the same approximation for

the two MMSEs appearing in the fixed-point equation for $\hat{Q}(v)$:

$$\hat{Q}(v) \approx \frac{v^2 Q(v)}{\gamma C(\gamma)} \frac{\tau^2}{\tau(\tau + \hat{R}_2)} \hat{R}_2 \\ + 2 \frac{\alpha}{P_v(v) \gamma} \partial_{Q(v)} \phi_{P_{\text{out}}}(K(R_2, \mathcal{Q}); K_d), \quad (\text{B64})$$

where we are neglecting multiplicative constants for brevity. From the last equation in Eq. (B41), we see that $\hat{R}_2$ cannot diverge more than $O(\alpha)$. Thanks to the above approximation and the first equation of Eq. (B41), this entails that $Q(v)$ is approaching 1 exponentially fast in α [$1 - Q(v) = O(e^{-c\alpha})$ for some $c > 0$] due to the discreteness of its prior, which in turn implies τ is diverging exponentially in α . As a consequence,

$$\frac{\tau^2}{\tau(\tau + \hat{R}_2)} \approx 1. \quad (\text{B65})$$

Furthermore, one also has

$$\frac{1}{\alpha} [\iota(\tau) - \iota(\tau + \hat{R}_2)] = -\frac{1}{4\alpha} \int_{\tau}^{\tau + \hat{R}_2} \text{MMSE}_S(t) dt \\ \approx -\frac{C(\gamma)}{4\alpha} \ln \left(1 + \frac{\hat{R}_2}{\tau} \right) \xrightarrow{\alpha \rightarrow \infty} 0, \quad (\text{B66})$$

as $\hat{R}_2/\tau$ vanishes with exponential speed in α .

Concerning the function ψ_{P_W} , given that it is related to a Bayes-optimal scalar Gaussian channel, and its SNRs $\hat{Q}(v)$ are all diverging, one can compute the integral by the saddle point, which is inevitably attained at the ground truth:

$$\psi_{P_W}(\hat{Q}(v)) = \frac{\hat{Q}(v) Q(v)}{2} \\ \approx \mathbb{E}_{w^0} \ln \int dP_W(w) \mathbb{1}(w = w^0) \\ + \mathbb{E} \left[\left(\sqrt{\hat{Q}(v)} \xi + \hat{Q}(v) w^0 \right) w^0 - \frac{\hat{Q}(v)}{2} (w^0)^2 \right] \\ - \frac{\hat{Q}(v) [1 - O(e^{-c\alpha})]}{2} = -H(W) + o_\alpha(1). \quad (\text{B67})$$

Considering that $\phi_{P_{\text{out}}}(K(R_2, \mathcal{Q}); K_d) \xrightarrow{\alpha \rightarrow \infty} \phi_{P_{\text{out}}}(K_d; K_d)$, and using Eq. (A20), it is then straightforward to check that our RS version of the MI saturates to the entropy of the prior P_W when $\alpha \rightarrow \infty$:

$$-\frac{\alpha}{\gamma} \text{extr}_{\text{RS}}^{(1)} + \frac{\alpha}{\gamma} \mathbb{E}_\lambda \int dy P_{\text{out}}(y|\lambda) \ln P_{\text{out}}(y|\lambda) \xrightarrow{\alpha \rightarrow \infty} H(W), \quad (\text{B68})$$

where the factor α/γ is due to the fact that the components of $\mathbf{W}$ are kd , and not n .

4. Extension of GAMP-RIE to arbitrary activation for $L=1$

For simplicity, let us consider $P_{\text{out}}(y|\lambda) = \exp\{-[1/(2\Delta)](y-\lambda)^2\}/\sqrt{2\pi\Delta}$, which entails

$$y_\mu | (\boldsymbol{\theta}^0, \mathbf{x}_\mu) \stackrel{d}{=} \frac{\mathbf{v}^\top}{\sqrt{k}} \sigma\left(\frac{\mathbf{W}^0 \mathbf{x}_\mu}{\sqrt{d}}\right) + \sqrt{\Delta} z_\mu, \quad \mu = 1, \dots, n, \quad (\text{B69})$$

where z_μ are i.i.d. standard Gaussian random variables, and $\stackrel{d}{=}$ means equality in law. Expanding σ in the Hermite polynomial basis, we have

$$y_\mu | (\boldsymbol{\theta}^0, \mathbf{x}_\mu) \stackrel{d}{=} \mu_0 \frac{\mathbf{v}^\top \mathbf{1}_k}{\sqrt{k}} + \mu_1 \frac{\mathbf{v}^\top \mathbf{W}^0 \mathbf{x}_\mu}{\sqrt{kd}} + \frac{\mu_2}{2} \frac{\mathbf{v}^\top}{\sqrt{k}} \text{He}_2\left(\frac{\mathbf{W}^0 \mathbf{x}_\mu}{\sqrt{d}}\right) + \dots + \sqrt{\Delta} z_\mu, \quad (\text{B70})$$

where ... represents the terms beyond second order. Without loss of generality, for this choice of output channel we can set $\mu_0 = 0$ as discussed in Appendix B 1 g. For low enough α , it is reasonable to assume that higher-order terms in ... cannot be learned given quadratically many samples and, as a result, play the role of effective noise, which we assume independent of the first three terms. We shall see that this reasoning actually applies to the extension of the GAMP-RIE we derive, which plays the role of a “smart” spectral algorithm, regardless of the value of α . Therefore, these terms accumulate in an asymptotically Gaussian noise thanks to the central limit theorem (it is a projection of a centered function applied entrywise to a vector with i.i.d. entries), with variance $g(1)$. We thus obtain the effective model

$$y_\mu | (\boldsymbol{\theta}^0, \mathbf{x}_\mu) \stackrel{d}{=} \mu_1 \frac{\mathbf{v}^\top \mathbf{W}^0 \mathbf{x}_\mu}{\sqrt{kd}} + \frac{\mu_2}{2} \frac{\mathbf{v}^\top}{\sqrt{k}} \text{He}_2\left(\frac{\mathbf{W}^0 \mathbf{x}_\mu}{\sqrt{d}}\right) + \sqrt{\Delta + g(1)} z_\mu. \quad (\text{B71})$$

The first term in this expression can be learned with vanishing error given quadratically many samples (Remark 9); hence, it can be ignored. This consideration allows to further simplify the model to

$$\begin{aligned} \bar{y}_\mu &:= y_\mu - \mu_1 \frac{\mathbf{v}^\top \mathbf{W}^0 \mathbf{x}_\mu}{\sqrt{kd}} \\ &\stackrel{d}{=} \frac{\mu_2}{2} \frac{\mathbf{v}^\top}{\sqrt{k}} \text{He}_2\left(\frac{\mathbf{W}^0 \mathbf{x}_\mu}{\sqrt{d}}\right) + \sqrt{\Delta + g(1)} z_\mu, \end{aligned} \quad (\text{B72})$$

where $\bar{y}_\mu$ is y_μ with the (asymptotically) perfectly learned linear term removed, and the last equality in distribution is again conditional on $(\boldsymbol{\theta}^0, \mathbf{x}_\mu)$. From the formula

$$\begin{aligned} \frac{\mathbf{v}^\top}{\sqrt{k}} \text{He}_2\left(\frac{\mathbf{W}^0 \mathbf{x}_\mu}{\sqrt{d}}\right) &= \text{Tr} \frac{\mathbf{W}^{0\top} \text{diag}(\mathbf{v}) \mathbf{W}^0}{d\sqrt{k}} \mathbf{x}_\mu \mathbf{x}_\mu^\top - \frac{\mathbf{v}^\top \mathbf{1}_k}{\sqrt{k}} \\ &\approx \frac{1}{\sqrt{kd}} \text{Tr}[(\mathbf{x}_\mu \mathbf{x}_\mu^\top - I_d) \mathbf{W}^{0\top} \text{diag}(\mathbf{v}) \mathbf{W}^0], \end{aligned} \quad (\text{B73})$$

where $\approx$ is exploiting the concentration $\text{Tr} \mathbf{W}^{0\top} \text{diag}(\mathbf{v}) \mathbf{W}^0 / (d\sqrt{k}) \rightarrow \mathbf{v}^\top \mathbf{1}_k / \sqrt{k}$, and the Gaussian equivalence property that $\mathbf{M}_\mu := (\mathbf{x}_\mu \mathbf{x}_\mu^\top - I_d) / \sqrt{d}$ behaves like a GOE sensing matrix, i.e., a symmetric matrix whose upper triangular part has i.i.d. entries from $\mathcal{N}(0, (1 + \delta_{ij})/d)$ [96], the model can be seen as a GLM with signal $\tilde{\mathbf{S}}_2 := \mathbf{W}^{0\top} \text{diag}(\mathbf{v}) \mathbf{W}^0 / \sqrt{kd}$:

$$y_\mu^{\text{GLM}} = \frac{\mu_2}{2} \text{Tr}[\mathbf{M}_\mu \tilde{\mathbf{S}}_2] + \sqrt{\Delta + g(1)} z_\mu. \quad (\text{B74})$$

Starting from this equation, the arguments of Appendix B 1 and Ref. [96], based on known results on the GLM [100] and matrix denoising [101–103], allow us to obtain the free entropy of such a matrix-sensing problem. The result is consistent with the $\mathcal{Q} \equiv 0$ solution of the saddle-point equations obtained from the replica method in Appendix B 1, which, as anticipated, corresponds to the case where the Hermite polynomial combinations of the signal following the second one are not learned.

Note that, as supported by the numerics, the model actually admits specialization when α is big enough; hence, the above equivalence cannot hold on the whole phase diagram at the information-theoretic level. In fact, if specialization occurs one cannot consider the ... terms in Eq. (B70) as noise uncorrelated with the first ones, as the model is aligning with the actual teacher’s weights, such that it learns all the successive terms at once.

We now assume that this mapping holds at the algorithmic level; namely, that we can process the data algorithmically as if they were coming from the identified GLM, and thus try to infer the signal $\tilde{\mathbf{S}}_2^0 = \mathbf{W}^{0\top} \text{diag}(\mathbf{v}) \mathbf{W}^0 / \sqrt{kd}$ and construct a predictor from it. Based on this idea, we propose Algorithm 1 that can indeed reach the performance predicted by the $\mathcal{Q} \equiv 0$ solution of our replica theory.

Remark 9. In the linear data regime, where n/d converges to a fixed constant α_1 , only the first term in Eq. (B70) can be learned while the rest behaves like noise. By the same argument as above, the model is equivalent to

$$y_\mu = \mu_1 \frac{\mathbf{v}^\top \mathbf{W}^0 \mathbf{x}_\mu}{\sqrt{kd}} + \sqrt{\Delta + \nu - \mu_0^2 - \mu_1^2} z_\mu, \quad (\text{B76})$$

Algorithm 1. GAMP-RIE for training shallow neural networks with arbitrary activation

Input: Fresh data point $\mathbf{x}_{\text{test}}$ with unknown associated response y_{test} , dataset $\mathcal{D} = \{(\mathbf{x}_\mu, y_\mu)\}_{\mu=1}^n$.
Output: Estimator $\hat{y}_{\text{test}}$ of y_{test} .
 Estimate $y^{(0)} := \mu_0 \mathbf{v}^\top \mathbf{1} / \sqrt{k}$ as

$$\hat{y}^{(0)} = \frac{1}{n} \sum_{\mu} y_{\mu};$$

Estimate $\langle \mathbf{W}^\top \mathbf{v} \rangle / \sqrt{k}$ using Eq. (B77).

Estimate the μ_1 term in the Hermite expansion (B70) as

$$\hat{y}_{\mu}^{(1)} = \mu_1 \frac{\langle \mathbf{v}^\top \mathbf{W} \rangle \mathbf{x}_{\mu}}{\sqrt{kd}};$$

Compute

$$\tilde{y}_{\mu} = \frac{y_{\mu} - \hat{y}_{\mu}^{(0)} - \hat{y}_{\mu}^{(1)}}{\mu_2/2}; \quad \tilde{\Delta} = \frac{\Delta + g(1)}{\mu_2^2/4};$$

Input $\{(\mathbf{x}_{\mu}, \tilde{y}_{\mu})\}_{\mu=1}^n$ and $\tilde{\Delta}$ into Algorithm 1 in Ref. [96] to estimate $\langle \mathbf{W}^\top \text{diag}(\mathbf{v}) \mathbf{W} \rangle$;

Output

$$\hat{y}_{\text{test}} = \hat{y}^{(0)} + \mu_1 \frac{\langle \mathbf{v}^\top \mathbf{W} \rangle \mathbf{x}_{\text{test}}}{\sqrt{kd}} + \frac{\mu_2}{2} \frac{1}{d\sqrt{k}} \text{Tr}[(\mathbf{x}_{\text{test}} \mathbf{x}_{\text{test}}^\top - \mathbb{I}) \langle \mathbf{W}^\top \text{diag}(\mathbf{v}) \mathbf{W} \rangle]. \quad (\text{B75})$$

where $\nu = \mathbb{E}_{z \sim \mathcal{N}(0,1)} \sigma(z)^2$. The above identifies again a GLM with signal $\mathbf{S}_1^0 = \mathbf{W}^\top \mathbf{v} / \sqrt{k}$ and Gaussian sensing vectors $\mathbf{x}_{\mu}$. Define q_1 as the limit of $\mathbf{S}_1^a \mathbf{S}_1^b / d$ where $\mathbf{S}_1^a, \mathbf{S}_1^b$ are drawn independently from the posterior. With $k \rightarrow \infty$, the signal converges in law to a standard Gaussian vector. Using known results on GLMs with Gaussian signal [100], we obtain the following equations characterizing q_1 :

$$q_1 = \frac{\hat{q}_1}{\hat{q}_1 + 1}, \quad \hat{q}_1 = \frac{\alpha_1}{1 + \Delta_1 - q_1}, \quad \text{where} \\ \Delta_1 = \frac{\Delta + \nu - \mu_0^2 - \mu_1^2}{\mu_1^2}.$$

In the quadratic data regime, as $\alpha_1 = n/d$ goes to infinity, the overlap q_1 converges to 1, and the first term in Eq. (B70) is learned with vanishing error.

Moreover, since $\mathbf{S}_1^0$ is asymptotically Gaussian, the linear problem (B76) is equivalent to denoising the Gaussian vector $(\mathbf{v}^\top \mathbf{W}^0 \mathbf{x}_{\mu} / \sqrt{kd})_{\mu=0}^n$ whose covariance is known as a function of $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$. This leads to the following simple MMSE estimator for $\mathbf{S}_1^0$:

$$\langle \mathbf{S}_1^0 \rangle = \frac{1}{\sqrt{d\Delta_1}} \left(\mathbf{I} + \frac{1}{d\Delta_1} \mathbf{X} \mathbf{X}^\top \right)^{-1} \mathbf{X} \mathbf{y}, \quad (\text{B77})$$

where $\mathbf{y} = (y_1, \dots, y_n)$. Note that the derivation of this estimator does not assume the Gaussianity of $\mathbf{x}_{\mu}$.

Remark 10. The same argument can be easily generalized for general P_{out} , leading to the following equivalent GLM in the universal $\mathcal{Q}^* \equiv 0$ phase of the quadratic data regime:

$$y_{\mu}^{\text{GLM}} \sim \tilde{P}_{\text{out}}(\cdot | \text{Tr}[\mathbf{M}_{\mu} \tilde{\mathbf{S}}_2^0]), \quad \text{where} \\ \tilde{P}_{\text{out}}(y|x) := \mathbb{E}_{z \sim \mathcal{N}(0,1)} P_{\text{out}}\left(y \left| \frac{\mu_2}{2} x + z \sqrt{g(1)} \right.\right), \quad (\text{B78})$$

and $\mathbf{M}_{\mu}$ are independent GOE sensing matrices.

Remark 11. One can show that the system of equations in Eq. (B43) with $\mathcal{Q}(\mathbf{v})$ all set to 0 (and consequently, $\tau = 0$) can be mapped onto the fixed point of the state evolution equations (92) and (94) of the GAMP-RIE in Ref. [96] up to changes of variables. This confirms that when such a system has a unique solution, which is the case in all our tests, the GAMP-RIE asymptotically matches our universal solution. Assuming the validity of the aforementioned effective GLM, a potential improvement for discrete weights could come from a generalization of GAMP which, in the denoising step, would correctly exploit the discrete prior over inner weights rather than using the RIE (which is prior independent). However, the results of Ref. [171] suggest that optimally denoising matrices with discrete entries is hard, and the RIE is the best efficient procedure to do so. Consequently, we tend to believe that improving GAMP-RIE in the case of discrete weights is out of reach without strong side information about the teacher, or exploiting non-polynomial-time algorithms (see Appendix B 5).

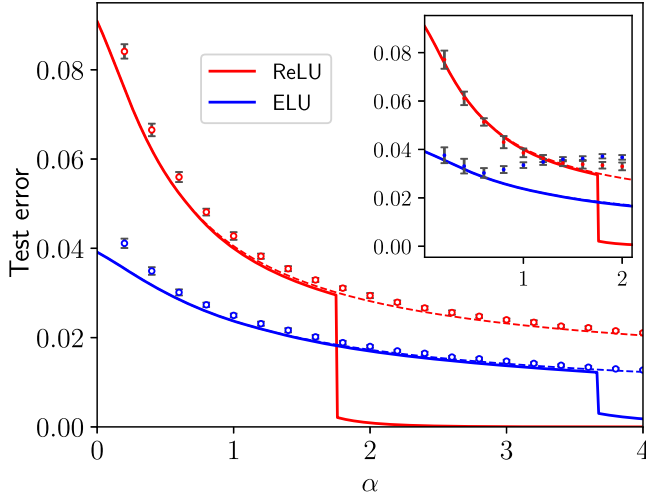


FIG. 24. Theoretical prediction (solid curves) of the Bayes-optimal mean-square generalization error for binary inner weights and ReLU, eLU activations, with $\gamma = 0.5$, $d = 150$, Gaussian label noise with $\Delta = 0.1$, and fixed readouts $\mathbf{v} = \mathbf{1}$. Dashed lines are obtained from the solution of the fixed-point equations (B41) with all $Q(\mathbf{v}) = 0$. Circles are the test error of GAMP-RIE [96] extended to generic activation. The MCMC points initialized uninformatively (inset) are obtained using Eq. (A29) to account for lack of equilibration due to glassiness, which prevents using Eq. (A31) (see Remark 7). Even in the possibly glassy region, the GAMP-RIE attains the universal branch performance. Data for GAMP-RIE and MCMC are averaged over 16 data instances, with error bars representing 1 standard deviation over instances. GAMP-RIE's performance follows the universal theoretical curve even in the α regime where MCMC sampling experiences a computationally hard phase with worse performance, and in particular, after α_{sp} .

5. Algorithmic complexity of finding the specialization solution for $L = 1$

We now provide empirical evidence concerning the computational complexity to attain specialization, namely, to have one of the $Q(\mathbf{v}) > 0$, or equivalently, to beat the universal performance [$Q(\mathbf{v}) = 0$ for all $\mathbf{v} \in V$] in terms of generalization error. We tested two algorithms that can find it in affordable computational time: ADAM with optimized batch size for every dimension tested (the learning rate is automatically tuned), and HMC, both trying to infer a two-layer teacher network with Gaussian inner weights. Both algorithms were tested with readout weights frozen to the teacher ones. We will later on discuss the case of learnable readouts.

a. ADAM

We focus on $\text{ReLU}(x)$ activation, with $\gamma = 0.5$, Gaussian output channel with low label noise ($\Delta = 10^{-4}$) and $\alpha = 5.0 > \alpha_{\text{sp}}$ ($= 0.22, 0.12, 0.02$ for homogeneous, Rademacher, and Gaussian readouts, respectively; thus, we are deep in the specialization phase in all the cases we

report), so that the specialization solution exhibits a very low generalization error. We test the learned model at each gradient update measuring the generalization error with a moving average of ten steps to smoothen the curves. Let us define ϵ^{uni} as the generalization error associated with the overlap $Q \equiv 0$, then fixing a threshold $\epsilon^{\text{opt}} < \bar{\epsilon} < \epsilon^{\text{uni}}$, we define $\bar{t}(d)$ the time (in gradient updates) needed for the algorithm to cross the threshold for the first time. We optimize over different batch sizes B_p as follows: We define them as $B_p = \lfloor n/2^p \rfloor$, $p = 2, 3, \dots, \lfloor \log_2(n) \rfloor - 1$. Then for each batch size, the student network is trained until the moving average of the test loss drops below $\bar{\epsilon}$ and thus outperforms the universal solution. We have checked that in such a scenario, the student ultimately gets close to the performance of the specialization solution. The batch size that requires the fewest gradient updates is selected. We used the ADAM routine implemented in PYTORCH.

We test different distributions for the readout weights (kept fixed to $\mathbf{v}$ during training of the inner weights). We report all the values of $\bar{t}(d)$ in Fig. 25 and 26 for various dimensions d at fixed (α, γ) , providing an exponential fit $\bar{t}(d) = \exp(ad + b)$ (left panel) and a power-law fit $\bar{t}(d) = ad^b$ (right panel). We report the χ^2 test for the fits in Table II. We observe that for homogeneous and Rademacher readouts, the exponential fit is more compatible with the experiments, while for Gaussian readouts the comparison is inconclusive.

