Bayes-optimal Learning of Deep Random Networks of Extensive-width

Hugo Cui [1]   Florent Krzakala [2]   Lenka Zdeborová [1]

Abstract

We consider the problem of learning a target function corresponding to a deep, extensive-width, non-linear neural network with random Gaussian weights. We consider the asymptotic limit where the number of samples, the input dimension and the network width are proportionally large. We propose a closed-form expression for the Bayes-optimal test error, for regression and classification tasks. We further compute closed-form expressions for the test errors of ridge regression, kernel and random features regression. We find, in particular, that optimally regularized ridge regression, as well as kernel regression, achieve Bayes-optimal performances, while the logistic loss yields a near-optimal test error for classification. We further show numerically that when the number of samples grows faster than the dimension, ridge & kernel methods become suboptimal, while neural networks achieve test error close to zero from quadratically many samples.

1. Introduction

Learning with neural networks has proven to be an extraordinarily versatile tool to approximate (learn) non-trivial functions from data. Many fundamental theoretical questions, however, remain open. For instance, the determination, for a given target function, of just *how many* training data samples are needed in order to learn the target to a given precision? This is tantamount to determining the minimal error that can be achieved from a training set of a given size.

While for a generic target function and generic training set this question is very challenging, valuable insight can be accessed by studying simplified settings with Gaussian input data and specific target functions with known functional forms. Of particular interest is the rich class of functions given by *random neural networks*. The lowest achievable test error is known to be obtained through Bayesian inference of the parameters of the target function, assuming (as we will) the distribution of the parameters is given. The *Bayes-optimal* test error corresponds to the information-theoretically minimal test error that any algorithm can achieve. In the context of Gaussian data, with target functions being *single-layer random neural networks* the problem was studied as early as in (Opper & Haussler, 1991; Seung et al., 1992; Watkin et al., 1993; Schwarze, 1993). More recently, (Barbier et al., 2017) provide a rigorous characterization of the Bayes-optimal error in the asymptotic proportional regime, where the number of samples is proportional to the input dimension and both of them are large with a fixed ratio α. These results were then extended in (Schwarze, 1993; Aubin et al., 2018) to neural networks with one *narrow* hidden layer, whose width remains of order one in the above limit of large dimension and a proportional number of samples.

In practice, neural networks are trained using Empirical Risk Minimization (ERM) methods, and it is hence also important to know whether those methods are able to achieve the Bayes-optimal error. (Thrampoulidis et al., 2018; Montanari et al., 2019; Hastie et al., 2019; Mei & Montanari, 2019; Aubin et al., 2020; Loureiro et al., 2021), between others, addressed this question for Gaussian data, providing closed-form formulas for the ERM test error for generalized *linear* models for target functions corresponding to single-layer neural network with random weights from a number of samples proportional to the dimension.

Here, we pursue these lines of work and study a target function given by a *deep non-linear neural network with random weights*, in the limit where the layers-widths and the input dimension are comparably large, hereafter referred to as the *extensive-width* regime. We call such target function the *deep extensive-width random network*. We consider Gaussian input data. Our main question is the characterization of the test error that can be achieved information-theoretically from a given number of samples, as well as its reachability with ERM approaches. While the assumptions of Gaussian input data and the prescribed target function seem far from current machine-learning practice, from a theoretical point

---

[1]Statistical Physics Of Computation lab, Institute of Physics, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland [2]Information Learning and Physics lab, Institute of Electrical Engineering, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland. Correspondence to: Hugo Cui <hugo.cui@epfl.ch>.

of view, these questions remain challenging and widely open even in such a simplified setting (even for a single hidden layer). It is hard to imagine that we could obtain a plausible theory of deep learning without being able to answer such questions first.

Main contributions   For the target function corresponding to the deep extensive-width random network and random Gaussian input data we obtain the following results:

• We conjecture a closed-form characterization for the asymptotic Bayes-optimal error, for regression and classification tasks, in the proportional regime where the number of samples n scales linearly with the input dimension d.
• A fundamental step in our derivation, of independent interest, is the deep (Bayes) Gaussian Equivalence Property conjecture (GEP) , which specifies the Gaussian statistics of the output of deep networks whose weights are Gaussian, or sampled from the Bayes posterior. We show how the GEP follows from Bayes theory and the asymptotic concentration of random variables in the proportional regime.
• We contrast the Bayes-optimal test error to test errors achieved by simple ERM methods. For regression, ridge and kernel regression are found to achieve the Bayes-optimal mean-squared error, provided they are optimally regularized. An explicit formula for optimal regularization is provided. These results establish that it is impossible to learn more than a linear estimator of the target extensive-width network from linearly many samples. In the case of classification, logistic and ridge classification are found to yield test errors close (but not equal) to the Bayes error.
• We provide a numerical exploration of the regime where the number of samples n tends to infinity *faster* than linearly with the input dimension d, in which the deep (Bayes) GEPs can no longer be employed. We show that ridge and kernel methods then cease to be optimal, while gradient-trained neural networks manage to almost perfectly learn the target, evidencing the superiority of neural nets.

A repository with the code employed in the present work can be found here.

1.1. Related works
Bayesian learning of neural networks   It is well known that Bayesian learning using networks of infinite width (i.e. width much larger than the number of samples and the input dimension) is equivalent to kernel regression (Neal, 1994; Lee et al., 2018; 2019; de G. Matthews et al., 2018; Hron et al., 2020). A theoretical analysis for extensive-width, however, proved for a long time a challenging endeavor. (Yaida, 2019; Roberts et al., 2021; Zavatone-Veth et al., 2021) computed (perturbative) first-order corrections to the mean test error with respect to the infinite width limit, but only accommodate a finite number of training samples. The recent work of (Zavatone-Veth et al., 2022; Hanin & Zlokapa, 2022) respectively provide an asymptotic and non-

asymptotic study of Bayesian learning, but are limited to linear activations. (Li & Sompolinsky, 2021) and (Ariosto et al., 2022) conjecture that in the proportional regime, i.e. $n \sim d$, the estimator yielded by extensive-width networks with ReLU or sign activations is still given by the associated Gaussian Process (GP) kernel, with the width only rescaling the variance term in the test error. We note that these works rely on a heuristic Gaussianity assumption and provide expressions depending explicitly on the entire dataset. Here instead we address specifically the Bayes-optimal performance for a random network target function and Gaussian inputs, which allows us to provide closed-form *scalar* formulae and leverage the principled GEP to characterize the statistics. Finally, while all the previously cited work study the case of Bayesian regression with a square log likelihood, the present work also covers classification settings.

Replica method in ML   The replica method has been applied in a sizeable body of work to access asymptotic characterizations of the test error (Bayes or ERM) in a variety of setups (Seung et al., 1992; Watkin et al., 1993). While being heuristic, its predictions have been proven rigorously in many cases, e.g. (Talagrand, 2006; Barbier et al., 2017). This toolbox has been successfully deployed to analyze architectures with one trainable layer, including generalized linear models (Advani & Ganguli, 2016; Aubin et al., 2020; Cui et al., 2019; Maillard et al., 2020; Loureiro et al., 2021), narrow networks with frozen readout (Aubin et al., 2018), random features (RF) (Gerace et al., 2020) and kernel methods (Canatar et al., 2020; Cui et al., 2021; 2022). Recently (Zavatone-Veth et al., 2022) studied the multiple layers case, in the framework of linear networks. Here we go a step further and analyze deep non-linear networks.

The proportional regime   The proportional $n \sim d$ regime has been investigated for shallow networks in a sizeable body of work, leveraging tools like the convex gaussian minimax theorem (Thrampoulidis et al., 2018; Aubin et al., 2020; Loureiro et al., 2021; Montanari et al., 2019), random matrix theory (El Karoui, 2008; Pennington & Worah, 2019; Louart et al., 2017) or approximate message-passing (Aubin et al., 2018; Gabrié, 2019), in addition to the replica method.

Gaussian Equivalence   The equivalence between the asymptotic test error of simple ERM algorithms with that of the associated problem where the data samples are replaced by Gaussian covariates with matching population covariance has been observed in many situations, starting with the seminal work (El Karoui, 2008) on kernel matrices. In particular, (Goldt et al., 2021; 2022; Montanari & Saeed, 2022; Hu & Lu, 2022a) have proven a *Gaussian Equivalence* principle that shows that, in the proportional regime, one can often replace projected data with Gaussian ones. Such equivalences were used, for instance, in (Loureiro et al., 2021) to characterize the ERM test error in a variety of setups, in terms

solely of the population covariances of the target/learner networks. Concomitant works (Schröder et al., 2023; Bosch et al., 2023) characterize the Gaussian universality of the test error of deep learners with fixed random weights and trainable readout.

2. Setting

We consider the problem of learning from a train set $\mathcal{D} = \{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^n$, with n independently sampled Gaussian covariates $\mathbf{x}^\mu \in \mathbb{R}^d \sim \mathcal{N}(0, \Sigma)$. The covariance Σ is assumed to admit a well-defined limiting spectral distribution μ as $d \to \infty$ with finite non-zero first and second moments. The labels y^μ are assumed to be generated by a L-layers deep network with random weights. Denoting $\xi \sim \mathcal{N}(0, \Delta)$ a Gaussian additive output noise, we have

$$y^\mu = f_\star \Big[\frac{1}{\sqrt{k_L}} \mathbf{a}_\star^\top \underbrace{(\varphi_L^\star \circ \cdots \circ \varphi_2^\star \circ \varphi_1^\star)}_{L}(\mathbf{x}^\mu) + \xi \Big], \quad (1)$$

$$\text{with layers } \varphi_\ell^\star(\mathbf{x}) = \sigma_\ell \left(\frac{1}{\sqrt{k_{\ell-1}}} W_\ell^\star \cdot \mathbf{x} \right).$$

$(\sigma_\ell)_{\ell=1}^L$ is a sequence of activation functions, which are assumed to be odd for simplifying technical reasons. The readout function $f_\star(\cdot)$ will be taken to be the identity function for regression, and the sign function for classification. The width of the ℓ-th layer is denoted k_ℓ, and the associated weight matrix is $W_\ell^\star \in \mathbb{R}^{k_\ell \times k_{\ell-1}}$, with elements sampled i.i.d from $\mathcal{N}(0, \Delta_\ell)$. Similarly, the readout weight vector $a_\star$ is sampled from $\mathcal{N}(0, \Delta_a \mathbb{I}_{k_L})$.

We wish to characterize the Bayes-optimal test errors when learning on data produced by the target function (1). We consider that all the hyperparameters $L, k_\ell, \sigma_\ell, \Delta_a, \Delta_\ell, \Delta$ of the architecture (1) are known, but the weights $a_\star, \{W_\ell^\star\}_{\ell=1}^L$ are not known to the learner.

Throughout Sec. 3 to 4, we consider the *proportional regime*: the high-dimensional asymptotic limit where $\forall \ell, \ n, d, k_\ell \to \infty$ with fixed $\mathcal{O}(1)$ ratios $\alpha \equiv {}^n/d$ and $\gamma_\ell^\star \equiv {}^{k_\ell}/d$. The parameters $L, \Delta_\ell, \Delta_a, \Delta$ are assumed to be $\mathcal{O}(1)$. The *quadratic regime* $n \sim d^2 \sim k_\ell^2$ is numerically explored in Sec. 5. It is known that learning a target of large width k (resp. using a network of large width) with a finite number of samples $n = \mathcal{O}_k(1)$ simplifies drastically to the problem of learning a Gaussian process (resp. kernel regression) (Neal, 1994; Lee et al., 2018; de G. Matthews et al., 2018). We consider here widths $\{k_\ell\}_{\ell=1}^L$ at most *comparable*, and not very large compared to, the input dimension d and the number of samples n, which makes for a non-trivial, and much richer, learning problem.

3. Bayes-optimal Error

The Bayes-optimal error for data generated using the target function (1) is achieved by sampling the weights $\mathbf{a}, \{W_\ell\}_\ell$ from a posterior measure involving a neural network of matching architecture. We thus define

$$\hat{y}(\mathbf{x}) = \frac{1}{\sqrt{k_L}} \mathbf{a}^\top \underbrace{(\varphi_L \circ \varphi_{L-1} \circ \cdots \circ \varphi_2 \circ \varphi_1)}_{L}(\mathbf{x}), \quad (2)$$

$$\text{with layers } \varphi_\ell(\mathbf{x}) = \sigma_\ell \left(\frac{1}{\sqrt{k_{\ell-1}}} W_\ell \cdot \mathbf{x} \right). \quad (3)$$

The Bayes-optimal Mean Squared Error (MSE) is then

$$\epsilon_{g,\text{reg}}^{\text{BO}} = \mathbb{E}_{\mathcal{D}, \{W_\ell^\star\}_{\ell=1}^L, \mathbf{a}_\star} \mathbb{E}_{\mathbf{x}, y} \Big[\big(y - \langle \hat{y}(\mathbf{x}) \rangle_{\mathbf{a}, \{W_\ell\}_{\ell=1}^L \sim \mathbb{P}} \big)^2 \Big] \quad (4)$$

where $\mathbf{x}, y$ should be understood as a test sample. The Bayes-optimal classification error (defined as the probability to wrongly classify a test sample) is given by

$$\epsilon_{g,\text{class}}^{\text{BO}} = \mathbb{E}_{\mathcal{D}, \{W_\ell^\star\}_{\ell=1}^L, \mathbf{a}_\star}$$
$$\mathbb{P}_{\mathbf{x}, y} \Big[y \neq \text{sign} \big(\langle \text{sign}(\hat{y}(\mathbf{x})) \rangle_{\mathbf{a}, \{W_\ell\}_{\ell=1}^L \sim \mathbb{P}} \big) \Big]. \quad (5)$$

In (4,5), the learner network is averaged over the posterior

$$\mathbb{P} \Big[\mathbf{a}, \{W_\ell\}_{\ell=1}^L | \mathcal{D} \Big] \propto e^{-\frac{||\mathbf{a}||^2}{2\Delta_a} - \sum_\ell \frac{||W_\ell||_F^2}{2\Delta_\ell}}$$
$$\times \prod_{\mu=1}^n \int \frac{d\xi e^{-\frac{1}{2\Delta}\xi^2}}{\sqrt{2\pi\Delta}} \delta \big[y^\mu - f_\star(\hat{y}(\mathbf{x}^\mu) + \xi) \big]. \quad (6)$$

The Bayes errors (4) and (5) provide information-theoretic lower bounds on the test error for learning the target (1), in the sense that no learning algorithm can reach better performance when learning from the dataset $\mathcal{D}$.

Accessing numerically the Bayes errors (4) and (5) requires sampling an $\mathcal{O}(d^2)$-dimensional distribution, a difficult task. It is on the other hand possible to theoretically derive closed-form formulas using the replica method (Parisi, 1979; Mézard & Montanari, 2002) that allows characterizing the Bayes error in terms of the moments of independent instances of $\hat{y}(\mathbf{x})$ (the eponymous replicas) drawn from the posterior eq. (6). In the replica calculation, one averages over the randomness in the model and in order to be able to carry out such averages in a closed form, the Gaussian equivalence property described in the next section is crucial.

3.1. The Bayesian Gaussian Equivalence Property

A seminal step in our analytical approach is the property that we can replace the statistics of the output $\hat{y}(\mathbf{x})$ with respect to the randomness of the input $\mathbf{x}$ by Gaussian, with a covariance depending linearly on the covariance of the weight matrices W_l. In fact, (Li & Sompolinsky, 2021; Ariosto et al., 2022) do rely on a related Gaussianity assumption, which (Ariosto et al., 2022) heuristically justify for $L = 1$ using the Breuer-Major theorem, for generic datasets. Since in the present work, we consider the specific Bayes-optimal

setting, we are in a position to state a more principled conjecture which follows from the GEP (Goldt et al., 2020) and the Nishimori identities (Nishimori, 2001; Iba, 1998).

Conjecture 3.1. *(Shallow Bayes GEP) Consider* $\mathbf{x}$ *a random Gaussian vector. Then for* $L = 1$*, in the extensive-width asymptotic limit* $d, k_1 \to \infty$ *with fixed* $\mathcal{O}(1)$ *ratio* $\gamma_1 = {}^{k_1}/d$*, any finite number of replicas* $\hat{y}^1(\mathbf{x}; W_1^1, \mathbf{a}^1), ..., \hat{y}^s(\mathbf{x}; W_1^s, \mathbf{a}^s)$ *independently drawn from the Bayes posterior* (6) *are jointly Gaussian. Furthermore, their correlation reads* $\mathbb{E}_\mathbf{x} \hat{y}^a(\mathbf{x}) \hat{y}^b(\mathbf{x}) = \mathbf{a}^{a\top} \Omega_1^{ab} \mathbf{a}^b / k_1$ *where* Ω_1^{ab} *is the population covariance of the last layer post-activations* $\mathbb{E}_\mathbf{x} \varphi_1^a(\mathbf{x}) \varphi_1^b(\mathbf{x})^\top$ *that reads*

$$\Omega_1^{ab} = \left(\kappa_1^{(1)}\right)^2 \frac{W_1^a \Sigma W_1^{b\top}}{d} + \delta_{a,b} \left(\kappa_*^{(1)}\right)^2 \mathbb{I}_{k_1} , \quad (7)$$

where $\kappa_1^{(1)} = {}^{\mathbb{E}_z [z\sigma_1(z)]}/r_1$ *and* $(\kappa_*^{(1)})^2 = \mathbb{E}_z \left[\sigma_1(z)^2\right] - r_1(\kappa_1^{(1)})^2$*, with* $r_1 = \Delta_1 {}^{\mathrm{Tr}\,\Omega_1}/d$ *and* $z \sim \mathcal{N}(0, r_1)$*.*

We now explain how Conjecture 3.1 is motivated. In the proportional regime, for $a = b$, conditional on the matrix W_1, the Gaussian Equivalence Theorem of (Goldt et al., 2021; Hu & Lu, 2022a; Montanari & Saeed, 2022) prove indeed that the model (2) for $L = 1$ shares the same second-order post-activation statistics as the noisy *linear* network $\hat{y} = \mathbf{a}^T(\kappa_1^{(1)} W_1 \mathbf{x}/\sqrt{d} + \kappa_*^{(1)} Z)$ (with Z a random Gaussian variable), thus leading to the covariance (7). This so-called one-dimensional central limit theorem (1dCLT) (Goldt et al., 2021; Hu & Lu, 2022a; Montanari & Saeed, 2022) holds under some strict assumptions on the weight matrix W_1, that are satisfied in particular for random matrices with independent entries.

In the Bayesian setting one needs to integrate over the posterior distribution of the matrix W_1, learned from the data. For Conjecture 3.1 to be valid, the conditions of the 1dCLT must be satisfied with high probability over the learned matrices, which is by no means a trivial requirement. This is where the properties of Bayes-optimal inference come in handy: indeed, a classic property of Bayesian learning (often called the Nishimori property (Nishimori, 2001; Iba, 1998; Zdeborová & Krzakala, 2015)) is that the statistics of weights drawn from the Bayes posterior is exactly the same as the one of the target network weights. This is a direct consequence of the Bayes formula (see e.g. section 1.2.3. in (Zdeborová & Krzakala, 2015)). As a consequence, the learned matrices are following Gaussian statistics as well (given this is the statistics of the target ones by definition), and thus respect the conditions of the 1dCLT.

