

PAPER • OPEN ACCESS

Generalization performance of narrow shallow neural networks in the teacher–student setting

To cite this article: Rodrigo Pérez Ortiz *et al* *J. Stat. Mech.* (2026) 043405

View the [article online](#) for updates and enhancements.

You may also like

- [Dynamics of stochastic gradient descent for two-layer neural networks in the teacher–student setup](#)
Sebastian Goldt, Madhu S Advani, Andrew M Saxe et al.
- [Asymptotic generalization error in the online learning of linearized two-layer neural networks](#)
Roman Worschech and Bernd Rosenow
- [Scaling description of generalization with number of parameters in deep learning](#)
Mario Geiger, Arthur Jacot, Stefano Spigler et al.

Generalization performance of narrow shallow neural networks in the teacher–student setting

Rodrigo Pérez Ortiz^{1,7,*} , Gibbs Nwemadji^{2,7} ,
Jean Barbier³ , Federica Gerace¹ ,
Alessandro Ingrosso⁴ , Clarissa Lauditi⁵ 
and Enrico M Malatesta⁶ 

¹ Alma Mater Studiorum–Università di Bologna (Unibo), Bologna, Italy

² International School of Advanced Studies (SISSA), Trieste, Italy

³ The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy

⁴ Radboud University, Nijmegen, The Netherlands

⁵ Harvard University, Cambridge, Massachusetts, United States of America

⁶ Department of Computing Sciences, Bocconi University, Milano, Italy

⁷ Equal contributions.

E-mail: rodrigoep.ortiz@gmail.com

Received 25 November 2025

Accepted for publication 22 March 2026

Published 21 April 2026



Online at stacks.iop.org/JSTAT/2026/043405

<https://doi.org/10.1088/1742-5468/ae5a23>

Abstract. Understanding the generalization properties of neural networks on simple input–output distributions is key to explaining their performance on real datasets. The classical teacher–student setting, where a network is trained on data generated by a teacher model, provides a canonical theoretical test bed. In this context, a complete theoretical characterization of fully connected one-hidden-layer networks with generic activation functions remains missing. In this work, we develop a general framework for such networks with large width, yet much smaller than the input dimension. Using methods from statistical physics,

* Author to whom any correspondence should be addressed.



Original Content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

we derive closed-form expressions for the typical performance of both finite-temperature (Bayesian) and empirical risk minimization estimators in terms of a small number of order parameters. We uncover a transition to a specialization phase, where hidden neurons align with teacher features once the number of samples becomes sufficiently large and proportional to the number of network parameters. Our theory accurately predicts the generalization error of networks trained on regression and classification tasks using either noisy full-batch gradient descent (GD) (Langevin dynamics) or deterministic full-batch GD.

Keywords: learning theory, machine learning, deep learning

Supplementary material for this article is available [online](#)

Contents

1. Introduction and related works	2
2. Setting and main results	5
2.1. ERM, and statistical physics formulation	5
2.2. Main results: closed form formulas for the free entropy, test and training errors	8
3. Numerical experiments and comparison with theoretical predictions ...	10
4. Phase diagram and extensive-width limit	15
5. Discussion and conclusion	18
Acknowledgments	19
References	19

1. Introduction and related works

Predicting theoretically the generalization abilities of complex neural architectures for generic tasks in terms of the number of parameters and training samples is a daunting task. The statistical physics approach has been successful in doing so in the past few decades [1–3], where the performance of a network is characterized by the typical generalization error over synthetic data generated by probabilistic models of increasing statistical complexity. The replica method from the physics of disordered systems [4, 5] is particularly effective for studying feed-forward networks solving supervised learning tasks in the high-dimensional regime. The two classical, complementary settings in this area are the *storage capacity problem* [6], where a network is tasked with memorizing a random input-output mapping, and the *teacher–student* scenario [7], where a rule

mapping labels to input data is instantiated also in a form of a similar random network architecture.

While the analysis of the typical performance of single-layer networks (perceptrons) is—at least in the case of continuous parameters—relatively straightforward and classic [8], the one-hidden layer case is much more challenging. In particular, research has been mainly focused on the case of committee machines, where the second-layer weights are not trained. A variety of works (detailed in the next section) analyzed the information-theoretic limits of supervised learning in committee machines, where a network is trained on P independent N -dimensional examples, with the sample-to-parameter ratio $\alpha := P/N = O_N(1)$ and $N \rightarrow \infty$. The general picture emerging is that generalization performance is controlled by the ratio of the dataset size and the number of network parameters. In particular, *specialization* of hidden neurons in the direction of the ones in the teacher can only occur when the number of samples is of the order of the number of network parameters, i.e. $P \propto NK$.

Classical results exist for the simpler case of *tree* architectures (with non-overlapping receptive fields of the hidden units) and for *fully connected* (FC) committees with sign activation function [7, 9–11]. While recent studies [12, 13] started addressing the case of arbitrary activations, a general theory of learning using empirical risk minimization (ERM) is not yet present in the statistical mechanics literature. This gap is particularly relevant for modeling more realistic machine-learning tasks, where the learner lacks access to the true data-generating distribution—unlike in Bayes-optimal (BO) settings. This framework allows for the exploration of mismatched priors and provides a pathway to understand how emergent neural representations adapt to the underlying structure of the data. In this work, we study the learning problem in the limit where the input dimension N diverge and the number of hidden units K is large, yet much smaller than N . More precisely, we consider the regime $1 \ll K \ll N$, taking $N \rightarrow \infty$ first and subsequently analyzing the large- K limit. While this asymptotic setting formally allows for joint scalings such as $K = o(N)$, our analysis is controlled in the sequential limit described above, and the joint-scaling regime should be understood as a consistent asymptotic extension. This scaling enables a typical-case analysis via the replica method in terms of a finite number of *order parameters* (i.e. sufficient statistics that characterize the generalization error).

We therefore study the typical performance of ERM estimators in a teacher–student setup with generic activation function and fixed (non-trainable) second-layer weights, investigating the emergence of learning phase transitions. We obtain closed form expressions for the weight statistics and the overlap between the teacher and student weight vectors, thus providing a theoretical prediction for the generalization error as a function of sample-to-parameter ratio $\tilde{\alpha} := \alpha/K$, expressed as the ratio of training set size and the total number of adjustable weights in the network. We show that our framework accurately predicts the performance of Langevin dynamics (LD), including in the vanishing-noise limit, and also captures the test-time behavior of networks trained with pure gradient descent (GD). Our analysis identifies distinct learning regimes, notably a computationally hard phase (i.e. a statistical–computational gap) in which posterior sampling methods such as LD exhibit exponentially slow convergence. To probe this regime, we consider an informed initialization of LD, where the student weights are

initialized in the vicinity of the teacher configuration, enabling the dynamics to reach specialized states that are otherwise algorithmically hard to sample.

Related works— Shallow neural networks have been extensively studied using a statistical physics framework [14], where a description of the typical learning and generalization behavior over a joint input-label ensemble can be provided in the limit of large input dimension $N \rightarrow \infty$. The specific phenomenology of such a problem is related to the number of hidden units.

In the $K = O_N(1)$ regime, a two-layer neural network falls into the category of *multi-index functions*. Classical works studied online learning dynamics in such a model, deriving a set of ordinary differential equations for the sufficient statistics of the hidden representations [15–19]. In the *teacher-student* setting, where all information about the teacher is available to the student except for the input-to-hidden-layer weights, a rigorous classification of the difficulty in recovering the target function via approximate message passing (AMP)/TAP-like algorithms has been established in [20]. The work [21] systematically classified SGD performance on two-layer neural networks for any class of multi-index functions based on the *leap complexity* of the target function.

While many questions have been resolved for finite index, the large-index K regime remains to be explored. The regime $1 \ll K \ll N$ has been addressed both in the tree [22–24] and the FC case [7, 10, 25–28]. These classical studies focused on *sign* (or erf [29]) activation function—with either binary or spherically constrained first-layer weights—and i.i.d. inputs. In the FC case, [30] studied the effect of L_2 regularization using an annealed approximation, corresponding to the *large-dataset limit*. A FC model with a bounded $\mathcal{C}^2$ activation function with bounded first and second derivatives was studied from a Bayesian-optimal perspective in [31], where the authors supported replica predictions with rigorous theorems, proposing an efficient inference algorithm based on AMP. In the ERM context, a simple kernel order parameter has been shown to control the storage capacity in the simplified case of a tree committee with generic activation functions [32–34]. Using similar methods, a recent work identified analogous kernel order parameters for generic activation function and a particular choice of second-layer weights [35]. We further note the works [12, 13], which analyzed networks with square-summable activation functions using Hermite expansions and a simplifying approximation in the limit of a very large random dataset. Finally, we also mention a recent burst of interest for the *extensive-width regime* $K = O(N)$ [36–39]. In particular, when comparing our work with [37], a key distinction in the extensive- K regime is the emergence of kernel-like learning in the small-data regime (which does not occur in our narrow-but-large- K setting). Conversely, both scalings converge on *specialization* in the high- α regime. We emphasize that, based on our theoretical observations, for large sample-to-parameter ratio, the Bayes-optimal generalization error (BOE) in our framework is consistent with the extensive width result of [37]. However, exact analytical agreement for all values of α requires $K/N \rightarrow 0$. In other cases, while the BOE may match, the underlying order parameters differ. Our approach remains analytically tractable by reducing the problem to a fixed set of scalar order parameters. This avoids the complex *functional matrix order parameters*— and the necessary integration of replica

theory with random matrix models, specifically the Harish Chandra–Itzykson–Zuber ‘spherical’ integral [40, 41]—required in the $K = O(N)$ case.