In Fig. 27, we report the test loss of ADAM as a function of the gradient updates used for training, for various dimensions and choice of the readout distribution (as before, the readouts are not learned but fixed to the teacher's). Here, we fix a batch size for simplicity. For both cases of homogeneous ($\mathbf{v} = \mathbf{1}$) and Rademacher readouts (left and center panels), the model experiences plateaux in performance increasing with the system size, in accordance with the observation of exponential complexity we reported above. The plateaux happen at values of the test loss comparable to twice the value for the Bayes error predicted by the universal branch of the theory (remember the relationship between Gibbs and Bayes errors reported in Appendix A 5). The curves are smoother for the case of Gaussian readouts.

b. Hamiltonian Monte Carlo

The experiment is performed for the polynomial activation $\sigma_3 = \text{He}_2/\sqrt{2} + \text{He}_3/6$ with parameters $\Delta = 0.1$ for the Gaussian noise in the linear readout $\gamma = 0.5$ and $\alpha = 1.0 > \alpha_{\text{sp}}$ ($= 0.26, 0.30, 0.02$ for homogeneous, Rademacher, and Gaussian readouts, respectively). Our HMC consists of 4000 iterations for homogeneous readouts, or 2000 iterations for Rademacher and Gaussian readouts. Each iteration is adaptive (with initial step size of 0.01) and uses ten leapfrog steps. Instead of measuring the Gibbs error, whose relationship with ϵ^{opt} holds only at

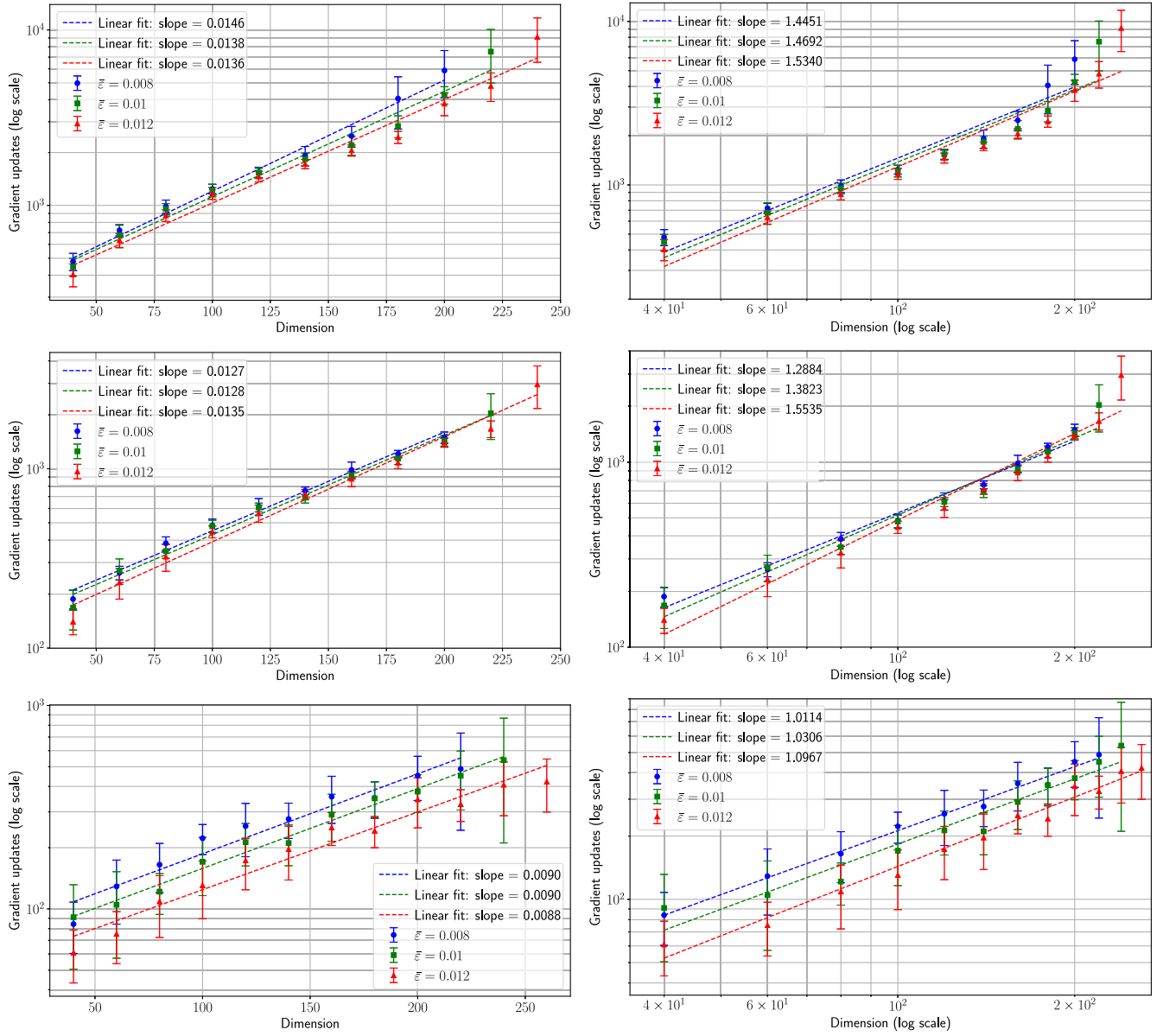


FIG. 25. Semilog (left) and log-log (right) plots of the number of gradient updates needed to achieve a test loss below the threshold $\bar{\epsilon} < \epsilon^{\text{uni}}$. Student network trained with ADAM with optimized batch size for each point. The dataset was generated from a teacher network with $\text{ReLU}(x)$ activation and parameters $\Delta = 10^{-4}$ for the Gaussian noise variance of the linear readout, $\gamma = 0.5$ and $\alpha = 5.0$ for which $\epsilon^{\text{opt}} - \Delta = 1.115 \times 10^{-5}$. Points are obtained averaging over ten teacher and data instances with error bars representing the standard deviation. Each row corresponds to a different distribution of the readouts kept fixed during training. Top: homogeneous readouts, for which the error of the universal branch is $\epsilon^{\text{uni}} - \Delta = 1.217 \times 10^{-2}$. Center: Rademacher readouts, for which $\epsilon^{\text{uni}} - \Delta = 1.218 \times 10^{-2}$. Bottom: Gaussian readouts, for which $\epsilon^{\text{uni}} - \Delta = 1.210 \times 10^{-2}$. The quality of the fits can be read from Table II.

TABLE II. χ^2 test for exponential and power-law fits for the time needed by ADAM to reach the thresholds $\bar{\epsilon}$, for various priors on the readouts. Fits are displayed in Fig. 25. Smaller values of χ^2 (in bold, for given threshold and readouts) indicate a better compatibility with the hypothesis.

Readouts	$\bar{\epsilon} =$	χ^2 exponential fit			χ^2 power-law fit		
		0.008	0.010	0.012	0.008	0.010	0.012
Homogeneous		5.57	9.00	21.1	32.3	26.5	61.1
Rademacher		4.51	6.84	12.7	12.0	17.4	16.0
Uniform $[-\sqrt{3}, \sqrt{3}]$		5.08	1.44	4.21	8.26	8.57	3.82
Gaussian		2.66	0.76	3.02	0.55	2.31	1.36

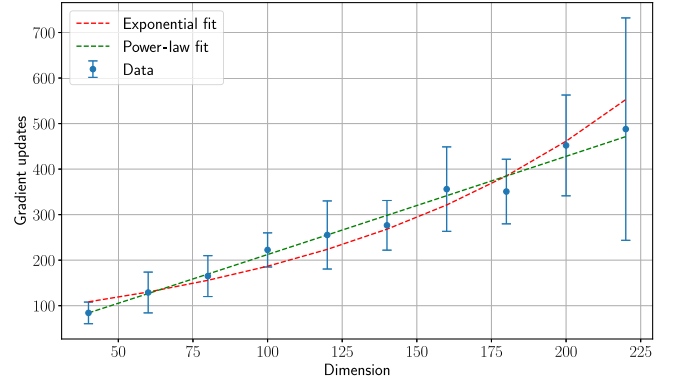
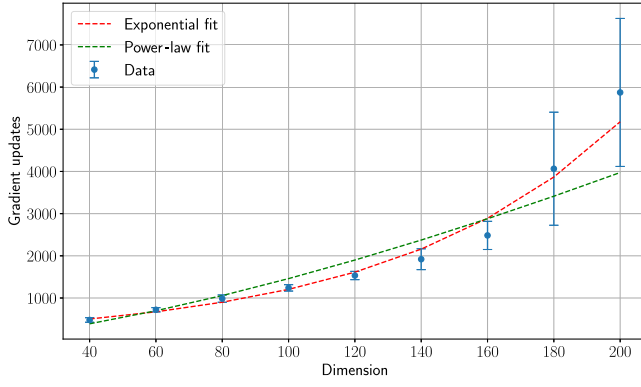


FIG. 26. Same as in Fig. 25, but in linear scale for better visualization, for homogeneous readouts (left) and Gaussian readouts (right), with threshold $\bar{\epsilon} = 0.008$.

equilibrium (see the last remark in Appendix A 5), we measured the teacher-student R_2 overlap which is meaningful at any HMC step and is informative about the learning. For a fixed threshold $\bar{R}_2$ and dimension d , we measure $\bar{t}(d)$ as the number of HMC iterations needed for the R_2 overlap between the HMC sample (obtained from uninformative initialization) and the teacher weights $\mathbf{W}^0$ to cross the threshold. This criterion is again enough to assess that the student outperforms the universal solution.

As before, we test homogeneous, Rademacher and Gaussian readouts, getting to the same conclusions: While for homogeneous and Rademacher readouts exponential time is more compatible with the observations, the

experiments remain inconclusive for Gaussian readouts (see Fig. 29). We report in Fig. 28 the values of the overlap R_2 measured along the HMC runs for different dimensions. Note that, with HMC steps, all R_2 curves saturate to a value that is off by approximately 1% with respect to that predicted by our theory for the selected values of α , γ , and Δ . Whether this is a finite-size effect or an effect not taken into account by the current theory is an interesting question requiring further investigation; see Appendix. B 2 b for possible directions.

c. Learnable readouts

As discussed in the main text, the static properties of the model remain unchanged whether the readout weights are

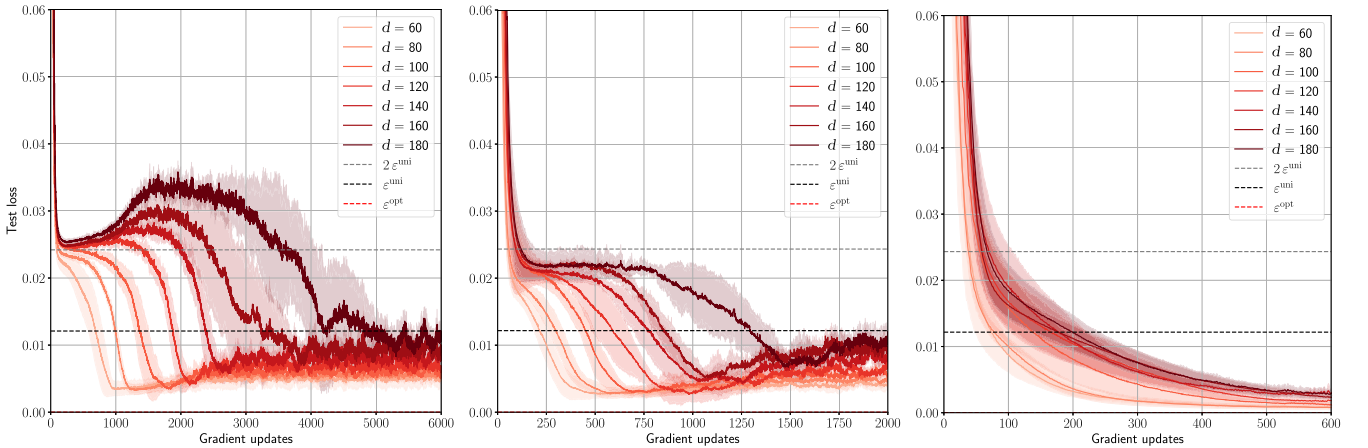


FIG. 27. Trajectories of the generalization error of neural networks trained with ADAM at fixed batch size $B = \lfloor n/4 \rfloor$, learning rate 0.05, for $\text{ReLU}(x)$ activation with parameters $\Delta = 10^{-4}$ for the linear readout, $\gamma = 0.5$ and $\alpha = 5.0 > \alpha_{\text{sp}}$ ($= 0.22, 0.12, 0.02$ for homogeneous, Rademacher, and Gaussian readouts, respectively). The error ϵ^{uni} is the mean-square generalization error associated with the universal solution with overlap $\mathcal{Q} \equiv 0$. Left: homogeneous readouts. Center: Rademacher readouts. Right: Gaussian readouts. Readouts are kept fixed (and equal to the teacher's) in all cases during training. Points on the solid lines are obtained by averaging over five teacher and data instances, and shaded regions around them correspond to 1 standard deviation.

quenched to the teacher values or learned during training. However, the dynamics can differ when the readouts are learnable. We verified that, for ADAM, the results regarding hardness are qualitatively unchanged when the readouts are learned. Although ADAM can achieve a lower test error in this case, the convergence time required to reach this solution increases substantially.

Specifically, when the readouts are fixed, specialization occurs after approximately 10^4 – 10^5 gradient updates for homogeneous priors (see Fig. 27, left). In contrast, as shown in Fig. 8, learning the readouts increases the number of gradient updates required for specialization by at least an order of magnitude.

For HMC, which is constrained to sample according to the prior over both the inner and readout weights, the behavior is essentially identical whether the readouts are fixed or learnable. The reasoning in Remark 3 therefore applies equally to HMC, as it is a posterior sampler.

6. A potential route for a proof for $L=1$

Here we provide an argument for a potential proof of our results based the adaptive interpolation technique introduced in Ref. [204], used as in Ref. [100]. In order to make the model more amenable to rigorous treatment, we select activation functions with $\mu_0 = \mu_1 = \mu_2 = 0$ and all-ones readouts $\mathbf{v} = \mathbf{1}$. While assumptions on μ_0, μ_1 are not that restrictive, $\mu_2 = 0$ is what induces the main simplifications as it erases from the analyses the role of the overlaps R_2 . It is also useful to write the replica potential we are targeting, which in the hypotheses listed above and with a little abuse of notation reads

$$f_{\text{RS}}^{(1)}(\mathcal{Q}, \hat{\mathcal{Q}}) = \frac{\gamma}{\alpha} \psi_{P_w}(\hat{\mathcal{Q}}) + \phi_{P_{\text{out}}}(g(\mathcal{Q}), g(1)) - \frac{\gamma}{2\alpha} \mathcal{Q} \hat{\mathcal{Q}}. \quad (\text{B79})$$

Apart from the presence of $g(\mathcal{Q})$ inside $\phi_{P_{\text{out}}}$ the above formula looks like the one of a standard generalized linear model [100]; we shall thus use a similar formalism. Let us define an interpolating model

$$s_{t\mu} := \frac{\sqrt{1-t}}{\sqrt{k}} \sum_{i=1}^k \varphi\left(\frac{1}{\sqrt{d}} \mathbf{W}_i^* \mathbf{x}_\mu\right) + \sqrt{G(t)} V_\mu + \sqrt{g(1)t - G(t)} U_\mu^*, \quad (\text{B80})$$

where $t \in [0, 1]$, $G(t)$ is a non-negative interpolating function, and $V_\mu, U_\mu^* \sim \mathcal{N}(0, 1)$. Note that U_μ^* is a starred variable; thus, we consider it as learnable. The labels for this interpolating model are then given by

$$Y_{t\mu} \sim P_{\text{out}}(\cdot | S_{t\mu}). \quad (\text{B81})$$

To complete the interpolation, we also need another Gaussian observation channel about $\mathbf{W}^*$:

$$Y_{t;ij}^G = \sqrt{R(t)} W_{ij}^* + Z_{ij}, \quad (\text{B82})$$

where $Z_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$. The interpolating functions $G(t)$, $R(t)$ have to be appropriately chosen later, in order to make some remainders vanish. The only requirement for now is that $G(0) = R(0) = 0$. Define $u_y(x) := \ln P_{\text{out}}(y|x)$, and denote for brevity $\mathbf{Y}_t = (Y_{t\mu})_{\mu \leq n}$, $\mathbf{Y}_t^G = (Y_{t;ij}^G)_{i \leq k, j \leq d}$, $\mathbf{U} = (U_\mu)_{\mu \leq n}$, $\mathbf{V} = (V_\mu)_{\mu \leq n}$, and

$$s_{t\mu} := \frac{\sqrt{1-t}}{\sqrt{k}} \sum_{i=1}^k \varphi\left(\frac{1}{\sqrt{d}} \mathbf{W}_i^* \mathbf{x}_\mu\right) + \sqrt{G(t)} V_\mu + \sqrt{g(1)t - G(t)} U_\mu. \quad (\text{B83})$$

Then, the above interpolating model induces a Hamiltonian that reads

$$\begin{aligned} -\mathcal{H}_t(\mathbf{X}, \mathbf{Y}_t, \mathbf{Y}_t^G, \mathbf{V}, \mathbf{U}, \mathbf{W}) \\ = \sum_{\mu=1}^n u_{Y_{t\mu}}(s_{t\mu}) - \frac{1}{2} \|\mathbf{Y}_t^G - \sqrt{R(t)} \mathbf{W}\|^2, \end{aligned} \quad (\text{B84})$$

and the corresponding quenched free entropy is

$$f_n(t) = \frac{1}{n} \mathbb{E}_{(t)} \ln \mathcal{Z}_t \quad (\text{B85})$$

with

$$\begin{aligned} \mathbb{E}_{(t)}[\cdot] &= \mathbb{E}_{\mathbf{X}, \mathbf{V}, \mathbf{U}^*, \mathbf{W}^*} \int d\mathbf{Y}_t d\mathbf{Y}_t^G e^{-\mathcal{H}_t(\mathbf{X}, \mathbf{Y}_t, \mathbf{Y}_t^G, \mathbf{V}, \mathbf{U}^*, \mathbf{W}^*)} [\cdot], \\ \mathcal{Z}_t &= \int dP_w(\mathbf{W}) D\mathbf{U} e^{-\mathcal{H}_t(\mathbf{X}, \mathbf{Y}_t, \mathbf{Y}_t^G, \mathbf{V}, \mathbf{U}, \mathbf{W})}. \end{aligned} \quad (\text{B86})$$

At the extrema of interpolation, we have

$$\begin{aligned} f_n(0) &= f_n - \frac{\gamma}{2\alpha}, \\ f_n(1) &= \frac{\gamma}{\alpha} \psi_{P_w}(R(1)) + \phi_{P_{\text{out}}}(G(1), g(1)) - \frac{\gamma}{2\alpha} [1 + R(1)]. \end{aligned} \quad (\text{B87})$$

The role of the interpolation is that of decoupling continuously the quenched disorder in the $\mathbf{x}_\mu$'s from the weights $\mathbf{W}^*$ and simultaneously linearizing the nonlinearity φ . We shall now convince the reader that there are choices for the functions G, R that produce Eq. (B79). To begin with, we need to control the t derivative of the interpolating free entropy:

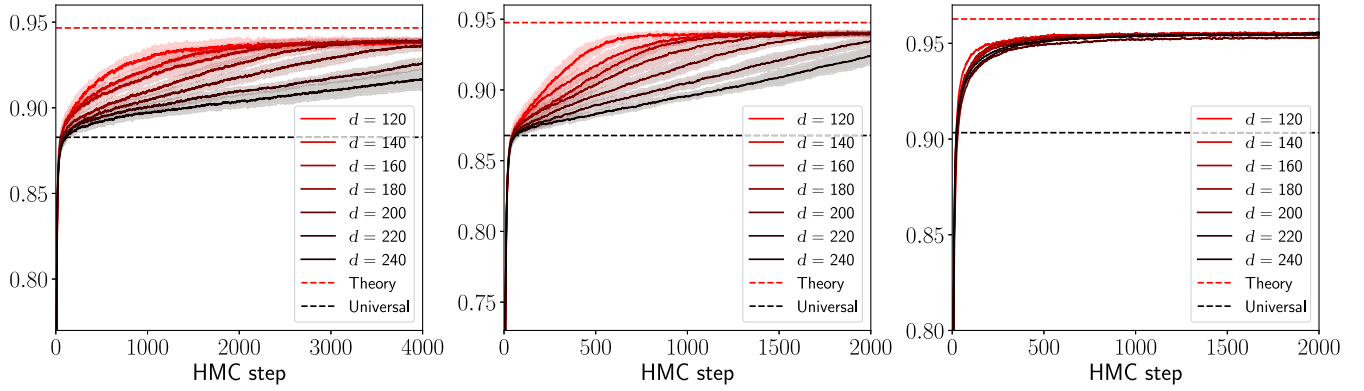


FIG. 28. Trajectories of the overlap R_2 in HMC runs initialized uninformatively for the polynomial activation $\sigma_3 = \text{He}_2/\sqrt{2} + \text{He}_3/6$ with parameters $\Delta = 0.1$ for the linear readout, $\gamma = 0.5$ and $\alpha = 1.0$. Left: homogeneous readouts. Center: Rademacher readouts. Right: Gaussian readouts. Points on the solid lines are obtained by averaging over ten teacher and data instances, and shaded regions around them correspond to 1 standard deviation. Notice that the y axes are limited for better visualization. For the left and center plots, any threshold (horizontal line in the plot) between the prediction of the $\mathcal{Q} \equiv 0$ branch of the theory (black dashed line) and its prediction for the R_2^{sp} (red dashed line, obtained with informative initialization) crosses the curves in points $\bar{t}(d)$ more compatible with an exponential fit (see Fig. 29 and Table III, where these fits are reported and χ^2 tested). For the cases of homogeneous and Rademacher readouts, the value of the overlap at which the dynamics slows down (predicted by the $\mathcal{Q} \equiv 0$ branch) is in quantitative agreement with the theoretical predictions (lower dashed line). The theory is instead off by approximately 1% for the values R_2 at which the runs ultimately converge.

$$\frac{d}{dt}f_n(t) = -\frac{1}{n}\mathbb{E}_{(t)}\frac{d}{dt}\mathcal{H}_t(\mathbf{X}, \mathbf{Y}_t, \mathbf{Y}_t^G, \mathbf{V}, \mathbf{U}^*, \mathbf{W}^*) \ln \mathcal{Z}_t - \frac{1}{n}\mathbb{E}_{(t)}\left\langle \frac{d}{dt}\mathcal{H}_t(\mathbf{X}, \mathbf{Y}_t, \mathbf{Y}_t^G, \mathbf{V}, \mathbf{U}, \mathbf{W}) \right\rangle_t, \quad (\text{B88})$$

where $\langle \cdot \rangle_t$ is the Gibbs measure associated with the Hamiltonian $\mathcal{H}_t$.