When considering different replicas ($a \neq b$), the Nishimori conditions ensure that one of the two replicas can be taken, without loss of generality, to be the target weight $W_1^\star$. Since W_1 is learnt and therefore generically correlated with the target weights $W_1^\star$, the assessment of the covariance Ω_1^{ab} is a

challenging task. However, the results of (Aubin et al., 2018) suggest that W_1 is asymptotically uncorrelated with $W_1^\star$ for sample complexities $\alpha \lesssim k_1$. Since we consider here $\alpha = \mathcal{O}(1) \ll k_1$, this motivates the following conjecture:

Conjecture 3.2. *(Non-specialization) for* $L = 1$*, in the asymptotic limit* $n, d, k_1 \to \infty$ *with fixed* $\mathcal{O}(1)$ *ratio* $\gamma_1 = {}^{k_1}/d$ *and* $\alpha = {}^n/d$*, let* W_1 *be sampled from the Bayes posterior* (6)*. Then with high probability* W_1 *has vanishing overlap with* $W_1^\star$*, i.e.*

$$\frac{1}{d} \max_{1 \leq i,j \leq k_1} \left(W_1^\star \Sigma W_1^\top\right)_{i,j} = \mathcal{O}\left({}^1/\sqrt{d}\right). \quad (8)$$

3.2 implies that the second term in the right-hand side of (7) is only present for $a = b$. The detailed derivation of (7) is presented in Appendix A

3.2. Deep (Bayesian) Gaussian Equivalence Property

We next discuss how these results generalize to deep networks ($L \geq 2$). While a sizeable body of work has been devoted to the distribution induced by the random weights for fixed inputs (Lee et al., 2018; de G. Matthews et al., 2018; Hanin & Zlokapa, 2022; Hanin, 2022; Yaida, 2019), little is known, in the deep case, for the distribution induced by the input distribution, for *fixed* weights. While there is no proof of the equivalence of the 1dCLT of (Goldt et al., 2021; Hu & Lu, 2022b) for $L \geq 2$, we provide, in App. A, numerical evidence of the following conjecture:

Conjecture 3.3. *The output* $\hat{y}(\mathbf{x})$ *of a deep random network, conditional on its Gaussian weights* $\{W_\ell\}_{\ell=1}^L$*,* $\mathbf{a}$*, in the extensive-width limit* $d \to \infty$ *and* $\forall \ell, k_\ell \to \infty$ *with fixed ratios* $\gamma_\ell = {}^{k_\ell}/d$*, is asymptotically Gaussian with respect to* $\mathbf{x}$*.*

Conjecture 3.3 thus extends the first part of 3.1 to the deep setting. This intuitively follows from the fact that higher order cumulants of the post-activations at intermediary layers are asymptotically suppressed (as shown in (Fischer et al., 2022)) and thus approximately Gaussian – allowing one to iterate 3.1. App. A further establishes a closed-form expression for the variance of $\hat{y}(\mathbf{x})$, which like the shallow case 3.1 is amenable to being interpreted in terms of an equivalent noisy network. We defer the discussion of the latter to the subsection 3.3. Finally, the Nishimori property again ensures that conjecture 3.3 transfers to weights sampled from the Bayes posterior (6). Defining the following recursion on $\{r_\ell\}_{\ell=1}^L$, $\{\kappa_1^{(\ell)}\}_{\ell=1}^L$ and $\{\kappa_*^{(\ell)}\}_{\ell=1}^L$:

$$r_{\ell+1} = \Delta_{\ell+1} \mathbb{E}_{z \sim \mathcal{N}(0, r_\ell)} \left[\sigma_\ell(z)^2\right] ,$$

$$\kappa_1^{(\ell)} = \frac{1}{r_\ell} \mathbb{E}_{z \sim \mathcal{N}(0, r_\ell)} [z\sigma_\ell(z)] ,$$

$$\kappa_*^{(\ell)} = \sqrt{\mathbb{E}_{z \sim \mathcal{N}(0, r_\ell)} [\sigma_\ell(z)^2] - r_\ell \left(\kappa_1^{(\ell)}\right)^2} , \quad (9)$$

with $r_1 \equiv \Delta_1 {}^{\mathrm{tr}\,\Sigma}/d$, the deep version of (3.1) and 3.2 reads:

Conjecture 3.4. *(Deep Bayes GEP) in the extensive width asymptotic limit $d \to \infty$ and $\forall \ell, k_\ell \to \infty$ with fixed ratios $\gamma_\ell = {}^{k_\ell}/d$, let $\hat{y}^1(\mathbf{x}), ..., \hat{y}^s(\mathbf{x})$ be any finite number of replicas independently drawn from the Bayes posterior eq. (6). Then $\hat{y}^1(\mathbf{x}), ..., \hat{y}^s(\mathbf{x})$ are jointly Gaussian with correlation $\mathbb{E}_\mathbf{x} \hat{y}^a(\mathbf{x}) \hat{y}^b(\mathbf{x}) = \mathbf{a}^{a\top} \Omega_L^{ab} \mathbf{a}^b / k_1$, where the population covariance $\Omega_L^{ab} \equiv \mathbb{E}_\mathbf{x} (\varphi_L^a \circ ... \varphi_1^a(\mathbf{x}))(\varphi_L^b \circ ... \varphi_1^b(\mathbf{x}))^\top$ is given by*

$$\Omega_\ell^{ab} = \left(\kappa_1^{(\ell)}\right)^2 \frac{W_\ell^a \Omega_{\ell-1}^{ab} W_\ell^{b\top}}{k_{\ell-1}} + \delta_{ab} \left(\kappa_*^{(\ell)}\right)^2 \mathbb{I}_{k_\ell}. \quad (10)$$

Finally, defining $\Omega_\ell^{a\star} \equiv \mathbb{E}_\mathbf{x}(\varphi_\ell^a \circ ... \varphi_1^a(\mathbf{x}))(\varphi_\ell^\star \circ ... \varphi_1^\star(\mathbf{x}))^\top$ for any a, there is no specialization, i.e. with high probability

$$\frac{1}{d} \max_{1 \le i, j \le k_\ell} \left(W_\ell^a \Omega_{\ell-1}^{a\star}(W_\ell^\star)^\top\right)_{i,j} = \mathcal{O}\left({}^1/\sqrt{d}\right). \quad (11)$$

In (10), $\Omega_0^{ab} \equiv \Sigma$. The derivation of (10) can be found in Appendix A, eqs (51) and (64). We precise that 3.4 holds for any sequence of activations $\{\sigma_\ell\}_{\ell=1}^L$ satisfying $\forall \ell, \mathbb{E}_{z \sim \mathcal{N}(0, r_\ell)}[\sigma_\ell(z)] = 0$. This is in particular always true for odd activations. We adopt in this work the latter (stronger) assumption for the sake of definiteness and clarity. An important note is that the population covariance between post-activations at *any* two layers $1 \le \ell, \ell' \le L$ (not just $\ell = \ell' = L$) can be generically computed. Because the post-activations result from the propagation of the Gaussian variable $\mathbf{x}$ through several non-linear layers, this computation (which we detail in App. A) is non-trivial for $L \ge 2$ and of independent interest.

3.3. Bayes-optimal errors

Conjectures 3.1 and 3.4 are used in equations (81) of App. B for shallow networks and (126) of App. C for deep networks, allowing to characterize the Bayes error in terms of the sole second-order statistic $q = \mathbb{E}_{\mathcal{D}, W_1^\star, \mathbf{a}_\star} \langle \mathbb{E}_\mathbf{x} \hat{y}^a(\mathbf{x}) \hat{y}^b(\mathbf{x}) \rangle_\mathbb{P}$. q is known as the *self overlap* in the statistical physics literature. The replica computation then proceeds in a rather standard way, provided one employs the so-called *replica-symmetry ansatz*, which is always correct in Bayes-optimal settings (see e.g. (Zdeborová & Krzakala, 2015)). One finally reaches the following characterizations in Apps. B and C.

For regression, the Bayes-optimal MSE reads

$$\epsilon_{g,\text{reg}}^{\text{BO}} = \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 \left(\Delta_a \left(\int z \, d\mu(z)\right) \prod_{\ell=1}^L \Delta_\ell - q\right) + \epsilon_r \quad (12)$$

with the self-overlap q satisfying the equation

$$q = \frac{1}{2} \int \frac{\alpha \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 z^2 \Delta_a^2 \prod_{\ell=1}^L \Delta_\ell^2}{\epsilon_{g,\text{reg}}^{\text{BO}} + \alpha \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 z \Delta_a \prod_{\ell=1}^L \Delta_\ell} \, d\mu(z). \quad (13)$$

We have denoted the residual error

$$\epsilon_r \equiv \sum_{\ell_0=1}^{L-1} \left(\kappa_*^{(\ell_0)}\right)^2 \Delta_a \prod_{\ell=\ell_0+1}^L \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell + \left(\kappa_*^{(L)}\right)^2 \Delta_a + \Delta. \quad (14)$$

For classification, the Bayes-optimal error reads

$$\epsilon_{g,\text{class}}^{\text{BO}} = \frac{1}{\pi} \arccos\left[\frac{\sqrt{\prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 q}}{\sqrt{\Delta_a \int z \, d\mu(z) \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell + \epsilon_r}}\right], \quad (15)$$

where the self-overlap q satisfies the system of equations

$$\begin{cases} q = \int \frac{\hat{q} \Delta_a^2 \prod_{\ell=1}^L \Delta_\ell^2 z^2}{\hat{q} z \Delta_a \prod_{\ell=1}^L \Delta_\ell + 1} d\mu(z) \\[2ex] \hat{q} = \frac{2\alpha \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2}{\Delta_a \int z \, d\mu(z) \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell + \epsilon_r - \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 q} \\[2ex] \int \frac{d\xi}{(2\pi)^{\frac{3}{2}}} \frac{2e^{-\frac{1}{2}\frac{\Delta_a \int z d\mu(z) \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell + \epsilon_r + \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 q}{\Delta_a \int z d\mu(z) \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell + \epsilon_r - \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 q} \xi^2}}{1 - \text{erf}\left(\frac{\prod_{\ell=1}^L \kappa_1^{(\ell)} \sqrt{q} \xi}{\sqrt{2\left(\Delta_a \int z d\mu(z) \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell + \epsilon_r - \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 q\right)}}\right)} \end{cases} .$$

Equivalent shallow network Remarkably, the Bayes errors (12) and (15) are equal to the Bayes errors of a simple single-layer target function with random weights

$$y^{\text{eq}}(\mathbf{x}) = f_\star\left(\frac{\sqrt{\rho}\boldsymbol{\theta}^\top \mathbf{x}}{\sqrt{d}} + \sqrt{\epsilon_r}\xi\right), \quad (16)$$

where $\xi \sim \mathcal{N}(0, 1)$ and $\boldsymbol{\theta}$ is a Gaussian weight vector with independent Gaussian entries of unit variance. We have defined the effective signal strength

$$\rho \equiv \Delta_a \prod_{\ell=1}^L \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell.$$

We refer the reader to App. D for a derivation of this equivalence. To gain intuition on this equivalence, observe that the deep Bayes GEP 3.4 applied to a single replica implies that the deep non-linear network (1) is characterized by the

same second-order activations statistics as a network with noisy *linear* layers

$$\varphi_\ell^{\text{eq.}}(\mathbf{x}) = \kappa_1^{(\ell)} \frac{1}{\sqrt{k_{\ell-1}}} W_\ell^\star \cdot \mathbf{x} + \kappa_*^{(\ell)} \mathcal{N}(0, \mathbb{I}_{k_\ell}). \quad (17)$$

In turn, this deep noisy network reduces equivalently to the shallow network (16). Interestingly, note that while the multilayer target (1) is deterministic for a given instance of the weights, the equivalent target (16) displays a stochastic output noise ξ. This noise subsumes the effect of the higher order terms introduced by the non-linearities, which are not learnt in the proportional regime (Mei et al., 2021).

Fig. 1 shows the Bayes MSE, eq. (12), for networks with tanh activation, with $L = 1, 2$ hidden layers. This is contrasted to the MSE achieved by an expressive neural network (NN) with twice the target width, optimized end-to-end with full batch gradient descent. Fig. 3 presents the same experiment in the classification setting. As expected, even this expressive learning algorithm cannot achieve a lower error than the information-theoretic lower bounds (12), (15).

4. ERM with Linear Methods

Eqs. (12) and (15) provide the information-theoretic minimal error for deep extensive-width targets (1). However, (Barbier et al., 2017; Aubin et al., 2018) evidenced that the Bayes error is not always attainable, in practice, by known polynomial time algorithms. In this section, we investigate the performance of some standard ERM methods. We provide a tight asymptotic characterization of the test error of each algorithm and show that for regression the Bayes error is, in fact, also achievable algorithmically. We address in succession: ridge regression, RF regression, kernel regression, logistic regression and ridge classification.

We give, for each of the considered ERM algorithms, a sharp asymptotic characterization of the associated test error. The fact that the deep non-linear target (1) shares the same Bayes error as the equivalent shallow model (16) suggests that the test error of ERM methods should also be identical. Applying Theorem 1 of (Loureiro et al., 2021) on the equivalent shallow target (16) thus leads to the formulas provided here. This heuristic line of reasoning is put on a firm basis in some settings in the concomitant work of (Schröder et al., 2023), where the formulas characterizing the performance of the considered ERM methods are derived. We discuss this further in App. E.

4.1. Ridge regression
We first consider ridge regression, corresponding to the minimization of the risk

$$\mathcal{R}(w) = \sum_{\mu=1}^{n} \left(y^\mu - \frac{w \cdot \mathbf{x}^\mu}{\sqrt{d}} \right)^2 + \frac{\lambda}{2} ||w||^2, \quad (18)$$
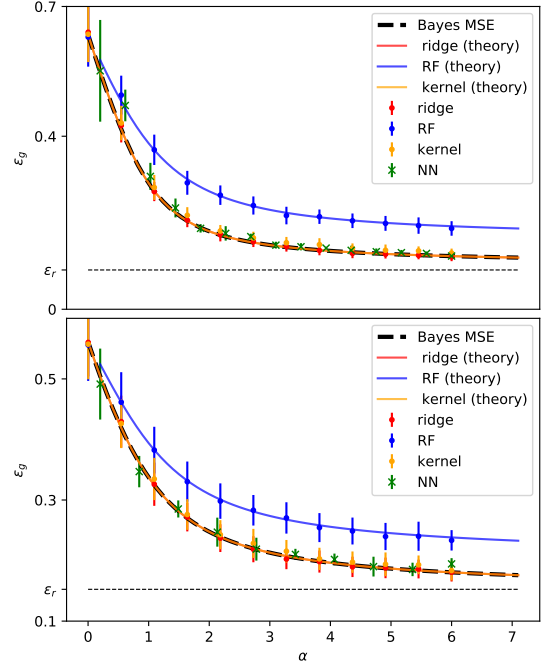

Figure 1. Targets (1) with $L = 1$ (top) and $L = 2$ (bottom) hidden layers, with $\sigma_{1,2}(x) = \tanh(2x)$ activation, widths $k_{1,2} = 700$, and $\Delta_a = \Delta_{1,2} = 1$, $\Delta = 0$, in dimension $d = 500$. The Bayes-optimal MSE (12) (dashed black) is contrasted to the replica predictions and simulations for optimally regularized ridge regression (19) (red), optimally regularized random features (23) with $\sigma(x) = \tanh(2x)$ non-linearity (blue), and optimally regularized kernel regression (24) with $\sigma(x) = \tanh(2x)$ non-linearity (orange). Green dots represent simulations for a one (top) and two (bottom) hidden layers neural network of width 1500, optimized with full-batch GD, learning rate $\eta = 8.10^{-3}$ and weight decay $\lambda = 0.1$. Dashed grey lines represent the residual error ϵ_r (14). Error bars represent one standard deviation over 30 trials.

with a ℓ_2 regularization term of strength λ. The associated test error can be computed by combining the deep GEP (9) and the theorem of (Loureiro et al., 2021), as

$$\epsilon_g = \rho \int z \mathrm{d}\mu(z) + q - 2 \prod_{\ell=1}^{L} \kappa_1^{(\ell)} m + \epsilon_r, \quad (19)$$

where m, q, V are the solutions of the system of equations

$$\begin{cases} \hat{V} = \frac{\alpha}{1+V} \\ \hat{q} = \alpha \frac{\epsilon_g}{(1+V)^2} \\ \hat{m} = \frac{\prod\limits_{\ell=1}^{L} \kappa_1^{(\ell)} \alpha}{1+V} \end{cases} \begin{cases} V = \int \frac{z}{\lambda+\hat{V}z} \mathrm{d}\mu(z) \\ q = \int \frac{\Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell \hat{m}^2 z^3 + \hat{q}z^2}{(\lambda+\hat{V}z)^2} \mathrm{d}\mu(z) \\ m = \Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell \hat{m} \int \frac{z^2}{\lambda+\hat{V}z} \mathrm{d}\mu(z) \end{cases} \quad (20)$$

The derivation and further discussion are provided in App. E.

Ridge regression is Bayes-optimal   The optimal regularization $\lambda^\star$ leading to minimal test error is shown in App. F
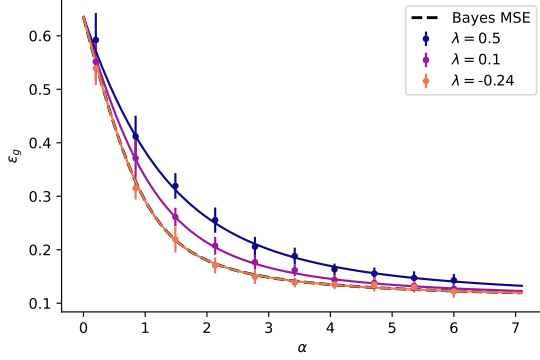
Figure 2. Arccosine kernel regression on a two-layers target of width $k_1 = 700$, $\Delta_a = \Delta_1 = 1$, $\Delta = 0$, with $\sigma_1(x) = \tanh(2x)$ activation, in dimension $d = 500$. Different curves correspond to different regularizations $\lambda = 0.5, 0.1$ and the optimal $\lambda^\star \approx -0.24$ (25). Solid lines correspond to the replica predictions (24). The learning curve for the optimal *negative* regularizer $\lambda^\star$ (red) superimposes with the Bayes-optimal MSE (12). Error bars represent one standard deviation over 30 trials.

to admit the compact expression

$$\lambda^\star = \frac{\epsilon_r}{\rho}. \tag{21}$$

The expression (21) mirrors the result of (Sahraee-Ardakan et al., 2022) for GP targets, see App. E. Eq. (21) intuitively corresponds to requiring that the regularization λ used in the ERM (18) should be equal to the true noise-to-signal ratio of the equivalent target (16).

For $\lambda = \lambda^\star$, the equation (20) reduces to (12), implying that optimally regularized ridge regression achieves the Bayes-optimal MSE. The solution of (19), (20), (21), with the corresponding numerical simulations, is plotted in Fig. 1, and can indeed be seen to exactly fall on the Bayes-optimal baseline (12). This implies that in the proportional high-dimensional limit $n \sim d$, no algorithm can learn more accurately the non-linear function (1) than optimally regularized linear regression. This echoes the claims of (Sahraee-Ardakan et al., 2022) when the target is a GP.