Our contributions— In this work, we develop a complete theoretical characterization of the typical performance of a one-hidden-layer neural network in the regime $1 \ll K \ll N$ —again formally interpreted as a sequential limit N large first then K large—in the presence of i.i.d. inputs and responses generated from a teacher network with the same structure.

Our contributions are the following:

- We compute the generalization error and the corresponding learning curves as a function of the number of samples, for narrow and shallow networks with generic activation functions and a generic loss, in the classic teacher–student setting. Unlike previous works, which used an *annealed* calculation [12, 13] for regression, our approach relies on a *quenched* computation of the free entropy using the replica method, whose predictions have been validated through Langevin-based simulations. While our results agree with the annealed computation for very large dataset regime, finite datasets can display a continuous-to-discontinuous behavior, as we report for the ReLU activation function. This computation is of independent interest to the machine learning community, and we anticipate that it will have further applications in learning theory.
- We study the limit in which the Bayesian posterior concentrates on minima of the empirical loss, and experimentally compare its predictions to solutions found by ERM using GD and LD. While LD can sample from the posterior when run long enough, GD is a deterministic optimizer and does not explore the full posterior. Remarkably, we find that GD converges to solutions whose generalization performance is equivalent to those obtained by LD, suggesting that optimization and sampling dynamics can yield statistically similar learned representations.
- We provide a phase diagram highlighting the role of the data-to-parameters ratio $\tilde{\alpha}$ and the L_2 regularization strength λ in shaping the properties of the equilibrium solution. This characterization identifies the optimal regularization required to approach the BO error.
- In the BO setting, taking the $K \rightarrow \infty$ limit after $N \rightarrow \infty$ recovers the BOE reported in [37] for activation functions without second Hermite polynomial in their Hermite decomposition. This suggests that for such activations, the large-width (K) and large-dimension (N) limits commute. For a general activation function, while both theories predict asymptotically (as $\tilde{\alpha} \rightarrow \infty$) the same generalization error, our results exactly match those of [37] at all $\tilde{\alpha}$ only when $\gamma := K/N \rightarrow 0$.

2. Setting and main results

2.1. ERM, and statistical physics formulation

Throughout the paper, we consider the standard supervised learning setup with a synthetic dataset $\mathcal{D} := \{(\mathbf{x}^\mu, y_*^\mu)\}_{\mu=1}^P$, with responses generated by a two-layer teacher neural

network. The samples are thus generated as follows: (i) Construct the target function by sampling entry-wise i.i.d. the first-layer teacher weights $\mathbf{W}^* = (\mathbf{w}_k^* \in \mathbb{R}^N)_{k=1}^K \in \mathbb{R}^{K \times N}$ from P_{W^*} , and the second-layer weights $\mathbf{A}^* = (A_k^*)_{k=1}^K$ from P_A , with finite first and second moments. (ii) The inputs $\mathbf{x}^\mu$ are i.i.d. standard Gaussian vectors: $\mathbf{x}^\mu \sim \mathcal{N}(0, \mathbf{I}_N)$. (iii) The responses/labels $(y_\star^\mu)_{\mu=1}^P$ are generated as

$$\begin{aligned} y_\star^\mu &= \varphi_{\mathbf{A}^*}(\mathbf{W}^* \mathbf{x}^\mu, z^\mu \sqrt{\Delta^*}) \\ &:= f\left(\frac{1}{\sqrt{K}} \sum_{k \leq K} A_k^* \sigma\left(\frac{\mathbf{w}_k^* \cdot \mathbf{x}^\mu}{\sqrt{N}}\right) - B^* \sqrt{K} + z^\mu \sqrt{\Delta^*}\right). \end{aligned} \quad (1)$$

The activation function $\sigma: \mathbb{R} \rightarrow \mathbb{R}$ acts entry-wise and is differentiable almost everywhere, and the readout function $f: \mathbb{R} \rightarrow \mathbb{R}$ defines the task; e.g. in the case of regression one has $f(x) = x$ whereas in the case of binary classification $f(\cdot) = \text{sign}(\cdot)$. The i.i.d. $z^\mu \sim \mathcal{N}(0, 1)$ are label noise whose standard deviation is controlled by Δ^* . For analytical convenience, we introduce the term B^* in the second layer, fixed to remove the mean of the second-layer pre-activation.

Given the dataset $\mathcal{D}$, we study the problem of learning the target in a teacher–student (realizable) setting under the ERM framework, using a student (trainable) model with same parametric form as the target. We focus on the case where the trainable model is already set with the correct readout weights (which are few compared to the inner ones) and we thus skip the dependency on those weights to lighten notations, i.e. $\mathbf{A}^* = \mathbf{A}$ and $\varphi_{\mathbf{A}^*}(\cdot, \cdot) = \varphi_{\mathbf{A}}(\cdot, \cdot)$; our main running example will be all-ones readouts, as is usually the case in the literature on committee machines. The student’s bias B , like B^* , is adjusted to remove the mean of the readout pre-activation. In practice, B is updated separately during learning once per epoch, based on the current state of the weights. The ERM with weight decay we consider reads

$$\begin{aligned} \mathbf{W}_{\text{erm}} &\in \operatorname{argmin} \mathcal{L}_P(\mathbf{W}), \quad \text{with} \\ \mathcal{L}_P(\mathbf{W}) &:= \sum_{\mu \leq P} \ell(y_\star^\mu, \varphi_{\mathbf{A}}(\mathbf{W} \mathbf{x}^\mu, 0)) + \frac{\lambda}{2} \|\mathbf{W}\|_{\text{F}}^2. \end{aligned} \quad (2)$$

We consider the standard notion of mean-square generalization error ϵ_g : for a test sample $(\mathbf{x}^{\text{new}}, y_\star^{\text{new}})$ with same law as the samples in the training data and a given $\mathbf{W}$, we define

$$\epsilon_g(\mathbf{W}) := \frac{1}{4^l} \mathbb{E}_{\mathbf{x}^{\text{new}}, y_\star^{\text{new}}} \left[(\hat{y}_{\mathbf{W}}(\mathbf{x}^{\text{new}}) - y_\star^{\text{new}})^2 \right] \quad (3)$$

with $l=0$ for regression and $l=1$ for classification, $\hat{y}_{\mathbf{W}}(\mathbf{x}) := \varphi(\mathbf{W} \mathbf{x}, 0)$ is the student prediction.

Statistical mechanics formulation— We analyze this setup in the high-dimensional regime where the network architecture (matching the target) has a large width, yet

vanishingly small compared to the input data dimension, and the number of training samples $P = P_N$ scales linearly with the number NK of trainable parameters:

$$N \rightarrow \infty, \text{ with } K \gg 1 \text{ (so that } N \gg K), \text{ and finite} \\ \tilde{\alpha} := \lim_{N \rightarrow \infty} \frac{P_N}{NK} \in (0, \infty). \quad (4)$$

In order to analyze the problem in this asymptotic limit using statistical mechanics, it is useful to introduce a temperature (later taken small) and the associated Gibbs-Boltzmann measure:

$$P_\beta(\mathbf{W} \mid \mathcal{D}) = \frac{1}{\mathcal{Z}_\beta} \exp(-\beta \mathcal{L}_P(\mathbf{W})), \quad (5)$$

where $\mathcal{Z}_\beta$ is the partition function (normalization factor), and $\beta > 0$ is an inverse temperature that controls the measure's concentration around solutions with small loss, going from Bayesian learning at finite β to ERM when it diverges.