Let us focus first on the second term on the rhs. Using the Nishimori identities on it, we readily get

$$\text{II} = \frac{1}{n}\mathbb{E}_{(t)}\frac{d}{dt}\mathcal{H}_t(\mathbf{X}, \mathbf{Y}_t, \mathbf{Y}_t^G, \mathbf{V}, \mathbf{U}^*, \mathbf{W}^*) = \frac{1}{n}\mathbb{E}_{(t)}\left[\sum_{\mu=1}^n u'_{Y_{\mu}}(S_{t\mu})\dot{S}_{t\mu} + \frac{\dot{R}(t)}{2\sqrt{R(t)}}\sum_{i,j}^{k,d} W_{ij}^*(Y_{t,ij}^G - \sqrt{R(t)}W_{ij}^*)\right]. \quad (\text{B89})$$

Considering that $Y_{t,ij}^G - \sqrt{R(t)}W_{ij}^* = Z_{ij}$, which is independent of W_{ij}^* , and that

$$\int dY_{t\mu} P_{\text{out}}(Y_{t\mu}|S_{t\mu}) u'_{Y_{t\mu}}(S_{t\mu}) = \partial_x \int dy P_{\text{out}}(y|x)|_{x=S_{t\mu}} = 0,$$

we have that $\text{II} = 0$ identically. Concerning the first term instead,

$$\text{I} = \mathbb{E}_{(t)} \ln \mathcal{Z}_t \frac{1}{n} \left[\sum_{\mu=1}^n u'_{Y_{\mu}}(S_{t\mu})\dot{S}_{t\mu} + \sum_{i,j}^{k,d} (Y_{t,ij}^G - \sqrt{R(t)}W_{ij}^*) \frac{\dot{R}(t)}{2\sqrt{R(t)}} W_{ij}^* \right]. \quad (\text{B90})$$

We start by the first term on the rhs which requires the most care. After replacing $\dot{S}_{t\mu}$ with its expression, we have

$$\mathbb{E}_{(t)} \ln \mathcal{Z}_t \frac{1}{2n} \sum_{\mu=1}^n u'_{Y_{\mu}}(S_{t\mu}) \left[-\frac{1}{\sqrt{(1-t)}} \lambda_{\mu}^* + \frac{\dot{G}(t)}{\sqrt{G(t)}} V_{\mu} + \frac{g(1) - \dot{G}(t)}{\sqrt{g(1)t - G(t)}} U_{\mu}^* \right], \quad (\text{B91})$$

where $\lambda_{\mu}^* = (1/\sqrt{k}) \sum_{i \leq k} \varphi[(1/\sqrt{d}) \mathbf{W}_i^{*\top} \mathbf{x}_{\mu}]$. In the GLM, one aims at integrating $\mathbf{x}_{\mu}$ by parts, but here it is not possible due to the presence of the nonlinearity. Hence, we need to make a Gaussian assumption to treat it.

Assumption 1. Defining $\lambda_{\mu} = (1/\sqrt{k}) \sum_{i \leq k} \varphi[(1/\sqrt{d}) \mathbf{W}_i^{\top} \mathbf{x}_{\mu}]$ and λ_{μ}^* as above, the following holds under the randomness of $\mathbf{x}_{\mu}$:

$$\lambda_\mu, \lambda_\mu^{*\text{iid}} \sim \mathcal{N}(\mathbf{0}, C(\mathbf{W}, \mathbf{W}^*)), \quad C_{ab} \equiv C_{ab}(\mathbf{W}, \mathbf{W}^*) = \sum_{\ell \geq 3} \frac{\mu_\ell^2}{\ell!} \frac{1}{k} \sum_{i,j=1}^k \left(\frac{\mathbf{W}^a \mathbf{W}^{b\top}}{d} \right)_{ij}^{\circ \ell}, \quad (\text{B92})$$

where $a, b = \cdot, *,$ and $\cdot$ labels a posterior sample.

When integrating by parts, one needs to take into account that the probability weights hidden in $\mathbb{E}_{(t)}$ also depend on $S_{t\mu}$. Bearing this in mind, integration by parts of $\lambda_\mu^*, U_\mu^*, V_\mu$ in Eq. (B91) yields

$$\begin{aligned} & -\mathbb{E}_{(t)} \ln \mathcal{Z}_t \frac{1}{2n} \sum_{\mu=1}^n [u''_{Y_{t\mu}}(S_{t\mu}) + [u'_{Y_{t\mu}}(S_{t\mu})]^2] C_{**} - \frac{1}{2} \mathbb{E}_{(t)} \left\langle \frac{1}{n} \sum_{\mu=1}^n u'_{Y_{t\mu}}(S_{t\mu}) u'_{Y_{t\mu}}(s_{t\mu}) C_{.*} \right\rangle_t \\ & + \mathbb{E}_{(t)} \ln \mathcal{Z}_t \frac{1}{2n} \sum_{\mu=1}^n [u''_{Y_{t\mu}}(S_{t\mu}) + [u'_{Y_{t\mu}}(S_{t\mu})]^2] \dot{G}(t) + \frac{\dot{G}(t)}{2} \mathbb{E}_{(t)} \left\langle \frac{1}{n} \sum_{\mu=1}^n u'_{Y_{t\mu}}(S_{t\mu}) u'_{Y_{t\mu}}(s_{t\mu}) \right\rangle_t \\ & + \mathbb{E}_{(t)} \ln \mathcal{Z}_t \frac{1}{2n} \sum_{\mu=1}^n [u''_{Y_{t\mu}}(S_{t\mu}) + [u'_{Y_{t\mu}}(S_{t\mu})]^2] [g(1) - \dot{G}(t)]. \end{aligned} \quad (\text{B93})$$

Considering that $u''_y(x) + [u'_y(x)]^2 = \partial_x^2 P_{\text{out}}(y|x)/P_{\text{out}}(y|x)$, by gathering all the previous terms together, Eq. (B91) becomes

$$\mathbb{E}_{(t)} \ln \mathcal{Z}_t \frac{1}{2n} \sum_{\mu=1}^n \frac{\partial_x^2 P_{\text{out}}(Y_{t\mu}|S_{t\mu})}{P_{\text{out}}(Y_{t\mu}|S_{t\mu})} [g(1) - C_{**}] + \frac{1}{2} \mathbb{E}_{(t)} \left\langle \frac{1}{n} \sum_{\mu=1}^n u'_{Y_{t\mu}}(S_{t\mu}) u'_{Y_{t\mu}}(s_{t\mu}) [\dot{G}(t) - C_{.*}] \right\rangle_t. \quad (\text{B94})$$

Concerning instead the second term on the rhs of I , it can be simplified via a standard integration by parts of the Gaussian random variable $Y_{t;ij}^G - \sqrt{R(t)} W_{ij}^* = Z_{ij}$. We thus report just the final result for I :

$$\begin{aligned} I &= \mathbb{E}_{(t)} \ln \mathcal{Z}_t \frac{1}{2n} \sum_{\mu=1}^n \frac{\partial_x^2 P_{\text{out}}(Y_{t\mu}|S_{t\mu})}{P_{\text{out}}(Y_{t\mu}|S_{t\mu})} [g(1) - C_{**}] + \frac{1}{2} \mathbb{E}_{(t)} \left\langle \frac{1}{n} \sum_{\mu=1}^n u'_{Y_{t\mu}}(S_{t\mu}) u'_{Y_{t\mu}}(s_{t\mu}) [\dot{G}(t) - C_{.*}] \right\rangle_t \\ & - \frac{\gamma}{2\alpha} \dot{R}(t) [1 - \mathcal{Q}(t)] - \frac{\gamma}{2\alpha} \dot{R}(t) \left[\mathcal{Q}(t) - \frac{1}{kd} \mathbb{E}_{(t)} \langle \text{Tr} \mathbf{W}^* \mathbf{W}^\top \rangle_t \right], \end{aligned} \quad (\text{B95})$$

where we have added and subtracted the term containing $\mathcal{Q}(t)$ in the second line. $\mathcal{Q}(t)$ here is an arbitrary non-negative function for the moment.

By a simple application of the fundamental theorem of integral calculus, we have thus proved the following *sum rule*: *Proposition 1.* (Sum rule) Assume the GEP in Eq. (B92) holds. Then,

$$f_n = f_n(0) + \frac{\gamma}{2\alpha} = f_n(1) + \frac{\gamma}{2\alpha} - \int_0^1 I(t) dt = \frac{\gamma}{\alpha} \psi_{P_w}(R(1)) + \phi_{P_{\text{out}}}(G(1), g(1)) - \frac{\gamma}{2\alpha} R(1) - \int_0^1 I(t) dt, \quad (\text{B96})$$

where we have stressed the t dependence of I .

It is now time to make some choices about our interpolating functions. First, we link $\mathcal{Q}(t)$ and $G(t)$ as follows: $G(t) = \int_0^t g(\mathcal{Q}(s)) ds$. Then, out of convenience we call $\hat{\mathcal{Q}}(t) = \dot{R}(t)$. Second, we need the following:

Assumption 2. The equation

$$\mathcal{Q}(t) = \frac{1}{kd} \mathbb{E}_{(t)} \langle \text{Tr} [\mathbf{W}^* \mathbf{W}^\top] \rangle_t \quad (\text{B97})$$

has a solution. Furthermore, assume

$$\mathbb{E}_{(t)} (g(1) - C_{**})^2 = o_n(1), \quad (\text{B98})$$

$$\mathbb{E}_{(t)} \langle [g(\mathcal{Q}(t)) - C_{.*}]^2 \rangle_t = o_n(1), \quad (\text{B99})$$

uniformly in t .

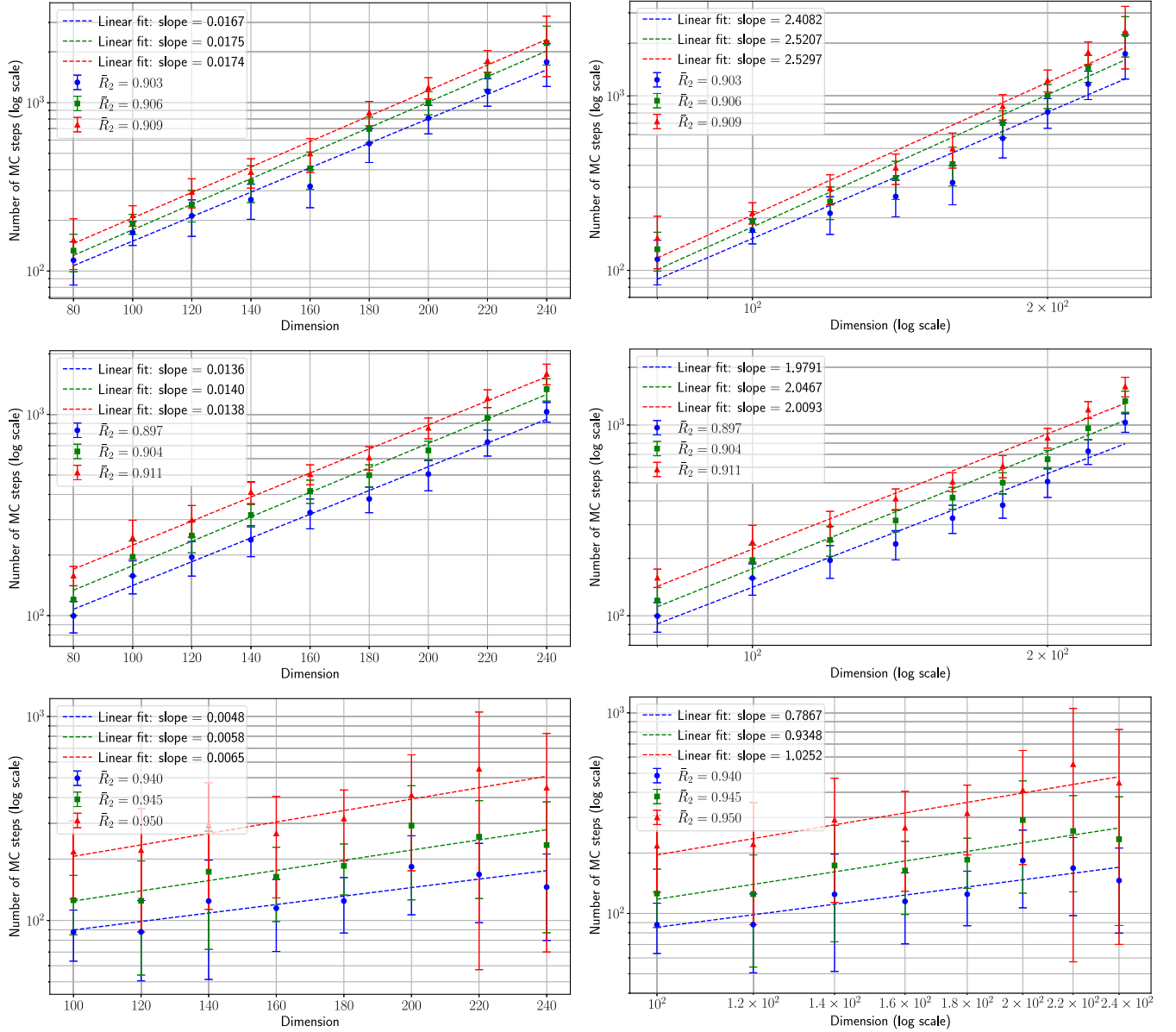


FIG. 29. Semilog (left) and log-log (right) plots of the number of Hamiltonian Monte Carlo steps needed to achieve an overlap $\bar{R}_2 > R_2^{\text{uni}}$ that certifies the universal solution is outperformed. The dataset was generated from a teacher with polynomial activation $\sigma_3 = \text{He}_2/\sqrt{2} + \text{He}_3/6$ and parameters $\Delta = 0.1$ for the linear readout, $\gamma = 0.5$ and $\alpha = 1.0 > \alpha_{\text{sp}}$ ($= 0.26, 0.30, 0.02$ for homogeneous, Rademacher, and Gaussian readouts, respectively). Student weights are sampled using HMC (initialized uninformatively) with 4000 iterations for homogeneous readouts (top row, for which $R_2^{\text{uni}} = 0.883$), or 2000 iterations for Rademacher (center row, with $R_2^{\text{uni}} = 0.868$) and Gaussian readouts (bottom row, for which $R_2^{\text{uni}} = 0.903$). Each iteration is adaptive (with initial step size of 0.01) and uses ten leapfrog steps. $R_2^{\text{sp}} = 0.941, 0.948, 0.963$ in the three cases. The readouts are kept fixed during training. Points are obtained averaging over ten teacher and data instances with error bars representing the standard deviation.

TABLE III. χ^2 test for exponential and power-law fits for the time needed by Hamiltonian Monte Carlo to reach the thresholds $\bar{R}_2$, for various priors on the readouts. For a given row, we report three values of the χ^2 test per hypothesis, corresponding with the thresholds $\bar{R}_2$ on the left, in the order given. Fits are displayed in Fig. 29. Smaller values of χ^2 (in bold, for given threshold and readouts) indicate a better compatibility with the hypothesis.

Readouts		χ^2 exponential fit			χ^2 power-law fit		
Homogeneous	$(\bar{R}_2 \in \{0.903, 0.906, 0.909\})$	2.22	1.47	1.14	8.01	7.25	6.35
Rademacher	$(\bar{R}_2 \in \{0.897, 0.904, 0.911\})$	1.88	2.12	1.70	8.10	7.70	8.57
Gaussian	$(\bar{R}_2 \in \{0.940, 0.945, 0.950\})$	0.66	0.44	0.26	0.62	0.53	0.39

Assumption (B97) is not trivial, as $G(t)$ and $Q(t)$ are now linked, and thus $Q(t)$ appears on both sides of the above equality (it is contained in the definition of $\mathbb{E}_{(t)}\langle \cdot \rangle_t$). A formal proof of Eq. (B98) is at reach with standard concentration of measure tools, whereas Eq. (B99) requires much more care. The proofs of Eqs. (B98) and (B99) are both left for future work. Both of them are enforcing that

$$\frac{1}{k} \sum_{i,j=1}^k \left(\frac{\mathbf{W}^a \mathbf{W}^{b\top}}{d} \right)_{ij}^{\circ \ell} \approx \frac{1}{k} \sum_{i=1}^k \left(\frac{\mathbf{W}_i^a \cdot \mathbf{W}_i^b}{d} \right)^\ell$$

under the $\mathbb{E}_{(t)}\langle \cdot \rangle_t$ measure for $\ell \geq 3$. Since there is permutation symmetry over the readout neurons when $\mathbf{v} = \mathbf{1}$, all the terms in the above equation are basically assumed to concentrate onto Eq. (B97).

Under Assumption 2, the sum rule reads

$$f_n = \frac{\gamma}{\alpha} \psi_{P_w} \left(\int_0^1 \hat{Q}(t) dt \right) + \phi_{P_{\text{out}}} \left(\int_0^1 g(Q(t)) dt, g(1) \right) - \frac{\gamma}{2\alpha} \int_0^1 \hat{Q}(t) Q(t) dt + o_n(1). \quad (\text{B100})$$

Observe that $\phi_{P_{\text{out}}}$, g , and ψ_{P_w} are all nondecreasing and convex functions of their arguments. Furthermore, the above estimate holds for any function $\hat{Q}(t)$, whereas $Q(t)$ has been fixed as the solution of Eq. (B97). We start by choosing $\hat{Q}(t) = \hat{Q} = \text{const}$, and we use Jensen's inequality on g :

$$f_n \geq \frac{\gamma}{\alpha} \psi_{P_w}(\hat{Q}) + \phi_{P_{\text{out}}}(g(Q), g(1)) - \frac{\gamma}{2\alpha} \hat{Q} Q + o_n(1) \geq \inf_Q f_{\text{RS}}^{(1)}(Q, \hat{Q}) + o_n(1) \quad (\text{B101})$$

with $Q = \int_0^1 Q(t) dt$. The above is then made tight by taking the sup.

The converse bound is instead obtained by using Jensen's inequality to take the $\int_0^1 dt$ out of the ψ functions, which yields

$$f_n \leq \int_0^1 f_{\text{RS}}^{(1)}(Q(t), \hat{Q}(t)) dt + o_n(1). \quad (\text{B102})$$

In order to make the bound tight, we now choose $\hat{Q}(t)$ as the solution of the optimization $\inf_{\hat{Q}} f_{\text{RS}}^{(1)}(Q(t), \hat{Q})$, which is unique by convexity. Therefore,

$$f_n \leq \int_0^1 \inf_{\hat{Q}} f_{\text{RS}}^{(1)}(Q(t), \hat{Q}) dt + o_n(1) \leq \sup_Q \inf_{\hat{Q}} f_{\text{RS}}^{(1)}(Q, \hat{Q}) + o_n(1). \quad (\text{B103})$$

To summarize,

$$\sup_Q \inf_{\hat{Q}} f_{\text{RS}}^{(1)}(Q, \hat{Q}) + o_n(1) \leq f_n \leq \sup_Q \inf_{\hat{Q}} f_{\text{RS}}^{(1)}(Q, \hat{Q}) + o_n(1). \quad (\text{B104})$$

Strictly speaking, the two variational principles on the two sides of these bounds are different, but for sure they have the same stationary points. Under suitable conditions (see, for instance, Corollary 7 in the Supplemental Material of Ref. [100]), they actually yield the same value, which would close the proof.

7. Generalization errors for learnable readouts

In the main text, we prove that from an information-theoretical point of view, having the readouts learnable or fixed to those of the teacher does not alter the problem. In particular, the generalization errors predicted by our theory should be the same for both cases.