4.2. Random Features

Random features learning (Rahimi & Recht, 2007) was initially introduced as a means to speed up kernel methods. They correspond to the ERM

$$\mathcal{R}(w) = \sum_{\mu=1}^{n} \left(y^\mu - \frac{1}{\sqrt{k}} w \cdot \sigma\left(\frac{F \cdot \mathbf{x}^\mu}{\sqrt{d}}\right) \right)^2 + \frac{\lambda}{2} \|w\|^2, \tag{22}$$

where σ is a nonlinearity with associated GEP coefficients (9) κ_1, κ_*, and $F \in \mathbb{R}^{k \times d}$ is the random feature matrix, assumed in the following to possess i.i.d Gaussian entries. RF learning corresponds formally to ridge regression in a $k-$dimensional space, often taken larger than the $d-$

dimensional input space to allow for overparametrization. Again, we consider the proportional regime $n, d, k \to \infty$ and introduce the width/dimension ratio $\gamma \equiv k/d$. In the following, for simplicity, we restrict ourselves to the case of isotropic covariates $\Sigma = \mathbb{I}_d$. Sharp asymptotics for the test error of such models have been provided in (Mei & Montanari, 2019; Gerace et al., 2020) for single-layer targets. We tie those results in with (9) in the case of the deep random target (1). The asymptotic test error is shown in App. E to be given again by (19), where q, m satisfy

$$\begin{cases} \hat{V} = \frac{\frac{\alpha}{\gamma}}{1+V} \\ \hat{q} = \frac{\alpha}{\gamma} \frac{\epsilon_g}{(1+V)^2} \\ \hat{m} = \sqrt{\Delta_a \prod_{\ell=1}^{L} \Delta_\ell} \sqrt{\gamma} \frac{\prod_{\ell=1}^{L} \kappa_1^{(\ell)} \frac{\alpha}{\gamma}}{1+V} \end{cases} \tag{23}$$

$$\begin{cases} V = \frac{1}{\hat{V}} - \frac{\lambda}{\hat{V}^2 \kappa_1^2} g\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right) \\ q = \frac{\hat{m}^2 + \hat{q}}{\hat{V}^2} - \frac{1}{\kappa_1^2 \hat{V}^2} \left(\frac{2\lambda(\hat{m}^2 + \hat{q})}{\hat{V}} + \hat{m}^2 \kappa_*^2\right) g\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right) \\ \quad + \frac{\lambda}{\kappa_1^4 \hat{V}^3} \left(\frac{\lambda(\hat{m}^2 + \hat{q})}{\hat{V}} + \hat{m}^2 \kappa_*^2\right) g'\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right) \\ m = \sqrt{\gamma} \frac{\hat{m}}{\hat{V}} \left[1 - \frac{1}{\kappa_1^2} \left(\frac{\lambda}{\hat{V}} + \kappa_*^2\right) g\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right)\right]. \end{cases}$$

$g(z)$ is the Stieljes transform of the Marchenko-Pastur distribution with aspect ratio γ. The solution of (23) is plotted in Fig. 1, with the regularization λ being numerically optimized over, and is observed to yield higher MSE than the optimal (12). Intuitively, this is because from (16) the target function (1) effectively acts as a linear function on the original space, while the RF transformation $\sigma(F\cdot)$ (22) introduces a mismatch between the original basis where the target acts and the features basis in which the linear regression readout is carried out. This intuition is formalized in App. F. This mismatch can be shown to be benign only in the infinite overparametrization (infinite width) $\gamma \to \infty$ limit, has discussed in the next subsection.

4.3. Kernels

In the $\gamma \to \infty$ limit, RF learning (22) becomes equivalent to kernel regression (Neal, 1994; Lee et al., 2018). The saddle-point equations (23) characterizing the generalization error (19) then reduce to

$$\begin{cases} \hat{V} = \frac{\alpha}{1+V} \\ \hat{q} = \alpha \frac{\epsilon_g}{(1+V)^2} \\ \hat{m} = \alpha \frac{\prod_{\ell=1}^{L} \kappa_1^{(\ell)}}{1+V} \end{cases} \begin{cases} V = \frac{\kappa_*^2}{\lambda} + \frac{\kappa_1^2}{\lambda + \hat{V}\kappa_1^2} \\ q = \frac{\Delta_a \prod_{\ell=1}^{L} \Delta_\ell \hat{m}^2 \kappa_1^4 + \hat{q}\kappa_1^4}{(\lambda + \hat{V}\kappa_1^2)^2} \\ m = \Delta_a \prod_{\ell=1}^{L^\star} \Delta_\ell \hat{m} \frac{\kappa_1^2}{\lambda + \hat{V}\kappa_1^2} \end{cases}. \tag{24}$$

A complete derivation of the eqs (24) is given in App. E.

Kernel regression is Bayes-optimal The regularization λ minimizing the test error (19) for kernel regression (24) can

be shown as in the ridge case to admit the simple expression

$$\lambda^\star = \kappa_1^2 \left(\frac{\epsilon_r}{\rho} - \frac{\kappa_*^2}{\kappa_1^2} \right). \qquad (25)$$

We discuss the connection of this formula with the one provided in (Sahraee-Ardakan et al., 2022) for GP targets in App. B. Again, the expression (25) also admits a very intuitive interpretation. An informal takeaway from (Mei et al., 2021; Hu & Lu, 2022b) is indeed that in the linear $n \sim d$ regime, kernel regression can be seen as effectively implementing ridge regression with an implicit ℓ_2 regularization equal to $\lambda + \kappa_*^2/\kappa_1^2$. Requiring this implicit regularization to match the true target noise-to-signal ratio (16) naturally leads to (25). The test error (19) evaluated at the solution of (24), at the optimal regularization (25), is plotted in Fig. 1, and can be seen to exactly superimpose with the Bayes-optimal baseline (12).

When the activation σ of the kernel is very non-linear (as quantified by the ratio κ_*^2/κ_1^2) and implements too strong an implicit regularization (Wu & Xu, 2020; Hastie et al., 2019) compared to the actual target noise, the optimal regularization (25) can be *negative*. This negative ridge phenomenon can for example be observed when learning a target $2-$layer network with $\tanh$ activation with a kernel using the more non-linear sign activation, see Fig. 2. Note that even in such cases where the optimal explicit regularization λ is negative, the risk remains convex due to the implicit ℓ_2 regularization.

4.4. Logistic and ridge classification

This last subsection addresses ERM in the classification setting. We consider two standard classification methods, namely logistic regression and ridge classification. They correspond to ERM on the risk

$$\mathcal{R}(w) = \sum_{\mu=1}^{n} \ell \left(y^\mu, \frac{w \cdot \mathbf{x}^\mu}{\sqrt{d}} \right) + \frac{\lambda}{2} \|w\|^2 \qquad (26)$$

with $\ell(y, z) = \ln(1 + e^{-yz})$ and $\ell(y, z) = 1/2(y - z)^2$ respectively. For simplicity, we assume the noiseless $\Delta = 0$ setting. An asymptotic expression for the test error of logistic regression can again be reached and is detailed in App. H. These theoretical predictions are plotted in Fig. 3, and contrasted to the Bayes-optimal baseline (15). At odds with the regression case, the learning curves of logistic regression and ridge classification do not exactly superimpose with the Bayes-optimal baseline but lie extremely close. Fig. 3 shows that for a $\sigma(x) = \tanh(2x)$ activation the difference is of order of 10^{-4} (for logistic regression) and 10^{-3} (for ridge classification). Such a gap has also been observed by (Aubin et al., 2020) for targets without hidden layer.

5. Beyond the Proportional Regime

Section 4 establishes that ridge regression and kernel regression are Bayes-optimal in the linear $n \sim d$ regime for
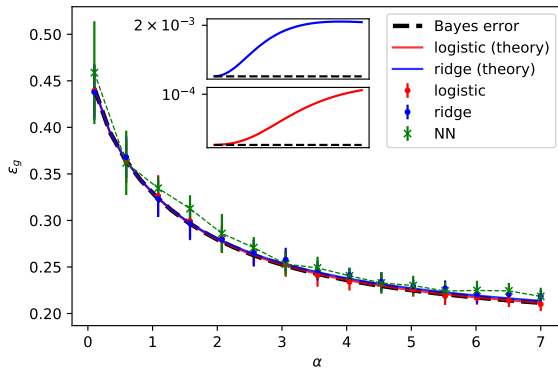

Figure 3. Learning curves for classification, with a three-layers target (1) with $\tanh(2\cdot)$ activation and width $k_{1,2} = 700$, $\Delta_a = \Delta_{1,2} = 1$, $\Delta = 0$, in dimension $d = 500$. The black dashed line represents the Bayes-optimal error (15). The theoretical learning curves for optimally regularized logistic regression (ridge classification) are shown in red (blue), alongside the corresponding numerical simulations. Green dots show the test error of a three layers fully connected network trained end-to-end with full-batch Adam, learning rate 0.003 and weight decay 0.01, after 2000 epochs. Error bars represent one standard deviation over 30 trials. (inset) Zoom in on the theoretical learning curves for logistic regression (red) and ridge classification (blue), with the Bayes-optimal baseline (15) subtracted. Logistic regression and ridge classification can be seen to be very close, but not equal, to the Bayes-optimal baseline (up to 10^{-4} and 10^{-3} respectively).

the target (1). These conclusions and the ones of (Sahraee-Ardakan et al., 2022) in the context of GP targets, and are reminiscent on a high level of the empirical observations of (Arora et al., 2020) that neural networks can only marginally outperform kernel methods for small train sets on several benchmark real datasets. Note that the study (Lee et al., 2020) also employs networks with comparable width to the dataset size & dimension, and reaches similar conclusions.

In fact, the information-theoretic optimality of linear/kernel ERM methods is essentially due to the fact that a proportional number of samples $n \sim d$ is not enough to learn features beyond linear approximations. The conclusions drawn in other scaling regimes are anticipated to be very different. In particular, beyond the linear $n \sim d$ regime, the test error should pick up, and depend on, non-Gaussian asymptotic corrections to the GEP (Goldt et al., 2021), Bayes GEP 3.1, and deep Bayes GEP 3.4. Besides, information-theoretic intuition suggests that since the target (1) is parametrized by $\mathcal{O}(d^2)$ real numbers, a quadratic number of samples $n \sim d^2$ is needed, and sufficient, to learn it perfectly. In other words, one expects the Bayes-optimal error to vanish in the quadratic regime. This means in particular that kernel methods (which are known to learn at best a quadratic approximation of the target in this scaling regime (Misiakiewicz, 2022; Xiao & Pennington, 2022; Hu & Lu, 2022b)), and ridge regression (which learns a linear approximation) will
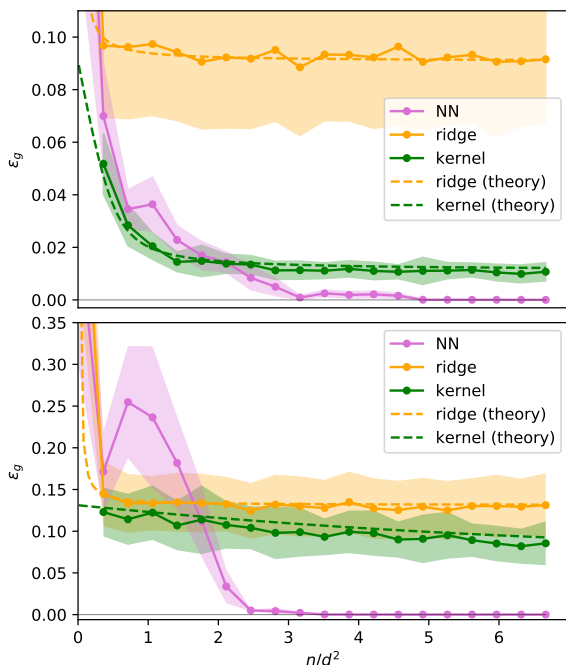
Figure 4. MSE of regression on a ReLU (top) erf($2x$) (bottom) $2-$layers target of width $k_1 = 20$, $\Delta_a = \Delta_1 = 1$, $\Delta = 0$, in dimension $d = 30$. Dots represent simulations for optimally regularized ridge regression (orange), optimally regularized arccosine (top), and arcsine (bottom) kernel (green). Dashed lines represent the theoretical predictions for the MSE of the kernel in polynomial regimes of (Hu & Lu, 2022a). Purple dots indicate the MSE of a $2-$layers fully connected neural network of width $k = 30$ trained end-to-end using Adam (purple), batch size $n/3$ and learning rate $\eta = 3.10^{-3}$, over 2000 epochs. Error bars represent one standard deviation over 10 trials. For a quadratic number of samples $n \sim d^2$ the network learns the target perfectly (up to errors of order 10^{-5}) while kernel regression only learns a quadratic approximation of the target and yields MSEs larger by an order of 10^3. Ridge regression can only learn the best linear approximation and leads to even higher MSE.

cease to be optimal. In this section, this intuition is further explored numerically.

Fig. 4 contrasts the MSE of an Adam-optimized neural network, optimally regularized ridge regression, and optimally regularized arcosine kernel regression, in the $n \sim d^2$ regime, for a ReLU network target with one hidden layer. For completeness, we also plot the theoretical predictions for the test error for kernel regression derived in (Hu & Lu, 2022b). While ridge (kernel) regression can only learn the best linear (quadratic) approximation of the non-linear target, the neural network manages to learn the target almost perfectly (up to an MSE of $\mathcal{O}(10^{-5})$). These results show the superiority of neural networks over kernel methods in the quadratic

regime for the multi-layer target (1), and complement similar superiority results (see e.g. (Ghorbani et al., 2019; 2020)) for single-layer targets.

Conclusion

We investigate the problem of learning a deep, non-linear, extensive-width random neural network. We propose asymptotic expressions for the Bayes-optimal error and the test error of linear / kernel ERM methods, for classification and regression. The technical backbone of the derivations is the deep Bayes GEP conjecture 3.4, and novel closed-form formulae for second-order population statistics of network post-activations. The conclusion is that kernel methods are optimal in this regime. We showed, however, that the situation is drastically different in the quadratic sample regime, and evidence the onset of feature learning leading to a vanishing test error for neural nets, while kernel methods can learn only quadratic approximation and thus become suboptimal. This marks a clear separation in the power of neural nets with respect to kernels as soon as $n \sim d^2$.

From a theoretical standpoint, a quantitative analysis of the learning curve for the quadratic regime is challenging. In particular, Gaussian universality results such as (Goldt et al., 2021; Hu & Lu, 2022a; Mei & Montanari, 2019), cease to hold outside of the proportional regime and would need to be extended. Building a theory for extensive-width networks in these superlinear sample regimes could unveil rich behavior and properties, and constitutes in the authors's point of view a crucial challenge in machine learning theory.

Acknowledgements

References

Advani, M. and Ganguli, S. Statistical mechanics of optimal convex inference in high dimensions. *Physical Review X*, 6(3):031034, 2016.

Ariosto, S., Pacelli, R., Pastore, M., Ginelli, F., Gherardi, M., and Rotondo, P. Statistical mechanics of deep learning beyond the infinite-width limit. *ArXiv*, abs/2209.04882, 2022.

Arora, S., Du, S. S., Li, Z., Salakhutdinov, R., Wang, R., and Yu, D. Harnessing the power of infinitely wide deep nets on small-data tasks. *Proc. Int. Conf. Learning Rep. (ICLR)*, 2020.

Aubin, B., Maillard, A., Barbier, J., Krzakala, F., Macris, N., and Zdeborová, L. The committee machine: computational to statistical gaps in learning a two-layers neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2019, 2018.

Aubin, B., Krzakala, F., Lu, Y. M., and Zdeborová, L. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. *Advances in Neural Information Processing Systems*, 33: 12199–12210, 2020.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences of the United States of America*, 116:5451 – 5460, 2017.

Bosch, D., Panahi, A., and Hassibi, B. Precise asymptotic analysis of deep random feature models. *ArXiv*, abs/2302.06210, 2023.

Canatar, A., Bordelon, B., and Pehlevan, C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12, 2020.

Clarté, L., Loureiro, B., Krzakala, F., and Zdeborová, L. A study of uncertainty quantification in overparametrized high-dimensional models. *ArXiv*, abs/2210.12760, 2022.

Cui, H., Saglietti, L., and Zdeborová, L. Large deviations for the perceptron model and consequences for active learning. *Mach. Learn. Sci. Technol.*, 2:45001, 2019.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In *NeurIPS*, 2021.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. Error rates for kernel classification under source and capacity conditions. *ArXiv*, abs/2201.12655, 2022.

d'Ascoli, S., Gabrié, M., Sagun, L., and Biroli, G. On the interplay between data structure and loss function in classification problems. In *NeurIPS*, 2021.

de G. Matthews, A. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. *NeurIPS Workshop on Advances in Approximate Bayesian Inference*, 2018.

El Karoui, N. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics*, 36(6):2757–2790, 2008.

Fan, Z. and Wang, Z. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *ArXiv*, abs/2005.11879, 2020.

Fischer, K., Ren'e, A., Keup, C., Layer, M., Dahmen, D., and Helias, M. Decomposing neural networks as mappings of correlation functions. *Physical Review Research*, 2022.

Gabrié, M. Mean-field inference methods for neural networks. *Journal of Physics A: Mathematical and Theoretical*, 53, 2019.

Gerace, F., Loureiro, B., Krzakala, F., Mézard, M., and Zdeborová, L. Generalisation error in learning with random features and the hidden manifold model. *Proceedings of Machine Learning Research*, 37:3452–3462, 2020.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. *ArXiv*, abs/1904.12191, 2019.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. When do neural networks outperform kernel methods? *Journal of Statistical Mechanics: Theory and Experiment*, 2021, 2020.

Goldt, S., Mézard, M., Krzakala, F., and Zdeborová, L. Modelling the influence of data structure on learning in neural networks. *Physical Review X*, 4:041044, 2020.

Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mézard, M., and Zdeborová, L. The gaussian equivalence of generative models for learning with shallow neural networks. In *MSML*, 2021.

Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mézard, M., and Zdeborová, L. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pp. 426–471. PMLR, 2022.

Hanin, B. Correlation functions in random fully connected neural networks at finite width. *ArXiv*, abs/2204.01058, 2022.

Hanin, B. and Zlokapa, A. Bayesian interpolation with deep linear networks. *ArXiv*, abs/2212.14457, 2022.

Hastie, T. J., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50 2:949–986, 2019.

Hron, J., Bahri, Y., Novak, R., Pennington, J., and Sohl-Dickstein, J. N. Exact posterior distributions of wide bayesian neural networks. *ArXiv*, abs/2006.10541, 2020.

Hu, H. and Lu, Y. M. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 2022a.

Hu, H. and Lu, Y. M. Sharp asymptotics of kernel ridge regression beyond the linear regime. *ArXiv*, abs/2205.06798, 2022b.

Iba, Y. The nishimori line and bayesian statistics. *Journal of Physics A*, 32:3875–3888, 1998.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. N. Deep neural networks as gaussian processes. *International Conference on Learning Representations*, 2018.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Novak, R., Sohl-Dickstein, J. N., and Pennington, J. S. Wide neural networks of any depth evolve as linear models under gradient descent. *Journal of Statistical Mechanics: Theory and Experiment*, 2020, 2019.

Lee, J., Schoenholz, S. S., Pennington, J., Adlam, B., Xiao, L., Novak, R., and Sohl-Dickstein, J. N. Finite versus infinite neural networks: an empirical study. *NeurIPS*, 2020.

Li, Q. and Sompolinsky, H. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Physical Review X*, 2021.

Louart, C., Liao, Z., and Couillet, R. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28:1190–1248, 2017.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mézard, M., and Zdeborová, L. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

Maillard, A., Loureiro, B., Krzakala, F., and Zdeborová, L. Phase retrieval in high dimensions: Statistical and computational phase transitions. *Advances in Neural Information Processing Systems*, 33:11071–11082, 2020.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75, 2019.

Mei, S., Misiakiewicz, T., and Montanari, A. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 2021.

Mézard, M. and Montanari, A. *Information,Physics and computation*. Oxford University Press, 2002.

Misiakiewicz, T. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *ArXiv*, abs/2204.10425, 2022.

Montanari, A. and Saeed, B. Universality of empirical risk minimization. *Conference on Learning Theory*, pp. 4310–4312, 2022.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *ArXiv*, abs/1911.01544, 2019.

Neal, R. M. Priors for infinite networks (tech. rep. no. crg-tr-94-1). *University of Toronto*, 1994.

Nishimori, H. *Statistical Physics of Spin Glasses and Information Processing*. Oxford:Clarendon, 2001.

Opper and Haussler. Generalization performance of bayes optimal classification algorithm for learning a perceptron. *Physical review letters*, 66 20:2677–2680, 1991.

Parisi, G. Towards a mean field theory for spin glasses. *Phys. Lett*, 73(A):203–205, 1979.

Parisi, G. Order parameter for spin glasses. *Phys. Rev. Lett*, 50:1946–1948, 1983.

Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019, 2019.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *NIPS*, 2007.