The starting point of the statistical physics approach to learning is to compute the log-partition function, the so-called free entropy: $\ln \mathcal{Z}_\beta / (NK)$. We expect it to be self-averaging, i.e. its variance w.r.t. the realization of the problem (teacher and data) vanishes as $N \rightarrow \infty$ (this is proven using standard concentration techniques, see e.g. [42]). This allows us to focus on its expectation value

$$\Phi_{\beta,K} := \lim_{N \rightarrow \infty} \frac{1}{NK} \mathbb{E}_{\mathbf{W}^*, \mathcal{D}, \mathbf{z}} \ln \mathcal{Z}_\beta, \quad (6)$$

where $\mathbb{E}_{\mathbf{z}}$ denotes the expectation w.r.t. to the label noise. Computing $\Phi_{\beta,K}$ yields the order parameters of the problem—sufficient statistics of interest—particularly those related to the generalization error, whose characterization is our main goal. The finite temperature equivalent of the generalization error is the average Gibbs error, defined as

$$\epsilon_{\text{Gibbs}}(\beta) := \frac{1}{4^l} \mathbb{E}_{\mathbf{W}^*, \mathcal{D}, \mathbf{z}, \mathbf{x}_*^{\text{new}}, y_*^{\text{new}}} \left[\left\langle (\hat{y}_{\mathbf{W}}(\mathbf{x}^{\text{new}}) - y_*^{\text{new}})^2 \right\rangle \right], \quad (7)$$

where $\langle \cdot \rangle$ denotes the average w.r.t. the Gibbs-Boltzmann measure (5). Again by the self-averaging property, in the aforementioned high-dimensional limit, the Gibbs error approaches $\epsilon_g(\mathbf{W}_{\text{typ}})$ where $\mathbf{W}_{\text{typ}}$ is a typical sample from the Gibbs measure. This latter error converges to $\epsilon_g(\mathbf{W}_{\text{erm}})$ as $\beta \rightarrow \infty$. This makes the finite (but low) temperature analysis suitable to study the performance of ERM. By the same argument we can approximate the training error of ERM using

$$\epsilon_{\text{tr}}(\beta) := \mathbb{E}_{\mathbf{W}^*, \mathcal{D}, \mathbf{z}} \left[\left\langle \frac{1}{P} \sum_{\mu \leq P} \ell(y_*^\mu, \varphi_{\mathbf{A}}(\mathbf{W} \mathbf{x}^\mu, 0)) \right\rangle \right]. \quad (8)$$

The asymptotic regime in equation (4) formally allows for joint scaling of the network width with the input dimension, such as $K \propto N^\delta$ with $\delta < 1$. However, the replica computation presented below is controlled in the sequential limit, where $N, P \rightarrow \infty$ a fix (but arbitrary) K . In this limit, the free entropy is obtained from an extremization over $K \times K$ overlap matrices. In a second step, we consider the regime of large but finite K . Here, we assume the overlap matrices are parameterized by a uniform diagonal and off-diagonal structure, as suggested in [11] and detailed in the following section. Upon rescaling the free entropy by K , the stationary equations admit a well-defined large- K limit, yielding asymptotically equivalent results. Consequently, the joint-scaling regime ($K \propto N^\delta$ with $\delta < 1$) should be interpreted a heuristic extension consistent with these asymptotic equations, rather than a rigorously derived limit within the current framework.

2.2. Main results: closed form formulas for the free entropy, test and training errors

Replica symmetric free entropy, and order parameters— In order to state our main results we need to introduce some definitions. Define the so-called *replica symmetric (RS) potential*

$$K\Phi_{\beta,K}^{\text{RS}}(m, q, v) := \mathcal{G}_{SI}(m, q, v) + \tilde{\alpha}K\mathcal{G}_E(m, q, v), \quad (9)$$

where m denotes a $K \times K$ matrix whose elements can be expressed in terms of two parameters m_d and m_a as $m = m_d \mathbf{I}_K + (m_a/K) \mathbf{1}_K \mathbf{1}_K^T$ and similarly for v, q . This parametrization of the overlap matrices can be justified in the regime $K \ll N$. By invoking statistical permutation symmetry—naturally suggested by the indistinguishable nature of the hidden units—we expect overlaps between distinct units to concentrate. This allows us to describe the system using a reduced set of scalar order parameters rather than tracking every individual pair. This choice is further constrained by asymptotic self-consistency: to ensure the network's output variance remain $\mathcal{O}(1)$ in the large K limit, these off-diagonal overlaps must scales as $\mathcal{O}(1/K)$. If they were $\mathcal{O}(1)$, the $K(K-1)$ correlations between different units sum would cause a divergence. Thus, the $1/K$ scaling represents a structurally stable state where units are collectively balanced, and where individual fluctuations around their mean become negligible in the thermodynamic limit.

Then, the ‘entropic potential’ is

$$\begin{aligned} \mathcal{G}_{SI}(m, q, v) := & \frac{K}{2} \left[1 + \frac{q_a - m_d^2}{v_d} - (q_d + v_d)\beta\lambda + \ln(2\pi v_d) \right] \\ & - \frac{1}{2} \log \left(\frac{v_d}{v_d + v_a} \right) + \frac{1}{2(v_d + v_a)} \left[q_a - 2m_d m_a \right. \\ & \left. - m_a^2 - \frac{v_a}{v_d} (q_d - m_d^2) - (v_d + v_a)(q_a + v_a)\beta\lambda \right], \end{aligned} \quad (10)$$

whereas the ‘energetic potential’ reads

$$\begin{aligned}
\mathcal{G}_E(m, q, v) &:= \mathbb{E}_{\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I}_K)} \int d\mathbf{y} Z_T(m, q, v; \mathbf{y}, \boldsymbol{\xi}, \mathbf{A}) \ln Z_S(m, q, v; \mathbf{y}, \boldsymbol{\xi}, \mathbf{A}), \\
Z_T &:= \mathbb{E}_{z \sim \mathcal{N}(0, 1)} \int_{\mathbb{R}^K} d\mathbf{T} \mathcal{N}(\mathbf{T}; mq^{-1/2}\boldsymbol{\xi}, \mathbf{I}_K - mq^{-1}\mathbf{m}) \\
&\quad \times \delta(y - \varphi_{\mathbf{A}}(\mathbf{T}, z\sqrt{\Delta^*})), \\
Z_S &:= \int_{\mathbb{R}^K} d\mathbf{Z} \mathcal{N}(\mathbf{Z}; q^{1/2}\boldsymbol{\xi}, v) \exp[-\beta \ell(y, \varphi_{\mathbf{A}}(\mathbf{Z}, 0))],
\end{aligned} \tag{11}$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \Omega)$, denotes the normal p.d.f. with mean vector $\boldsymbol{\mu} \in \mathbb{R}^K$ and covariance matrix Ω . The RS potential allows us to approximate the free entropy through its extremization. Most importantly, the stationary equations constraining its extremizers give us directly access to the sufficient statistics and, consequently, the training and test errors. The RS approximation to the free entropy thus reads

$$\Phi_{\beta, K} = \text{extr}_{m, q, v} \Phi_{\beta, K}^{\text{RS}}(m, q, v). \tag{12}$$

The extremization extr selects a solution (m^*, q^*, v^*) of the stationary equations (called ‘RS equations’), obtained from $\nabla \Phi_{\beta, K}^{\text{RS}} = 0$, maximizing the RS potential. The replica method predicts that with high probability, in the high-dimensional limit the order parameters (called overlaps) have the following limits:

$$\frac{\mathbf{W}^* \mathbf{W}_a^\top}{N} \rightarrow m^*, \quad \frac{\mathbf{W}_a \mathbf{W}_b^\top}{N} \rightarrow q^*, \quad \frac{\mathbf{W}_a \mathbf{W}_a^\top}{N} \rightarrow q_{\text{self}}^* := q^* + v^*, \tag{13}$$

where $\mathbf{W}_a$ and $\mathbf{W}_b$ are conditionally (on $\mathcal{D}$) i.i.d. samples from the Gibbs measure. The sufficient statistics are thus three $K \times K$ matrices whose physical meaning are: q^* , the student-student overlap, quantifies the average correlation between two different weight samples from the Gibbs measure (5); q_{self}^* , the self-student overlap, measures the correlation matrix between learned weight vectors of a single configuration; and finally m^* , the teacher–student overlap, describes the alignment between the learned weights and the teacher’s. In the high-dimensional limit, our theory assumes that these matrices have constant diagonal and off-diagonal entries (up to irrelevant fluctuations), with the latter scaling differently with K . This assumption, along with the application of the central limit theorem following the approach in tree-like committee machines [32, 33], leads to a substantial simplification of the RS entropy and enables the treatment of general activation functions for large K , see the supplementary material (SM) for the complete derivation.