This is indeed numerically verified. In Figs. 30 and 31, we show that HMC posterior samples, in the case of learnable readouts, yield the generalization error predicted by our theory.

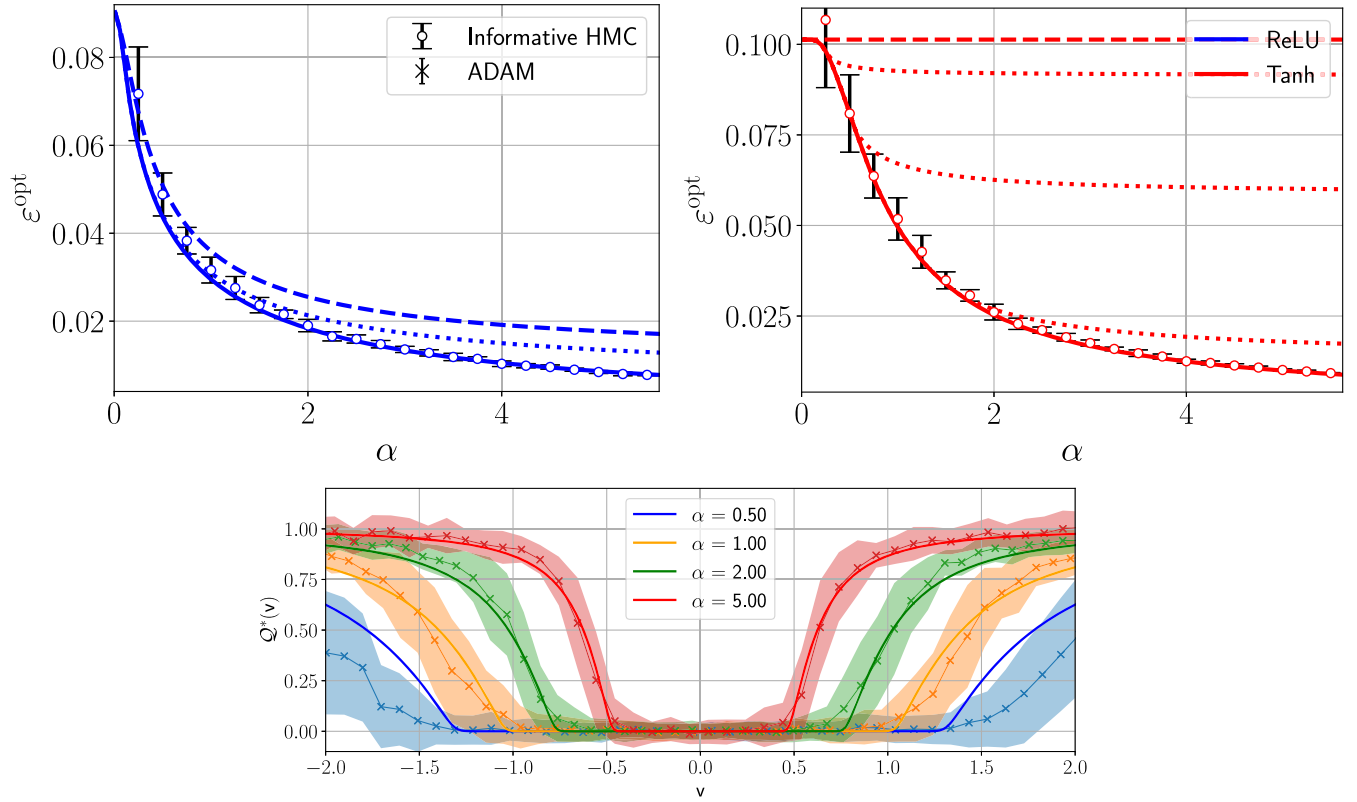


FIG. 30. Top: theoretical prediction (solid curves) of the specialization mean-square generalization error ϵ^{sp} for Gaussian inner weights with ReLU(x) activation (blue curves) and tanh($2x$) activation (red curves), $d = 200$, $\gamma = 0.5$, with linear readout and Gaussian label noise of variance $\Delta = 0.1$. The dashed lines show the theoretical prediction associated with the universal branch of our theory ϵ^{uni} . Markers are for the Hamiltonian Monte Carlo with informative initialization on the target (empty circles). Each point is averaged over 12 teacher and training-set instances; error bars denote the sample standard deviation across instances. Generalization errors are numerically evaluated as half Gibbs errors, assuming the validity of Nishimori identities on metastable states as in the main text [see also Appendix A 5 and Eq. (A33)]. The empirical average over test inputs is computed from 10^5 i.i.d. test samples. Bottom: theoretical prediction (solid curves) of the overlap for different sampling ratios α for Gaussian inner weights, $\sigma(x) = \text{ReLU}(x)$, $d = 200$, $\gamma = 0.5$, $\Delta = 0.1$, and Gaussian readouts. The shaded curves were obtained from informed HMC. Using a single posterior sample $\mathbf{W}$ (per α and data instance), $Q(v)$ is evaluated numerically by dividing the interval $[-2, 2]$ into bins and then computing the value of the overlap associated with the readout value in that bin. Each point has been averaged over 100 instances of the training set, and shaded regions around them correspond to 1 standard deviation. Note: in these plots, the readouts are learnable and drawn from a Gaussian prior $P_v = \mathcal{N}(0, 1)$.

APPENDIX C: DEEP MLP

1. Details of the replica calculation

Let us take a multilayer perceptron with $L = O(1)$ hidden layers. In order to be more general let us take at each layer a different activation function

$$\sigma^{(l)}(z) = \sum_{\ell \neq 0,2} \frac{\mu_{\ell}^{(l)}}{\ell!} \text{He}_{\ell}(z),$$

so that the MLP we will study is

$$\mathcal{F}_{\theta^0}^{(L)}(\mathbf{x}) := \frac{\mathbf{v}^{0\top}}{\sqrt{k^L}} \sigma^{(L)} \left\{ \frac{\mathbf{W}^{0(L)}}{\sqrt{k_{L-1}}} \sigma^{(L-1)} \left[\frac{\mathbf{W}^{0(L-1)}}{\sqrt{k_{L-2}}} \dots \sigma^{(1)} \left(\frac{\mathbf{W}^{0(1)}}{\sqrt{k_0}} \mathbf{x} \right) \dots \right] \right\},$$

with θ^0 denoting the whole collection of the teacher's parameters. To make the equations lighter, we take centered and normalized activations $\mathbb{E}_z \sigma^{(l)}(z)^2 = 1$, $\mu_0^{(l)} = 0$, and we allow for different priors over the inner weights $\mathbf{W}^{(l)} \sim P_{W_l}$.

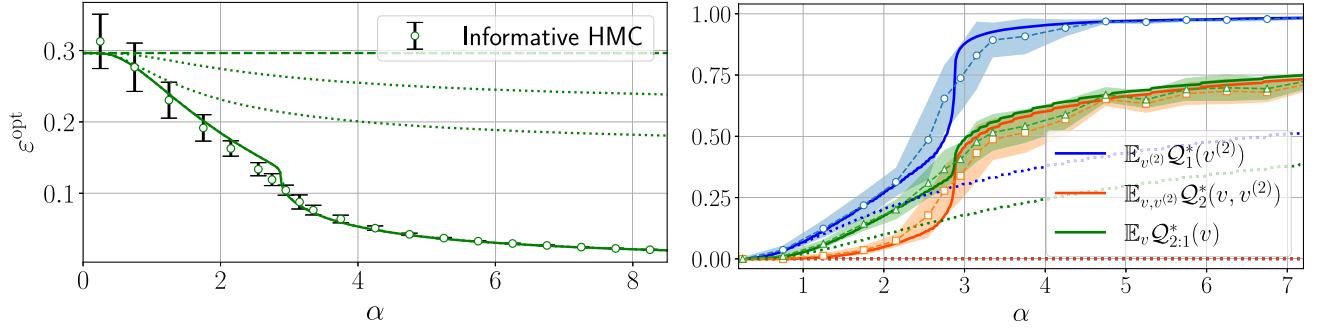


FIG. 31. Left: theoretical prediction (green solid curve) of the Bayes-optimal mean-square generalization error for $L = 2$ with Gaussian inner weights, $\sigma(x) = \tanh(2x)/\sigma_{\tanh}$, $d = 200$, $\gamma_1 = \gamma_2 = 0.5$, $\Delta = 0.2$, and different P_v laws. The dashed and dotted lines have the same meaning as in Fig. 13. Points are obtained with Hamiltonian Monte Carlo with informative initialization. Each point has been averaged over 20 instances of the data, with error bars representing 1 standard deviation. The generalization error is computed empirically from 10^4 i.i.d. test samples. Right: solid and dotted curves represent, respectively, the mean of different overlaps at equilibrium and in metastable specialized states, as a function of the sampling ratio α for $L = 2$ with Gaussian inner weights, $\sigma(x) = \tanh(2x)/\sigma_{\tanh}$, $d = 200$, $\gamma_1 = \gamma_2 = 0.5$, $\Delta = 0.2$. The shaded curves were obtained from informed HMC. Each point has been averaged over 20 instances of the training set, with 1 standard deviation depicted. Note: in these plots, the readouts are learnable and drawn from a Gaussian prior $P_v = \mathcal{N}(0, 1)$.

Importantly, we assume $\mu_2^{(l)} = 0$. Treating terms associated with the second Hermite coefficient would require spherical integration and a measure relaxation analogous to the approach used in the shallow case. We leave this extension for future work.

Let us define the pre- and postactivations, respectively, as

$$\left\{ \mathbf{h}^{(l)a} := \frac{1}{\sqrt{k_{l-1}}} \mathbf{W}^{(l)a} \mathbf{x}^{(l-1)a}, \quad \mathbf{x}^{(l)a} := \sigma^{(l)}(\mathbf{h}^{(l)a}) \right\}_{a=0}^s,$$

where $\mathbf{x}^{a(0)} := \mathbf{x}$, $\forall a = 0, \dots, s$ represents the input data, and

$$\left\{ \lambda^a(\boldsymbol{\theta}^a) := \frac{1}{\sqrt{k_L}} \mathbf{v}^{0\top} \sigma^{(L)}(\mathbf{h}^{(L)a}) \right\}_{a=0}^s$$

are the replicated readouts.

As for the shallow case, the key assumption is the joint Gaussianity of $\{\lambda^a(\boldsymbol{\theta}^a)\}_{0 \leq a \leq s}$ under the common input randomness $\mathbf{x}$. Since they are centered (recall $\mu_0^{(l)} = 0$), in order to characterize their distribution it suffices to evaluate their covariance, which for analogy with the shallow case shall be denoted as

$$K^{ab} := \mathbb{E}_{\mathbf{x}} \lambda^a(\boldsymbol{\theta}^a) \lambda^b(\boldsymbol{\theta}^b) = \frac{1}{k_L} \sum_{i,j=1}^{k_L} v_i^0 v_j^0 \mathbb{E}_{\mathbf{x}} \sigma^{(L)}(h_i^{(L)a}) \sigma^{(L)}(h_j^{(L)b}). \quad (\text{C1})$$

To further simplify the above expectation, we need to use recursively Mehler's formula (see Appendix A 2) from the first preactivation on, as follows. To begin with, define

$$\Omega_{i_1 j_1}^{(1)ab} := \mathbb{E}_{\mathbf{x}} h_{i_1}^{(1)a} h_{j_1}^{(1)b} = \mathbb{E}_{\mathbf{x}} \left(\frac{\mathbf{W}_{i_1}^{(1)a} \cdot \mathbf{x}}{\sqrt{d}} \right) \left(\frac{\mathbf{W}_{j_1}^{(1)b} \cdot \mathbf{x}}{\sqrt{d}} \right) = \frac{\mathbf{W}_{i_1}^a \cdot \mathbf{W}_{j_1}^b}{d}. \quad (\text{C2})$$

This allows us to compute the covariance of the second-layer preactivations under the same randomness:

$$\Omega_{i_2 j_2}^{(2)ab} := \mathbb{E}_{\mathbf{x}} h_{i_2}^{(2)a} h_{j_2}^{(2)b} = \mathbb{E}_{\mathbf{x}} \left(\frac{\mathbf{W}_{i_2}^{(2)a} \cdot \mathbf{x}^{(1)a}}{\sqrt{k_1}} \right) \left(\frac{\mathbf{W}_{j_2}^{(2)b} \cdot \mathbf{x}^{(1)b}}{\sqrt{k_1}} \right). \quad (\text{C3})$$

The expectation is resolved once one computes the covariance of the first-layer postactivations by means of Mehler's formula, yielding

$$\Omega_{i_2 j_2}^{(2)ab} = \frac{1}{k_1} \mathbf{W}_{i_2}^{(2)a\top} [(\mu_1^{(1)})^2 \Omega^{(1)ab} + g^{(1)}(\Omega^{(1)ab})] \mathbf{W}_{j_2}^{(2)b} \quad (\text{C4})$$

where the function $g^{(1)}$ defined as in Eq. (B6) is applied elementwise to the matrix argument. From this moment on, we assume that preactivations are Gaussian at each layer under the common randomness $\mathbf{x}$, as they are always expressed in terms of rescaled sums. Under this assumption, we can thus infer a generic recursion for the preactivation covariances:

$$\Omega_{i_{l+1} j_{l+1}}^{(l+1)ab} := \mathbb{E}_{\mathbf{x}} h_{i_{l+1}}^{(l+1)a} h_{j_{l+1}}^{(l+1)b} = \frac{1}{k_l} \mathbf{W}_{i_{l+1}}^{(l+1)a\top} [(\mu_1^{(l)})^2 \Omega^{(l)ab} + g^{(l)}(\Omega^{(l)ab})] \mathbf{W}_{j_{l+1}}^{(l+1)b}, \quad (\text{C5})$$

which naturally leads to the covariance we actually need for K^{ab} , i.e., $\Omega_{ij}^{(L)ab} = \mathbb{E}_{\mathbf{x}} h_i^{(L)a} h_j^{(L)b}$:

$$K^{ab} = \frac{(\mu_1^{(L)})^2}{k_L} \mathbf{v}^{0\top} [\Omega^{(L)ab} + g^{(L)}(\Omega^{(L)ab})] \mathbf{v}^0. \quad (\text{C6})$$

Let us define the following set of vectors and matrices for future convenience:

$$\begin{aligned} \mathbf{W}^{(l': l)a} &= \frac{1}{\sqrt{k_{l'-1} k_{l'-2} \dots k_l}} \mathbf{W}^{(l')a} \mathbf{W}^{(l'-1)a} \dots \mathbf{W}^{(l)a}, \\ \mathbf{v}^{(l)a} &= \frac{1}{\sqrt{k_l k_{l+1} \dots k_L}} \mathbf{W}^{(l)a\top} \mathbf{W}^{(l+1)a\top} \dots \mathbf{W}^{(L)a\top} \mathbf{v}^0. \end{aligned} \quad (\text{C7})$$

They will emerge from the computation due to the linear term in the Hermite expansion of the activation functions. The objects in (C7) represent effective readout vectors and weight matrices that the student can learn independently from the actual weights and readouts. With this notation, the postactivation covariance reads (recall $k_0 = d$)

$$\begin{aligned} K^{ab} &= \frac{(\mu_1^{(L)} \mu_1^{(L-1)} \dots \mu_1^{(1)})^2}{k_0} \mathbf{v}^{a(1)\top} \mathbf{v}^{b(1)} + \frac{(\mu_1^{(L-1)} \mu_1^{(L-2)} \dots \mu_1^{(1)})^2}{k_1} \mathbf{v}^{a(2)\top} g^{(1)}(\Omega^{ab(1)}) \mathbf{v}^{b(2)} \\ &+ \dots + \frac{(\mu_1^{(1)})^2}{k_{L-1}} \mathbf{v}^{a(L)\top} g^{(L-1)}(\Omega^{ab(L-1)}) \mathbf{v}^{b(L)} + \frac{1}{k_L} \mathbf{v}^{0\top} g^{(L)}(\Omega^{ab(L)}) \mathbf{v}^0. \end{aligned} \quad (\text{C8})$$

As in Eq. (B3), we will assume that for all $l = 1, \dots, L$,

$$\Omega_{i_l i_l}^{ab(l)} = O(1), \quad \Omega_{i_l j_l}^{ab(l)} = O\left(\frac{1}{\sqrt{k_{l-1}}}\right) \quad \text{for } i_l \neq j_l. \quad (\text{C9})$$

Therefore, only the diagonal elements of the matrix $g^{(l)}(\Omega^{ab(l)})$ will contribute in the thermodynamic limit.

Note that the overlap $(1/k_0) \mathbf{v}^{(1)a\top} \mathbf{v}^{(1)b}$ is analogous to the OP described by Ref. [109] in the deep setting; $\mathbf{v}^{(1)0}$ is the only feature of the target function that is learnable in the $n \propto d$ regime, which we will consider as known; namely, all the overlaps between $\mathbf{v}^{(1)a}$'s are set to 1. Analogously to what happens in the shallow case, the components of the other $\mathbf{v}^{(l)a}$'s enter trivially the energetic term. Specifically, only those components of $\mathbf{v}^{(l)a}$ that are perfectly reconstructible by the student would enter the energy, namely, those for which the associated $\Omega_{i_l i_l}^{(l)ab} = O(1)$ and not smaller. Hence, without loss of generality one can assume that all the $\mathbf{v}^{(l)a}$ are set to $\mathbf{v}^{(l)0}$ as in the shallow case. This allows us to considerably simplify the equations and leads to

$$\begin{aligned}
K^{ab} = & (\mu_1^{(L)} \mu_1^{(L-1)} \dots \mu_1^{(1)})^2 + \frac{(\mu_1^{(L-1)} \mu_1^{(L-2)} \dots \mu_1^{(1)})^2}{k_1} \sum_{i_1}^{k_1} (v_{i_1}^{(2)})^2 g^{(1)} \left(\Omega_{i_1 i_1}^{ab(1)} \right) \\
& + \dots + \frac{(\mu_1^{(1)})^2}{k_{L-1}} \sum_{i_{L-1}}^{k_{L-1}} (v_{i_{L-1}}^{(L)})^2 g^{(L-1)} \left(\Omega_{i_{L-1} i_{L-1}}^{ab(L-1)} \right) + \frac{1}{k_L} \sum_{i_L}^{k_L} (v_{i_L}^0)^2 g^{(L)} \left(\Omega_{i_L i_L}^{ab(L)} \right)
\end{aligned} \quad (C10)$$

with, additionally,

$$\Omega_{i_{l+1} i_{l+1}}^{ab(l+1)} \approx \frac{(\mu_1^{(l)})^2}{k_l} \mathbf{W}_{i_{l+1}}^{a(l+1)\top} \Omega^{ab(l)} \mathbf{W}_{i_{l+1}}^{b(l+1)} + \frac{1}{k_l} \sum_{i_l}^{k_l} W_{i_{l+1} i_l}^{(l+1)a} W_{i_{l+1} i_l}^{(l+1)b} g^{(l)} \left(\Omega_{i_l i_l}^{(l)ab} \right). \quad (C11)$$

Here is the difference with the shallow case: The way the replicas $\mathbf{W}^{(l+1)a}$ align with one another may depend not only on the “left” indices i_{l+1} in the formula above, but can be affected also by the reconstruction performance from previous layers encoded in $\Omega_{i_l i_l}^{(l)ab}$. The values of $\Omega_{i_l i_l}^{(l)ab}$ are themselves driven by the values of the associated $v_{i_l}^{(l+1)0}$ that appear coupled to them in the energy through K^{ab} as above. Hence, we can choose to label the values of $\Omega_{i_l i_l}^{(l)ab}$ through those of $v_{i_l}^{(l+1)0}$. We denote these values as $\mathbf{v}^{(l+1)}$, collected in the sets $\mathbf{V}^{(l+1)}$, and $\mathcal{I}_{\mathbf{V}^{(l+1)}} = \{i_l | v_{i_l}^{(l+1)0} = \mathbf{v}^{(l+1)}\}$.