Roberts, D. A., Yaida, S., and Hanin, B. The principles of deep learning theory. *ArXiv*, abs/2106.10165, 2021.

Sahraee-Ardakan, M., Emami, M., Pandit, P., Rangan, S., and Fletcher, A. K. Kernel methods and multi-layer perceptrons learn linear models in high dimensions. *ArXiv*, abs/2201.08082, 2022.

Schröder, D., Cui, H., Dmitriev, D., and Loureiro, B. Deterministic equivalent and error universality of deep random features learning. *ArXiv*, abs/2302.00401, 2023.

Schwarze, H. Learning a rule in a multilayer neural network. *Journal of Physics A: Mathematical and General*, 26(21): 5781, 1993.

Seung, H. S., Sompolinsky, H., and Tishby, N. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

Talagrand, M. The parisi formula. *Annals of mathematics*, pp. 221–263, 2006.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

Watkin, T. L., Rau, A., and Biehl, M. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499, 1993.

Wu, D. and Xu, J. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Xiao, L. and Pennington, J. Precise learning curves and higher-order scaling limits for dot product kernel regression. *ArXiv*, abs/2205.14846, 2022.

Yaida, S. Non-gaussian processes and neural networks at finite widths. In *Mathematical and Scientific Machine Learning*, 2019.

Zavatone-Veth, J. A., Canatar, A., and Pehlevan, C. Asymptotics of representation learning in finite bayesian neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2022, 2021.

Zavatone-Veth, J. A., Tong, W., and Pehlevan, C. Contrasting random and learned features in deep bayesian linear regression. *Physical review. E*, 105 6-1:064118, 2022.

Zdeborová, L. and Krzakala, F. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65:453 – 552, 2015.

A. Deep Gaussian Equivalence Principles

In this Appendix, we provide a discussion of the deep GEP 3.4. We first provide a derivation for the closed form expressions of the population covariances of the network activations, from which (10) can be deduced as a particular case. We then provide strong numerical support for the 1dCLT 3.1.

A.1. Setting

In this section, we provide a reminder of the notation. Since the matrix identities are applied also in section 4, where the learner network does not necessarily possess the same architecture as the target (1), we generically study the correlations between any pair of activations of *two* networks. For clarity, in this appendix, we will denote all target hyperparameters by a star $\star$, to distinguish them with the parameters of the learner.

We consider a $L^\star$ layers target (teacher) function

$$y(x) = \frac{1}{\sqrt{k_{L^\star}^\star}} a_\star^\top \underbrace{\left(\varphi_{L^\star}^\star \circ \varphi_{L^\star-1}^\star \circ \cdots \circ \varphi_2^\star \circ \varphi_1 \right)}_{L^\star} (x) \tag{27}$$

where the ℓ-th layer is

$$\varphi_\ell^\star(x) = \sigma_\ell^\star \left(\frac{1}{\sqrt{k_{\ell-1}^\star}} W_\ell^\star \cdot x \right) \tag{28}$$

$(\sigma_\ell^\star)_{\ell=1}^{L^\star}$ is a sequence of ***odd*** activation functions. The width of the ℓ-th layer is denoted k_ℓ, and the associated weight matrix is $W_\ell^\star \in \mathbb{R}^{k_\ell \times k_{\ell-1}}$. In contrast to the main text, for clarity, we use a star $\star$ superscript for all the parameters related to the target function. Unstarred variables shall conversely refer to the learner (hereafter referred to as the *student*) network.

The student network is also taken to be a random network, with trainable last layer:

$$y(x) = \frac{1}{\sqrt{k_L}} a^\top \underbrace{\left(\varphi_L \circ \varphi_{L-1} \circ \cdots \circ \varphi_2 \circ \varphi_1 \right)}_{L} (x), \tag{29}$$

with layers

$$\varphi_\ell(x) = \sigma_\ell \left(\frac{1}{\sqrt{k_{\ell-1}}} W_\ell \cdot x \right). \tag{30}$$

In the main text, we consider a student network with the same architecture as the target in section 3 ($L = L^\star$ and $\forall \ell, \ k_\ell = k_\ell^\star$), where the Bayes errors are derived. In section 4 on the other hand, the student shall correspond to a linear single layer ($L = 0$) network (for ridge regression, logistic regression and ridge classification) or a one-hidden-layer ($L = 1$) network with frozen Gaussian first layer (for random features). We denote the ℓ-th layer post-activation $h_\ell(x) = (\varphi_\ell \circ ... \circ \varphi_1)(x)$ (for student) $h_\ell^\star(x) = (\varphi_\ell^\star \circ ... \circ \varphi_1^\star)(x)$ (for the teacher), and introduce the covariances

$$\Omega_\ell = \langle h_\ell(x) h_\ell^\top(x) \rangle_x, \qquad \Psi_\ell = \langle h_\ell^\star(x) h_\ell^{\star\top}(x) \rangle_x \qquad \Phi_{\ell\star\ell} = \langle h_{\ell\star}^\star(x) h_\ell^\top(x) \rangle_x, \tag{31}$$

which we shall characterize for any ℓ. Note that we could also compute correlations of the form $\langle h_\ell(x) h_{\ell'}^\top(x) \rangle_x, \langle h_\ell^\star(x) h_{\ell'}^{\star\top}(x) \rangle_x$ for generic $\ell \neq \ell'$, i.e. population covariances between activations of the same network, but at different layers. Since these matrices however never enter our main discussion, and are furthermore a straightforward extension away from the results presented here, we choose to skip the discussion thereof for clarity purposes.

We further suppose the data x is i.i.d Gaussian

$$x \sim \mathcal{N}(0, \Sigma), \tag{32}$$

with a covariance Σ of extensive Frobenius norm and trace, i.e. there exists constant c, c' so that asymptotically (noting $k_0 = d$)

$$c < \frac{1}{d} \operatorname{tr} \Sigma^2 = \frac{1}{d} ||\Sigma||_F^2 < c' < \infty, \qquad\qquad c < \frac{1}{d} \operatorname{tr} \Sigma < c' < \infty. \tag{33}$$

In terms of the limiting spectral density μ, these assumptions imply that the first and second moments are finite and non zero. Finally, all weights are assumed to be Gaussian:

$$W_\ell^\star \overset{i.i.d.}{\sim} \mathcal{N}(0, \Delta_\ell^\star), \qquad\qquad W_\ell \overset{i.i.d.}{\sim} \mathcal{N}(0, \Delta_\ell). \tag{34}$$

Note that this last assumption is made for simplicity and definiteness. An exhaustive study of the most generic distribution for the weights under which (10) holds is a research line of independent interest, and falls out of the scope of the present manuscript.

A.2. Closed-form formulas for second order population statistics

We now re-state the recursion relation allowing to compute the covariances (31). Consider the sequence of variances defined by the recurrence

$$r_{\ell+1}^{(\star)} = \Delta_{\ell+1}^{(\star)} \mathbb{E}_z^{\mathcal{N}(0, r_\ell)} \left[\sigma_\ell^{(\star)}(z)^2 \right], \tag{35}$$

with the initial condition

$$r_1^{(\star)} = \Delta_1^{(\star)} \frac{1}{d} \operatorname{tr} \Sigma. \tag{36}$$

$r_\ell^{(\star)}$ physically corresponds to the average squared norm of the post-activations after layer ℓ. Then introduce the coefficients

$$\kappa_1^{\ell(\star)} = \frac{1}{r_\ell^{(\star)}} \mathbb{E}_z^{\mathcal{N}(0, r_\ell)} \left[z \sigma_\ell^{(\star)}(z) \right], \qquad\qquad \kappa_*^{\ell(\star)} = \sqrt{\mathbb{E}_z^{\mathcal{N}(0, r_\ell)} \left[\sigma_\ell^{(\star)}(z)^2 \right] - r_\ell^{(\star)} \left(\kappa_1^{\ell(\star)} \right)^2}. \tag{37}$$

For the first layer ($\ell = 1$), these coefficient correspond to the coefficients introduced by the seminal works of (Pennington & Worah, 2019; Louart et al., 2017), sometimes referred to as GET coefficients (Goldt et al., 2020; Gerace et al., 2020). These coefficients also appear in the characterization of the spectrum of the conjugate kernel of random multi-layer networks in (Fan & Wang, 2020). The layer-wise covariances Ω, Ψ (31) are then defined by the following recursion

$$\Omega_{\ell+1} = \kappa_1^{\ell 2} \frac{W_{\ell+1} \Omega_\ell W_{\ell+1}^\top}{k_\ell} + \kappa_*^{\ell 2} I_{k_{\ell+1}}, \tag{38}$$

$$\Psi_{\ell+1} = \kappa_1^{\star \ell 2} \frac{W_{\ell+1}^\star \Psi_\ell W_{\ell+1}^{\star\top}}{k_\ell^\star} + \kappa_*^{\star \ell 2} I_{k_{\ell+1}^\star}, \tag{39}$$

with initialization

$$\Omega_0 = \Psi_0 = \Sigma. \tag{40}$$

The teacher/teacher student correlation Φ admits the closed form expression

$$\Phi_{\ell^\star \ell} = \prod_{r=1}^{\ell} \prod_{s=1}^{\ell^\star} \kappa_1^r \kappa_1^{s\star} \times \frac{W_{\ell^\star} \cdot \ldots \cdot W_1^\star \cdot \Sigma \cdot W_1^\top \cdot \ldots \cdot W_\ell^\top}{\prod\limits_{r=0}^{\ell-1} \prod\limits_{s=0}^{\ell^\star-1} \sqrt{k_r k_s^\star}}. \tag{41}$$

Example for $L = 2$ For definiteness, let us provide an example for $L^\star = 1, L = 2$ (one hidden layer teacher, two hidden layers student). The recursions (38)(41) then translate to the compact formulas

$$\Omega_2 = (\kappa_1^1)^2 (\kappa_1^2)^2 \frac{W_2 W_1 \Sigma W_1^\top W_2^\top}{k_1 d} + (\kappa_1^2)^2 (\kappa_*^1)^2 \frac{W_2 W_2^\top}{k_1} + (\kappa_*^2)^2 I_{k_1}, \tag{42}$$

$$\Psi_1 = (\kappa_1^{1\star})^2 \frac{W_1^\star \Sigma W_1^{\star\top}}{d} + (\kappa_*^{1\star})^2 I_{k_1^\star}, \tag{43}$$

$$\Phi_{1,2} = \kappa_1^1 \kappa_1^2 \kappa_1^{1\star} \frac{W_1^\star \Sigma W_2^\top W_1^\top}{d \sqrt{k_1}}. \tag{44}$$

A.3. Derivation for Ω, Ψ

In this subsection, we detail the derivation of the recursion (10). We first derive a preliminary results, describing the evolution of the data covariance through propagation of a single layer, and subsequently iterate thereupon to reach (10). Note that because the student and teacher share the same architecture and model definition, the treatment of the teacher/teacher covariance Ψ and the student/student covariance Ω is identical. In the following, we therefore only discuss the student/student covariance Ω.

Preliminary: propagating through one layer Consider the auxiliary single-layer problem

$$h(x) = \sigma\left(\frac{1}{\sqrt{d}} W \cdot x\right) \tag{45}$$

with $x \sim \mathcal{N}(0, \Sigma)$. The covariance of the post-activation $h(x)$ reads

$$\Omega_{ij} = \langle h_i(x) h_j(x) \rangle_x = \int \frac{e^{-\frac{1}{2}\begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} \frac{w_i^\top \Sigma w_i}{d} & \frac{w_i^\top \Sigma w_j}{d} \\ \frac{w_i^\top \Sigma w_j}{d} & \frac{w_j^\top \Sigma w_j}{d} \end{pmatrix}^{-1} \begin{pmatrix} u \\ v \end{pmatrix}}}{\sqrt{\det 2\pi \begin{pmatrix} \frac{w_i^\top \Sigma w_i}{d} & \frac{w_i^\top \Sigma w_j}{d} \\ \frac{w_i^\top \Sigma w_j}{d} & \frac{w_j^\top \Sigma w_j}{d} \end{pmatrix}}} \sigma(u)\sigma(v). \tag{46}$$

Note that the random variable $\frac{w^\top \Sigma w}{d}$ concentrates, as w is i.i.d Gaussian of variance Δ. We indeed have that

$$\mathbb{E}_w \frac{w^\top \Sigma w}{d} = \frac{\Delta}{d} \operatorname{tr} \Sigma \equiv r \tag{47}$$

and (diagonalizing $\Sigma = U\Lambda U^\top$) and noting that $U^\top w$ is still Gaussian with independent entries

$$\mathbb{V}_w \left[\frac{w^\top \Sigma w}{d} \right] = \frac{1}{d^2} \sum_{i=1}^{d} \lambda_i^2 \mathbb{V}_w \left[(U^\top w)_i^2 \right] = \frac{2\Delta}{d^2} \operatorname{tr} \Sigma^2 = \frac{2\Delta}{d} \frac{||\Sigma||_F^2}{d} = \mathcal{O}\left(\frac{1}{d}\right) \tag{48}$$

provided $||\Sigma||_F^2/d$ is finite. We used the fact that the variance of a $1-$degree of freedom χ^2 variable is 2. Plugging the definition of r into the above yields, for an off-diagonal entry $i \neq j$:

$$
\begin{aligned}
\Omega_{ij} &= \int \frac{e^{-\frac{1}{2}\frac{1}{r^2 - \mathcal{O}\left(\frac{1}{d}\right)}(ru^2 + rv^2)} e^{\frac{1}{r^2 - \mathcal{O}\left(\frac{1}{d}\right)} \frac{w_i^\top \Sigma w_j}{d} uv}}{2\pi \sqrt{r^2 - \mathcal{O}\left(\frac{1}{d}\right)}} \sigma(u)\sigma(v) \\
&= \left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}} \sigma(z) \right)^2 + \frac{1}{r} \frac{w_i^\top \Sigma w_j}{d} \left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}} z\sigma(z) \right)^2 + \mathcal{O}\left(\frac{1}{d}\right) \\
&= \kappa_1^2 \times \frac{w_i^\top \Sigma w_j}{d}.
\end{aligned} \tag{49}
$$

On the diagonal,

$$\Omega_{ii} = \int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}} \sigma(z)^2 = \kappa_*^2 + r\kappa_1^2. \tag{50}$$

Therefore,

$$\Omega = \kappa_1^2 \frac{W\Sigma W^\top}{d} + \kappa_*^2 I_k, \tag{51}$$

where

$$\kappa_1 = \frac{1}{r} \mathbb{E}_z^{\mathcal{N}(0,r)} [z\sigma(z)] \qquad\qquad \kappa_*^2 = \mathbb{E}_z^{\mathcal{N}(0,r)} [\sigma(z)^2] - r \times \kappa_1^2 \tag{52}$$

This extends the single-layer GET (Gerace et al., 2020) to arbitrary input covariances. A similar generalization was reported in (d'Ascoli et al., 2021) in the special case of Σ which spectral density is equal to a finite sum of Dirac atoms.

Iterating layer to layer (9) and (10) follow by straightforward recursion from the single-layer results (52) and (51), by making the strong assumption *that all post-activations can be treated as Gaussian variables*. This assumption is supported, at a physics level of rigor, by the work of (Fischer et al., 2022), which establishes that cumulants of order ≥ 4 are asymptotically suppressed at least as $1/k_\ell^2$. One then just needs to connect (35) to the single-layer variance r (47).

$$
\begin{aligned}
r_{\ell+1} &= \Delta_\ell \frac{1}{k_\ell} \operatorname{tr} \Omega_\ell \\
&= \Delta_{\ell+1} \left(\frac{1}{k_\ell} \left(\kappa_1^\ell \right)^2 \operatorname{tr} \left[\frac{W_\ell \Omega_{\ell-1} W_\ell^\top}{k_{\ell-1}} \right] + \left(\kappa_*^\ell \right)^2 \right) \\
&= \Delta_\ell \left(\left(\kappa_1^\ell \right)^2 r_\ell + \left(\kappa_*^\ell \right)^2 \right) \\
&= \Delta_\ell \mathbb{E}_z^{\mathcal{N}(0,r_\ell)} \left[\sigma(z)^2 \right].
\end{aligned}
\tag{53}
$$

We used

$$
\begin{aligned}
\frac{1}{k_\ell} \operatorname{tr} \left[\frac{W_\ell \Omega_{\ell-1} W_\ell^\top}{k_{\ell-1}} \right] &= \frac{1}{k_{\ell-1}} \sum_{i=1}^{k_{\ell-1}} \lambda_i^{\ell-1} \frac{1}{k_\ell} \left(U^\top W_\ell^\top W_\ell U \right)_{ii} \\
&= \frac{1}{k_{\ell-1}} \sum_{i=1}^{k_{\ell-1}} \lambda_i^{\ell-1} \Delta_\ell \\
&= \Delta_\ell \frac{1}{k_{\ell-1}} \operatorname{tr} \Omega_{\ell-1} = r_\ell
\end{aligned}
\tag{54}
$$

Finally, one must check that the assumption on Σ that $\|\Sigma\|_F^2/d,\ ^{\operatorname{tr}\Sigma}/d = \mathcal{O}(1)$ carries over to Ω. Because $W\Sigma W^\top$ is positive semi definite it is straightforward that

$$
\frac{1}{k} \left\| \kappa_1^2 \frac{W\Sigma W^\top}{d} + \kappa_*^2 I_k \right\|_F^2 \geq \kappa_*^2 > 0.
\tag{55}
$$

The upper bound can be established using the triangle inequality and the submultiplicativity of the Frobenius norm, as

$$
\begin{aligned}
\frac{1}{k} \left\| \kappa_1^2 \frac{W\Sigma W^\top}{d} + \kappa_*^2 I_k \right\|_F^2 &\leq \frac{1}{k} \left\| \kappa_1^2 \frac{W\Sigma W^\top}{d} \right\|_F^2 + \kappa_*^2 \\
&\leq \kappa_*^2 + \frac{\|W\|_F^4}{d^2} \frac{\|\Sigma\|_F^2}{k} \\
&\leq \kappa_*^2 + c' < \infty.
\end{aligned}
\tag{56}
$$

We used that $\|W\|_F^2/dk = 1$ almost surely asymptotically. Moving on to the trace,

$$
\frac{1}{k} \operatorname{Tr} \left[\kappa_1^2 \frac{W\Sigma W^\top}{d} + \kappa_*^2 I_k \right] = \kappa_*^2 + \frac{\kappa_1^2}{kd} \operatorname{Tr} \left[\Sigma W^\top W \right].
\tag{57}
$$

Bounding

$$
0 \leq \frac{\kappa_1^2}{kd} \operatorname{Tr} \left[\Sigma W^\top W \right] = \frac{\kappa_1^2}{kd} \sum_{i=1}^k w_i^\top \Sigma w_i = \kappa_1^2 \frac{1}{d} \operatorname{Tr} \{ \Sigma \} \leq \kappa_1^2 c',
\tag{58}
$$

where the last bound holds asymptotically almost surely. As discussed in the main text, this derivation carries over to the case where W is sampled from the Bayes posterior (5), with the Nishimori identities (Nishimori, 2001; Iba, 1998) ensuring its statistics are Gaussian.

A.4. Derivation sketch for Φ

This subsection deals with the cross-covariance Φ, which is needed to connect with the results of (Loureiro et al., 2021) to compute closed-form sharp asymptotics for the test error of linear ERM procedures. We first derive two preliminary results, corresponding to propagations through single layers, and conclude by constructing the straightforward recursion.