We note that the overlap structure assumed here is similar to that used in the capacity calculations for this architecture [1, 25] and in the teacher–student setting with spherical constraint [11] in the large- K limit. Nevertheless, the absence of the spherical constraint, together with the introduction of the bias term in a non-Bayes optimal setting, distinguishes the present derivation. These additions are crucial to

obtain well-defined RS equations for general activation functions and to explicitly analyze the specialization transition. Moreover, this setting is arguably relevant from a machine-learning perspective, as it corresponds to ERM rather than Bayesian learning.

Generalization error— Let μ_A, ν_A be the first and second moments of P_A , respectively. Let also (t, s) be a centered Gaussian vector with covariance elements $\mathbb{E}[t^2] = \mathcal{E}(1, 1, 1, 0) + \Delta^*$, $\mathbb{E}[s^2] = \mathcal{E}(q_{\text{self},d}^*, q_{\text{self},d}^*, q_{\text{self},d}^*, q_{\text{self},a}^*)$ and $\mathbb{E}[ts] = \mathcal{E}(1, q_{\text{self},d}^*, m_d^*, m_a^*)$ where, in the large width limit $K \rightarrow \infty$ (taken *after* $N \rightarrow \infty$), the function $\mathcal{E}(x_d, y_d, z_d, z_a)$ is

$$\begin{aligned} \mathcal{E}(x_d, y_d, z_d, z_a) &= \nu_A \mathcal{K}_0(x_d, y_d, z_d) - \mu_A^2 \mathcal{K}_0(x_d, y_d, 0) - \frac{z_a \mu_A^2}{x_d y_d} \mathcal{K}_1(x_d, y_d, 0), \\ \mathcal{K}_p(d_1, d_2, a) &:= \mathbb{E}_{(x_1, x_2) \sim \mathcal{N}(0, \Omega)} [(x_1 x_2)^p \sigma(x_1) \sigma(x_2)] \quad \text{with} \quad \Omega = \begin{pmatrix} d_1 & a \\ a & d_2 \end{pmatrix}. \end{aligned} \quad (14)$$

Given (m^*, q^*, v^*) , a solution of the RS equations maximizing the RS potential (9), the RS approximation to the (Gibbs) generalisation error $\lim_{N \rightarrow \infty} \epsilon_{\text{Gibbs}}$ is given by $\mathbb{E}_{(t,s)}[(f(t) - f(s))^2]$. Notice that $\mathcal{K}_0(d_1, d_2, a)$ coincides with the NNGP kernel [43, 44] of large-width neural networks at initialization. Similarly, the RS approximation to the training error $\lim_{N \rightarrow \infty} \epsilon_{\text{tr}}$ reads $\partial_\beta \mathcal{G}_E|_{m^*, q^*, v^*}$.

3. Numerical experiments and comparison with theoretical predictions

As our analysis describes a Bayesian posterior over the first-layer weights with a β -dependent likelihood, our primary algorithm for testing is LD, due to its explicit temperature dependence and theoretical guarantee of sampling the Gibbs measure when it reaches equilibrium. We consider two distinct initializations: *LD Planted Init*, where the student parameters are initialized near the ground truth, and *LD Random Init*, with i.i.d. $\mathcal{N}(0, 1)$ initial weights. Across data regimes, at least one of these LD variants consistently displays a test error matching well the theoretical prediction extracted from a solution of the RS equations that *locally* maximizes $\Phi_{\beta, K}^{\text{RS}}$.

We perform numerical simulations at $K = 10$ and $N = 500$. While the theoretical equations predict a slight shift in the specialization threshold as K increases towards the asymptotic limit, $K = 10$ and $N = 500$ is sufficient to capture the relevant physical features—such as the specialization transition and the emergence of a hard phase—while allowing for extensive LD sampling within reasonable computational limits.

The *global* maximizer corresponds to the ‘equilibrium branch’ of the theory, describing typical student weights dominating the Gibbs measure, which may coexist with other solutions associated with metastable states, as described below. We also compare our $\beta \rightarrow \infty$ theory to GD, identifying regimes where GD performance aligns well with the theory, despite lacking guarantees for solving the non-convex ERM problem. Although neither GD nor LD consistently reach the minimal generalization error of ERM across all $\tilde{\alpha}$, our theoretical framework reliably describes their performance. A single GD run took at most 16 h, while LD required up to 58 h, including GPU and CPU time. Total compute time for all figures is 42 effective GPU hours (see SM).

Generalization performance of narrow shallow neural networks in the teacher-student setting

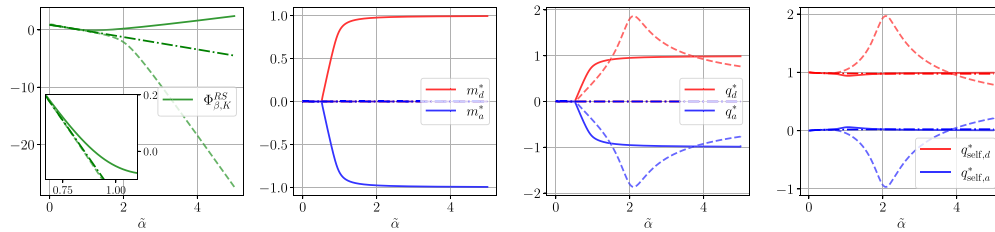


Figure 1. Free entropy $\Phi_{\beta,K}^{\text{RS}}$ (first panel) and order parameters associated with the stationary solutions of the RS equations, as a function of the sample-to-parameter ratio $\tilde{\alpha}$, for quadratic activation function $\sigma(x) = x^2$, MSE loss, weight decay $\lambda = 0.1$, number of hidden units $K = 10$ and $\beta = 10$. Blue curves represent off-diagonal components of the teacher–student overlap m^* , the student–student overlap q^* , and the self-student overlap q_{self}^* . Red curves represent the dominant diagonal components of these quantities. The dotted-dashed lines correspond to the *permutation-symmetric branch* (where $m_d^* = q_d^* = 0$), the solid lines to the *specialization branch* ($m_d^* > 0, q_d^* > 0$) and the dashed one to the *memorization branch* ($m_d^* = 0, q_d^* > 0$).

While our theory extends to broader settings, for the sake of illustration we focus on regression with a mean-square (MSE) loss and classification with Hinge loss. Given the non-convex nature of the problem, our theory yields multiple stationary solutions of the RS free entropy potential, characterized by the order parameters (m^*, q^*, v^*) . We describe them in the next paragraph and figure 1. Once this picture is further detailed, we then provide numerical validations for all predicted observables for regression at finite β (figure 2) and ERM (figure 3). Finally, we numerically validate our theory adapted to study the BO setting for classification (figure 4).

Solutions of the replica symmetric equations and the specialization transition— To begin illustrating our theory, we present a complete description of all stationary solutions of the RS equations in figure 1, for regression with MSE loss $\ell(y, x) = (y - x)^2/2$, and quadratic activation $\sigma(x) = x^2$, for which the solutions are well separated. The phenomenology discussed here remains broadly general for different activation functions. The first panel in figure 1 shows the value of the free entropy, while the order parameters corresponding to each stationary solution are shown in the remaining panels (with matching style of curve). In the particular setting shown here, a unique stationary solution exists up to $\tilde{\alpha} \approx 0.7$, which we refer to as *permutation-symmetric (PS)*. There, the diagonal overlap $q_d^* = 0$, implying that the solution exhibits global permutation symmetry: permuting the hidden units in a typical weight configuration sampled from the Gibbs measure yields an equivalent configuration. At the same time, as we show in the second panel, in this phase each hidden unit of the student is equally correlated with every hidden unit of the teacher, resulting in $m_d^* = 0$. We refer to this as the *PS branch* (dotted-dashed lines). Notice that this PS branch persists for all values of $\tilde{\alpha}$. For $\tilde{\alpha} \gtrsim 0.7$, two new solutions of the saddle point equations appear. The first one, that we name *specialized branch* (solid lines), corresponds to a phase where the hidden units in the student begin to align with specific teacher units as $\tilde{\alpha}$ increases ($m_d^* > 0$, second panel), leading to PS breaking ($q_d^* > 0$, third panel). The second type

of solution (dashed lines) is what we call the *memorization branch*, in analogy to the storage capacity problem: despite exhibiting PS breaking ($q_d^* > 0$, third panel), in this phase the student does not align with the teacher, resulting in poor generalization abilities. As we show in the first panel, for $\tilde{\alpha} > 0.7$, the solution with the largest free entropy corresponds to the *specialized branch*, meaning that the student finally starts learning the teacher rule.