This brings us to defining the following overlaps:

$$\mathcal{Q}_l^{ab}(i_l, \mathbf{v}^{(l)}) = \frac{1}{|\mathcal{I}_{\mathbf{V}^{(l)}}|} \sum_{i_{l-1} \in \mathcal{I}_{\mathbf{V}^{(l)}}} W_{i_l i_{l-1}}^{(l)a} W_{i_l i_{l-1}}^{(l)b}, \quad (C12)$$

$$\mathcal{Q}_{l':l}^{ab}(i_{l'}, \mathbf{v}^{(l)}) = \frac{1}{|\mathcal{I}_{\mathbf{V}^{(l)}}|} \sum_{i_{l-1} \in \mathcal{I}_{\mathbf{V}^{(l)}}} W_{i_{l'} i_{l-1}}^{(l':l)a} W_{i_{l'} i_{l-1}}^{(l':l)b}. \quad (C13)$$

Using Eq. (C5), the diagonal elements of each $\Omega^{(l)ab}$ in terms of those overlaps read

$$\Omega_{i_{l+1} i_{l+1}}^{ab(l+1)} \approx \frac{(\mu_1^{(l)})^2}{k_l} \mathbf{W}_{i_{l+1}}^{a(l+1)\top} \Omega^{ab(l)} \mathbf{W}_{i_{l+1}}^{b(l+1)} + \sum_{\mathbf{v}^{(l+1)} \in \mathbf{V}^{(l+1)}} \frac{|\mathcal{I}_{\mathbf{V}^{(l+1)}}|}{k_l} \mathcal{Q}_{l+1}^{ab}(i_{l+1}, \mathbf{v}^{(l+1)}) \frac{1}{|\mathcal{I}_{\mathbf{V}^{(l+1)}}|} \sum_{i_l \in \mathcal{I}_{\mathbf{V}^{(l+1)}}} g^{(l)} \left(\Omega_{i_l i_l}^{ab(l)} \right). \quad (C14)$$

This is the first step of the recursion; in order to express everything in terms of the overlaps $\mathcal{Q}$, we need to express the first term as well in terms of diagonal elements, keeping in mind that $\Omega_{i_1 i_1}^{(1)ab} = \mathbf{W}_{i_1}^{(1)a\top} I_d \mathbf{W}_{i_1}^{(1)b} / d =: \mathcal{Q}_1^{ab}(i_1)$. In other words, the first overlap is not labeled by any other index, as the $\mathbf{W}^{(1)}$'s here do not sandwich any matrix other than the identity. Hence, no inhomogeneity can arise. An analogous reasoning holds for $\mathcal{Q}_{l+1:l}^{ab}(i_{l+1})$.

At a generic step of the recursion, we have

$$\begin{aligned}
\Omega_{i_{l+1} i_{l+1}}^{ab(l+1)} \approx & (\mu_1^{(l)} \dots \mu_1^{(1)})^2 \mathcal{Q}_{l+1:l}^{ab}(i_{l+1}) + (\mu_1^{(l)} \dots \mu_1^{(2)})^2 \sum_{\mathbf{v}^{(2)} \in \mathbf{V}^{(2)}} \frac{|\mathcal{I}_{\mathbf{V}^{(2)}}|}{k_1} \mathcal{Q}_{l+1:2}^{ab}(i_{l+1}, \mathbf{v}^{(2)}) \frac{1}{|\mathcal{I}_{\mathbf{V}^{(2)}}|} \sum_{i_l \in \mathcal{I}_{\mathbf{V}^{(2)}}} g^{(1)} \left(\Omega_{i_l i_l}^{(1)ab} \right) \\
& + \dots + (\mu_1^{(l)})^2 \sum_{\mathbf{v}^{(l)} \in \mathbf{V}^{(l)}} \frac{|\mathcal{I}_{\mathbf{V}^{(l)}}|}{k_{l-1}} \mathcal{Q}_{l+1:l}^{ab}(i_{l+1}, \mathbf{v}^{(l)}) \frac{1}{|\mathcal{I}_{\mathbf{V}^{(l)}}|} \sum_{i_{l-1} \in \mathcal{I}_{\mathbf{V}^{(l)}}} g^{(l-1)} \left(\Omega_{i_{l-1} i_{l-1}}^{(l-1)ab} \right) \\
& + \sum_{\mathbf{v}^{(l+1)} \in \mathbf{V}^{(l+1)}} \frac{|\mathcal{I}_{\mathbf{V}^{(l+1)}}|}{k_l} \mathcal{Q}_{l+1}^{ab}(i_{l+1}, \mathbf{v}^{(l+1)}) \frac{1}{|\mathcal{I}_{\mathbf{V}^{(l+1)}}|} \sum_{i_l \in \mathcal{I}_{\mathbf{V}^{(l+1)}}} g^{(l)} \left(\Omega_{i_l i_l}^{(l)ab} \right),
\end{aligned} \quad (C15)$$

where $\Omega_{i_1 i_1}^{ab(1)} = \mathcal{Q}_1^{ab}(i_1)$. This defines the full recursion, which allows one to compute the covariance K^{ab} only in terms of the abovementioned overlaps.

As the derivation may become very cumbersome, we specialize to the $L = 2$ setting in the following section.

a. Two hidden layers $L=2$

In the case of two-hidden-layer networks, the equation for the covariance is

$$K^{ab} = (\mu_1^{(2)} \mu_1^{(1)})^2 + \frac{(\mu_1^{(2)})^2}{k_1} \sum_{i_1}^{k_1} (v_{i_1}^{(2)})^2 g^{(1)} \left(\mathcal{Q}_1^{ab}(i_1) \right) + \frac{1}{k_2} \sum_{i_2}^{k_2} (v_{i_2}^{(2)})^2 g^{(2)} \left((\mu_1^{(1)})^2 \mathcal{Q}_{2:1}^{ab}(i_2) + \sum_{\mathbf{v}^{(2)} \in \mathcal{V}^{(2)}} \frac{|\mathcal{I}_{\mathbf{v}^{(2)}}|}{k_1} \mathcal{Q}_2^{ab}(i_2, \mathbf{v}^{(2)}) \frac{1}{|\mathcal{I}_{\mathbf{v}^{(2)}}|} \sum_{i_1 \in \mathcal{I}_{\mathbf{v}^{(2)}}} g^{(1)}(\mathcal{Q}_1^{ab}(i_1)) \right). \quad (\text{C16})$$

Importantly, the index i_1 is linked only to the vector $\mathbf{v}^{(2)}$ and the overlap $\mathcal{Q}_1^{ab}$, while the index i_2 to the vector $\mathbf{v}^{(0)}$ and the overlaps $\mathcal{Q}_2^{ab}$ and $\mathcal{Q}_{2:1}^{ab}$. We can relabel the values of $\mathcal{Q}_1$ with those of $\mathbf{v}^{(2)}$: $\mathcal{Q}_1(\mathbf{v}^{(2)}) = \mathcal{Q}_1(i_1)$ for all $i_1 \in \mathcal{I}_{\mathbf{v}^{(2)}}$. An analogous relabeling can be carried out for $\mathcal{Q}_2(i_2, \mathbf{v}^{(2)})$ in the index i_2 based on the values of $v_{i_2}^{(2)}$. After the relabeling, one could also redefine the overlaps through partial traces in order to mimic the notation from the main text, as follows:

$$\mathcal{Q}_l^{ab}(\mathbf{v}^{(l+1)}, \mathbf{v}^{(l)}) = \frac{1}{|\mathcal{I}_{\mathbf{v}^{(l+1)}}| |\mathcal{I}_{\mathbf{v}^{(l)}}|} \sum_{i_l \in \mathcal{I}_{\mathbf{v}^{(l+1)}}} \sum_{i_{l-1} \in \mathcal{I}_{\mathbf{v}^{(l)}}} W_{i_l i_{l-1}}^{(l)a} W_{i_l i_{l-1}}^{(l)b}, \quad (\text{C17})$$

$$\mathcal{Q}_{l':l}^{ab}(\mathbf{v}^{(l'+1)}, \mathbf{v}^{(l)}) = \frac{1}{|\mathcal{I}_{\mathbf{v}^{(l'+1)}}| |\mathcal{I}_{\mathbf{v}^{(l)}}|} \sum_{i_{l'} \in \mathcal{I}_{\mathbf{v}^{(l'+1)}}} \sum_{i_{l-1} \in \mathcal{I}_{\mathbf{v}^{(l)}}} W_{i_{l'} i_{l-1}}^{(l':l)a} W_{i_{l'} i_{l-1}}^{(l':l)b}. \quad (\text{C18})$$

Consider also that $(|\mathcal{I}_{\mathbf{v}^{(2)}}|/k_1) \rightarrow P_{\mathbf{v}^{(2)}}(\mathbf{v}^{(2)})$ and $(|\mathcal{I}_{\mathbf{v}}|/k_2) \rightarrow P_{\mathbf{v}}(\mathbf{v})$. This allows us to recast K^{ab} in the asymptotic limit as

$$K^{ab} = (\mu_1^{(2)} \mu_1^{(1)})^2 + (\mu_1^{(2)})^2 \mathbb{E}_{v^{(2)} \sim P_{v^{(2)}}} (v^{(2)})^2 g^{(1)}(\mathcal{Q}_1^{ab}(v^{(2)})) + \mathbb{E}_{v \sim P_v} (v)^2 g^{(2)} \left((\mu_1^{(1)})^2 \mathcal{Q}_{2:1}^{ab}(v) + \mathbb{E}_{v^{(2)} \sim P_{v^{(2)}}} \mathcal{Q}_2^{ab}(v, v^{(2)}) g^{(1)}(\mathcal{Q}_1^{ab}(v^{(2)})) \right). \quad (\text{C19})$$

Notice that as soon as the covariance of the postactivation is written in terms of overlaps, this fully determines the energetic part appearing in the free entropy. Indeed, in the RS *Ansatz*, $\mathbf{K}$ appears as in Eq. (B11), where

$$K^{(2)}(\bar{\mathcal{Q}}) = (\mu_1^{(2)} \mu_1^{(1)})^2 + (\mu_1^{(2)})^2 \mathbb{E}_{v^{(2)} \sim P_{v^{(2)}}} (v^{(2)})^2 g^{(1)}(\mathcal{Q}_1(v^{(2)})) + \mathbb{E}_{v \sim P_v} (v)^2 g^{(2)} \left((\mu_1^{(1)})^2 \mathcal{Q}_{2:1}(v) + \mathbb{E}_{v^{(2)} \sim P_{v^{(2)}}} \mathcal{Q}_2(v, v^{(2)}) g^{(1)}(\mathcal{Q}_1(v^{(2)})) \right), \quad K_d = 1. \quad (\text{C20})$$

while $\rho_K = K_d$ and $m_K = K$ using the Nishimori identities. Therefore, the energetic term will be equal to Eq. (B15), with $K = K^{(2)}(\bar{\mathcal{Q}})$ and $K_d = 1$ as defined in the last equations.

Let us now discuss the entropic contribution associated with the OPs defined above. This can be written as

$$\begin{aligned}
e^{F_S} &= \int d\mathbf{W}^{a(2:1)} \int \prod_{a=0}^s \prod_{l=1}^2 dP_{W_l}(\mathbf{W}^{a(l)}) \delta\left(\mathbf{W}^{a(2:1)} - \frac{\mathbf{W}^{a(2)} \mathbf{W}^{a(1)}}{\sqrt{k_1}}\right) \\
&\times \prod_{a \leq b}^{0,s} \prod_{\mathbf{v} \in \mathbf{V}} \prod_{\mathbf{v}^{(2)} \in \mathbf{V}^{(2)}} \delta\left(|\mathcal{I}_{\mathbf{v}}| |\mathcal{I}_{\mathbf{v}^{(2)}}| \mathcal{Q}_2^{ab}(\mathbf{v}, \mathbf{v}^{(2)}) - \sum_{i_2 \in \mathcal{I}_{\mathbf{v}}} \sum_{i_1 \in \mathcal{I}_{\mathbf{v}^{(2)}}} W_{i_2 i_1}^{(2)a} W_{i_2 i_1}^{(2)b}\right) \\
&\times \prod_{a \leq b}^{0,s} \prod_{\mathbf{v}^{(2)} \in \mathbf{V}^{(2)}} \delta\left(d |\mathcal{I}_{\mathbf{v}^{(2)}}| \mathcal{Q}_1^{ab}(\mathbf{v}^{(2)}) - \sum_{i_1 \in \mathcal{I}_{\mathbf{v}^{(2)}}} \mathbf{W}_{i_1}^{a(1)\top} \mathbf{W}_{i_1}^{b(1)}\right) \\
&\times \prod_{a \leq b}^{0,s} \prod_{\mathbf{v} \in \mathbf{V}} \delta\left(d |\mathcal{I}_{\mathbf{v}}| \mathcal{Q}_{2:1}^{ab}(\mathbf{v}) - \sum_{i_2 \in \mathcal{I}_{\mathbf{v}}} \mathbf{W}_{i_2}^{a(2:1)\top} \mathbf{W}_{i_2}^{b(2:1)}\right). \tag{C21}
\end{aligned}$$

Besides the labeling of $\mathcal{Q}_2$ in terms of two indices $\mathbf{v}, \mathbf{v}^{(2)}$, another important difference with respect to the shallow case is the presence of the overlap $\mathcal{Q}_{2:1}(\mathbf{v})$ between the replicated matrices $\mathbf{W}^{(2:1)a} = \mathbf{W}^{a(2)} \mathbf{W}^{a(1)} / \sqrt{k_1}$. This overlap will depend on the alignment between the first- and second-layer weights. In order to encode this dependence, we use a similar relaxation of the measure as the one described in Sec. B 1 b. We thus define

$$\begin{aligned}
dP((\mathbf{W}^{(2:1)a}) | \mathcal{Q}_1, \mathcal{Q}_2) &\propto \prod_{a=0}^s d\mathbf{W}^{(2:1)a} \int \prod_{a=0}^s \prod_{l=1}^2 dP_{W_l}(\mathbf{W}^{(l)a}) \delta\left(\mathbf{W}^{(2:1)a} - \frac{\mathbf{W}^{(2)a} \mathbf{W}^{(1)a}}{\sqrt{k_1}}\right) \\
&\times \prod_{a \leq b}^{0,s} \prod_{\mathbf{v} \in \mathbf{V}} \prod_{\mathbf{v}^{(2)} \in \mathbf{V}^{(2)}} \delta\left(|\mathcal{I}_{\mathbf{v}}| |\mathcal{I}_{\mathbf{v}^{(2)}}| \mathcal{Q}_2^{ab}(\mathbf{v}, \mathbf{v}^{(2)}) - \sum_{i_2 \in \mathcal{I}_{\mathbf{v}}} \sum_{i_1 \in \mathcal{I}_{\mathbf{v}^{(2)}}} W_{i_2 i_1}^{(2)a} W_{i_2 i_1}^{(2)b}\right) \\
&\times \prod_{a \leq b}^{0,s} \prod_{\mathbf{v}^{(2)} \in \mathbf{V}^{(2)}} \delta\left(d |\mathcal{I}_{\mathbf{v}^{(2)}}| \mathcal{Q}_1^{ab}(\mathbf{v}^{(2)}) - \sum_{i_1 \in \mathcal{I}_{\mathbf{v}^{(2)}}} \mathbf{W}_{i_1}^{a(1)\top} \mathbf{W}_{i_1}^{b(1)}\right), \tag{C22}
\end{aligned}$$

where the normalization constant is implicit. The aim is now to relax this measure to the one of a product of two Ginibre matrices with a proper tilt given by the coupling between replicas. The relaxation we choose is the one with a matched second moment $[1/(d|\mathcal{I}_{\mathbf{v}}|)] \sum_{i_2 \in \mathcal{I}_{\mathbf{v}}} \mathbb{E}[\mathbf{W}_{i_2}^{(2:1)a\top} \mathbf{W}_{i_2}^{(2:1)b} | \mathcal{Q}_1, \mathcal{Q}_2]$ where $\mathbb{E}[\cdot | \mathcal{Q}_1, \mathcal{Q}_2]$ denotes the expectation with respect to the conditional measure defined above. As done in Appendix B 1 c, by rewriting Dirac deltas in Fourier form, the measure decouples and the calculation goes through, yielding asymptotically

$$\frac{1}{d|\mathcal{I}_{\mathbf{v}}|} \sum_{i_2 \in \mathcal{I}_{\mathbf{v}}} \mathbb{E}[\mathbf{W}_{i_2}^{(2:1)a\top} \mathbf{W}_{i_2}^{(2:1)b} | \mathcal{Q}_1, \mathcal{Q}_2] \approx \mathbb{E}_{\mathbf{v}^{(2)} \sim P_{\mathbf{v}^{(2)}}} \mathcal{Q}_2^{ab}(\mathbf{v}, \mathbf{v}^{(2)}) \mathcal{Q}_1^{ab}(\mathbf{v}^{(2)}). \tag{C23}$$

In order to fix this moment in our relaxation, we thus need a Lagrange multiplier for each value $\mathbf{v} \in \mathbf{V}$:

$$d\bar{P}((\mathbf{W}^{(2:1)a}) | \mathcal{Q}_1, \mathcal{Q}_2) = \prod_{\mathbf{v} \in \mathbf{V}} V(\boldsymbol{\tau}_{\mathbf{v}})^{-1} \prod_{a=0}^s d\mathcal{N}(\mathbf{U}_{\mathbf{v}}^a) d\mathcal{N}(\mathbf{V}^a) e^{\sum_{a < b, 0}^s \boldsymbol{\tau}_{\mathbf{v}}^{ab} \text{Tr} \mathbf{U}_{\mathbf{v}}^a \mathbf{V}^a (\mathbf{U}_{\mathbf{v}}^b \mathbf{V}^b)^{\top}}, \tag{C24}$$

where $\mathbf{U}_{\mathbf{v}}^a \in \mathbb{R}^{|\mathcal{I}_{\mathbf{v}}| \times k_1}$, $\mathbf{V}^a \in \mathbb{R}^{k_1 \times d}$ are matrices with i.i.d. Gaussian elements (their factorized measure being synthetically denoted by $\mathcal{N}$), $\mathbf{W}^{(2:1)a} = \mathbf{U}_{\mathbf{v}}^a \mathbf{V}^a$, and $\boldsymbol{\tau}_{\mathbf{v}} = (\tau_{\mathbf{v}}^{ab})_{a < b = 0, \dots, s}$.

With this relaxation, we have that

$$\begin{aligned}
e^{F_S} &= V_{W_1}^{k_1 d}(\mathcal{Q}_1) V_{W_2}^{k_1 d}(\mathcal{Q}_2) \int d\hat{\mathcal{Q}}_{2:1} \int \prod_{\mathbf{v} \in \mathbf{V}} V(\boldsymbol{\tau}_{\mathbf{v}})^{-1} \prod_{a=0}^s d\mathcal{N}(\mathbf{U}_{\mathbf{v}}^a) d\mathcal{N}(\mathbf{V}^a) e^{\sum_{a < b, 0}^s [\boldsymbol{\tau}_{\mathbf{v}}^{ab} + \hat{\mathcal{Q}}_{2:1}^{ab}(\mathbf{v})] \text{Tr} \mathbf{U}_{\mathbf{v}}^a \mathbf{V}^a (\mathbf{U}_{\mathbf{v}}^b \mathbf{V}^b)^{\top}} \\
&\times e^{-d \sum_{a < b, 0}^s \sum_{\mathbf{v} \in \mathbf{V}} |\mathcal{I}_{\mathbf{v}}| \hat{\mathcal{Q}}_{2:1}^{ab}(\mathbf{v}) \mathcal{Q}_{2:1}^{ab}(\mathbf{v})}. \tag{C25}
\end{aligned}$$

Where not specified, integrals over OPs and their Fourier conjugates run over all replica indices and $\mathbf{v}$ values. Standard steps as the ones in Appendix B 1, after taking the 0 replica limit, yield

$$\begin{aligned} f_{\text{RS}}^{(2)} := & \phi_{P_{\text{out}}}(K^{(2)}(\bar{\mathcal{Q}}); 1) + \frac{\gamma_1}{\alpha} \mathbb{E}_{v^{(2)} \sim P_{v^{(2)}}} \left[\psi_{P_1}(\hat{\mathcal{Q}}_1(v^{(2)})) - \frac{1}{2} \mathcal{Q}_1(v^{(2)}) \hat{\mathcal{Q}}_1(v^{(2)}) \right] \\ & + \frac{\gamma_1 \gamma_2}{\alpha} \mathbb{E}_{v \sim P_v, v^{(2)} \sim P_{v^{(2)}}} \left[\psi_{P_2}(\hat{\mathcal{Q}}_2(v, v^{(2)})) - \frac{1}{2} \mathcal{Q}_2(v, v^{(2)}) \hat{\mathcal{Q}}_2(v, v^{(2)}) \right] \\ & + \frac{\gamma_2}{\alpha} \mathbb{E}_{v \sim P_v} \left[\frac{1}{2} \hat{\mathcal{Q}}_{2:1}(v) (1 - \mathcal{Q}_{2:1}(v)) - \iota_v(\tau_v + \hat{\mathcal{Q}}_{2:1}(v)) + \iota_v(\tau_v) \right], \end{aligned} \quad (\text{C26})$$

where $\iota_v(x)$ is the MI of the following matrix denoising problem:

$$\mathbf{Y}_v(x) = \sqrt{x} \frac{\mathbf{U}_v^0 \mathbf{V}^0}{\sqrt{k_1}} + \mathbf{Z}_v \in \mathbb{R}^{|\mathcal{I}_v| \times d} \quad (\text{C27})$$

with $\mathbf{U}_v^0 \in \mathbb{R}^{|\mathcal{I}_v| \times k_1}$, $\mathbf{V}^0 \in \mathbb{R}^{k_1 \times d}$, and $\mathbf{Z}_v \in \mathbb{R}^{|\mathcal{I}_v| \times d}$ three Ginibre matrices. Furthermore, we assume $|\mathcal{I}_v|/k_2 \rightarrow P_v(\mathbf{v})$, $k_2/d \rightarrow \gamma_2$, $k_1/d \rightarrow \gamma_1$, and $n/d^2 \rightarrow \alpha$. Hence, $|\mathcal{I}_v|/d \rightarrow P_v(\mathbf{v})\gamma_2$ and

$$\iota_v(x) := \lim_{d \rightarrow \infty} \frac{x \mathbb{E} \|\mathbf{U}_v^0 \mathbf{V}^0\|^2}{2k_1 |\mathcal{I}_v| d} - \frac{1}{|\mathcal{I}_v| d} \mathbb{E} \ln \int_{\mathbb{R}^{|\mathcal{I}_v| \times k_1}} d\mathcal{N}(\mathbf{U}_v) \int_{\mathbb{R}^{k_1 \times d}} d\mathcal{N}(\mathbf{V}) \exp \text{Tr} \left[\sqrt{\frac{x}{k_1}} \mathbf{Y}_v(x) (\mathbf{U}\mathbf{V})^\top - \frac{x}{2k_1} \mathbf{U}\mathbf{V}(\mathbf{U}\mathbf{V})^\top \right]. \quad (\text{C28})$$