Two Gaussians propagating through two layers Consider two jointly Gaussian variables $u \in \mathbb{R}^d$, $v \in \mathbb{R}^k$

$$(u,v) \sim \mathcal{N}\left(\begin{array}{cc} \Psi & \Phi \\ \Phi^\top & \Omega \end{array}\right) \tag{59}$$

each independently propagated through a non-linear layer

$$h^\star(u) = \sigma_\star\left(\frac{1}{\sqrt{d_\star}}W_\star \cdot u\right), \qquad\qquad h(v) = \sigma\left(\frac{1}{\sqrt{d}}W \cdot v\right). \tag{60}$$

The weights $W_\star \in \mathbb{R}^{k_\star \times d_\star}$ and $W \in \mathbb{R}^{k \times d}$ have independently sampled Gaussian entries, with respective variance $\Delta_\star$ and Δ. The $i,j-$th element of the cross-covariance Φ^h can be expressed as

$$\Phi^h_{ij} = \langle h_i^\star(u)h_j(v)\rangle_{u,v} = \int \frac{e^{-\frac{1}{2}\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} \frac{w_i^{\star\top}\Sigma w_i^\star}{d_\star} & \frac{w_i^{\star\top}\Sigma w_j}{\sqrt{d_\star d}} \\ \frac{w_i^{\star\top}\Sigma w_j}{\sqrt{d_\star d}} & \frac{w_j^\top \Sigma w_j}{d}\end{pmatrix}^{-1}\begin{pmatrix} x \\ y \end{pmatrix}}}{\sqrt{\det 2\pi \begin{pmatrix} \frac{w_i^{\star\top}\Sigma w_i^\star}{d_\star} & \frac{w_i^{\star\top}\Sigma w_j}{\sqrt{d_\star d}} \\ \frac{w_i^{\star\top}\Sigma w_j}{\sqrt{d_\star d}} & \frac{w_j^\top \Sigma w_j}{d}\end{pmatrix}}}\sigma_\star(x)\sigma(y). \tag{61}$$

As before, the random variables $w_i^{\star\top}\Sigma w_i^\star/d_\star$ and $w_j^\top \Sigma w_j/d$ concentrate around their mean value

$$r_\star \equiv \frac{\Delta_\star}{d_\star}\operatorname{tr}\Psi, \qquad\qquad r \equiv \frac{\Delta}{d}\operatorname{tr}\Omega. \tag{62}$$

Plugging these definitions into the above:

$$\begin{aligned}
\Phi^h_{ij} &= \int \frac{e^{-\frac{1}{2}\frac{1}{r_\star r - \mathcal{O}\left(\frac{1}{d}\right)}(rx^2 + r_\star y^2)}e^{\frac{1}{r_\star r - \mathcal{O}\left(\frac{1}{d}\right)}\frac{w_i^{\star\top}\Phi w_j}{\sqrt{d_\star d}}xy}}{2\pi\sqrt{r_\star r - \mathcal{O}\left(\frac{1}{d}\right)}}\sigma_\star(x)\sigma(y) \\
&= \left(\int \frac{e^{-\frac{1}{2r_\star}z^2}}{\sqrt{2\pi r_\star}}\sigma_\star(z)\right)\left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}}\sigma(z)\right) + \frac{1}{r_\star r}\frac{w_i^{\star\top}\Phi w_j}{\sqrt{d_\star d}}\left(\int \frac{e^{-\frac{1}{2r_\star}z^2}}{\sqrt{2\pi r_\star}}z\sigma_\star(z)\right)\left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}}z\sigma(z)\right) + \mathcal{O}\left(\frac{1}{d}\right) \\
&:= \kappa_1\kappa_1^\star \times \frac{w_i^{\star\top}\Phi w_j}{\sqrt{d_\star d}},
\end{aligned} \tag{63}$$

yielding

$$\Phi^h = \kappa_1\kappa_1^\star\frac{W_\star\Phi W^\top}{\sqrt{d_\star d}}, \tag{64}$$

with

$$\kappa_1 = \frac{1}{r}\mathbb{E}_z^{\mathcal{N}(0,r)}\left[z\sigma(z)\right] \qquad\qquad \kappa_1 = \frac{1}{r_\star}\mathbb{E}_z^{\mathcal{N}(0,r_\star)}\left[z\sigma_\star(z)\right]. \tag{65}$$

One Gaussian propagating through one layer Consider two jointly Gaussian variables $u \in \mathbb{R}^{d_\star}$, $v \in \mathbb{R}^d$

$$(u,v) \sim \mathcal{N}\left(\begin{array}{cc} \Psi & \Phi \\ \Phi^\top & \Omega \end{array}\right) \tag{66}$$

with *only* v being propagated through a non linear layer

$$h(v) = \sigma\left(\frac{1}{\sqrt{k}}W \cdot v\right). \tag{67}$$

17

The entries $W \in \mathbb{R}^{k \times d}$ are independently sampled from a Gaussian distribution with variance Δ. The $i, j-$th element of the cross-covariance Φ between $h(v)$ and u can be expressed as

$$\Phi_{ij}^h = \langle u_i h_j(v) \rangle_{u,v} = \int \frac{e^{-\frac{1}{2} \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} \Psi_{ii} & \frac{\Phi_i w_j}{\sqrt{k}} \\ \frac{\Phi_i w_j}{\sqrt{k}} & \frac{w_j^\top \Sigma w_j}{k} \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix}}}{\sqrt{\det 2\pi \begin{pmatrix} \Psi_{ii} & \frac{\Phi_i w_j}{\sqrt{k}} \\ \frac{\Phi_i w_j}{\sqrt{k}} & \frac{w_j^\top \Sigma w_j}{k} \end{pmatrix}}} x\sigma(y). \tag{68}$$

Again, the random variable $w_j^\top \Sigma w_j / k$ concentrate around its mean value

$$r \equiv \frac{\Delta}{k} \operatorname{tr} \Omega. \tag{69}$$

Plugging this definition into the above:

$$\Phi_{ij}^h = \int \frac{e^{-\frac{1}{2} \frac{1}{\Psi_{ii} r - \mathcal{O}\left(\frac{1}{d}\right)}(rx^2 + \Psi_{ii} y^2)} e^{\frac{1}{\Psi_{ii} r - \mathcal{O}\left(\frac{1}{d}\right)} \frac{\Phi_i w_j}{\sqrt{k}} xy}}{2\pi \sqrt{\Psi_{ii} r - \mathcal{O}\left(\frac{1}{d}\right)}} x\sigma(y)$$

$$= \left(\int \frac{e^{-\frac{1}{2\Psi_{ii}} z^2}}{\sqrt{2\pi \Psi_{ii}}} z \right) \left(\int \frac{e^{-\frac{1}{2r} z^2}}{\sqrt{2\pi r}} \sigma(z) \right) + \frac{1}{\Psi_{ii} r} \frac{\Phi_i w_j}{\sqrt{k}} \left(\int \frac{e^{-\frac{1}{2\Psi_{ii}} z^2}}{\sqrt{2\pi \Psi_{ii}}} z^2 \right) \left(\int \frac{e^{-\frac{1}{2r} z^2}}{\sqrt{2\pi r}} z\sigma(z) \right) + \mathcal{O}\left(\frac{1}{d}\right)$$

$$:= \kappa_1 \times \frac{\Phi_i w_j}{\sqrt{k}} \tag{70}$$

yielding

$$\Phi^h = \kappa_1 \frac{\Phi W^\top}{\sqrt{k}} \tag{71}$$

with

$$\kappa_1 = \frac{1}{r} \mathbb{E}_z^{\mathcal{N}(0,r)} [z\sigma(z)]. \tag{72}$$

Iterating Iterating (71) L times under Gaussianity assumptions on the post-activations yields the cross-covariance between the last layer post-activations of the teacher and the input of a linear ERM (in the present work, ridge regression/classification, and logistic regression), which can in turn be plugged into (Loureiro et al., 2021) to access sharp error asymptotics. The exact expressions for the covariances Ψ, Φ and Ω used are detailed on a case-per-case basis in Appendix E and H.

Applying (64) and then (71) $L - 1$ times yields the cross covariance between the last layer post-activations of the teacher and the random features in the framework of random features regression. This extends the results of (Gerace et al., 2020) to a multilayer target function.

This result carries over to the case where $W, W_\star$ are sampled from the Bayes posterior (or, because of the Nishimori identites (Nishimori, 2001; Iba, 1998), when $W_\star$ is a the target weight), assuming non-specialization 3.2. The latter ensures that the term $w_i^{\star\top} \Phi w_j / \sqrt{dd_\star}$ stay of order $1/\sqrt{d}$ for all i, j.

A.5. Numerical evidence of the closed-form recursions

We close the discussion on the closed-form formulae for the population covariances by providing numerical evidence of their validity. Figures 5, 6 and 7 show the first 10 rows and columns of the numerically estimated population covariance $\Omega^{\text{emp.}}$, estimated from the sample covariance of $N = 10^5$ Gaussian samples x, and the closed-form formula (10), for a rectangular network in dimension $d = 500$ and tanh/sign/erf activations. These covariances concern the activations at layers $\ell = 4$ and $\ell = 7$. In all these plots, for clarity, a matrix $t_\ell \mathbb{I}_{k_\ell}$ was substracted, with

$$t_\ell = \sum_{\ell^0=1}^{\ell-1} \left(\kappa_*^{\ell^0}\right)^2 \prod_{\ell'=\ell^0+1}^{L} \left(\kappa_1^{\ell'}\right)^2 \Delta_{\ell'} + \left(\kappa_*^L\right)^2 + \int z d\mu(z) \prod_{\ell^0=1}^{\ell} \left(\kappa_1^{\ell^0}\right)^2 \Delta_{\ell^0}. \tag{73}$$
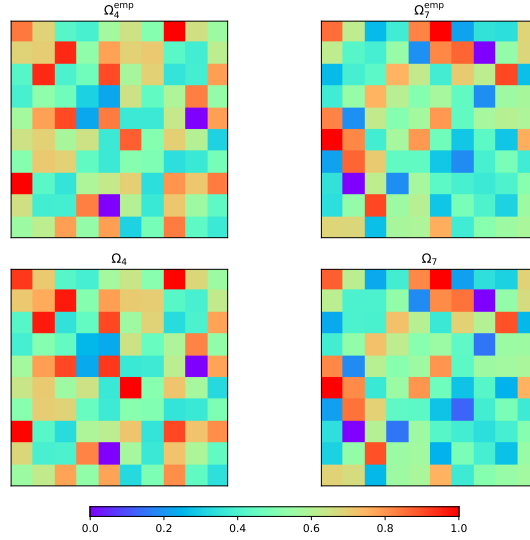
Figure 5. First 10 rows of the population covariance matrix of activations at layers $\ell = 4$ (left) and $\ell = 7$(right), for a random network with tanh activation and architecture $\forall \ell,\ \gamma_\ell = 1$, in dimension $d = 500$.Top: numerically computed population covariance $\Omega_\ell^{\mathrm{emp.}}$ estimated from $N = 10^5$ samples. Bottom: theoretical Ω_ℓ, computed from the closed-form formula (10).

t_ℓ is the theoretical average value of $^{\mathrm{tr}}\Omega_\ell/d$. The visual agreement between the numerically estimated $\Omega_\ell^{\mathrm{emp.}}$ and the theoretical Ω_ℓ is very good, with small relative distances $||\Omega_\ell^{\mathrm{emp.}} - \Omega_\ell||_F^2/||\Omega^{\mathrm{emp.}}||_F^2 \approx 0.005$ for tanh, $||\Omega_\ell^{\mathrm{emp.}} - \Omega_\ell||_F^2/||\Omega^{\mathrm{emp.}}||_F^2 \approx 0.008$ for sign, and $||\Omega_\ell^{\mathrm{emp.}} - \Omega_\ell||_F^2/||\Omega^{\mathrm{emp.}}||_F^2 \approx 0.004$ for erf.

A.6. Numerical test of the 1dCLT

We provide numerical evidence of the 1dCLT 3.3, namely that the output $\hat{y}(x)$ of a deep random Gaussian network is Gaussian. Note that from the closed-form formula (10), the theoretical variance of the output reads

$$\check{q} = \sum_{\ell=1}^{L-1} \left(\kappa_*^\ell\right)^2 \prod_{\ell'=\ell+1}^{L} \left(\kappa_1^{\ell'}\right)^2 \Delta_{\ell'} + \left(\kappa_*^L\right)^2 + \int z d\mu(z) \prod_{\ell=1}^{L} \left(\kappa_1^\ell\right)^2 \Delta_\ell. \tag{74}$$

The empirical histograms of the scaled output $\hat{y}(x)/\sqrt{\check{q}}$ of networks with $L = 3$ and $L = 9$ hidden layers are plotted in Fig. 9 (for tanh activation) and Fig. 8 (for sign activation). In all cases, the widths are taken equal to the input dimension d. Several sizes d are plotted. As supplementary confirmation, quantile-quantile (QQ) plots are also included. Fig. 10 further shows, for a depth $L = 3$ network with tanh activation, that the higher order cumulants of the output distribution is asymptotically suppressed.. These numerical simulation provide compelling evidence that as $d \to \infty$ the distribution of $\hat{y}(x)/\sqrt{\check{q}}$ converges to a normal distribution.

A.7. Towards rigorous proofs

It would be desirable to rigorously prove the conjectured formulae discussed in this paper. There are a few difficulties that we discuss now briefly.

Concerning the single layer conjecture 3.1(**Shallow Bayes GEP**), a full theorem can probably be within reach of current techniques with, however, one difficulty, summarized in conjecture 3.2. Indeed, the Gaussian equivalence theorem applied to two neural nets with weights W_1 and W_2 implies that both of them are effectively linear so that, in terms of second-order statistics, $\sigma(W_1 x) \approx \kappa_1^{(1)} W_1 X + \kappa_*^{(1)} Z_1$ and $\sigma(W_2 x) \approx \kappa_1^{(1)} W_2 X + \kappa_*^{(1)} Z_2$, with $Z_{1,2}$ Gaussian-distributed. When W_1 and W_2 are independent, so are Z_1 and Z_2. However, when W_1 and W_2 are correlated then Z_1 and Z_2 becomes correlated Gaussians as well, thus complicating the rigorous analysis. As noted in the main text, the work of (Aubin et al., 2018) shows that this is not expected to happens, but a rigorous control remains so far elusive, and is left for future work.

The deep conjecture requires significantly more work, as an additional difficulty arises. While conjecture 3.4 should also

Figure 6. First 10 rows of the population covariance matrix of activations at layers $\ell = 4$ (left) and $\ell = 7$(right), for a random network with sign activation and architecture $\forall \ell$, $\gamma_\ell = 1$, in dimension $d = 500$.Top: numerically computed population covariance $\Omega_\ell^{\text{emp.}}$ estimated from $N = 10^5$ samples. Bottom: theoretical Ω_ℓ, computed from the closed-form formula (10). Due to the nature of the sign activation, the finite size effects are slightly stronger than for the tanh activation (Fig.5).