Comparison with experiments— In figure 2, we illustrate the remarkable agreement between the analytical results for the overlaps, equation (13), and the generalization error described in section 2.2.2, with numerical results obtained from LD. Notably, systems with as few as $K=10$ hidden units are already well described by our large-width theory. Previous work using a simplified *annealed* calculation [12] showed that the choice of activation function influences the nature of the learning phase transition. In a similar regression setting, the authors found that networks with ReLU activation undergo a continuous phase transition at a critical value of $\tilde{\alpha}$. We recover this result, as shown in the upper-left panel in figure 2, where we observe a *specialization transition* at $\tilde{\alpha}_{\text{spec}}$. In this large-dataset regime, the quenched computation effectively reproduces the annealed prediction, which corresponds to the small β limit. However, our more accurate *quenched* computation predicts that for sufficiently large β , i.e. close to the ERM setting, ReLU networks exhibit a discontinuous phase transition (see left panel in figure 3). In contrast, for networks with an Erf activation, we observe a discontinuous transition, independently of β (upper-right panel of figure 2 and right panel in figure 3). Discontinuous transitions are associated with the presence of metastable thermodynamic states. Depending on the initialization, these states can trap LD, preventing it from sampling the Gibbs measure. We identify the presence of metastability by running LD with both random and teacher-close initializations of the weights. In the latter case, LD follows the specialized branch for $\tilde{\alpha} \geq \tilde{\alpha}_s$, where the teacher–student overlap increases with the training set size. Conversely, with random initialization, LD approaches the specialization branch at a value $\tilde{\alpha} > \tilde{\alpha}_{\text{spec}}$ and not at the theoretical transition. This signals the possible presence of a *hard phase* for LD, where it fails to equilibrate when the weights are initialized randomly. A theoretical explanation for this may require a finer analysis taking into account the effects of ‘replica symmetry breaking’ in physics jargon [4, 45].

The specific branches that becomes stable around the specialization transition $\tilde{\alpha}_{\text{spec}}$ depends on β . At intermediate temperatures (e.g. $\beta = 10$), the transition occurs between the permutation-symmetric and specialized branches (see right-upper panel of figure 2). In contrast, in the $\beta \rightarrow \infty$ limit, the transition shifts to one between the memorization and specialized branches (see right panel of figure 3). This indicates that at low temperatures the PS branch loses stability to the memorization branch.

The results obtained using LD with random initializations support this; in the $\beta \rightarrow \infty$ limit, both the generalization error and order parameters converge to the corresponding memorization branch for $\tilde{\alpha} < \tilde{\alpha}_{\text{spec}}$ (figure 3). Notably, we observe in the same figure that GD with random initialization recovers the statistics of typical ERM

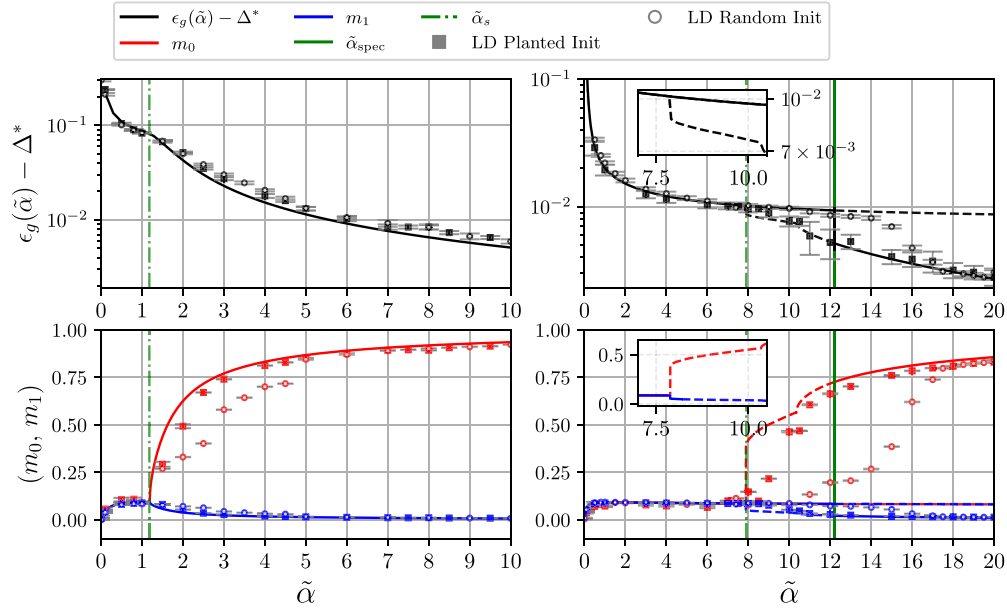


Figure 2. Generalization error ϵ_g and teacher–student overlap for noise-free regression task ($\Delta^* = 0$) as functions of the sample-to-parameter ratio $\tilde{\alpha}$. Results are shown for inverse temperature $\beta = 10$, weight decay $\lambda = 0.1$, and activation functions $\sigma(x) = \text{ReLU}(x)$ (**Left**), which exhibit a continuous specialization transition, and $\sigma(x) = \text{Erf}(x/\sqrt{2})$ (**Right**) which exhibit a discontinuous specialization transition. Theoretical curves are derived from the RS equations. The teacher–student overlaps are parametrized as $m_0^* = m_d^* + m_a^*/K$ (diagonal) and $m_1^* = m_a^*/K$ (off-diagonal). The solid lines represent the equilibrium RS solutions (globally maximizing free entropy), while the dashed lines for the Erf case denote metastable solutions. The specialized branch becomes the equilibrium at $\tilde{\alpha} \geq \tilde{\alpha}_{\text{spec}}$; for Erf, this branch emerges at a spinodal point $\tilde{\alpha}_s < \tilde{\alpha}_{\text{spec}}$. Markers represent averages over 500 LD samples with $K = 10$ and $N = 500$, for a single dataset. Generalization error is computed on 1000 test samples; error bars show standard deviations. Notably, for $\tilde{\alpha} > \tilde{\alpha}_{\text{spec}}$, LD with random initialization fails to follow the RS equilibrium branch, suggesting metastability effects or dynamical barriers not fully captured by the RS equilibrium analysis.

minimizers. Such observation aligns with recent work suggesting that (S)GD concentrates on solutions with probabilities close to the Bayesian posterior [46–49], providing a possible explanation for why GD reliably finds typical ERM minimizers despite non-convex landscape. However, at finite system size, initializing at the planted solution alongside incomplete equilibration may trap the dynamics in atypical low-error states. Consequently, the observed generalization error can dip below the asymptotic BO bound, reflecting the lack of full posterior sampling.

Generalization performance of narrow shallow neural networks in the teacher-student setting

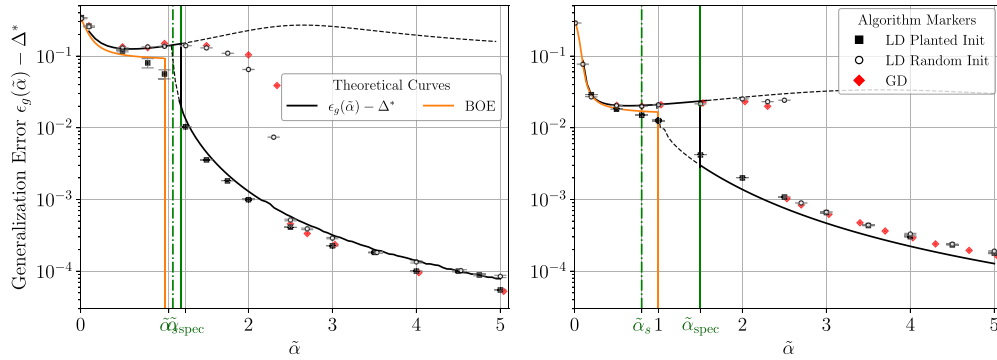


Figure 3. Generalization error ϵ_g for noise-free regression ($\Delta^* = 0$) as a function of the sample-to-parameter ratio $\tilde{\alpha}$. **(Left)** $\sigma(x) = \text{ReLU}(x)$; **(Right)** $\sigma(x) = \text{Erf}(x/\sqrt{2})$. At large β both activations exhibit a discontinuous specialization transition. Results corresponds to ERM ($\beta \rightarrow \infty$ limit) with weight decay $\lambda = 0.01$, where typical configurations are sampled from the memorization state for $\tilde{\alpha} < \tilde{\alpha}_{\text{spec}}$. The equilibrium and metastable RS branches are denoted by solid and dashed black lines, respectively, while the orange line represents the BO error (BOE), which vanishes for $\tilde{\alpha} \geq 1$. Numerical results for a single dataset with $K = 10$ and $N = 500$, are obtained via LD at $\beta = 10^4$ (averaging over 500 samples) and a single GD estimator with random initialization. Generalization error is computed on 1000 test samples; error bars indicate standard deviation. Notably, for $\tilde{\alpha} < \tilde{\alpha}_{\text{spec}}$, LD results with planted initialized can fall below the Bayes-optimal prediction. This apparent violation is likely due to initialization bias and slow relaxation dynamics, which prevent the system from reaching the true equilibrium within the simulation time.