The above matrix integral can be solved by means of the rectangular spherical integral, whose asymptotics is studied in Ref. [185]. Since we will not be needing this expression explicitly, we report only the one for the associated MMSE (x) function, as derived in Ref. [165]:

$$\begin{aligned} \text{MMSE}_v(x) := & 2 \frac{d}{dx} \iota_v(x) = \lim_{d \rightarrow \infty} \frac{1}{k_1 |\mathcal{I}_v| d} \mathbb{E} \|\mathbf{U}_v^0 \mathbf{V}^0 - \langle \mathbf{U}_v \mathbf{V} \rangle\|^2 \\ = & \frac{1}{x} \left[1 - P_v(\mathbf{v}) \gamma_2 \left(\frac{1}{P_v(\mathbf{v}) \gamma_2} - 1 \right)^2 \int \frac{\rho_{\mathbf{Y}_v(x)}(y)}{y^2} dy - P_v(\mathbf{v}) \gamma_2 \frac{\pi^2}{3} \int \rho_{\mathbf{Y}_v(x)}^3(y) dy \right]. \end{aligned} \quad (\text{C29})$$

Here, $\rho_{\mathbf{Y}_v(x)}$ is the singular value density of $\mathbf{Y}_v(x)/\sqrt{|\mathcal{I}_v|}$ obtained from a rectangular free convolution as described in the main text. The moment-matching condition (C23) in its replica-symmetric version thus reads

$$\text{MMSE}_v(\tau_v) = 1 - \mathbb{E}_{v^{(2)} \sim P_{v^{(2)}}} \mathcal{Q}_2(\mathbf{v}, v^{(2)}) \mathcal{Q}_1(v^{(2)}). \quad (\text{C30})$$

Recalling that $\tau_v = \tau_v(\mathcal{Q}_2, \mathcal{Q}_1)$, the saddle-point equations are obtained by equating the gradient of $f_{\text{RS}}^{(2)}$ with respect to the order parameters to 0:

$$\begin{aligned} \mathcal{Q}_1(\mathbf{v}^{(2)}) &= \mathbb{E}_{w_1^0 \sim P_{w_1}, \xi \sim \mathcal{N}(0,1)} [w_1^0 \langle w_1 \rangle_{\hat{\mathcal{Q}}_1(\mathbf{v}^{(2)})}], \\ P_{v^{(2)}}(\mathbf{v}^{(2)}) \hat{\mathcal{Q}}_1(\mathbf{v}^{(2)}) &= 2 \frac{\alpha}{\gamma_1} \partial_{\mathcal{Q}_1(\mathbf{v}^{(2)})} \phi_{P_{\text{out}}}(K^{(2)}(\bar{\mathcal{Q}}); 1) + \frac{\gamma_2}{\gamma_1} \mathbb{E}_{v \sim P_v} [\mathcal{Q}_{2:1}(v) - \mathbb{E}_{v^{(2)} \sim P_{v^{(2)}}} \mathcal{Q}_2(v, v^{(2)}) \mathcal{Q}_1(v^{(2)})] \partial_{\mathcal{Q}_1(\mathbf{v}^{(2)})} \tau_v, \\ \mathcal{Q}_2(\mathbf{v}, \mathbf{v}^{(2)}) &= \mathbb{E}_{w_2^0 \sim P_{w_2}, \xi \sim \mathcal{N}(0,1)} [w_2^0 \langle w_2 \rangle_{\hat{\mathcal{Q}}_2(\mathbf{v}, \mathbf{v}^{(2)})}], \\ P_v(\mathbf{v}) P_{v^{(2)}}(\mathbf{v}^{(2)}) \hat{\mathcal{Q}}_2(\mathbf{v}, \mathbf{v}^{(2)}) &= 2 \frac{\alpha}{\gamma_1 \gamma_2} \partial_{\mathcal{Q}_2(\mathbf{v}, \mathbf{v}^{(2)})} \phi_{P_{\text{out}}}(K^{(2)}(\bar{\mathcal{Q}}); 1) \\ &+ \frac{P_v(\mathbf{v})}{\gamma_1} [\mathcal{Q}_{2:1}(\mathbf{v}) - \mathbb{E}_{v^{(2)} \sim P_{v^{(2)}}} \mathcal{Q}_2(\mathbf{v}, v^{(2)}) \mathcal{Q}_1(v^{(2)})] \partial_{\mathcal{Q}_2(\mathbf{v}, \mathbf{v}^{(2)})} \tau_v, \\ \mathcal{Q}_{2:1}(\mathbf{v}) &= 1 - \text{MMSE}_v[\tau_v + \hat{\mathcal{Q}}_{2:1}(\mathbf{v})], \\ P_v(\mathbf{v}) \hat{\mathcal{Q}}_{2:1}(\mathbf{v}) &= 2 \frac{\alpha}{\gamma_2} \partial_{\mathcal{Q}_{2:1}(\mathbf{v})} \phi_{P_{\text{out}}}(K^{(2)}(\bar{\mathcal{Q}}); 1). \end{aligned} \quad (\text{C31})$$

b. Three or more hidden layers

To extend the derivations to an arbitrary number of layers, one has to find a way to write the entropic contributions of the overlaps entering the energetic part; see, for example, Eq. (C15). The challenging part is due to the overlaps $\mathcal{Q}_{l', l}$ defined in Eq. (C13). Indeed, the analog of the measure (C22) over the matrices $(\mathbf{W}^{(l':l)a})$ should be conditioned over all the overlaps defined from subsets of the indices $\{l', l' - 1, \dots, l\}$, which encode the possibility of all possible partial reconstructions [of $(\mathbf{W}^{(l':l'')0})$]. We leave this challenge for future works. Here, we focus on the case of activations with $\mu_1^{(l)} = 0$ (in addition to $\mu_0^{(l)} = \mu_2^{(l)} = 0$). In this case, the postactivation covariance (C10) is easy to write:

$$K^{ab} = \frac{1}{k_L} \sum_{i_L}^{k_L} (v_{i_L}^0)^2 g^{(L)} \left(\Omega_{i_L i_L}^{ab(L)} \right),$$

$$\Omega_{i_{l+1} i_{l+1}}^{ab(l+1)} \approx \sum_{\mathbf{v}^{(l+1)} \in \mathbf{V}^{(l+1)}} \frac{|\mathcal{I}_{\mathbf{v}^{(l+1)}}|}{k_l} \mathcal{Q}_{l+1}^{ab}(i_{l+1}, \mathbf{v}^{(l+1)}) \frac{1}{|\mathcal{I}_{\mathbf{v}^{(l+1)}}|} \sum_{i_l \in \mathcal{I}_{\mathbf{v}^{(l+1)}}} g^{(l)} \left(\Omega_{i_l i_l}^{(l)ab} \right). \quad (\text{C32})$$

In this recursion, only single-layer overlaps are entering, which have a simple entropic contribution. Moreover, no effective readout $\mathbf{v}^{(l)}$ is entering this equation, which means that nothing distinguishes neurons in hidden layers $l < L$. By our exchangeability hypothesis on neurons connected to readouts with the same amplitude, the order parameters can then be written as

$$\mathcal{Q}_l^{ab} = \frac{1}{k_l k_{l-1}} \text{Tr} \mathbf{W}^{(l)a} \mathbf{W}^{(l)b\top} \quad \text{for } l = 1, \dots, L-1, \quad (\text{C33})$$

$$\mathcal{Q}_L^{ab}(\mathbf{v}) = \frac{1}{|\mathcal{I}_{\mathbf{v}}| k_{L-1}} \sum_{i \in \mathcal{I}_{\mathbf{v}}} (\mathbf{W}^{(L)a} \mathbf{W}^{(L)b\top})_{ii}, \quad (\text{C34})$$

where we used a noncalligraphic symbol for $\mathcal{Q}_l$ to emphasize that these are just scalars, not functions of readout values. In terms of the order parameters, the recursion can be solved as

$$K^{ab} = \mathbb{E}_{v \sim P_v} v^2 g^{(L)}(\mathcal{Q}_L^{ab}(v) g^{(L-1)}\{\mathcal{Q}_{L-1}^{ab} g^{(L-2)}[\dots \mathcal{Q}_2^{ab} g^{(1)}(\mathcal{Q}_1^{ab}) \dots]\}). \quad (\text{C35})$$

The energetic term follows as before. For the entropic part, we notice that the contributions for each order parameter is factorized,

$$e^{Fs} = V_{W_L}^{k_L k_{L-1}}(\mathcal{Q}_L) \prod_{l=1}^{L-1} V_{W_l}^{k_l k_{l-1}}(\mathcal{Q}_l), \quad (\text{C36})$$

where, as before,

$$\lim_{s \rightarrow 0^+} \lim_{ns} \frac{1}{ns} \ln V_{W_l}^{k_l k_{l-1}}(\mathcal{Q}_l) = \frac{\gamma_l \gamma_{l-1}}{\alpha} \text{extr} \left[-\frac{\hat{\mathcal{Q}}_l \mathcal{Q}_l}{2} + \psi_{P_W}(\hat{\mathcal{Q}}_l) \right], \quad (\text{C37})$$

$$\lim_{s \rightarrow 0^+} \lim_{ns} \frac{1}{ns} \ln V_{W_L}^{k_L k_{L-1}}(\mathcal{Q}_L) = \frac{\gamma_L \gamma_{L-1}}{\alpha} \mathbb{E}_{v \sim P_v} \text{extr} \left[-\frac{\hat{\mathcal{Q}}_L(v) \mathcal{Q}_L(v)}{2} + \psi_{P_W}(\hat{\mathcal{Q}}_L(v)) \right]. \quad (\text{C38})$$

Denote by $K^{(L)}(\bar{\mathcal{Q}})$ the off-diagonal of the matrix $\mathbf{K} = K^{ab})_{a,b=0}^s$ in the RS Ansatz. The free entropy follows

$$f_{\text{RS}}^{(L)} = \phi_{P_{\text{out}}}(K^{(L)}(\bar{\mathcal{Q}}); 1) + \frac{\gamma_L \gamma_{L-1}}{\alpha} \mathbb{E}_{v \sim P_v} \left[\psi_{P_{W_L}}(\hat{\mathcal{Q}}_L(v)) - \frac{1}{2} \mathcal{Q}_L(v) \hat{\mathcal{Q}}_L(v) \right] + \sum_{l=1}^{L-1} \frac{\gamma_l \gamma_{l-1}}{\alpha} \left[\psi_{P_{W_l}}(\hat{\mathcal{Q}}_l) - \frac{1}{2} \mathcal{Q}_l \hat{\mathcal{Q}}_l \right]. \quad (\text{C39})$$

TABLE IV. Parameters for the experiments.

Figure	Tool	Burn-in steps	Sampling steps	No. of instances
2, 5	NUTS NUMPYRO	5000–8000	1	12
6	MCMC	Highly varied ^a	1/10 no. of burn-in steps	16
7	NUTS NUMPYRO	1000–8000	1–20	12–100
11, 12	HMC TENSORFLOW	4000	500	9
13	NUTS NUMPYRO	7000–25 000	500	20
14, 15	NUTS NUMPYRO	7000–25000	20	20
17	HMC TENSORFLOW	50 000, 150 000, 25 000 ^b	1	100
19	HMC TENSORFLOW	2500–8000	500	9
30	NUTS NUMPYRO	1000–8000	1–500	12–100
31	NUTS NUMPYRO	7000–25000	20–500	20

^aTo reduce the computational cost, we use a different number of MCMC steps for each α , based on how long it takes to reach the equilibrium; for some values of α , this number can be up to 500 times higher than for others. These data are provided in a spreadsheet on the GitHub repository.

^bRespectively, for $\alpha = 1.75, 2.75, 3.75$.

2. Structured data: Quenching the first-layer weights

In this subsection, we show consistency between the computations for the $L = 2$ case and the structured data setting. In fact, the latter is equivalent to taking a two-hidden-layer NN where the first activation is $\sigma^{(1)} = \text{Id}$, and the first set of weights are quenched and given to the student. In the notation of the previous section, this implies directly $Q_1(v^{(2)}) = 1$ for all $v^{(2)}$'s. Furthermore, since $g^{(1)} = 0$ in the definition (C20) of $K^{(2)}$, Q_2 disappears from the energetic part, letting entropy win. We thus conclude right away that $Q_2(v, v^{(2)}) = 0$ for all $v, v^{(2)}$'s. This implies in turn that $\tau_v = 0$ for all v 's. The formula for the free entropy thus simplifies to

$$f_{\text{RS}}^{(2)} := \phi_{P_{\text{out}}}(K^{(2)}(\bar{Q}); 1) + \frac{\gamma_2}{\alpha} \mathbb{E}_{v \sim P_v} \left[\frac{1}{2} \hat{Q}_{2:1}(v) (1 - Q_{2:1}(v)) - \iota_v(\hat{Q}_{2:1}(v)) \right]. \quad (\text{C40})$$

The correct way to think of ι_v is now that of Eq. (C28) where the annealed variables $\mathbf{V}$ mimicking the original $\mathbf{W}^{(1)}$ are fixed to the ground truth value $\mathbf{V}^0$ as they are given to the statistician. In that case, Eq. (C27) reduces to a set of $|\mathcal{I}_v|$ decoupled random linear estimation problems with a Gaussian prior on $\mathbf{U}_v$. The latter can be integrated via Gaussian integration yielding precisely

$$\iota_v(x) = \frac{1}{2} \int \ln(1 + xs) \rho_{\text{MP}}(s; 1/\gamma_1) ds \quad (\text{C41})$$

with $\rho_{\text{MP}}(s; 1/\gamma_1)$ the asymptotic spectral density of the Wishart matrix $\mathbf{V}^0 \mathbf{V}^0 / k_1$, namely, a Marchenko-Pastur of parameter $d/k_1 = 1/\gamma_1$. Note that $\int s \rho_{\text{MP}}(s; 1/\gamma_1) ds = 1$. Hence,

$$f_{\text{RS}}^{(2)} := \phi_{P_{\text{out}}}(K^{(2)}(\bar{Q}); 1) + \frac{\gamma_2}{2\alpha} \mathbb{E}_{v \sim P_v} \left\{ -\hat{Q}_{2:1}(v) Q_{2:1}(v) + \int [\hat{Q}_{2:1}(v)s] - \ln[1 + \hat{Q}_{2:1}(v)s] d\mu_C(s) \right\}, \quad (\text{C42})$$

the last integral being exactly $\psi_{P_w}(\hat{Q}_{2:1}(v))$ in Eq. (22).

APPENDIX D: DETAILS ON THE NUMERICAL PROCEDURES

In this appendix, we detail the implementation of the various numerical experiments involving algorithms such as the HMC, MCMC, and ADAM. Most of these algorithms were employed through their standard implementations available in NUMPY, TENSORFLOW, and PYTORCH PYTHON libraries. As already discussed, the GAMP-RIE algorithm introduced in Ref. [96] and publicly released in Ref. [193] was adapted to accommodate generic activation functions and inhomogeneous readouts. The only algorithm implemented entirely from scratch is the MCMC procedure used to sample from the posterior distribution with a Rademacher prior on the inner weights.

1. Sampling algorithms

Depending on the setting of each experiment, different algorithms and libraries are used to sample from the posterior distribution (see Table IV):

- (i) MCMC for the Rademacher prior
- (ii) HMC for the Gaussian prior. HMC augments the parameter space with auxiliary momenta and simulates Hamiltonian dynamics to propose distant moves with high acceptance probability. HMC is implemented in different PYTHON libraries: HMC package in TENSORFLOW.PROBABILITY [228] and the No-U-Turn Sampler (NUTS) [229] implemented in NUMPYRO [230]. This is an advanced version of HMC that automatically adapts the trajectory length to avoid redundant retracing

Let θ_t^a be the parameter sample obtained by running one of these sampling algorithm in t steps. The experimental Bayes-optimal error evaluated at θ_t^a is

$$\varepsilon_t^{\text{exp}} = \frac{1}{n_{\text{test}}} \sum_{\mu=1}^{n_{\text{test}}} \frac{1}{2} [\lambda(\theta_t^a, \mathbf{x}_\mu) - \lambda(\theta^0, \mathbf{x}_\mu)]^2,$$

where $\mathbf{x}_\mu$ are data from a test set of size 10^4 – 10^5 . Important experimental parameters include the following:

- (i) Burn-in steps. During the burn-in period, the sampler is run for a sufficiently large number of steps to reach a stationary state, either from informative or uninformative initialization. Stationarity can be assessed from the plot of $\varepsilon_t^{\text{exp}}$ versus t , where it fluctuates around a constant value. This differs from the case where the sampler gets stuck, in which $\varepsilon_t^{\text{exp}}$ remains constant after a certain number of steps.
- (ii) Sampling steps. In the sampling period, the sampler continues to run after the burn-in period. The Bayes-optimal error is computed as the average of $\varepsilon_t^{\text{exp}}$ for each value of time step t in this period. This averaging helps reduce the effect of dynamical fluctuations in the trajectory of $\varepsilon_t^{\text{exp}}$. Without it, the estimated Bayes-optimal errors obtained from half Gibbs errors exhibit larger standard deviations.

The following parameters are library specific:

- (i) Acceptance rate for NUTS NUMPYRO. This parameter was set between 0.6 and 0.7 in all experiments.
- (ii) Tree depth for NUTS NUMPYRO. This parameter specifies the maximum number of binary-doubling expansions of the Hamiltonian trajectory corresponding to a maximum of $2^{\text{depth}} - 1$ leapfrog steps per iteration. The depth was chosen between 7 and 8 in all experiments.
- (iii) Initial step size for HMC TENSORFLOW. This parameter is fixed to 0.01 in all experiments.
- (iv) Number of adaptation steps for HMC TENSORFLOW. This parameter is set to be the total number of steps (burn-in steps plus sampling steps). In other words,

every HMC step is adaptive, so the initial step size matters little: It will automatically adjust during the HMC trajectory to optimize sampling efficiency.

- (v) Number of leapfrog steps for HMC TENSORFLOW. Leapfrog steps control how long HMC simulates Hamiltonian dynamics before making a proposal. This parameter is fixed to 10 in all experiments

The following techniques are used to reduce finite-dimensional effects:

- (i) Averaging over sampling steps. This is discussed in the bullet point on sampling steps above.
- (ii) Reducing readout fluctuations. For experiments with fixed readouts, as k is typically of the order 10^2 , the empirical readout density can considerably differ from the true one. This finite-size effect increases the variance of the Bayes-optimal-error estimate. We reduce this variance as follows. For instance, binary readouts are generated with equal numbers of 1 and -1 . The same idea applies to other discrete readouts. For readouts with a continuous density such as Gaussian, we generate many (10^2 – 10^4) readout samples, sort their entries in increasing order, and average over the sorted vectors. This way of generating readouts yields more accurate estimates for Bayes-optimal error with fewer teacher instances.

2. ADAM-based optimization

ADAM is a first-order stochastic optimizer that adapts per-parameter learning rates using running estimates of the first and second moments of the gradients. In contrast to HMC, which requires a fully specified probabilistic model and prior, ADAM is a practical optimization algorithm widely used to train large neural networks. We therefore employ it to estimate the generalization error achieved by student networks trained with standard optimization methods on datasets generated by teachers with the same architecture.

In our experiments, we examine the generalization error of student networks trained with ADAM as a function of the number of gradient updates. Networks are initialized randomly from independent standard-normal draws. Typical optimizer settings are no weight decay, learning rates in the range 10^{-2} – 10^{-3} , large minibatches (typically between $\lfloor n/4 \rfloor$ and $\lfloor n/8 \rfloor$), and up to 3×10^5 gradient steps. During optimization, we record the predictive performance (mean-squared error) at regular intervals as a function of the gradient steps; these test-loss trajectories are averaged across independent teacher runs and reported in Figs. 9, 18, and 27.

3. Random feature model trained by ridge regression

We also study student networks trained as random feature models, where the student does not learn its hidden weights but instead fixes them at random and trains only a

linear readout via ridge regression. In this setting, the student network builds its feature matrix $\Phi^{\text{RF}} = \sigma(\mathbf{W}^{\text{RF}}\mathbf{X}/\sqrt{d})/\sqrt{\beta kd}$ using randomly drawn standard normal weights $\mathbf{W}^{\text{RF}} \in \mathbb{R}^{\beta kd \times d}$ with $\beta = O(1)$, which are independent of the teacher, and then learns only the readout weights $\mathbf{a}$ by solving the ridge-regularized least-squares problem $\min_{\mathbf{a}} \|\Phi^{\text{RF}}\mathbf{a} - \mathbf{y}\|^2 + t\|\mathbf{a}\|^2$. In Fig. 2, we fix $\beta = 3$ and sweep over different dataset sizes, drawing independent training and test sets for each realization. The regularization strength t is selected through a lightweight validation procedure over a small set of candidate values, while large-scale problems are solved efficiently with standard conjugate gradient procedures.