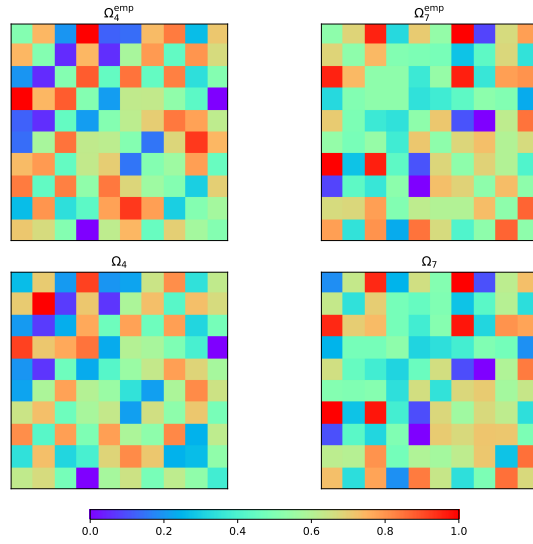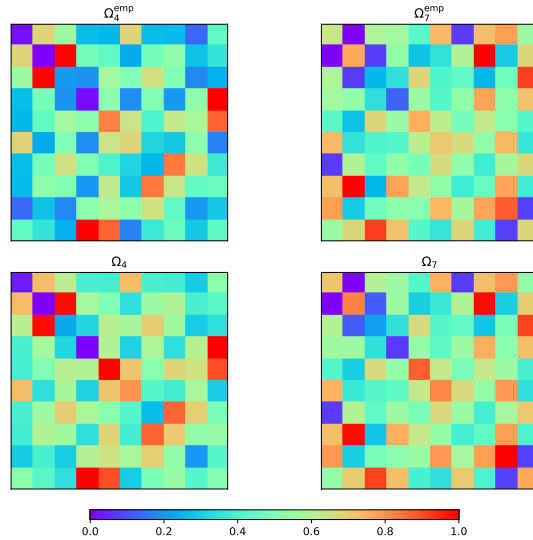*Figure 7.* First 10 rows of the population covariance matrix of activations at layers $\ell = 4$ (left) and $\ell = 7$(right), for a random network with erf activation and architecture $\forall \ell$, $\gamma_\ell = 1$, in dimension $d = 500$.Top: numerically computed population covariance $\Omega_\ell^{\text{emp.}}$ estimated from $N = 10^5$ samples. Bottom: theoretical Ω_ℓ, computed from the closed-form formula (10).
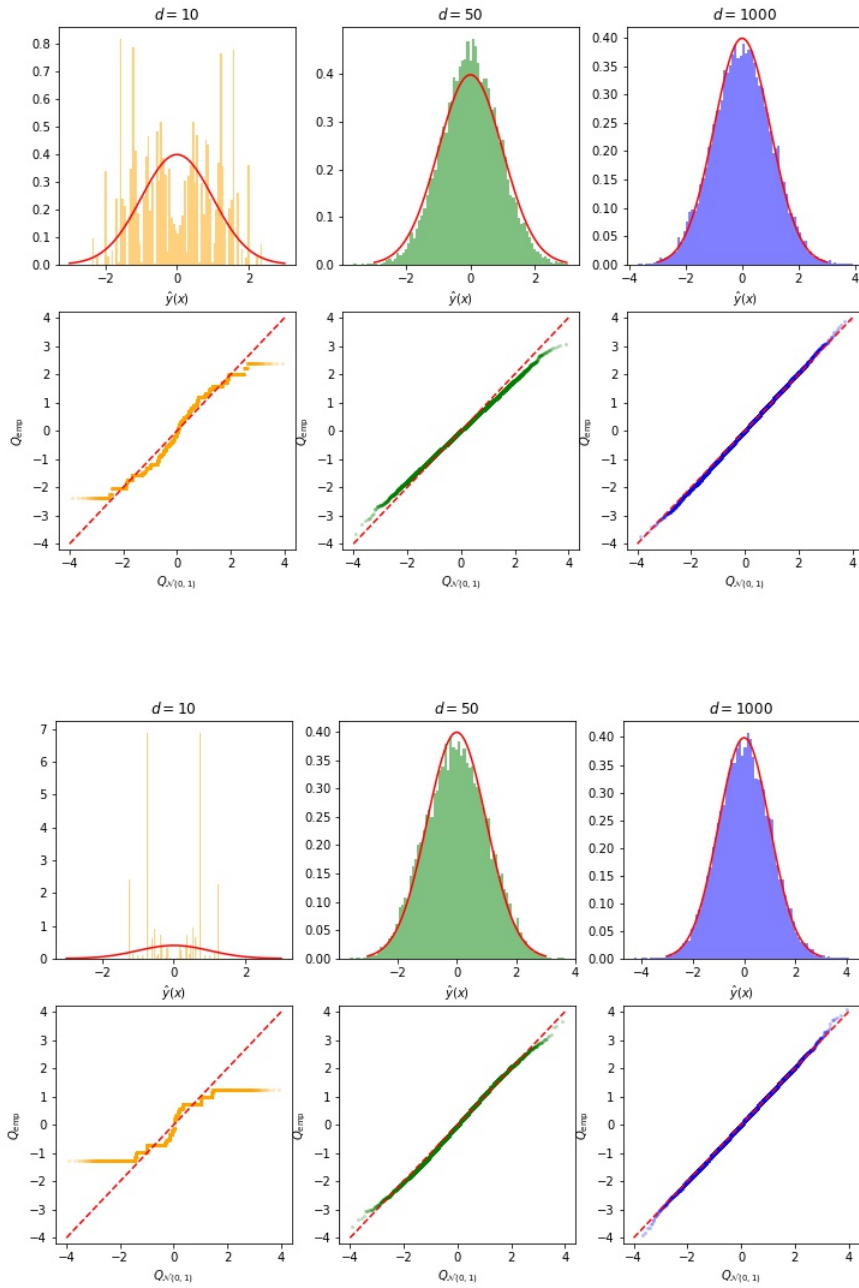
Figure 8. We study the rescaled output $\hat{y}(x)/\sqrt{q}$ of a random Gaussian network, $L = 3$ (top) and $L = 9$ (bottom) hidden layers, sign activations, and widths $k_1 = ... = k_L = d$. From left to right, the dimension d is varied from $d = 10$ to $d \sim 1000$. (top) Histogram of the output. The red line represents a normal distribution. (bottom) Quantile-Quantile plot. The x axis represent the theoretical quantiles of a normal distribution $Q_{\mathcal{N}(0,1)}$, while the empirical quantiles are plotted on the y axis.
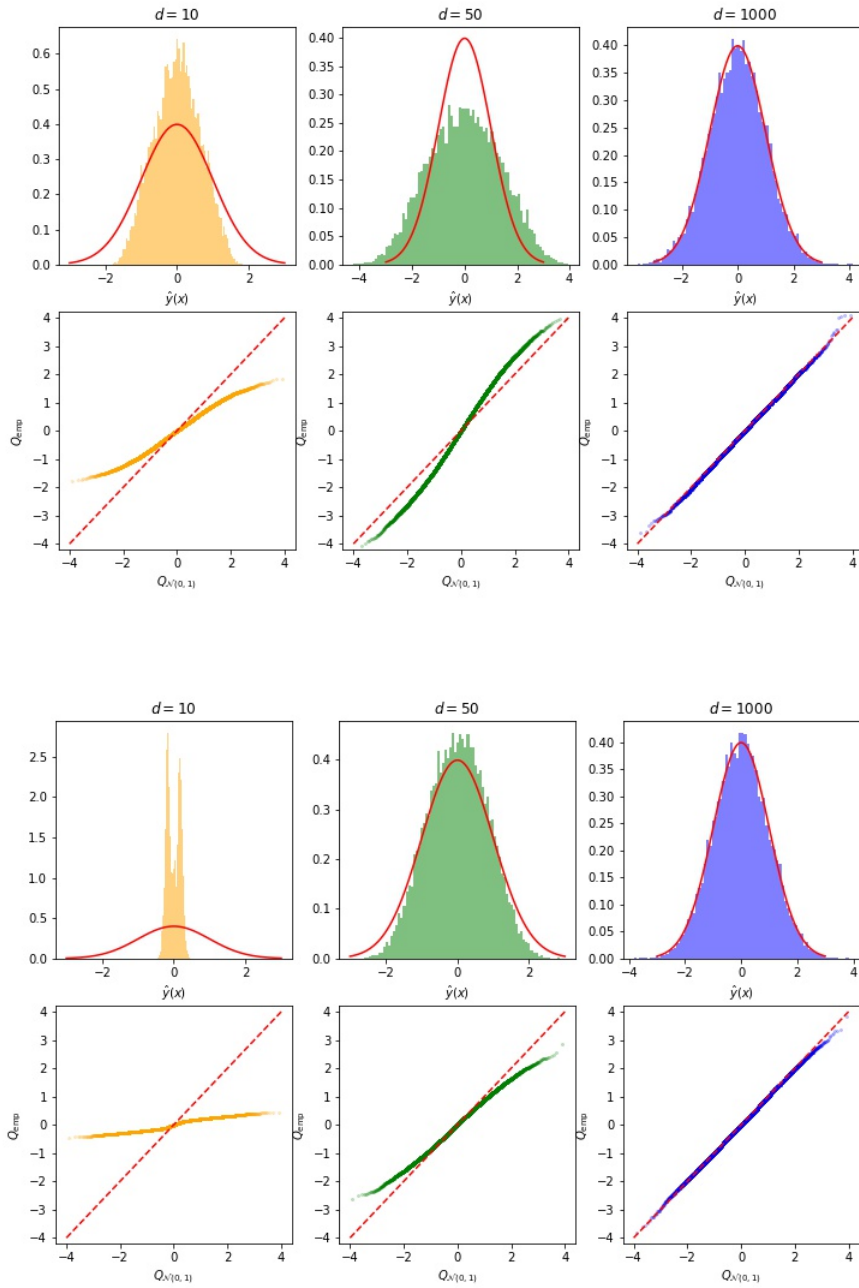
Figure 9. We study the output $\hat{y}(x)/\sqrt{q}$ of a random Gaussian network, $L = 3$ (top) and $L = 9$ (bottom) hidden layers, tanh activations, and widths $k_1 = ... = k_L = d$. From left to right, the dimension d is varied from $d = 10$ to $d \sim 1000$. (top) Histogram of the output. The red line represents a normal distribution. (bottom) Quantile-Quantile plot. The x axis represent the theoretical quantiles of a normal distribution $Q_{\mathcal{N}(0,1)}$, while the empirical quantiles are plotted on the y axis.
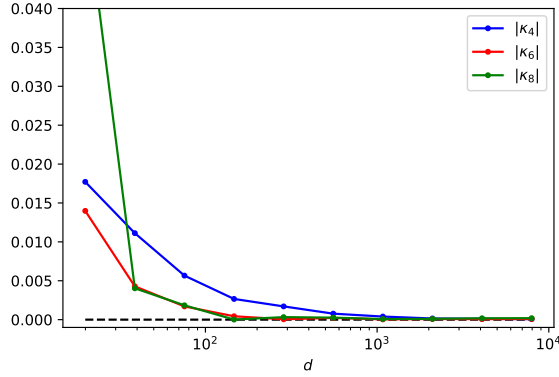
Figure 10. Order $4, 6, 8$ of the output $\hat{y}(x)/\sqrt{q}$ of a random Gaussian network of depth $L = 3$ with tanh activations, and widths $k_1 = \ldots = k_L = d$. The cumulants were evaluated from $N = 5.10^4$ samples, for one realization of the weights. Note that because of the fact that the activation is odd, all odd cumulants are vanishing.

be amenable to a rigorous treatment similar to the one of conjecture 3.1, we are still lacking a control of a deep GET as expressed in 3.3 (10). While much progress has been attained on the topic, giving very strong indication for its validity (see e.g. (Fan & Wang, 2020; Schröder et al., 2023; Bosch et al., 2023)) this is still an open problem mathematically.

## B. Replica computation for shallow networks

This appendix provides a detailed derivation of the asymptotic formula for the test error of the Bayes optimal MSE using the replica method (Parisi, 1979; Zdeborová & Krzakala, 2015; Gabrié, 2019). The replica method has been leveraged in a very large body of works (Seung et al., 1992; Watkin et al., 1993; Talagrand, 2006; Barbier et al., 2017; Advani & Ganguli, 2016; Aubin et al., 2020; Cui et al., 2019; Maillard et al., 2020; Loureiro et al., 2021; Aubin et al., 2018; Gerace et al., 2020; Canatar et al., 2020; Cui et al., 2021; 2022; Zavatone-Veth et al., 2022) to access exact asymptotic characterizations of accuracy metrics in simple learning tasks, for both Bayesian learning and ERM.

In this appendix, we consider the case of a two-layer target

$$y^\mu \sim P_{\text{out}}\left(\cdot \mid \frac{1}{k_\star} a_\star^\top \sigma_\star \left(\frac{1}{\sqrt{d}} W_\star^\top x^\mu\right)\right)$$

and study Bayesian learning using a two-layer network

$$\hat{y}(x) = \frac{1}{k} a_\star^\top \sigma\left(\frac{1}{\sqrt{d}} W^\top x^\mu\right).$$

Note that the large part of the computation is detailed for the generic case where the student architecture is not required to match that of the target, i.e. $k \neq k_\star$. We denote by P_{out} the Gaussian measure of variance Δ, corresponding to the Gaussian additive channel (1). Similarly, we adopt the notation $P_w(\cdot)$ (resp. $P_a(\cdot)$) for the Gaussian distribution of variance Δ_w (resp. Δ_a). In the last sections, we specialize to the Bayes optimal setting $\sigma = \sigma_\star$ and $k = k_\star$ to derive the Bayes-optimal MSE (12). The general case with deep target and student is discussed in Appendix C.

The exact computation of the Bayes posterior measure (6) is generically intractable. However, it is possible to compute the associated free energy

$$f \equiv -\mathbb{E}_{a_\star, \{W_\ell^\star\}} \mathbb{E}_{\mathcal{D}} \ln Z$$

where Z is the partition function associated to the measure (6), i.e.

$$Z = \int \mathrm{d}a \int \mathrm{d}W \, P_w(W) P_a(a) P_{\text{out}}(y^\mu | \hat{y}(x^\mu)). \tag{75}$$

We introduced the priors

$$P_w(W) \propto e^{-\frac{1}{2\Delta_w}||W_\ell||_F^2}, \qquad\qquad P_a(a) \propto e^{-\frac{1}{2\Delta_a}||a||^2}, \qquad\qquad (76)$$

and the posteriors

$$P_{\text{out}}(y|\hat{y}) \propto e^{-\frac{1}{2\Delta}(y-\hat{y})^2} \qquad\qquad (77)$$

The computation of the free energy yields asymptotic characterizations of the overlaps

$$q_w = \mathbb{E}\left\|\left\langle \frac{a^\top W}{\sqrt{kd}} \right\rangle\right\|^2, \qquad q_a = \mathbb{E}\left\langle \left\|\frac{a}{\sqrt{k}}\right\|^2 \right\rangle, \qquad m_w = \mathbb{E}\frac{\langle a^\top W \rangle W_\star^\top a_\star}{\sqrt{kk_\star d}},$$

which fully describe the asymptotic average test error. In order to compute the free energy, the *replica trick* (Parisi, 1979; 1983)

$$\ln Z = \lim_{s\to 0} \frac{Z^s - 1}{s}$$

can be employed.

B.1. Replica trick

The replicated partition function reads

$$Z^s = \int \prod_{a=1}^s dW^a da^a P_w(W^a) P_a(a^a) \underbrace{\mathbb{E}_{\{x^\mu\}_\mu} \int \prod_{\mu=1}^n P_{\text{out}}^\star\left(y^\mu|\frac{1}{\sqrt{k_\star}}a_\star^\top \sigma_\star\left(\frac{1}{\sqrt{d}}W_\star x^\mu\right)\right) \prod_{\mu=1}^n \prod_{a=1}^s P_{\text{out}}\left(y^\mu|\frac{1}{\sqrt{k}}a_a^\top \sigma_\star\left(\frac{1}{\sqrt{d}}W_a x^\mu\right)\right)}_{*}.$$

$$(78)$$

The expectation with respect to the train set in the energetic part can be carried out as:

$$* = \prod_\mu \left[\mathbb{E}_{\lambda_\star, \{\lambda_a\}_{a=1}^s} \int dy P_{\text{out}}^\star(y|\lambda_\star) \prod_{a=1}^s P_{\text{out}}(y|\lambda_a) \right]. \qquad\qquad (79)$$

We have introduced the local fields

$$\lambda_\star \equiv \frac{1}{\sqrt{k_\star}}a_\star^\top \sigma_\star\left(\frac{1}{\sqrt{d}}W_\star x^\mu\right), \qquad\qquad \lambda_a = \frac{1}{\sqrt{k}}a_a^\top \sigma_\star\left(\frac{1}{\sqrt{d}}W_a x^\mu\right). \qquad\qquad (80)$$

B.2. GET-linearization of the hidden layer

The shallow Bayes GEP (3.1) implies that these $s+1$ fields are jointly Gaussian and have statistics

$$\begin{cases} \langle \lambda_a \lambda_b \rangle = \kappa_1^2 \frac{1}{kd}a_a^\top W_a \Sigma W_b^\top a_b + \delta_{ab}\kappa_*^2 \frac{1}{k}a_a^\top a_b \\ \langle \lambda_a \lambda_\star \rangle = \kappa_1 \kappa_1^\star \frac{1}{dk}a_a^\top W_a \Sigma W_\star^\top a_\star \\ \langle \lambda_\star \lambda_\star \rangle = \kappa_1^{\star 2} \frac{1}{kd}a_\star^\top W_\star \Sigma W_\star^\top a_\star + \kappa_*^{\star 2} \frac{1}{k}a_\star^\top a_\star \end{cases}, \qquad\qquad (81)$$

where

$$\begin{cases} \kappa_1 = \frac{1}{\frac{\text{Tr}\,\Sigma}{\Delta_w^{-1}d}}\mathbb{E}_z^{\mathcal{N}\left(0,\frac{\text{Tr}\,\Sigma}{\Delta_w^{-1}d}\right)}[z\sigma(z)] \\ \kappa_* = \sqrt{\mathbb{E}_z^{\mathcal{N}\left(0,\frac{\text{Tr}\,\Sigma}{\Delta_w^{-1}d}\right)}[\sigma(z)^2] - \kappa_1^2 \frac{\text{Tr}\,\Sigma}{\Delta_w^{-1}d}} \end{cases} \qquad \begin{cases} \kappa_1^\star = \frac{1}{\Delta_w \frac{\text{Tr}\,\Sigma}{d}}\mathbb{E}_z^{\mathcal{N}\left(0,\Delta_w \frac{\text{Tr}\,\Sigma}{d}\right)}[z\sigma(z)] \\ \kappa_*^\star = \sqrt{\mathbb{E}_z^{\mathcal{N}\left(0,\frac{\text{Tr}\,\Sigma}{\Delta_w^{-1}d}\right)}[\sigma(z)^2] - \kappa_1^2 \Delta_w \frac{\text{Tr}\,\Sigma}{d}} \end{cases}, \qquad (82)$$

see (9).

We introduce the order parameters:

$$q_{ab}^w = \frac{1}{kd} a_a^\top W_a \Sigma W_b^\top a_b \qquad m_a^w = \frac{1}{kd} a_\star^\top W_\star \Sigma W_\star^\top a_\star \qquad q_{ab}^a = \delta_{ab} \frac{1}{k} a_a^\top a_b \tag{83}$$

$$\rho_w = \frac{1}{kd} a_\star^\top W_\star \Sigma W_\star^\top a_\star \qquad \rho_a = \frac{1}{k} a_\star^\top a_\star \tag{84}$$

Introducing Diracs in their Fourier transform form to enforce those definitions, the replicated partition function reads

$$Z^s = \int \prod_{a \leq b} dq_{ab}^w dq_{ab}^a \prod_a dm_a^w \underbrace{e^{-\sum_a \hat{m}_a^w m_a^w + \sum_{a \leq b}(\hat{q}_{ab}^w q_{ab}^w + \hat{q}_{ab}^a q_{ab}^a)}}_{e^{sd\Psi_t}}$$

$$\underbrace{\int \prod_{a=1}^s dW_a da_a P_w(W_a) P_a(a_a) e^{\sum_{a \leq b} \hat{q}_{ab}^w \frac{a_a^\top W_a \Sigma W_b^\top a_b}{k} + \sum_{a \leq b} \hat{q}_{ab}^a \frac{a_a^\top a_b}{\gamma} + \sum_a \hat{m}_a^w \frac{a_a^\top W_a \Sigma W_\star^\top a_\star}{\sqrt{kk^\star}}}}_{e^{sd\Psi_w}}$$

$$\underbrace{\prod_\mu \left[\mathbb{E}_{\lambda_\star, \{\lambda_a\}_{a=1}^s} \int dy P_{\text{out}}^\star(y|\lambda_\star) \prod_{a=1}^s P_{\text{out}}(y|\lambda_a) \right]}_{e^{sd\Psi_y}} \tag{85}$$

B.3. Entropic potential

We first compute the entropic potential Ψ_w.

$$e^{sd\Psi_w} = \int \prod_{a=1}^s dW_a da_a P_w(W_a) P_a(a_a) e^{\sum_{a \leq b} \hat{q}_{ab}^w \frac{a_a^\top W_a \Sigma W_b^\top a_b}{k} + \sum_{a \leq b} \hat{q}_{ab}^a \frac{a_a^\top a_b}{\gamma} + \sum_a \hat{m}_a^w \frac{a_a^\top W_a \Sigma W_\star^\top a_\star}{\sqrt{kk^\star}}} \tag{86}$$

$$= \underbrace{\int \prod_{a=1}^s da_a P_a(a_a) e^{\sum_{a \leq b} \hat{q}_{ab}^a \frac{a_a^\top a_b}{\gamma}}}_{(a)} \prod_{j=1}^d \underbrace{\left[\int \prod_{a=1}^s dw_a P_w(w_a) e^{\sum_{a \leq b} \sigma_j \hat{q}_{ab}^w \frac{(a_a^\top w_a)(a_b^\top w_b)}{k} + \sum_a \sigma_j \hat{m}_a \frac{(a_a^\top W_a)(a_\star^\top w_j^\star)}{\sqrt{kk^\star}}} \right]}_{(b)} \tag{87}$$

We have used the fact that the prior $P_w(\cdot)$, being Gaussian, factorizes over the d columns $\{w_j\}_{j=1}^d$ of the first layer weights $W, W^\star$. Also, Σ has been without loss of generality supposed to be diagonal, i.e. $\Sigma = \text{diag}(\sigma_1, ..., \sigma_d)$. Note that all the final expression hold for generic Σ, and that one can carry out the computation in full detail without this shortcut by introducing $\Sigma = U \text{diag}(\sigma_1, ..., \sigma_d) U^\top$ and $U-$ rotating $W, W^\star$. Defining

$$\eta_a = \frac{a_a^\top w_a}{\sqrt{k}} \tag{88}$$

and remembering $P_w(w) = e^{-\frac{\Delta_w^{-1}}{2} w^\top w}$, it follows that the variables $\{\eta_a\}_{a=1}^s$ are jointly Gaussian with statistics

$$\begin{cases} \langle \eta_a \rangle = 0 \\ \langle \eta_a \eta_b \rangle = \delta_{ab} \frac{q_{aa}^a}{\Delta_w^{-1}}. \end{cases} \tag{89}$$

It then follows that

$$(b) = \int \left(\prod_{a=1}^s \frac{d\eta_a}{\sqrt{2\pi \frac{q_{aa}^a}{\Delta_w^{-1}}}} e^{-\frac{1}{2} \frac{\Delta_w^{-1}}{q_{aa}^a} \eta_a^2} \right) e^{\sigma_j \sum_{a \leq b} \hat{q}_{ab}^w \eta_a \eta_b + \sigma_j \frac{a_\star^\top w_j^\star}{\sqrt{k^\star}} \sum_a \hat{m}_a \eta_a} \tag{90}$$

$$= e^{\frac{1}{2} \sigma_j^2 \left(\frac{a_\star^\top w_j^\star}{\sqrt{k^\star}} \right)^2 \hat{m}^\top \left[\text{diag}\left(\frac{\Delta_w^{-1}}{q_{aa}^a} \right) - \sigma_j \times (2I_s) \odot \hat{Q}_w \right]^{-1} \hat{m} - \frac{1}{2} \ln \det \left[I_s - \sigma_j \text{diag}\left(\frac{\Delta_w^{-1}}{q_{aa}^a} \right) (2I_s) \odot \hat{Q}_w \right]}. \tag{91}$$