A similar behavior was observed for two-layer neural networks in the extensive-width regime in the BO setting with $\sigma(x) = x^2$ [36]⁸. In our case, we observed that GD reaches the specialization branch at values significantly larger than $\tilde{\alpha}_{\text{spec}}$ —specifically, at $\tilde{\alpha}_{\text{GD}}^{\text{ReLU}} \simeq 2.3$ and $\tilde{\alpha}_{\text{GD}}^{\text{Erf}} \simeq 2.5$. This discrepancy likely stems from subdominant, atypical configurations that attract GD but are not captured by our theory. While the BOE [31] remains a strict lower bound in the thermodynamic limit, finite-size effects and biased (planted) initialization can lead to apparent violations in LD simulations at small $\tilde{\alpha}$ due to incomplete equilibration. Nevertheless, as discussed in the next section, ERM can closely approach the BOE when the regularization λ is properly tuned.

Classification— Analogous comparisons between theory and simulations can be done in the classification case (where $f = \text{sign}$). In figure 4 we show the accuracy ($1 - \epsilon_g$) as a function of $\tilde{\alpha}$ for both ReLU and Erf activations in the noiseless case ($\Delta^* = 0$) found by doing ERM on the hinge loss, in the BO setting (obtained by setting $\lambda\beta = 1$ and $\beta \rightarrow \infty$). In both cases the theory predicts that the equilibrium is given by the specialized branch. This is in contrast with the sign activation, which was investigated

⁸ Maillard *et al* [36] observed a gap between the performance of GD with small regularization and the BO one in the noisy teacher-student regression setting with $\sigma(x) = x^2$. Interestingly, in their paper GD follows a smooth trajectory, which may reflect the memorization branch we described here, a branch that could still exist in the extensive-width regime $K = O(N)$ they consider.

Generalization performance of narrow shallow neural networks in the teacher-student setting

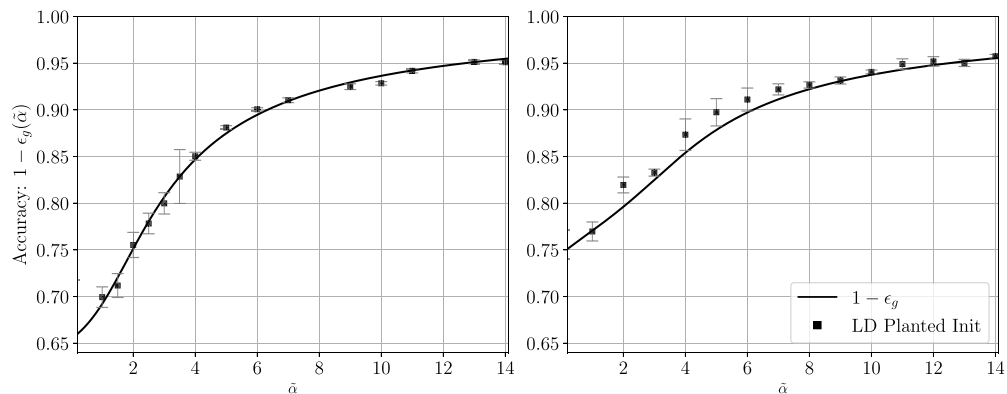


Figure 4. Accuracy as a function of $\tilde{\alpha}$ for classification with $f(x) = \text{sign}(x)$ and no label noise ($\Delta^* = 0$). Activations: **(Left)** $\sigma(x) = \text{ReLU}(x)$ and **(Right)** $\sigma(x) = \text{Erf}(x/\sqrt{2})$. Solid lines: ERM theory with Hinge loss in the Bayes-optimal setting (i.e. $\beta\lambda = 1$ and $\beta \rightarrow \infty$); markers: LD Planted Init. Accuracy is the average fraction of correctly classified samples by the student. Each point averages 500 LD-sampled equilibria at $\beta = 10^4$, $K = 10$, $N = 500$ for one dataset. Student weights are projected onto the unit sphere at each update, and teacher weights have unit norm. Accuracy is computed on 1000 test samples; error bars show standard deviations.

in [10, 26], for which one has a discontinuous transition from the permutation-symmetric to the specialized branch. A study of the small data regime where the size of the dataset is proportional to the size of the input (i.e. $\alpha = P/N = O(1)$), shows that in both those cases the transition from the PS solution is continuous (see SM for additional plots and details).

4. Phase diagram and extensive-width limit

Understanding how the regularization strength and the data-to-parameter ratio shape the equilibrium solution is a central question in learning theory. In this section, we characterize the resulting phase diagram and investigate how the limit $K \rightarrow \infty$ in our setting relates to the $N \rightarrow \infty$ limit, clarifying the interplay between width and input dimension.

Phase diagram— This diagram is a key outcome of the replica analysis: it identifies which of the PS, Memorization, and Specialization learning phases is the equilibrium state, and which are metastable, as a function of training set size and the inverse of the regularization strength. Specifically, it shows when the system admits multiple coexisting phases—some close to yielding BO performance, and others only locally stable—, and moreover allows us to determine the optimal regularization strength in terms of generalization performance as a function of $\tilde{\alpha}$.

The optimal regularization strength depends on the sample-to-parameter ratio $\tilde{\alpha}$, reducing overfitting in low-data regimes while enabling alignment with the teacher in high-data regimes. Contrary to convex optimization problems [50], the replica analysis reveals a discontinuous behavior of the optimal regularization as a function of the training set size, a behavior generally difficult to grasp using statistical learning bounds. In particular, for $\tilde{\alpha} < \tilde{\alpha}_{\text{sp}}^{\text{BO}}$ there coexist two learning phases: Memorization and PS. Although neither describes a phase where the student aligns with the teacher, the latter is still preferable in terms of generalization performance since it matches the BO performance for small $\tilde{\alpha}$. In this regime, a large regularization strength is optimal: it keeps the system in the PS phase, avoiding overfitting and preventing harmful memorization.

For $\tilde{\alpha} > \tilde{\alpha}_{\text{sp}}^{\text{BO}}$, the student starts having enough samples to align with the teacher, causing the specialization phase to emerge and gradually become stable. To enter this phase, λ must be drastically reduced: small enough to allow alignment with the teacher, yet large enough to avoid entering the Memorization phase, which can coexist with specialization in the large-data regime and give rises to metastable configurations where algorithms like GD or LD can easily get trapped. In this regime, the replica analysis shows that the best regularization is found near the boundary in the $(\lambda^{-1}, \tilde{\alpha})$ plane where Memorization loses global stability and Specialization becomes the only stable learning phase. This can be better observed in the right panel of figure 5, where no single value of λ saturates the BOE across all $\tilde{\alpha}$. Based on these fixed- λ curves, we reconstruct in red the effective learning curve that a student trained with the λ_{opt} across $\tilde{\alpha}$ would follow.

The replica analysis thus predicts non-trivial, discontinuous behavior of the optimal regularization as a function of training set size, due to the presence of distinct learning phases—PS, Memorization, and Specialization—across data regimes. This phase diagram provides key theoretical insights into the algorithmic limitations of standard training procedures such as GD and LD, illustrating when the specialized phase is metastable, i.e. exists but hard to reach via common optimization methods.

The extensive width limit— It is natural to ask how the following two limits are related: the one treated in the present paper, namely $K \rightarrow \infty$ *after* $N \rightarrow \infty$, and the small-but-extensive width limit considered in [37], where $N, K \rightarrow \infty$ jointly with $K = \gamma N$ and $\gamma \ll 1$ but independent of N . In figure 6, we address this question in the BO setting by plotting the BOE for three activation functions, from left to right:

$$\sigma(x) = \text{Erf}\left(\frac{x}{\sqrt{2}}\right), \quad \sigma(x) = He_3(x), \quad \sigma(x) = He_2(x) + \frac{1}{\sqrt{6}}He_3(x),$$

where we use the normalized probabilist's Hermite polynomials $He_n(x) := \frac{(-1)^n}{\sqrt{n!}} e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2}$. Across the panels, the solid lines for different values of γ correspond to the extensive-width prediction of [37], while the dashed lines represent the results obtained from our theory of narrow-network. In the left and middle panels of figure 6, the two theories coincide for all values of the sample-to-parameter ratio $\tilde{\alpha}$. In

Generalization performance of narrow shallow neural networks in the teacher-student setting