-
- [1] P. L. Bartlett, A. Montanari, and A. Rakhlin, *Deep learning: A statistical viewpoint*, *Acta Numer.* **30**, 87 (2021).
 - [2] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, *Nature (London)* **521**, 436 (2015).
 - [3] D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Spin-glass models of neural networks*, *Phys. Rev. A* **32**, 1007 (1985).
 - [4] E. Gardner, *The space of interactions in neural network models*, *J. Phys. A* **21**, 257 (1988).
 - [5] E. Gardner and B. Derrida, *Three unfinished works on the optimal storage capacity of networks*, *J. Phys. A* **22**, 1983 (1989).
 - [6] H. S. Seung, M. Oppen, and H. Sompolinsky, *Query by committee*, in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92 (Association for Computing Machinery, New York, USA, 1992), pp. 287–294.
 - [7] A. Engel, H. M. Köhler, F. Tschepe, H. Vollmayr, and A. Zippelius, *Storage capacity and learning algorithms for two-layer neural networks*, *Phys. Rev. A* **45**, 7590 (1992).
 - [8] K. Kang, J.-H. Oh, C. Kwon, and Y. Park, *Generalization in a two-layer neural network*, *Phys. Rev. E* **48**, 4805 (1993).
 - [9] D. O’Kane and O. Winther, *Learning to classify in large committee machines*, *Phys. Rev. E* **50**, 3201 (1994).
 - [10] H. Schwarze and J. Hertz, *Generalization in fully connected committee machines*, *Europhys. Lett.* **21**, 785 (1993).
 - [11] R. Urbanczik, *Storage capacity of the fully-connected committee machine*, *J. Phys. A* **30**, L387 (1997).
 - [12] O. Winther, B. Lautrup, and J.-B. Zhang, *Optimal learning in multilayer neural networks*, *Phys. Rev. E* **55**, 836 (1997).
 - [13] H. Schwarze and J. Hertz, *Generalization in a large committee machine*, *Europhys. Lett.* **20**, 375 (1992).
 - [14] H. Schwarze, M. Oppen, and W. Kinzel, *Generalization in a two-layer neural network*, *Phys. Rev. A* **46**, R6185 (1992).
 - [15] G. Mato and N. Parga, *Generalization properties of multilayered neural networks*, *J. Phys. A* **25**, 5047 (1992).
 - [16] R. Monasson and R. Zecchina, *Weight space structure and internal representations: A direct approach to learning and generalization in multilayer neural networks*, *Phys. Rev. Lett.* **75**, 2432 (1995).
 - [17] B. Schottky, *Phase transitions in the generalization behaviour of multilayer neural networks*, *J. Phys. A* **28**, 4515 (1995).
 - [18] A. Engel, *Correlation of internal representations in feed-forward neural networks*, *J. Phys. A* **29**, L323 (1996).
 - [19] D. Malzahn, A. Engel, and I. Kanter, *Storage capacity of correlated perceptrons*, *Phys. Rev. E* **55**, 7369 (1997).
 - [20] D. Malzahn and A. Engel, *Correlations between hidden units in multilayer neural networks and replica symmetry breaking*, *Phys. Rev. E* **60**, 2097 (1999).
 - [21] H. Sompolinsky, N. Tishby, and H. S. Seung, *Learning from examples in large neural networks*, *Phys. Rev. Lett.* **65**, 1683 (1990).
 - [22] G. Györgyi, *First-order transition to perfect generalization in a neural network with binary synapses*, *Phys. Rev. A* **41**, 7097 (1990).
 - [23] R. Meir and J. F. Fontanari, *Learning from examples in weight-constrained neural networks*, *J. Phys. A* **25**, 1149 (1992).
 - [24] D. M. L. Barbato and J. F. Fontanari, *The effects of lesions on the generalization ability of a perceptron*, *J. Phys. A* **26**, 1847 (1993).
 - [25] A. Engel and L. Reimers, *Reliability of replica symmetry for the generalization problem of a toy multilayer neural network*, *Europhys. Lett.* **28**, 531 (1994).
 - [26] G. J. Bex, R. Serneels, and C. Van den Broeck, *Storage capacity and generalization error for the reversed-wedge Ising perceptron*, *Phys. Rev. E* **51**, 6309 (1995).
 - [27] E. Barkai, D. Hansel, and H. Sompolinsky, *Broken symmetries in multilayered perceptrons*, *Phys. Rev. A* **45**, 4146 (1992).
 - [28] H. Schwarze, *Learning a rule in a multilayer neural network*, *J. Phys. A* **26**, 5781 (1993).
 - [29] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, England, 2001).
 - [30] H. Cui, *High-dimensional learning of narrow neural networks*, *J. Stat. Mech.* (2025) 023402.
 - [31] J. Bruna and D. Hsu, *Survey on Algorithms for Multi-Index Models*, *Stat. Sci.* **40**, 378 (2025).
 - [32] G. B. Arous, R. Gheissari, and A. Jagannath, *Online stochastic gradient descent on non-convex losses from high-dimensional inference*, *J. Mach. Learn. Res.* **22** (2021).
 - [33] A. Damian, L. Pillaud-Vivien, J. Lee, and J. Bruna, *Computational-statistical gaps in Gaussian single-index models (extended abstract)*, in *Proceedings of Thirty Seventh Conference on Learning Theory*, Proceedings of Machine Learning Research Vol. 247, edited by S. Agrawal and A. Roth (PMLR, 2024), pp. 1262–1262.
 - [34] E. Abbe, E. Boix-Adserà, M. Brennan, G. Bresler, and D. Nagaraj, *The staircase property: How hierarchical structure can guide deep learning*, in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21 (Curran Associates Inc., Red Hook, USA, 2021).
 - [35] E. Abbe, E. B. Adserà, and T. Misiakiewicz, *SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics*, in *Proceedings of Thirty Sixth Conference on Learning Theory*, Proceedings of Machine Learning

- Research, Vol. 195, edited by G. Neu and L. Rosasco (PMLR, 2023), pp. 2552–2623.
- [36] E. Troiani, Y. Dandi, L. Defilippis, L. Zdeborova, B. Loureiro, and F. Krzakala, *Fundamental computational limits of weak learnability in high-dimensional multi-index models*, in *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research Vol. 258, edited by Y. Li, S. Mandt, S. Agrawal, and E. Khan (PMLR, 2025), pp. 2467–2475.
- [37] R. M. Neal, *Priors for infinite networks*, in *Bayesian Learning for Neural Networks* (Springer, New York, 1996), pp. 29–53.
- [38] C. Williams, *Computing with infinite networks*, in *Advances in Neural Information Processing Systems*, edited by M. Mozer, M. Jordan, and T. Petsche (MIT Press, Cambridge, MA, 1996), Vol. 9.
- [39] 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).
- [40] A. G. D. 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).
- [41] B. Hanin, *Random neural networks in the infinite width limit as Gaussian processes*, *Ann. Appl. Probab.* **33**, 4798 (2023).
- [42] H. Yoon and J.-H. Oh, *Learning of higher-order perceptrons with tunable complexities*, *J. Phys. A* **31**, 7771 (1998).
- [43] R. Dietrich, M. Opper, and H. Sompolinsky, *Statistical mechanics of support vector networks*, *Phys. Rev. Lett.* **82**, 2975 (1999).
- [44] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová, *Generalisation error in learning with random features and the hidden manifold model*, *J. Stat. Mech.* (2021) 124013.
- [45] 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.
- [46] A. Canatar, B. Bordelon, and C. Pehlevan, *Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks*, *Nat. Commun.* **12**, 2914 (2021).
- [47] L. Xiao, H. Hu, T. Misiakiewicz, Y. M. Lu, and J. Pennington, *Precise learning curves and higher-order scaling limits for dot-product kernel regression*, *J. Stat. Mech.* (2023) 114005.
- [48] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, *Linearized two-layers neural networks in high dimension*, *Ann. Stat.* **49**, 1029 (2021).
- [49] A. Rahimi and B. Recht, *Random features for large-scale kernel machines*, in *Advances in Neural Information Processing Systems*, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (Curran Associates, Inc., 2007), Vol. 20.
- [50] A. Jacot, F. Gabriel, and C. Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), Vol. 31.
- [51] L. Chizat, E. Oyallon, and F. Bach, *On lazy training in differentiable programming*, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019), Vol. 32.
- [52] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, *When do neural networks outperform kernel methods?*, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., 2020), Vol. 33, pp. 14820–14830.
- [53] M. Refinetti, S. Goldt, F. Krzakala, and L. Zdeborova, *Classifying high-dimensional Gaussian mixtures: Where kernel methods fail and neural networks succeed*, in *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 139, edited by M. Meila and T. Zhang (PMLR, 2021), pp. 8936–8947.
- [54] E. Dyer and G. Gur-Ari, *Asymptotics of wide networks from Feynman diagrams*, in *International Conference on Learning Representations* (2020).
- [55] 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.
- [56] G. Naveh, O. Ben David, H. Sompolinsky, and Z. Ringel, *Predicting the outputs of finite deep neural networks trained with noisy gradients*, *Phys. Rev. E* **104**, 064301 (2021).
- [57] 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*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021), Vol. 34, pp. 24765–24777.
- [58] K. T. Grosvenor and R. Jefferson, *The edge of chaos: Quantum field theory and deep neural networks*, *SciPost Phys.* **12**, 081 (2022).
- [59] K. Fischer, J. Lindner, D. Dahmen, Z. Ringel, M. Krämer, and M. Helias, *Critical feature learning in deep neural networks*, in *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 235, edited by R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp (PMLR, 2024), pp. 13660–13690.
- [60] I. Banta, T. Cai, N. Craig, and Z. Zhang, *Structures of neural network effective theories*, *Phys. Rev. D* **109**, 105007 (2024).
- [61] M. Guillen, P. Misof, and J. E. Gerken, *Finite-width neural tangent kernels from Feynman diagrams*, [arXiv:2508.11522](https://arxiv.org/abs/2508.11522).

- [62] Y. Bahri, B. Hanin, A. Brossollet, V. Erba, C. Keup, R. Pacelli, and J. B. Simon, *Les Houches lectures on deep learning at large and infinite width**, *J. Stat. Mech.* (2024) 104012.
- [63] Z. Ringel, N. Rubin, E. Mor, M. Helias, and I. Seroussi, *Applications of statistical field theory in deep learning*, *arXiv:2502.18553*.
- [64] S. Mei, A. Montanari, and P.-M. Nguyen, *A mean field view of the landscape of two-layer neural networks*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, E7665 (2018).
- [65] S. Mei, T. Misiakiewicz, and A. Montanari, *Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit*, in *Proceedings of the Thirty-Second Conference on Learning Theory*, Proceedings of Machine Learning Research Vol. 99, edited by A. Beygelzimer and D. Hsu (PMLR, 2019), pp. 2388–2464.
- [66] G. Yang and E. J. Hu, *Tensor programs IV: Feature learning in infinite-width neural networks*, in *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 139, edited by M. Meila and T. Zhang (PMLR, 2021), pp. 11727–11737.
- [67] G. Rotskoff and E. Vanden-Eijnden, *Trainability and accuracy of artificial neural networks: An interacting particle system approach*, *Commun. Pure Appl. Math.* **75**, 1889 (2022).
- [68] J. Sirignano and K. Spiliopoulos, *Mean field analysis of neural networks: A central limit theorem*, *Stoch. Proc. Appl.* **130**, 1820 (2020).
- [69] B. Bordelon and C. Pehlevan, *Self-consistent dynamical field theory of kernel evolution in wide neural networks*, in *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., 2022), Vol. 35, pp. 32240–32256.
- [70] P.-M. Nguyen and H. T. Pham, *A rigorous framework for the mean field limit of multilayer neural networks*, *Math. Stat. Learn.* **6**, 201 (2023).
- [71] I. Seroussi, G. Naveh, and Z. Ringel, *Separation of scales and a thermodynamic description of feature learning in some CNNs*, *Nat. Commun.* **14**, 908 (2023).
- [72] F. Bassetti, M. Gherardi, A. Ingrosso, M. Pastore, and P. Rotondo, *Feature learning in finite-width Bayesian deep linear networks with multiple outputs and convolutional layers*, *J. Mach. Learn. Res.* **26**, 1 (2025).
- [73] N. Rubin, Z. Ringel, I. Seroussi, and M. Helias, *A unified approach to feature learning in Bayesian neural networks*, in *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning* (2024).
- [74] A. van Meegen and H. Sompolinsky, *Coding schemes in neural networks learning classification tasks*, *Nat. Commun.* **16**, 3354 (2025).
- [75] C. Lauditi, B. Bordelon, and C. Pehlevan, *Adaptive kernel predictors from feature-learning infinite limits of neural networks*, *arXiv:2502.07998*.
- [76] A. X. Yang, M. Robeyns, E. Milsom, B. Anson, N. Schoots, and L. Aitchison, *A theory of representation learning gives a deep generalisation of kernel methods*, in *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 202, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (PMLR, 2023), pp. 39380–39415.
- [77] N. Rubin, I. Seroussi, and Z. Ringel, *Grokking as a first order phase transition in two layer networks*, in *The Twelfth International Conference on Learning Representations* (2024).
- [78] A. M. Saxe, J. McClelland, and S. Ganguli, *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*, in *Proceedings of the International Conference on Learning Representations 2014* (2014).
- [79] Q. Li and H. Sompolinsky, *Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization*, *Phys. Rev. X* **11**, 031059 (2021).
- [80] L. Aitchison, *Why bigger is not always better: On finite and infinite neural networks*, in *International Conference on Machine Learning* (PMLR, 2020), pp. 156–164.
- [81] B. Hanin and A. Zlokapa, *Bayesian interpolation with deep linear networks*, *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2301345120 (2023).
- [82] 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).
- [83] B. Neyshabur, R. Tomioka, and N. Srebro, *Norm-based capacity control in neural networks*, in *Proceedings of The 28th Conference on Learning Theory*, Proceedings of Machine Learning Research Vol. 40, edited by P. Grünwald, E. Hazan, and S. Kale (PMLR, Paris, France, 2015), pp. 1376–1401.
- [84] S. Pesme and N. Flammarion, *Saddle-to-saddle dynamics in diagonal linear networks*, in *Advances in Neural Information Processing Systems*, edited by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Curran Associates, Inc., 2023), Vol. 36, pp. 7475–7505.
- [85] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, *The implicit bias of gradient descent on separable data*, *J. Mach. Learn. Res.* **19**, 1 (2018).
- [86] S. Pesme, L. Pillaud-Vivien, and N. Flammarion, *Implicit bias of SGD for diagonal linear networks: A provable benefit of stochasticity*, in *Advances in Neural Information Processing Systems*, edited by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (2021).
- [87] R. Berthier, *Incremental learning in diagonal linear networks*, *J. Mach. Learn. Res.* **24**, (2023), <http://jmlr.org/papers/v24/22-1395.html>.
- [88] H. Labarrière, C. Molinari, L. Rosasco, S. Villa, and C. Vega, *Optimization insights into deep diagonal linear networks*, *arXiv:2412.16765*.
- [89] S. Du and J. Lee, *On the power of over-parametrization in neural networks with quadratic activation*, in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 80, edited by J. Dy and A. Krause (PMLR, 2018), pp. 1329–1338.
- [90] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, *Theoretical insights into the optimization landscape of over-parameterized shallow neural networks*, *IEEE Trans. Inf. Theory* **65**, 742 (2019).

- [91] L. Venturi, A. S. Bandeira, and J. Bruna, *Spurious valleys in one-hidden-layer neural network optimization landscapes*, *J. Mach. Learn. Res.* **20**, 1 (2019).
- [92] S. Sarao Mannelli, E. Vanden-Eijnden, and L. Zdeborová, *Optimization and generalization of shallow neural networks with quadratic activation functions*, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., 2020), Vol. 33, pp. 13445–13455.
- [93] D. Gamarnik, E. C. Kızıldağ, and I. Zadik, *Stationary points of a shallow neural network with quadratic activations and the global optimality of the gradient descent algorithm*, *Math. Oper. Res.* **50**, 209 (2024).
- [94] S. Martin, F. Bach, and G. Biroli, *On the impact of overparameterization on the training of a shallow neural network in high dimensions*, in *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research Vol. 238, edited by S. Dasgupta, S. Mandt, and Y. Li (PMLR, 2024), pp. 3655–3663.
- [95] Y. Arjevani, J. Bruna, J. Kileel, E. Polak, and M. Trager, *Geometry and optimization of shallow polynomial networks*, [arXiv:2501.06074](https://arxiv.org/abs/2501.06074).
- [96] A. Maillard, E. Troiani, S. Martin, L. Zdeborová, and F. Krzakala, *Bayes-optimal learning of an extensive-width neural network from quadratically many samples*, in *Advances in Neural Information Processing Systems*, edited by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Curran Associates, Inc., 2024), Vol. 37, pp. 82085–82132.
- [97] Y. Xu, A. Maillard, L. Zdeborová, and F. Krzakala, *Fundamental limits of matrix sensing: Exact asymptotics, universality, and applications* (2025).
- [98] V. Erba, E. Troiani, L. Zdeborová, and F. Krzakala, *The nuclear route: Sharp asymptotics of ERM in overparametrized quadratic networks*, [arXiv:2505.17958](https://arxiv.org/abs/2505.17958).
- [99] G. Ben Arous, M. A. Erdogdu, N. M. Vural, and D. Wu, *Learning quadratic neural networks in high dimensions: SGD dynamics and scaling laws*, [arXiv:2508.03688](https://arxiv.org/abs/2508.03688).
- [100] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, *Optimal errors and phase transitions in high-dimensional generalized linear models*, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5451 (2019).
- [101] J. Barbier and N. Macris, *Statistical limits of dictionary learning: Random matrix theory and the spectral replica method*, *Phys. Rev. E* **106**, 024136 (2022).
- [102] A. Maillard, F. Krzakala, M. Mézard, and L. Zdeborová, *Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising*, *J. Stat. Mech.* (2022) 083301.
- [103] F. Pourkamali, J. Barbier, and N. Macris, *Matrix inference in growing rank regimes*, *IEEE Trans. Inf. Theory* **70**, 8133 (2024).
- [104] G. Semerjian, *Matrix denoising: Bayes-optimal estimators via low-degree polynomials*, *J. Stat. Phys.* **191**, 139 (2024).
- [105] R. Pacelli, S. Ariosto, M. Pastore, F. Ginelli, M. Gherardi, and P. Rotondo, *A statistical mechanics framework for Bayesian deep neural networks beyond the infinite-width limit*, *Nat. Mach. Intell.* **5**, 1497 (2023).
- [106] P. Baglioni, R. Pacelli, R. Aiudi, F. Di Renzo, A. Vezzani, R. Burioni, and P. Rotondo, *Predictive power of a Bayesian effective action for fully connected one hidden layer neural networks in the proportional limit*, *Phys. Rev. Lett.* **133**, 027301 (2024).
- [107] A. Ingrosso, R. Pacelli, P. Rotondo, and F. Gerace, *Statistical mechanics of transfer learning in fully connected networks in the proportional limit*, *Phys. Rev. Lett.* **134**, 177301 (2025).
- [108] B. Hanin and A. Zlokapa, *Gibbs measures from deep shaped multilayer perceptrons*, *Phys. Rev. Lett.* **136**, 067301 (2026).
- [109] H. Cui, F. Krzakala, and L. Zdeborová, *Bayes-optimal learning of deep random networks of extensive-width*, in *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 202, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (PMLR, 2023), pp. 6468–6521.
- [110] F. Camilli, D. Tiepova, and J. Barbier, *Fundamental limits of overparametrized shallow neural networks for supervised learning*, *Boll. Unione Mat. Ital.* (2025), [10.1007/s40574-025-00506-2](https://doi.org/10.1007/s40574-025-00506-2).
- [111] F. Camilli, D. Tiepova, E. Bergamin, and J. Barbier, *Information-theoretic reduction of deep neural networks to linear models in the overparametrized proportional regime*, in *Proceedings of Thirty Eighth Conference on Learning Theory*, Proceedings of Machine Learning Research Vol. 291, edited by N. Haghtalab and A. Moitra (PMLR, 2025), pp. 757–798.
- [112] 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*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021), Vol. 34, pp. 21352–21364.
- [113] R. Aiudi, R. Pacelli, P. Baglioni, A. Vezzani, R. Burioni, and P. Rotondo, *Local kernel renormalization as a mechanism for feature learning in overparametrized convolutional neural networks*, *Nat. Commun.* **16**, 568 (2025).
- [114] H. Yoshino, *From complex to simple: Hierarchical free-energy landscape renormalized in deep neural networks*, *SciPost Phys. Core* **2**, 005 (2020).
- [115] H. Yoshino, *Spatially heterogeneous learning by a deep student machine*, *Phys. Rev. Res.* **5**, 033068 (2023).
- [116] G. Huang, L. S. Chan, H. Yoshino, G. Zhang, and Y. Jin, *Liquid and solid layers in a thermal deep learning machine*, [arXiv:2506.06789](https://arxiv.org/abs/2506.06789).
- [117] J. Yao, Y. Yacoby, B. Coker, W. Pan, and F. Doshi-Velez, *An empirical analysis of the advantages of finite- v.s. infinite-width Bayesian neural networks* (2022).
- [118] J. Lee, S. S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein, *Finite versus infinite neural networks: An empirical study*, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20 (Curran Associates Inc., Red Hook, NY, USA, 2020).
- [119] L. Zdeborová, *Understanding deep learning is also a job for physicists*, *Nat. Phys.* **16**, 602 (2020).