B.4. Replica-symmetric ansatz

Since the integral for Z^s involves exponentials with arguments of order d, Laplace method suggests that the integral concentrates over the maximizer of the exponent $\Psi_t + \Psi_w + \Psi_y$. Since explicitly carrying out the optimization over the order parameters $q_{ab}^w, q_{ab}^a, m_{ab}^w$ is hard, we look for solutions of the form

$$\hat{q}_{ab}^w = \delta_{ab}\left(-\frac{\hat{r}_w}{2} - \hat{q}_w\right) + \hat{q}_w \tag{92}$$

$$\hat{q}_{ab}^a = -\delta_{ab}\frac{\hat{r}_a}{2} \tag{93}$$

$$\hat{m}_a^w = \hat{m}_w \tag{94}$$

and

$$q_{ab}^w = \delta_{ab}(r_w - q_w) + q_w \tag{95}$$

$$q_{ab}^a = \delta_{ab}r_a \tag{96}$$

$$m_a^w = m_w. \tag{97}$$

This ansatz is standardly known as the *replica symmetric* ansatz. For convenience, we further define the variances

$$V_w = r_w - q_w, \qquad\qquad V_a = r_a, \qquad\qquad \hat{V}_w = \hat{r}_w + \hat{q}_w, \qquad\qquad \hat{V}_a = \hat{r}_a. \tag{98}$$

B.5. RS free energy

Entropy potential We first recall as a preliminary statement two linear algebra results for matrices of the form $(\delta_{ab}(r - q) + q)_{1 \le a, b \le s}$:

$$(\delta_{ab}(r - q) + q)_{ab}^{-1} = \left(\delta_{ab}\left(\frac{r + (s-2)q}{(r-q)^2 + sq(r-q)} + \frac{q}{(r-q)^2 + sq(r-q)}\right) - \frac{q}{(r-q)^2 + sq(r-q)}\right)_{ab} \tag{99}$$

and

$$\det(\delta_{ab}(r - q) + q)_{ab}^{-1} = (r + (s-1)q)(r - q)^{s-1} \tag{100}$$

These identities can be leveraged to simplify the two terms in (89) as

$$\hat{m}^\top\left[\operatorname{diag}\left(\frac{\Delta_w^{-1}}{q_{aa}^a}\right) - \sigma_j \times (2I_s) \odot \hat{Q}_w\right]^{-1}\hat{m} = \sigma_j^2 s\hat{m}^2 \frac{1}{\sigma_j \hat{V}_w + \frac{\Delta_w^{-1}}{r_a}} \tag{101}$$

and

$$\ln\det\left[I_s - \sigma_j\operatorname{diag}\left(\frac{\Delta_w^{-1}}{q_{aa}^a}\right)(2I_s) \odot \hat{Q}_w\right] = s\ln\left(\frac{r_a}{\Delta_w^{-1}}\hat{V}_w\sigma_j + 1\right) - s\frac{\sigma_j \hat{q}_w}{\sigma_j \hat{V}_w + \frac{\Delta_w^{-1}}{r_a}}. \tag{102}$$

Thus

$$\prod_{j=1}^d (b) = e^{\frac{s\hat{m}^2}{2}\operatorname{Tr}\left[\frac{\Sigma^2 \frac{W_\star^\top a_\star a_\star^\top W_\star}{k_\star}}{\hat{V}_w\Sigma + \frac{\Delta_w^{-1}}{r_a}I_d}\right] - \frac{s}{2}\ln\det\left[\frac{r_a}{\Delta_w^{-1}}\hat{V}_w\Sigma + I_d\right] + \frac{s}{2}\operatorname{Tr}\left[\frac{\hat{q}_w\Sigma}{\hat{V}_w\Sigma + \frac{\Delta_w^{-1}}{r_a}I_d}\right]} \tag{103}$$

Computing now the (a) term:

$$(a) = \prod_{j=1}^k\left[\int \prod_{a=1}^s \frac{da_a}{\sqrt{2\pi\frac{1}{\Delta_a^{-1}}}} e^{-\frac{\Delta_a^{-1}}{2}\sum_{a=1}^s a_a^2 - \frac{\hat{r}_a}{2\gamma}\sum_{a=1}^s a_a^2}\right] = e^{-\frac{ks}{2}\ln\left(1 + \frac{\hat{V}_a}{\gamma\Delta_a^{-1}}\right)} \tag{104}$$

The entropic potential thus reads

$$\Psi_w = \frac{1}{ds} \ln \left[(a) \prod_{j=1}^{d} \right] (b) = -\frac{\gamma}{2} \ln \left(1 + \frac{\hat{V}_a}{\Delta_a^{-1}} \right) \tag{105}$$

$$+ \frac{\hat{m}_w^2}{2d} \operatorname{Tr} \left[\frac{\Sigma^2 \frac{W_\star^\top a_\star a_\star^\top W_\star}{k_\star}}{\hat{V}_w \Sigma + \frac{\Delta_w^{-1}}{r_a} I_d} \right] - \frac{1}{2d} \ln \det \left[\frac{r_a}{\Delta_w^{-1}} \hat{V}_w \Sigma + I_d \right] + \frac{1}{2d} \operatorname{Tr} \left[\frac{\hat{q}_w \Sigma}{\hat{V}_w \Sigma + \frac{\Delta_w^{-1}}{r_a} I_d} \right] \tag{106}$$

where for readability we redefined $\hat{V}_a \leftarrow \hat{V}_a/\gamma$.

Trace terms   It is straightforward to obtain

$$\Psi_t = \frac{1}{2} (\hat{r}_w r_w + \hat{q}_w q_w) + \frac{\gamma}{2} (\hat{r}_a r_a) - \hat{m}_w m_w \tag{107}$$

Loss potential   The loss potential is identical to (Loureiro et al., 2021), provided the relevant overlap are plugged in:

$$\Psi_y = \int D\xi \int dy \mathcal{Z}_\star \left(y, \frac{\kappa_1 \kappa_1^\star m_w}{\sqrt{\kappa_1^2 q_w}} \xi, \kappa_1^{\star 2} \rho_w + \kappa_\star^{\star 2} \rho_a - \frac{(\kappa_1 \kappa_1^\star m_w)^2}{\kappa_1^2 q_w} \right) \ln \mathcal{Z}_\ell \left(y, \sqrt{\kappa_1^2 q_w} \xi, \kappa_1^2 V_w + \kappa_\star^2 V_a \right) \tag{108}$$

with

$$\mathcal{Z}_{\ell/\star}(y, \omega, V) = \mathbb{E}_x^{\mathcal{N}(\omega, V)} P_{\text{out}}^{/\star}(y|x). \tag{109}$$

B.6. Finite temperature free energy

The free energy thus reads

$$f = \underset{V_w, q_w, m_w, V_a, \hat{V}_w, \hat{q}_w, \hat{m}_w, \hat{V}_a}{\operatorname{extr}} -\frac{1}{2} (\hat{V}_w V_w + \hat{V}_w q_w - \hat{q}_w V_w) - \frac{\gamma}{2} \hat{V}_a V_a + \hat{m}_w m_w + \frac{\gamma}{2} \ln \left(1 + \frac{\hat{V}_a}{\Delta_a^{-1}} \right)$$

$$- \frac{1}{2d} \mathbb{E}_{a_\star, W_\star} \operatorname{Tr} \left[\frac{\hat{m}_w^2 \Sigma^2 \frac{W_\star^\top a_\star a_\star^\top W_\star}{k_\star} + \hat{q}_w \Sigma}{\hat{V}_w \Sigma + \frac{\Delta_w^{-1}}{V_a} I_d} \right] + \frac{1}{2d} \ln \det \left[\frac{V_a}{\Delta_w^{-1}} \hat{V}_w \Sigma + I_d \right]$$

$$- \alpha \int D\xi \int dy \mathcal{Z}_\star \left(y, \frac{\kappa_1 \kappa_1^\star m_w}{\sqrt{\kappa_1^2 q_w}} \xi, \kappa_1^{\star 2} \rho_w + \kappa_\star^{\star 2} \rho_a - \frac{(\kappa_1 \kappa_1^\star m_w)^2}{\kappa_1^2 q_w + \kappa_\star^2 q_a} \right) \ln \mathcal{Z}_\ell \left(y, \sqrt{\kappa_1^2 q_w} \xi, \kappa_1^2 V_w + \kappa_\star^2 V_a \right). \tag{110}$$

In the next section, we specialize this expression to the Bayes optimal setting, for regression. Appendix H provides in parallel the specialization for the Bayes optimal classification case.

B.7. Bayes-Optimal setting

For the linear readout $f_\star = id$ and Gaussian additive noise considered in the regression case, $\mathcal{Z}(y, \omega, V)$ admits a compact expression

$$\mathcal{Z}(y, \omega, V) = \int \frac{dx}{\sqrt{2\pi V} \sqrt{2\pi \Delta}} e^{-\frac{(x-\omega)^2}{2V} - \frac{1}{2\Delta}(y-x)^2} = \frac{1}{\sqrt{2\pi(\Delta + V)}} e^{-\frac{(\omega-y)^2}{2(\Delta+V)}}. \tag{111}$$

B.8. Nishimori identities

Since we consider the Bayes optimal case, the Nishimori identities hold, yielding

$$r_a = \rho_a = \Delta_a, \qquad\qquad r_w = \rho_w = \frac{\Delta_w \Delta_a}{d} \operatorname{Tr} \Sigma, \qquad\qquad \hat{r}_{a,w} = 0 \tag{112}$$

$$q_{a,w} = m_{a,w}, \qquad\qquad\qquad\qquad \hat{q}_{a,w} = \hat{m}_{a,w}, \tag{113}$$

causing the free energy to simplify to

$$f = \underset{q_w \hat{q}_w}{\text{extr}} \left[\frac{1}{2} q_w \hat{q}_w - \frac{1}{2d} \text{Tr} \left[\frac{\Delta_w \Delta_a \hat{q}_w^2 \Sigma^2 + \hat{q}_w \Sigma}{\hat{q}_w \Sigma + \frac{1}{\Delta_w \Delta_a} I_d} \right] + \frac{1}{2d} \ln \det \left[\Delta_w \Delta_a \hat{q}_w \Sigma + I_d \right] + \frac{\alpha}{2} \ln \left(\Delta + \kappa_1^2 (\rho_w - q_w) + \kappa_*^2 \rho_a \right) \right]$$

(114)

i.e.

$$f = \underset{q_w \hat{q}_w}{\text{extr}} \left[\frac{1}{2} q_w \hat{q}_w - \frac{1}{2} \hat{q}_w \rho_w + \frac{1}{2d} \ln \det \left[\Delta_w \Delta_a \hat{q}_w \Sigma + I_d \right] + \frac{\alpha}{2} \ln \left(\Delta + \kappa_1^2 (\rho_w - q_w) + \kappa_*^2 \rho_a \right) \right]$$

(115)

B.9. Bayes-optimal saddle points

The extremization in (115) can be carried out by imposing that the gradient with respect to each optimization parameter be vanishing, yielding

$$\begin{cases} \hat{q}_w = \frac{\alpha \kappa_1^2}{\Delta + \kappa_1^2 (\rho_w - q_w) + \kappa_*^2 \rho_a} \\ q_w = \frac{1}{d} \text{Tr} \left[\frac{\Delta_w^2 \Delta_a^2 \hat{q}_w \Sigma^2}{I_d + \Delta_w \Delta_a \hat{q}_w \Sigma} \right] \end{cases} .$$

(116)

The equation (116) can be recovered by plugging the first line into the second, and remembering that the excess prediction error can be evaluated as

$$\epsilon_g - \Delta = \kappa_1^2 (\rho_w - q_w) + \kappa_*^2 \rho_a.$$

(117)

Indeed, the Bayes optimal estimator is given by the averaged output of the student network over the posterior measure (6)

$$\left\langle \frac{1}{\sqrt{k}} a^\top \sigma \left(\frac{1}{\sqrt{d}} W x \right) \right\rangle_{a,W \sim \mathbb{P}}$$

and the corresponding test error reads

$$\begin{aligned}
\epsilon_g &= \mathbb{E}_{a_\star, W_\star, \mathcal{D}} \mathbb{E}_{x, \xi} \left(\frac{1}{\sqrt{k}} a_\star^\top \sigma \left(\frac{1}{\sqrt{d}} W_\star x \right) + \sqrt{\Delta} \xi - \left\langle \frac{1}{\sqrt{k}} a^\top \sigma \left(\frac{1}{\sqrt{d}} W x \right) \right\rangle \right)^2 \\
&= E_\mathcal{D} E_{a_\star, W_\star} \frac{1}{k} a_\star^\top \mathbb{E}_x \left[\sigma \left(\frac{1}{\sqrt{d}} W_\star x \right) \sigma \left(\frac{1}{\sqrt{d}} W_\star x \right)^\top \right] a_\star \\
&\quad + E_\mathcal{D} E_{a_\star, W_\star} \frac{1}{k} \left\langle a^\top \mathbb{E}_x \left[\sigma \left(\frac{1}{\sqrt{d}} W x \right) \sigma \left(\frac{1}{\sqrt{d}} W' x \right)^\top \right] a' \right\rangle_{a,a',W,W' \overset{iid}{\sim} \mathbb{P}} \\
&\quad - 2 E_\mathcal{D} E_{a_\star, W_\star} \frac{1}{k} \left\langle a_\star^\top \mathbb{E}_x \left[\sigma \left(\frac{1}{\sqrt{d}} W_\star x \right) \sigma \left(\frac{1}{\sqrt{d}} W x \right)^\top \right] a' \right\rangle + \Delta \\
&= E_\mathcal{D} E_{a_\star, W_\star} \frac{1}{k} a_\star^\top \mathbb{E}_x \left[\kappa_1^2 \frac{W_\star \Sigma W_\star^\top}{d} + \kappa_*^2 I_d \right] a_\star \\
&\quad + E_\mathcal{D} E_{a_\star, W_\star} \frac{1}{k} \left\langle a^\top \mathbb{E}_x \left[\kappa_1^2 \frac{W \Sigma W'^\top}{d} \right] a' \right\rangle_{a,a',W,W' \overset{iid}{\sim} \mathbb{P}} \\
&\quad - 2 E_\mathcal{D} E_{a_\star, W_\star} \frac{1}{k} \left\langle a_\star^\top \mathbb{E}_x \left[\kappa_1^2 \frac{W_\star \Sigma W^\top}{d} \right] a' \right\rangle + \Delta \\
&= \kappa_1^2 \rho_w + \kappa_*^2 \rho_a + \kappa_1^2 q_w - 2\kappa_1^2 m_w + \Delta = \kappa_1^2 (\rho_w - q_w) + \kappa_*^2 \rho_a.
\end{aligned}$$

(118)

In going from the second line to the third, we used the covariance identities (10).

C. Replica computation for deep networks

C.1. Generalization to multi-layer nets

In this section we generalize the discussion of Appendix B to deep networks with $L \geq 2$ hidden layers. We use the same notations as the main text (1)(2). To treat simultaneously regression and classification, it is useful to introduce generic notations for the priors

$$P_w(W_\ell) \propto e^{-\frac{1}{2\Delta_\ell}||W_\ell||_F^2}, \qquad\qquad P_a(a) \propto e^{-\frac{1}{2\Delta_a}||a||^2}, \qquad (119)$$

and the posteriors

$$P_{\text{out}}(y|\hat{y}) \propto e^{-\frac{1}{2\Delta}(y-\hat{y})^2} \qquad (120)$$

for regression. In Appendix H, for classification, the output channel will be

$$P_{\text{out}}(y|\hat{y}) \propto \int \frac{d\xi e^{-\frac{1}{2\Delta}\xi^2}}{\sqrt{2\pi\Delta}} \Theta\left(y \times \text{sign}(\hat{y}+\xi)\right). \qquad (121)$$

With a slight abuse of notations, we omit for readability the layer index ℓ in the prior P_w, with the correct variance Δ_ℓ of the Gaussian distribution being implied by the argument W_ℓ thereof. Besides, we keep the notation P_w also for the marginal distribution of the rows of the weight matrices.

C.2. Tower of order parameters

Replicated partition function By the same token, we employ the replica trick, and write the replicated multi-layer partition function as

$$Z^s = \int \prod_{a=1}^{s}\prod_{\ell=1}^{L} dW_\ell^a P_w(W_\ell^a) da^a P_a(a^a) \, \underbrace{\mathbb{E}_{\{x^\mu\}_\mu} \int \prod_{\mu=1}^{n} P_{\text{out}}^\star\left(y^\mu|\hat{y}_\star^a(x^\mu)\right) \prod_{\mu=1}^{n}\prod_{a=1}^{s} P_{\text{out}}\left(y^\mu|\hat{y}^a(x^\mu)\right)}_{*}. \qquad (122)$$

Carrying out the expectation with respect to the train set $\mathcal{D}$:

$$* = \prod_\mu \left[\mathbb{E}_{\lambda_\star,\{\lambda_a\}_{a=1}^s} \int dy P_{\text{out}}(y|\lambda_\star) \prod_{a=1}^{s} P_{\text{out}}(y|\lambda_a) \right]. \qquad (123)$$

We have introduced the local fields

$$\lambda_\star \equiv \frac{1}{\sqrt{k_{L^\star}}} a_\star^\top \underbrace{\left(\varphi_{L^\star}^\star \circ \varphi_{L^\star-1}^\star \circ \cdots \circ \varphi_2^\star \circ \varphi_1^\star\right)}_{L^\star}(x), \qquad \lambda_a = \frac{1}{\sqrt{k_L}} a^\top \underbrace{\left(\varphi_L^a \circ \varphi_{L-1}^a \circ \cdots \circ \varphi_2^a \circ \varphi_1^a\right)}_{L}(x). \qquad (124)$$

φ_ℓ^a denotes the a−th replica of the ℓ−th layer, i.e.

$$\varphi_\ell^a(x) = \sigma_\ell\left(\frac{1}{\sqrt{k_{\ell-1}}} W_\ell^a \cdot x\right) \qquad (125)$$

Local fields statistics The second order statistics are given by an application of the deep Bayes GEP (3.4) as

$$\langle \lambda_a \lambda_b \rangle = \prod_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \frac{a_a^\top \left(\prod_{\ell=L}^{1} W_\ell^a\right) \Sigma \left(\prod_{\ell=1}^{L} W_\ell^{b\top}\right) a_b}{\prod_{\ell=0}^{L} k_\ell}$$

$$+ \delta_{ab} \left[\sum_{\ell_0=1}^{L-1} \left(\kappa_\star^{(\ell_0)}\right)^2 \prod_{\ell=\ell_0+1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \frac{a_a^\top \left(\prod_{\ell=L}^{\ell_0+1} W_\ell^a\right) \Sigma \left(\prod_{\ell=\ell_0+1}^{L} W_\ell^{b\top}\right) a_b}{\prod_{\ell=\ell_0}^{L} k_\ell} + \left(\kappa_\star^{(L)}\right)^2 \frac{a_a^\top a_b}{k_L} \right], \qquad (126)$$

and

$$\langle \lambda_a \lambda_\star \rangle = \prod_{\ell=1}^{L} \kappa_1^{(\ell)} \prod_{\ell=1}^{L^\star} \kappa_1^{(\ell)\star} \frac{a_a^\top \left(\prod_{\ell=L}^{1} W_\ell^a \right) \Sigma \left(\prod_{\ell=1}^{L^\star} W_\ell^{\star\top} \right) a_\star}{\prod_{\ell=0}^{L} \sqrt{k_\ell} \prod_{\ell=0}^{L} \sqrt{k_\ell}^\star}. \tag{127}$$

By the same token the teacher variance reads

$$\langle \lambda_\star^2 \rangle = \prod_{\ell=1}^{L^\star} \left(\kappa_1^{\star(\ell)} \right)^2 \frac{a_\star^\top \left(\prod_{\ell=L^\star}^{1} W_\ell^\star \right) \Sigma \left(\prod_{\ell=1}^{L^\star} W_\ell^{\star\top} \right) a_\star}{\prod_{\ell=0}^{L} k_\ell^\star}$$

$$+ \sum_{\ell_0=1}^{L^\star-1} \left(\kappa_\star^{\star(\ell_0)} \right)^2 \prod_{\ell=\ell_0+1}^{L^\star} \left(\kappa_1^{\star(\ell)} \right)^2 \frac{a_\star^\top \left(\prod_{\ell=L^\star}^{\ell_0+1} W_\ell^\star \right) \Sigma \left(\prod_{\ell=\ell_0+1}^{L^\star} W_\ell^{\star\top} \right) a_\star}{\prod_{\ell=\ell_0}^{L^\star} k_\ell^\star} + \left(\kappa_\star^{\star(L)} \right)^2 \frac{a_\star^\top a_\star}{k_{L^\star}^\star} \tag{128}$$

These statistics warrant the introduction of the order parameters $m_a, \{ \left(q_\ell^L \right)_{ab} \}_{\ell=1}^L, q_{ab}^a, \{ \rho_\ell^{L^\star} \}_{\ell=1}^{L^\star}, \rho_a$ as

$$\begin{cases} m_a = \dfrac{a_a^\top \left(\prod_{\ell=L}^{1} W_\ell^a \right) \Sigma \left(\prod_{\ell=1}^{L^\star} W_\ell^{\star\top} \right) a_\star}{\prod_{\ell=0}^{L} \sqrt{k_\ell} \prod_{\ell=0}^{L} \sqrt{k_\ell}^\star} \\[6pt] \left(q_1^L \right)_{ab} = \dfrac{a_a^\top \left(\prod_{\ell=L}^{\ell_0} W_\ell^a \right) \Sigma \left(\prod_{\ell=\ell_0}^{L} W_\ell^{b\top} \right) a_b}{\prod_{\ell=\ell_0-1}^{L} k_\ell} \\[6pt] \forall \ell_0 \geq 2, \ \left(q_{\ell_0}^L \right)_{ab} = \delta_{ab} \dfrac{a_a^\top \left(\prod_{\ell=L}^{\ell_0} W_\ell^a \right) \Sigma \left(\prod_{\ell=\ell_0}^{L} W_\ell^{b\top} \right) a_b}{\prod_{\ell=\ell_0-1}^{L} k_\ell} \\[6pt] q_{ab}^a = \delta_{ab} \dfrac{a_a^\top a_b}{k_L} \end{cases} \qquad \begin{cases} \rho_{\ell_0}^{L^\star} = \dfrac{a_\star^\top \left(\prod_{\ell=L^\star}^{\ell_0} W_\ell^\star \right) \Sigma \left(\prod_{\ell=\ell_0}^{L^\star} W_\ell^{\star\top} \right) a_\star}{\prod_{\ell=\ell_0-1}^{L^\star} k_\ell^\star} \\[6pt] \rho_a = \dfrac{a_\star^\top a_\star}{k_{L^\star}^\star} \end{cases} \tag{129}$$

Note that $q_{\ell_0}^L$s corresponds to the overlap between the products of weights from layer ℓ_0 (included) to layer L (included, i.e. the last layer before readout), with similar intuition holding for the $\rho_{\ell_0}^{L^\star}$s.