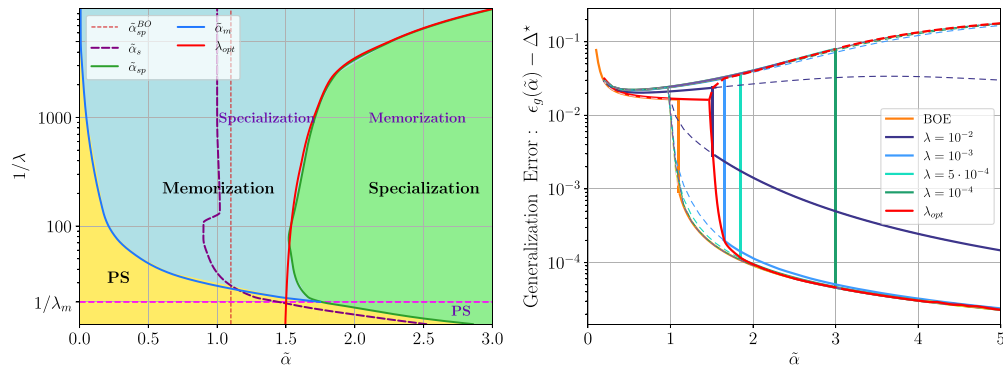


Figure 5. Phase diagram (**Left**) and generalization error (**Right**) of a two-layer student neural network with activation function $\text{Erf}(x/\sqrt{2})$, trained on a regression task via ERM ($\beta \rightarrow \infty$) with MSE loss $\ell(x, y) = (x - y)^2/2$, as a function of the inverse regularization λ^{-1} and the sample-to-parameter ratio $\tilde{\alpha}$. The training data are generated by a teacher network with matching architecture, label noise $\Delta^* = 10^{-4}$, and $K = 10$ hidden units. (**Left**) Phase diagram in the $(\tilde{\alpha}, \lambda^{-1})$ plane showing three learning regimes. In the **permutation-symmetric (PS)** phase, hidden units across different student networks—initialized differently but trained on data generated by the same teacher—remain uncorrelated ($q_d^* \simeq 0$); this occurs for small $\tilde{\alpha}$ and $\lambda > \lambda_m$. In the **memorization** phase, hidden units across different student networks—initialized differently but trained on data generated by the same teacher—become mutually correlated, yet remain uncorrelated with the teacher’s hidden units ($q_d^* > 0$, $m_d^* \simeq 0$); this corresponds to intermediate $\tilde{\alpha}$ and $\lambda < \lambda_m$. The bold blue line at $\tilde{\alpha}_m$ marks the onset of stability of this solution. In the **specialization** phase, student units align with teacher units ($m_d^* > 0$). The bold green line $\tilde{\alpha}_{sp}$ marks the stability threshold, while the dashed purple line $\tilde{\alpha}_s$ indicates where the specialized solution first appears (metastable until crossing the green boundary). The dashed vertical red line $\tilde{\alpha}_{sp}^{BO}$ is the BO specialization threshold, and the horizontal dashed pink line λ_m marks where the regularization becomes beneficial as a network can escape to go from the **PS** to **memorization** region. Bold purple scripts indicates that these solutions are present only as metastable states. (**Right**) Bayesian optimal error (BOE, orange) and Gibbs generalization error as functions of $\tilde{\alpha}$. Solid curves denote equilibrium solutions, while dashed curves represent metastable ones. Across different $\tilde{\alpha}$ (highlighted in red in both panels), λ_{opt} denotes the value of the regularization strength for which the student’s generalization error best approaches the BOE.

these cases, the collapse of the predictions for different γ in [37] is a consequence of the absence of the second Hermite coefficient in the activation functions. In particular, the specialization transition matches that of the extensive-width case. In the right panel, we observe that as $\gamma \rightarrow 0$ the two theories agree across all data-to-parameter ratios (here, taking $\gamma = 1/50$ already yields a quasi-perfect match), while for non-vanishing γ they (quickly) converge to the same limit as $\tilde{\alpha} \rightarrow \infty$. These observations suggest that

Generalization performance of narrow shallow neural networks in the teacher-student setting

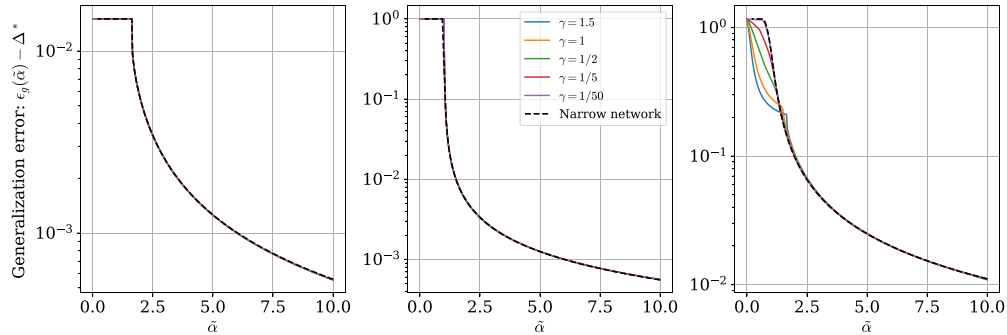


Figure 6. Comparison of the generalization error $\epsilon_g - \Delta^*$ as a function of the sample-to-parameter ratio $\tilde{\alpha} = P/(NK)$ in the Bayes-optimal setting, between our theory for narrow-networks (dashed black line, corresponding to the limit $K \rightarrow \infty$ taken after $N \rightarrow \infty$) and the extensive-width theory of [37] (colored lines, corresponding to the limit $K, N \rightarrow \infty$ with $K = \gamma N$). **(Left)** $\sigma(x) = \text{Erf}(x/\sqrt{2})$ and $\Delta^* = 0.005$, **(Middle)** $\sigma(x) = He_3(x)$ and $\Delta^* = 0.005$, **(Right)** $\sigma(x) = He_2(x) + \frac{1}{\sqrt{6}}He_3(x)$ and $\Delta^* = 0.1$.

the large- N and large- K limits commute for any $\tilde{\alpha}$ when the activation function has zero second Hermite coefficient, or ‘quasi commute’ when $\tilde{\alpha}$ diverges.

5. Discussion and conclusion

We developed a theory for the typical generalization performance of a narrow one-hidden-layer network in the teacher–student setting, under both finite-temperature Bayesian learning and ERM. Our theory highlights the presence of distinct phases and thermodynamic states, including metastable ones that can trap the dynamics of full-batch GD, both with and without explicit noise. We further showed that the activation function is responsible for the appearance of different types of transitions in a Bayesian setting, in a manner that depends on the temperature parameter. Interestingly, we identified a memorization phase in which the student network breaks permutation symmetry among its hidden units without aligning to the teacher’s weights. This study differs from the recent work of [12, 13], which focuses on the large-temperature, infinite-data regime, and from [35], which investigates the storage capacity problem. It also goes beyond classical committee-machine analysis [7, 10, 25–29] by offering a theoretical framework valid for arbitrary inner and outer activation functions with regularized weights.

Our work paves the way for additional investigations of more complex network architectures and input distributions. A key limitation of present analysis is the focus on i.i.d. inputs, which considerably simplifies analytical derivations. Extending the framework to structured inputs is crucial to model the learning performance on real datasets and is left for future work. For instance, we expect our analytical approach to be generalizable to input distributions modeled as mixtures of Gaussian with generic mean vectors and covariance matrices.

Learning in two-layer networks in the $K \ll N$ regime has been recently analyzed using dynamical mean field theory for gradient flow [51]. The authors consider both first- and second-layer learnable weights and identify different dynamical regimes using a separation of time-scales argument. It would be interesting to characterize the steady states of such learning dynamics, in relation to the equilibrium distribution induced by Langevin-based sampling of the weight posterior. Doing so would require extending our framework to the case of plastic second-layer weights, which we treated as known (quenched) for simplicity. This extension also help in fully clarifying the difference between the $K \ll N$ case and the so-called extensive-width regime, i.e. with $K = O(N)$ and $P = O(N)$, see [51, 52] and references therein.

Acknowledgments

F.G. is supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU. F.G. also acknowledges GNFM-Indam. E.M.M. acknowledges the MUR-Prin 2022 funding Prot. 20229T9EAT, financed by the European Union (Next Generation EU). J.B. and R.P. were funded by the European Union (ERC, CHORAL, project number 101039794). Views and opinions expressed are however 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. C.L. is supported by DARPA Award DIAL-FP-038, and The William F. Milton Fund from Harvard University. G.N. thanks Rudy Skerk for insightful discussions and comparative results for the extensive width setting.