- [120] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, *Statistical mechanics of deep learning*, *Annu. Rev. Condens. Matter Phys.* **11**, 501 (2020).
- [121] J. Hoffmann *et al.*, *Training compute-optimal large language models*, in *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., 2022), Vol. 35, pp. 30016–30030.
- [122] J. Ni, Q. Liu, C. Du, L. Dou, H. Yan, Z. Wang, T. Pang, and M. Q. Shieh, *Training optimal large diffusion language models*, *arXiv:2510.03280*.
- [123] M. Lan, P. Torr, A. Meek, A. Khakzar, D. Krueger, and F. Barez, *Quantifying feature space universality across large language models via sparse autoencoders*, *arXiv:2410.06981*.
- [124] Z. Li, C. Fan, and T. Zhou, *Grokking in LLM pretraining? Monitor memorization-to-generalization without test*, *arXiv:2506.21551*.
- [125] M. Mondelli and A. Montanari, *On the connection between learning two-layer neural networks and tensor decomposition*, in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research Vol. 89, edited by K. Chaudhuri and M. Sugiyama (PMLR, 2019), pp. 1051–1060.
- [126] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1986).
- [127] C. Itzykson and J. Zuber, *The planar approximation. II*, *J. Math. Phys. (N.Y.)* **21**, 411 (1980).
- [128] A. Matytsin, *On the large- N limit of the Itzykson-Zuber integral*, *Nucl. Phys.* **B411**, 805 (1994).
- [129] A. Guionnet and O. Zeitouni, *Large deviations asymptotics for spherical integrals*, *J. Funct. Anal.* **188**, 461 (2002).
- [130] A. Guionnet, *First order asymptotics of matrix integrals; a rigorous approach towards the understanding of matrix models*, *Commun. Math. Phys.* **244**, 527 (2004).
- [131] J.-B. Zuber, *The large- N limit of matrix integrals over the orthogonal group*, *J. Phys. A* **41**, 382001 (2008).
- [132] V. A. Kazakov, *Solvable matrix models*, *arXiv:hep-th/0003064*.
- [133] E. Brézin, S. Hikami *et al.*, *Random Matrix Theory with an External Source* (Springer, New York, 2016).
- [134] D. Anninos and B. Mühlmann, *Notes on matrix models (matrix musings)*, *J. Stat. Mech.* (2020) 083109.
- [135] J. Bun, J. P. Bouchaud, S. N. Majumdar, and M. Potters, *Instanton approach to large N Harish-Chandra-Itzykson-Zuber integrals*, *Phys. Rev. Lett.* **113**, 070201 (2014).
- [136] M. Potters and J.-P. Bouchaud, *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists* (Cambridge University Press, Cambridge, England, 2020).
- [137] J. Hussion and J. Ko, *Spherical integrals of sublinear rank*, *Probab. Theory Relat. Fields* **193**, 1 (2025).
- [138] G. Parisi and M. Potters, *Mean-field equations for spin models with orthogonal interaction matrices*, *J. Phys. A* **28**, 5267 (1995).
- [139] M. Oppen and O. Winther, *Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling*, *Phys. Rev. E* **64**, 056131 (2001).
- [140] M. Oppen, B. Çakmak, and O. Winther, *A theory of solving TAP equations for Ising models with general invariant random matrices*, *J. Phys. A* **49**, 114002 (2016).
- [141] Z. Fan, Y. Li, and S. Sen, *TAP equations for orthogonally invariant spin glasses at high temperature*, *arXiv:2202.09325*.
- [142] J. Barbier and M. Sáenz, *Marginals of a spherical spin glass model with correlated disorder*, *Electron. Commun. Probab.* **27**, 1 (2022).
- [143] Z. Fan and Y. Wu, *The replica-symmetric free energy for Ising spin glasses with orthogonally invariant couplings*, *Probab. Theory Relat. Fields* **190**, 1 (2024).
- [144] Y. Kabashima, *Inference from correlated patterns: A unified theory for perceptron learning and linear vector channels*, *J. Phys. Conf. Ser.* **95**, 012001 (2008).
- [145] M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborová, *Entropy and mutual information in models of deep neural networks*, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), Vol. 31.
- [146] K. Takeda, S. Uda, and Y. Kabashima, *Analysis of CDMA systems that are characterized by eigenvalue spectrum*, *Europhys. Lett.* **76**, 1193 (2006).
- [147] A. Tulino, G. Caire, S. Shamai, and S. Verdú, *Support recovery with sparsely sampled free random matrices*, in *2011 IEEE International Symposium on Information Theory Proceedings* (2011), pp. 2328–2332.
- [148] T. Hou, Y. Liu, T. Fu, and J. Barbier, *Sparse superposition codes under VAMP decoding with generic rotational invariant coding matrices*, in *2022 IEEE International Symposium on Information Theory (ISIT)* (2022), pp. 1372–1377.
- [149] J. Barbier, F. Camilli, Y. Xu, and M. Mondelli, *Information limits and Thouless-Anderson-Palmer equations for spiked matrix models with structured noise*, *Phys. Rev. Res.* **7**, 013081 (2025).
- [150] S. Rangan, P. Schniter, and A. K. Fletcher, *Vector approximate message passing*, *IEEE Trans. Inf. Theory* **65**, 6664 (2019).
- [151] J. Ma and L. Ping, *Orthogonal AMP*, *IEEE Access* **5**, 2020 (2017).
- [152] A. Maillard, L. Foini, A. L. Castellanos, F. Krzakala, M. Mézard, and L. Zdeborová, *High-temperature expansions and message passing algorithms*, *J. Stat. Mech.* (2019) 113301.
- [153] L. Liu, S. Huang, and B. M. Kurkoski, *Memory AMP*, *IEEE Trans. Inf. Theory* **68**, 8015 (2022).
- [154] K. Takeuchi, *On the convergence of orthogonal/vector AMP: Long-memory message-passing strategy*, in *2022 IEEE International Symposium on Information Theory (ISIT)* (2022), pp. 1366–1371.
- [155] T. Takahashi and Y. Kabashima, *Macroscopic analysis of vector approximate message passing in a model-mismatched setting*, *IEEE Trans. Inf. Theory* **68**, 5579 (2022).
- [156] Z. Fan, *Approximate Message Passing algorithms for rotationally invariant matrices*, *Ann. Stat.* **50**, 197 (2022).
- [157] J. Barbier, N. Macris, A. Maillard, and F. Krzakala, *The mutual information in random linear estimation beyond*

- i.i.d. matrices*, in *2018 IEEE International Symposium on Information Theory (ISIT)* (2018), pp. 1390–1394.
- [158] C. Gerbelot, A. Abbata, and F. Krzakala, *Asymptotic errors for high-dimensional convex penalized linear regression beyond Gaussian matrices*, in *Proceedings of Thirty Third Conference on Learning Theory*, Proceedings of Machine Learning Research Vol. 125, edited by J. Abernethy and S. Agarwal (PMLR, 2020), pp. 1682–1713.
- [159] C. Gerbelot, A. Abbata, and F. Krzakala, *Asymptotic errors for teacher-student convex generalized linear models (or: How to prove Kabashima's replica formula)*, *IEEE Trans. Inf. Theory* **69**, 1824 (2023).
- [160] R. Dudeja, Y. M. Lu, and S. Sen, *Universality of approximate message passing with semirandom matrices*, *Ann. Probab.* **51**, 1616 (2023).
- [161] J. Barbier, F. Camilli, M. Mondelli, and M. Sáenz, *Fundamental limits in structured principal component analysis and how to reach them*, *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2302028120 (2023).
- [162] R. Dudeja, S. Liu, and J. Ma, *Optimality of approximate message passing algorithms for spiked matrix models with rotationally invariant noise*, [arXiv:2405.18081](https://arxiv.org/abs/2405.18081).
- [163] O. Ledoit and S. Péché, *Eigenvectors of some large sample covariance matrix ensembles*, *Probab. Theory Relat. Fields* **151**, 233 (2011).
- [164] J. Bun, R. Allez, J.-P. Bouchaud, and M. Potters, *Rotational invariant estimator for general noisy matrices*, *IEEE Trans. Inf. Theory* **62**, 7475 (2016).
- [165] F. Pourkamali and N. Macris, *Rectangular rotational invariant estimator for general additive noise matrices*, in *2023 IEEE International Symposium on Information Theory (ISIT)* (2023), pp. 2081–2086.
- [166] E. Troiani, V. Erba, F. Krzakala, A. Maillard, and L. Zdeborová, *Optimal denoising of rotationally invariant rectangular matrices*, in *Proceedings of Mathematical and Scientific Machine Learning*, Proceedings of Machine Learning Research, Vol. 190, edited by B. Dong, Q. Li, L. Wang, and Z.-Q. J. Xu (PMLR, 2022), pp. 97–112.
- [167] H. C. Schmidt, *Statistical physics of sparse and dense models in optimization and inference*, Ph.D. thesis, IPHT—Institut de Physique Théorique, 2018.
- [168] A. Sakata and Y. Kabashima, *Statistical mechanics of dictionary learning*, *Europhys. Lett.* **103**, 28008 (2013).
- [169] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata, and L. Zdeborová, *Phase transitions and sample complexity in Bayes-optimal matrix factorization*, *IEEE Trans. Inf. Theory* **62**, 4228 (2016).
- [170] V. Erba, E. Troiani, L. Biggio, A. Maillard, and L. Zdeborová, *Bilinear sequence regression: A model for learning from long sequences of high-dimensional tokens*, *Phys. Rev. X* **15**, 021092 (2025).
- [171] J. Barbier, F. Camilli, J. Ko, and K. Okajima, *Phase diagram of extensive-rank symmetric matrix denoising beyond rotational invariance*, *Phys. Rev. X* **15**, 021085 (2025).
- [172] Y. Ren, E. Nichani, D. Wu, and J. D. Lee, *Emergence and scaling laws in SGD learning of shallow neural networks*, [arXiv:2504.19983](https://arxiv.org/abs/2504.19983).
- [173] A. Bodin and N. Macris, *Gradient flow on extensive-rank positive semi-definite matrix denoising*, in *2023 IEEE Information Theory Workshop (ITW)* (IEEE, 2023), pp. 365–370.
- [174] M.-T. Nguyen and R. Skerk, *Statistical physics of deep learning* (2026), <https://github.com/Minh-Toan/statphys-deep-NN>.
- [175] S. Mei and A. Montanari, *The generalization error of random features regression: Precise asymptotics and the double descent curve*, *Commun. Pure Appl. Math.* **75**, 667 (2022).
- [176] 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).
- [177] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, *Surprises in high-dimensional ridgeless least squares interpolation*, *Ann. Stat.* **50**, 949 (2022).
- [178] S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mezard, and L. Zdeborová, *The Gaussian equivalence of generative models for learning with shallow neural networks*, in *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, Proceedings of Machine Learning Research Vol. 145, edited by J. Bruna, J. Hesthaven, and L. Zdeborová (PMLR, 2022), pp. 426–471.
- [179] H. Hu and Y. M. Lu, *Universality laws for high-dimensional learning with random features*, *IEEE Trans. Inf. Theory* **69**, 1932 (2023).
- [180] A. Montanari and B. N. Saeed, *Universality of empirical risk minimization*, in *Proceedings of Thirty Fifth Conference on Learning Theory*, Proceedings of Machine Learning Research Vol. 178, edited by P.-L. Loh and M. Raginsky (PMLR, 2022), pp. 4310–4312.
- [181] G. G. Wen, H. Hu, Y. M. Lu, Z. Fan, and T. Misiakiewicz, *When does Gaussian equivalence fail and how to fix it: Non-universal behavior of random features with quadratic scaling*, [arXiv:2512.03325](https://arxiv.org/abs/2512.03325).
- [182] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Oxford University Press, New York, 2001).
- [183] L. Zdeborová and F. Krzakala, *Statistical physics of inference: Thresholds and algorithms*, *Adv. Phys.* **65**, 453 (2016).
- [184] D. Guo, S. Shamai, and S. Verdú, *Mutual information and minimum mean-square error in Gaussian channels*, *IEEE Trans. Inf. Theory* **51**, 1261 (2005).
- [185] A. Guionnet and J. Huang, *Asymptotics of rectangular spherical integrals*, *J. Funct. Anal.* **285**, 110144 (2023).
- [186] J. Barbier and D. Panchenko, *Strong replica symmetry in high-dimensional optimal Bayesian inference*, *Commun. Math. Phys.* **393**, 1199 (2022).
- [187] J. T. Parker, P. Schniter, and V. Cevher, *Bilinear generalized approximate message passing—Part I: Derivation*, *IEEE Trans. Signal Process.* **62**, 5839 (2014).
- [188] F. Krzakala, M. Mézard, and L. Zdeborová, *Phase diagram and approximate message passing for blind calibration and dictionary learning*, in *2013 IEEE International Symposium on Information Theory* (2013), pp. 659–663.
- [189] B. Aubin, A. Maillard, J. Barbier, F. Krzakala, N. Macris, and L. Zdeborová, *The committee machine: Computational to statistical gaps in learning a two-layers neural*

- network, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), Vol. 31.
- [190] C. Baldassi, E. M. Malatesta, and R. Zecchina, *Properties of the geometry of solutions and capacity of multilayer neural networks with rectified linear unit activations*, *Phys. Rev. Lett.* **123**, 170602 (2019).
- [191] J. Barbier, F. Gerace, A. Ingrosso, C. Lauditi, E. M. Malatesta, G. Nwemadji, and R. P. Ortiz, *Generalization performance of narrow one-hidden layer networks in the teacher-student setting*, [arXiv:2507.00629](https://arxiv.org/abs/2507.00629).
- [192] J. Barbier, *Overlap matrix concentration in optimal Bayesian inference*, *Inf. Inference* **10**, 597 (2020).
- [193] A. Maillard, E. Troiani, S. Martin, F. Krzakala, and L. Zdeborová, *Github repository ExtensiveWidthQuadraticSamples*, <https://github.com/SPOC-group/ExtensiveWidthQuadraticSamples> (2024).
- [194] T. Tao and V. Vu, *Random matrices: Universality of local eigenvalue statistics up to the edge*, *Commun. Math. Phys.* **298**, 549 (2010).
- [195] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [196] M. Hennick and S. D. Baerdemacker, *Almost Bayesian: The fractal dynamics of stochastic gradient descent*, [arXiv:2503.22478](https://arxiv.org/abs/2503.22478).
- [197] C. Mingard, G. Valle-Pérez, J. Skalse, and A. A. Louis, *Is SGD a Bayesian sampler? Well, almost*, *J. Mach. Learn. Res.* **22**, 1 (2021).
- [198] S. L. Smith, D. Duckworth, S. Rezhikov, Q. V. Le, and J. Sohl-Dickstein, *Stochastic natural gradient descent draws posterior samples in function space*, [arXiv:1806.09597](https://arxiv.org/abs/1806.09597).
- [199] S. Mandt, M. D. Hoffman, and D. M. Blei, *Stochastic gradient descent as approximate Bayesian inference*, *J. Mach. Learn. Res.* **18**, 1 (2017).
- [200] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, *SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability*, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), Vol. 30.
- [201] F. Cagnetta, L. Petrini, U. M. Tomasini, A. Favero, and M. Wyart, *How deep neural networks learn compositional data: The random hierarchy model*, *Phys. Rev. X* **14**, 031001 (2024).
- [202] F. Aguirre-López, S. Franz, and M. Pastore, *Random features and polynomial rules*, *SciPost Phys.* **18**, 039 (2025).
- [203] H. Hu, Y. M. Lu, and T. Misiakiewicz, *Asymptotics of random feature regression beyond the linear scaling regime*, [arXiv:2403.08160](https://arxiv.org/abs/2403.08160).
- [204] J. Barbier and N. Macris, *The adaptive interpolation method: A simple scheme to prove replica formulas in Bayesian inference*, *Probab. Theory Relat. Fields* **174**, 1133 (2019).
- [205] The work [206] concurrent with ours proposes an annealed approximation of the free entropy (rather than the more accurate quenched computation proposed in the present paper) to study shallow MLPs of extensive width with ReLU activation.
- [206] A. Afanah and B. Rosenow, *Unified description of learning dynamics in the soft committee machine from finite to ultra-wide regimes*, [arXiv:2512.16556](https://arxiv.org/abs/2512.16556).
- [207] R. Monasson, *Properties of neural networks storing spatially correlated patterns*, *J. Phys. A* **25**, 3701 (1992).
- [208] 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*, in *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021), Vol. 34, pp. 18137–18151.
- [209] Del Giudice, P., Franz, S., and M. A. Virasoro, *Perceptron beyond the limit of capacity*, *J. Phys. (Les Ulis, Fr.)* **50**, 121 (1989).
- [210] B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, and L. Zdeborová, *Learning Gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions*, in *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021), Vol. 34, pp. 10144–10157.
- [211] B. Lopez, M. Schroder, and M. Opper, *Storage of correlated patterns in a perceptron*, *J. Phys. A* **28**, L447 (1995).
- [212] S. Y. Chung, D. D. Lee, and H. Sompolinsky, *Classification and geometry of general perceptual manifolds*, *Phys. Rev. X* **8**, 031003 (2018).
- [213] P. Rotondo, M. Pastore, and M. Gherardi, *Beyond the storage capacity: Data-driven satisfiability transition*, *Phys. Rev. Lett.* **125**, 120601 (2020).
- [214] M. Pastore, P. Rotondo, V. Erba, and M. Gherardi, *Statistical learning theory of structured data*, *Phys. Rev. E* **102**, 032119 (2020).
- [215] A. Sclocchi, A. Favero, and M. Wyart, *A phase transition in diffusion models reveals the hierarchical nature of data*, *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2408799121 (2025).
- [216] D. Saad and S. A. Solla, *On-line learning in soft committee machines*, *Phys. Rev. E* **52**, 4225 (1995).
- [217] D. Saad and S. Solla, *Dynamics of on-line gradient descent learning for multilayer neural networks*, in *Advances in Neural Information Processing Systems*, edited by D. Touretzky, M. Mozer, and M. Hasselmo (MIT Press, Cambridge, MA, 1995), Vol. 8.
- [218] S. Goldt, M. S. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová, *Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup*, *J. Stat. Mech.* (2020) 124010.
- [219] L. F. Cugliandolo, *Recent applications of dynamical mean-field methods*, *Annu. Rev. Condens. Matter Phys.* **15**, 177 (2024).
- [220] A. Montanari and P. Urbani, *Dynamical decoupling of generalization and overfitting in large two-layer networks*, [arXiv:2502.21269](https://arxiv.org/abs/2502.21269).
- [221] B. Bordon, A. Atanasov, and C. Pehlevan, *A dynamical model of neural scaling laws*, in *Proceedings of the 41st International Conference on Machine Learning*, *Proceedings of Machine Learning Research* Vol. 235, edited by

- R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp (PMLR, 2024), pp. 4345–4382.
- [222] E. Paquette, C. Paquette, L. Xiao, and J. Pennington, *4 + 3 phases of compute-optimal neural scaling laws*, in *Advances in Neural Information Processing Systems*, edited by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Curran Associates, Inc., 2024), Vol. 37, pp. 16459–16537.
- [223] L. Lin, J. Wu, S. M. Kakade, P. L. Bartlett, and J. D. Lee, *Scaling laws in linear regression: Compute, parameters, and data*, in *Advances in Neural Information Processing Systems*, edited by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Curran Associates, Inc., 2024), Vol. 37, pp. 60556–60606.
- [224] K. Oko, Y. Song, T. Suzuki, and D. Wu, *Learning sum of diverse features: Computational hardness and efficient gradient-based training for ridge combinations*, in *Proceedings of Thirty Seventh Conference on Learning Theory*, Proceedings of Machine Learning Research, Vol. 247, edited by S. Agrawal and A. Roth (PMLR, 2024), pp. 4009–4081.
- [225] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra, *Grokking: Generalization beyond overfitting on small algorithmic datasets*, [arXiv:2201.02177](https://arxiv.org/abs/2201.02177).
- [226] N. Rubin, O. Davidovich, and Z. Ringel, *Mitigating the curse of detail: Scaling arguments for feature learning and sample complexity*, [arXiv:2512.04165](https://arxiv.org/abs/2512.04165).
- [227] T. M. Cover, *Elements of Information Theory* (John Wiley & Sons, New York, 1999).
- [228] M. Abadi *et al.*, *TENSORFLOW: Large-scale machine learning on heterogeneous systems* (2015), software available from tensorflow.org.
- [229] M. D. Hoffman and A. Gelman, *The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo*, *J. Mach. Learn. Res.* **15**, 1593 (2014).
- [230] E. Bingham *et al.*, *Pyro: Deep universal probabilistic programming*, *J. Mach. Learn. Res.* **20**, 28:1 (2019).