ORCID iDs

Rodrigo Pérez Ortiz  0009-0003-5653-9363

Gibbs Nwemadji  0000-0001-7994-4810

Jean Barbier  0000-0002-2652-6727

Federica Gerace  0000-0002-8992-5506

Alessandro Ingrosso  0000-0001-5430-7559

Clarissa Lauditi  0000-0001-8472-7562

Enrico M Malatesta  0000-0001-8558-6175

References

- [1] Seung H S, Sompolinsky H and Tishby N 1992 Statistical mechanics of learning from examples *Phys. Rev. A* **45** 6056
- [2] Engel A 2001 *Statistical Mechanics of Learning* (Cambridge University Press)
- [3] Zdeborová L 2020 Understanding deep learning is also a job for physicists *Nat. Phys.* **16** 602–4
- [4] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* vol 9 (World Scientific Publishing Company)
- [5] Charbonneau P, Marinari E, Parisi G, Ricci-terseghi F, Sicuro G, Zamponi F and Mezard M 2023 *Spin Glass Theory and far Beyond: Replica Symmetry Breaking After 40 Years* (World Scientific)
- [6] Gardner E and Derrida B 1988 Optimal storage properties of neural network models *J. Phys. A: Math. Gen.* **21** 271

- [7] Schwarze H and Hertz J 1992 Statistical mechanics of learning in a large committee machine *Advances in Neural Information Processing Systems* p 5
- [8] Gardner E 1988 The space of interactions in neural network models *J. Phys. A: Math. Gen.* **21** 257
- [9] Barkai E, Hansel D and Sompolinsky H 1992 Broken symmetries in multilayered perceptrons *Phys. Rev. A* **45** 4146–61
- [10] Schwarze H 1993 Learning a rule in a multilayer neural network *J. Phys. A: Math. Gen.* **26** 5781
- [11] Schwarze H and Hertz J 1992 Generalization in a large committee machine *Europhys. Lett.* **20** 375
- [12] Oostwal E, Straat M and Biehl M 2021 Hidden unit specialization in layered neural networks: ReLU vs. sigmoidal activation *Physica A* **564** 125517
- [13] Citton O, Richert F and Biehl M 2025 Phase transition analysis for shallow neural networks with arbitrary activation functions *Physica A* **660** 130356
- [14] Cui H 2025 High-dimensional learning of narrow neural networks *J. Stat. Mech.* 023402
- [15] Saad D and Solla S A 1995 On-line learning in soft committee machines *Phys. Rev. E* **52** 4225–43
- [16] Saad D and Solla S A 1995 Exact solution for on-line learning in multilayer neural networks *Phys. Rev. Lett.* **74** 4337–40
- [17] Riegler P and Biehl M 1995 On-line backpropagation in two-layered neural networks *J. Phys. A: Math. Gen.* **28** L507
- [18] Biehl M and Schwarze H 1995 Learning by on-line gradient descent *J. Phys. A: Math. Gen.* **28** 643
- [19] Goldt S, Advani M, Saxe A M, Krzakala F and Zdeborová L 2019 Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup *Advances in Neural Information Processing Systems* vol 32, ed H Wallach, H Larochelle, A Beygelzimer, F d' Alché-Buc, E Fox and R Garnett (Curran Associates, Inc.)
- [20] Troiani E, Dandi Y, Defilippis L, Zdeborová L, Loureiro B, and Krzakala F 2024 Fundamental limits of weak learnability in high-dimensional multi-index models (arXiv:2405.15480)
- [21] Abbe E, Adsera E B and Misiakiewicz T 2023 Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics *The 36th Annual Conf. on Learning Theory* (PMLR) pp 2552–623
- [22] Monasson R and Zecchina R 1995 Learning and generalization theories of large committee-machines *Mod. Phys. Lett. B* **9** 1887–97
- [23] Sollich P 1996 Learning from minimum entropy queries in a large committee machine *Phys. Rev. E* **53** R2060
- [24] Urbanczik R 1994 Storage capacity of the tree-structured committee machine with discrete weights *Europhys. Lett.* **26** 233
- [25] Engel A, Köhler H M, Tschepke F, Vollmayr H and Zippelius A 1992 Storage capacity and learning algorithms for two-layer neural networks *Phys. Rev. A* **45** 7590
- [26] Schwarze H and Hertz J 1993 Discontinuous generalization in large committee machines *Advances in Neural Information Processing Systems* p 6
- [27] Schwarze H and Hertz J 1993 Learning from examples in fully connected committee machines *J. Phys. A: Math. Gen.* **26** 4919
- [28] Urbanczik R 1997 Storage capacity of the fully-connected committee machine *J. Phys. A: Math. Gen.* **30** L387
- [29] Ahr M, Biehl M and Urbanczik R 1999 Statistical physics and practical training of soft-committee machines *Eur. Phys. J. B* **10** 583–8
- [30] Ahr M, Biehl M and Schlösser E 1999 Weight-decay induced phase transitions in multilayer neural networks *J. Phys. A: Math. Gen.* **32** 5003
- [31] Aubin B, Maillard A, Krzakala F, Macris N and Zdeborová L 2018 The committee machine: Computational to statistical gaps in learning a two-layers neural network *Advances in Neural Information Processing Systems* p 31
- [32] Baldassi C, Malatesta E M and Zecchina R 2019 Properties of the Geometry of Solutions and Capacity of Multilayer Neural Networks with Rectified Linear Unit Activations *Phys. Rev. Lett.* **123** 170602
- [33] Zavatone-Veth J A and Pehlevan C 2021 Activation function dependence of the storage capacity of treelike neural networks *Phys. Rev. E* **103** L020301
- [34] Annesi B L, Malatesta E M and Zamponi F 2025 Exact full-RSB SAT/UNSAT transition in infinitely wide two-layer neural networks *SciPost Phys.* **18** 118
- [35] Nishiyama S and Ohzeki M 2025 Solution space and storage capacity of fully connected two-layer neural networks with generic activation functions *J. Phys. Soc. Japan* **94** 014802
- [36] Maillard A, Troiani E, Martin S, Zdeborová L and Krzakala F 2024 Bayes-optimal learning of an extensive-width neural network from quadratically many samples *Advances in Neural Information Processing Systems* vol 37, ed A Globerson, L Mackey, D Belgrave, A Fan, U Paquet, J Tomczak and C Zhang (Curran Associates, Inc) pp 82085–132

- [37] Barbier J, Camilli F, Nguyen M-T, Pastore M, and Skerk R 2025 Statistical physics of deep learning: optimal learning of a multi-layer perceptron near interpolation (arXiv:[2510.24616](#))
- [38] Montanari A and Urbani P 2025 Dynamical decoupling of generalization and overfitting in large two-layer networks (arXiv:[2502.21269](#))
- [39] Ren Y, Nichani E, Wu D, and Lee J D 2025 Emergence and scaling laws in sgd learning of shallow neural networks (arXiv:[2504.19983](#))
- [40] Itzykson C and Zuber J B 1986 The planar approximation II **21** 411–21
- [41] Guionnet A and Zeitouni O 2002 Large deviations asymptotics for spherical integrals *J. Funct. Anal.* **188** 461–515
- [42] Barbier J, Krzakala F, Macris N, Miolane L and Zdeborová L 2019 Optimal errors and phase transitions in high-dimensional generalized linear models *Proc. Natl Acad. Sci.* **116** 5451–60
- [43] Neal R M 1996 *Priors for Infinite Networks* (Springer) pp 29–53
- [44] Williams C 1996 Computing with infinite networks *Advances in Neural Information Processing Systems* vol 9, ed M C Mozer, M Jordan and T Petsche (MIT Press)
- [45] Antenucci F, Franz S, Urbani P and Zdeborová L 2019 Glassy nature of the hard phase in inference problems *Phys. Rev. X* **9** 011020
- [46] Hennick M and De Baerdemacker S 2025 Almost Bayesian: the fractal dynamics of stochastic gradient descent (arXiv:[2503.22478](#))
- [47] Mingard C, Valle-Pérez G, Skalse J and Louis A A 2021 Is sgd a Bayesian sampler? well, almost *J. Mach. Learn. Res.* **22** 1–64
- [48] Smith S L, Duckworth D, Rezchikov S, Le Q V, and Sohl-Dickstein J 2018 Stochastic natural gradient descent draws posterior samples in function space (arXiv:[1806.09597](#))
- [49] Mandt S, Hoffman M D and Blei D M 2017 Stochastic gradient descent as approximate Bayesian inference *J. Mach. Learn. Res.* **18** 1–35
- [50] Krogh A and Hertz J 1991 A simple weight decay can improve generalization *Advances in Neural Information Processing Systems* P 4
- [51] Li Q and Sompolinsky H 2021 Statistical mechanics of deep linear neural networks: the backpropagating kernel renormalization *Phys. Rev. X* **11** 031059
- [52] Pacelli R, Ariosto S, Pastore M, Ginelli F, Gherardi M and Rotondo P 2023 A statistical mechanics framework for Bayesian deep neural networks beyond the infinite-width limit *Nat. Mach. Intell.* **5** 1497–507