C.3. Replica Symmetry

Similarly to the two-layers case, we assume replica symmetry. The order parameters are taken of the form

$$\left(\hat{q}_1^L \right)_{ab} = \delta_{ab} \left(-\frac{\hat{r}_1^L}{2} - \hat{q}_1^L \right) + \hat{q}_1^L, \tag{130}$$

$$\forall \ell \geq 2, \ \left(\hat{q}_\ell^L \right)_{ab} = -\delta_{ab} \frac{\hat{r}_\ell^L}{2}, \tag{131}$$

$$\hat{q}_{ab}^a = -\delta_{ab} \frac{\hat{r}_a}{2}, \tag{132}$$

$$\hat{m}_a = \hat{m}, \tag{133}$$

while the non-hatted overlap read

$$\left(q_1^L\right)_{ab} = \delta_{ab}\left(r_1^L - q_1^L\right) + q_1^L, \tag{134}$$

$$\forall \ell \geq 2, \ \left(q_\ell^L\right)_{ab} = \delta_{ab} r_\ell^L, \tag{135}$$

$$q_{ab}^a = \delta_{ab} r_a, \tag{136}$$

$$m_a = m. \tag{137}$$

$$\tag{138}$$

C.4. Multilayer Loss potential

The computation of the loss potential Ψ_y follows rather tightly the one detailed in (Loureiro et al., 2021) and leads to

$$\Psi_y = \int D\xi \int dy \, \mathcal{Z}_\star\left(y, \frac{\prod_{\ell=1}^{L}\kappa_1^{(\ell)}\prod_{\ell=1}^{L^\star}\kappa_1^{\star(\ell)}m}{\sqrt{\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q_1^L}}\xi, \prod_{\ell=1}^{L^\star}\left(\kappa_1^{\star(\ell)}\right)^2\rho_1^L + \sum_{\ell_0=1}^{L^\star-1}\kappa_*^{\star(\ell_0)}\prod_{\ell=\ell_0+1}^{L^\star}\left(\kappa_1^{\star(\ell)}\right)^2\rho_{\ell_0+1}^{L^\star} + \left(\kappa_*^{\star(L^\star)}\right)^2\rho_a - \frac{\left(\prod_{\ell=1}^{L}\kappa_1^{(\ell)}\prod_{\ell=1}^{L^\star}\kappa_1^{\star(\ell)}m\right)^2}{\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q_1^L}\right)$$

$$\times \ln \mathcal{Z}_\ell\left(y, \sqrt{\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q_1^L}\xi, \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 V_1^L + \sum_{\ell_0=1}^{L-1}\kappa_*^{(\ell_0)}\prod_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2 r_{\ell_0+1}^L + \left(\kappa_*^L\right)^2 r_a\right). \tag{139}$$

C.5. Layer-wise entropy

The computation of the entropy potential Ψ_w follows the same lines as for the two-layer case, see Appendix B. In practice, one has to first integrate the first layer, then the middle layers and the readout. For clarity purposes, we decompose Ψ_w into the following series of terms

$$\Psi_w = \Psi_1(\hat{q}_1^L, \hat{r}_1^L, q_2^L) + \sum_{\ell=2}^{L}\Psi_\ell(\hat{r}_\ell^L, q_{\ell+1}^L) + \Psi_a(\hat{r}_a) \tag{140}$$

and proceed to evaluate each sequentially.

First layer   We focus on $\ell = 1$ first:

$$e^{sd\Psi_1} = \prod_{j=1}^{d}\int\prod_{a=1}^{s} dw_a^1 P_w(w_a^1) e^{\sum_{a\leq b}\sigma_j\left(\hat{q}_1^L\right)_{ab}\frac{\left(w_a^{1\top}\prod_{\ell=2}^{L}W_a^\ell a_a\right)\times\left(w_b^{1\top}\prod_{\ell=2}^{L}W_b^\ell a_b\right)}{\prod_{\ell=1}^{L}k_\ell} + \sum_a \sigma_j \hat{m}\frac{\left(w_a^{1\top}\prod_{\ell=2}^{L}W_a^\ell a_a\right)\times\left(w_\star^{1\top}\prod_{\ell=2}^{L^\star}W_\star^\ell a_\star\right)}{\prod_{\ell=1}^{L}\sqrt{k_\ell}\prod_{\ell=1}^{L^\star}\sqrt{k_\ell^\star}}}. \tag{141}$$

Introducing the local fields

$$\eta_a \equiv \frac{w_a^{1\top}\prod_{\ell=2}^{L}W_a^\ell a_a}{\prod_{\ell=1}^{L}\sqrt{k_\ell}}, \tag{142}$$

which inherit from $P_w(\cdot)$ the statistics

$$\langle\eta_a\eta_b\rangle = \delta_{ab}\frac{\left(q_2^L\right)_{aa}}{\Delta_1^{-1}}, \tag{143}$$

it follows that

$$e^{sd\Psi_1} = \prod_{j=1}^{d} e^{\frac{1}{2}\sigma_j^2\left(\frac{a_\star^\top\prod_{\ell=L^\star}^{2}W_\ell^{\star\top}(w_1^\star)_j}{\prod_{\ell=1}^{L^\star}\sqrt{k_\ell^\star}}\right)^2 \hat{m}^\top\left[\text{diag}\left(\frac{\Delta_1^{-1}}{\left(q_2^L\right)_{aa}}\right) - \sigma_j\times(2I_s)\odot\hat{Q}_1^L\right]^{-1}\hat{m} - \frac{1}{2}\ln\det\left[I_s - \sigma_j\text{diag}\left(\frac{\Delta_1^{-1}}{\left(q_2^L\right)_{aa}}\right)(2I_s)\odot\hat{Q}_1^L\right]}. \tag{144}$$

Finally,

$$\Psi_1 = \frac{1}{d} \operatorname{Tr} \left[\frac{\hat{q}_1^L \Sigma + \hat{m}^2 \Sigma^2 \dfrac{\prod\limits_{\ell=L^\star}^{1} W_\ell^{\star\top} a_\star a_\star^\top \prod\limits_{\ell=1}^{L^\star} W_\ell^\star}{\prod\limits_{\ell=1}^{L^\star} k_\ell^\star}}{\hat{V}_1^L \Sigma + \dfrac{\Delta_1^{-1}}{r_2^L}} \right] - \frac{1}{2d} \ln \det \left[\frac{r_2^L}{\Delta_1^{-1}} \hat{V}_1^L \Sigma + I_d \right] \tag{145}$$

Middle layers We know provide the derivation for $\ell \geq 2$ (middle layers). The corresponding potential reads

$$e^{sd\Psi_{\ell_0}} = \int \prod_{a=1}^{s} dW_{\ell_0}^a P_w(W_{\ell_0}^a) e^{\sum\limits_{a=1}^{s} (\hat{q}_\ell^L)_{aa} \frac{a_a^\top \prod\limits_{\ell=L}^{\ell_0} W_\ell^{a\top} \prod\limits_{\ell=\ell_0}^{L} W_\ell^a a_a}{\prod\limits_{\ell=\ell_0}^{L} k_\ell}} \tag{146}$$

$$= \prod_{a=1}^{s} \left[\int dw e^{-\frac{\Delta_{\ell_0}^{-1}}{2} w^\top w - \frac{1}{2} \hat{r}_{\ell_0}^L \left(\frac{w^\top \prod\limits_{\ell=\ell_0+1}^{L} W_\ell^a a_a}{\prod\limits_{\ell=\ell_0}^{L} \sqrt{k_\ell}} \right)^2} \right]^{k_{\ell_0}-1}. \tag{147}$$

Introducing the local fields

$$\eta_a = \frac{w^\top \prod\limits_{\ell=\ell_0+1}^{L} W_\ell a_a}{\prod\limits_{\ell=\ell_0}^{L} \sqrt{k_\ell}}, \tag{148}$$

it follows that

$$\Psi_\ell = -\frac{\gamma_{\ell-1}}{2} \ln \left(1 + \frac{r_{\ell+1}^L \hat{r}_\ell^L}{\Delta_\ell^{-1}} \right). \tag{149}$$

Readout The readout layer entropy can be computed as in the two layer case yielding

$$\Psi_a = -\frac{\gamma_L}{2} \ln \left(1 + \frac{\hat{r}_a}{\Delta_a^{-1}} \right), \tag{150}$$

see Appendix B

Summary: multilayer entropic potential Assembling these results, the total entropic action reads

$$\Psi_w = -\frac{\gamma_L}{2} \ln \left(1 + \frac{\hat{r}_a}{\Delta_a^{-1}} \right) - \sum_{\ell=2}^{L} \frac{\gamma_{\ell-1}}{2} \ln \left(1 + \frac{r_{\ell+1}^L \hat{r}_\ell^L}{\Delta_\ell^{-1}} \right) + \frac{1}{d} \operatorname{Tr} \left[\frac{\hat{q}_1^L \Sigma + \hat{m}^2 \Sigma^2 \dfrac{\prod\limits_{\ell=L^\star}^{1} W_\ell^{\star\top} a_\star a_\star^\top \prod\limits_{\ell=1}^{L^\star} W_\ell^\star}{\prod\limits_{\ell=1}^{L^\star} k_\ell^\star}}{\hat{V}_1^L \Sigma + \dfrac{\Delta_1^{-1}}{r_2^L}} \right]$$

$$- \frac{1}{2d} \ln \det \left[\frac{r_2^L}{\Delta_1^{-1}} \hat{V}_1^L \Sigma + I_d \right]. \tag{151}$$

C.6. Finite T multilayer free energy

Finally, the complete free energy at finite temperature reads

$$
f = \operatorname*{extr}_{q_1^L, \hat{q}_1^L, V_1^L, \hat{V}_1^L, m, \hat{m}, \{r_\ell^L, \hat{r}_\ell^L\}, r_a, \hat{r}_a} -\frac{1}{2}(\hat{V}_1^L V_1^L + \hat{V}_1^L q_1^L - \hat{q}_1^L V_1^L) - \frac{1}{2}\sum_{\ell=2}^{L} \gamma_{\ell-1} \hat{r}_\ell^L r_\ell^L - \frac{\gamma_L}{2}\hat{r}_a r_a
$$

$$
+ \hat{m}m + \frac{\gamma_L}{2}\ln\left(1 + \frac{\hat{r}_a}{\Delta_a^{-1}}\right)
$$

$$
+ \sum_{\ell=2}^{L}\frac{\gamma_{\ell-1}}{2}\ln\left(1 + \frac{r_{\ell+1}^L \hat{r}_\ell^L}{\Delta_\ell^{-1}}\right) - \frac{1}{d}\operatorname{Tr}\left[\frac{\hat{q}_1^L \Sigma + \hat{m}^2 \Sigma^2 \frac{\prod_{\ell=L^\star}^{1} W_\ell^{\star\top} a_\star a_\star^\top \prod_{\ell=1}^{L^\star} W_\ell^\star}{\prod_{\ell=1}^{L^\star} k_\ell^\star}}{\hat{V}_1^L \Sigma + \frac{\Delta_1^{-1}}{r_2^L}}\right] + \frac{1}{2d}\ln\det\left[\frac{r_2^L}{\Delta_1^{-1}}\hat{V}_1^L \Sigma + I_d\right]
$$

$$
- \alpha \int D\xi \int dy \mathcal{Z}_\star\left(y, \frac{\prod_{\ell=1}^{L}\kappa_1^{(\ell)}\prod_{\ell=1}^{L^\star}\kappa_1^{\star(\ell)} m}{\sqrt{\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q_1^L}}\xi, \prod_{\ell=1}^{L^\star}\left(\kappa_1^{\star(\ell)}\right)^2\rho_1^L + \sum_{\ell_0=1}^{L^\star-1}\kappa_*^{\star(\ell_0)}\prod_{\ell=\ell_0+1}^{L^\star}\left(\kappa_1^{\star(\ell)}\right)^2\rho_{\ell_0+1}^{L^\star} + \left(\kappa_*^{\star(L^\star)}\right)^2\rho_a - \frac{\left(\prod_{\ell=1}^{L}\kappa_1^{(\ell)}\prod_{\ell=1}^{L^\star}\kappa_1^{\star(\ell)} m\right)^2}{\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q_1^L}\right)
$$

$$
\times \ln\mathcal{Z}_\ell\left(y, \sqrt{\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q_1^L}\xi, \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 V_1^L + \sum_{\ell_0=1}^{L-1}\kappa_*^{(\ell_0)}\prod_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2 r_{\ell_0+1}^L + \left(\kappa_*^L\right)^2 r_a\right). \tag{152}
$$

C.7. Bayes-Optimal setting

Nishimori identities For matching teacher/student the Nishimori identities guarantee that

$$
\begin{cases} q_1^L = m \\ \hat{q}_1^L = \hat{m} \end{cases} \qquad\qquad \forall\ell, \begin{cases} r_\ell^L = \rho_\ell^L = \Delta_a\prod_\ell^L \Delta_\ell \times \left(\frac{\operatorname{tr}\Sigma}{d}\right)^{\delta_{\ell,1}} \\ \hat{r}_\ell^L = 0. \end{cases} \tag{153}
$$

The Bayes-Optimal free energy accordingly simplifies to

$$
f = \operatorname*{extr}_{q,\hat{q}}\frac{1}{2}\hat{q}q - \frac{1}{d}\operatorname{Tr}\left[\frac{\hat{q}\Sigma + \hat{q}^2\Sigma^2\Delta_a\prod_{\ell=1}^{L}\Delta_\ell}{\hat{q}\Sigma + \frac{1}{\Delta_a\prod_{\ell=1}^{L}\Delta_\ell}}\right] + \frac{1}{2d}\ln\det\left[\Delta_a\prod_{\ell=1}^{L}\Delta_\ell\hat{q}\Sigma + I_d\right]
$$

$$
- \alpha\int D\xi\int dy \mathcal{Z}_\star\left(y, \sqrt{\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q}\xi, \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 V + \sum_{\ell_0=1}^{L-1}\kappa_*^{(\ell_0)}\prod_{\ell=\ell_0+1}^{L^\star}\left(\kappa_1^{(\ell)}\right)^2\rho_{\ell_0+1}^L + \left(\kappa_*^{\star(L)}\right)^2\rho_a\right)
$$

$$
\times \ln\mathcal{Z}_\ell\left(y, \sqrt{\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q}\xi, \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 V + \sum_{\ell_0=1}^{L-1}\kappa_*^{(\ell_0)}\prod_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2\rho_{\ell_0+1}^L + \left(\kappa_*^L\right)^2\rho_a\right). \tag{154}
$$

For a linear readout, and a noisy teacher channel with Gaussian additive noise of variance Δ (1):

$$
f = \operatorname*{extr}_{q,\hat{q}}\frac{1}{2}\hat{q}q - \frac{1}{d}\operatorname{Tr}\left[\frac{\hat{q}\Sigma + \hat{q}^2\Sigma^2\Delta_a\prod_{\ell=1}^{L}\Delta_\ell}{\hat{q}\Sigma + \frac{1}{\Delta_a\prod_{\ell=1}^{L}\Delta_\ell}}\right] + \frac{1}{2d}\ln\det\left[\Delta_a\prod_{\ell=1}^{L}\Delta_\ell\hat{q}\Sigma + I_d\right]
$$

$$
+ \frac{\alpha}{2}\ln\left(\Delta + \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2(\rho_1^L - q) + \sum_{\ell_0=1}^{L-1}\kappa_*^{(\ell_0)}\prod_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2\rho_{\ell_0+1}^L + \left(\kappa_*^L\right)^2\rho_a\right) \tag{155}
$$

C.8. Bayes-Optimal Saddle point equations

The saddle-point equations corresponding to the extremization (155) read

$$
\begin{cases}
\hat{q} = \alpha \dfrac{\prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2}{\Delta + \prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2(\rho_1^L - q) + \sum\limits_{\ell_0=1}^{L-1} \kappa_*^{(\ell_0)} \prod\limits_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2 \rho_{\ell_0+1}^L + (\kappa_*^L)^2 \rho_a} \\[2em]
q = \dfrac{1}{2d}\,\mathrm{Tr}\left[\dfrac{\hat{q}\Sigma^2 \prod\limits_{\ell=1}^{L}\Delta_\ell^2}{1+\hat{q}\Sigma \prod\limits_{\ell=1}^{L}\Delta_\ell}\right].
\end{cases}
\tag{156}
$$

The Bayes optimal MSE can be computed along the same lines as the two-layers case. Recall that $h_L^\star(x)$ and $h_L(x)$ designate the post activations at the last layer for an input x, respectively for the teacher (1) and student (2) networks. The Bayes optimal estimator is then $\langle a^\top h_L(x)\rangle/\sqrt{k_L}$, where the expectation is carried out over the Bayes posterior measure (6). The MSE then reads

$$
\begin{aligned}
\epsilon_g &= \mathbb{E}_{\mathcal{D},a_\star,\{W_\ell^\star\}_\ell}\mathbb{E}_{x,\xi}\left(\frac{1}{\sqrt{k_L}}a_\star^\top h_L^\star(x) + \xi - \left\langle\frac{1}{\sqrt{k_L}}a^\top h_L(x)\right\rangle\right)^2 \\
&= \mathbb{E}_{\mathcal{D},a_\star,\{W_\ell^\star\}_\ell}\left[\tfrac{1}{k_L}a_\star^\top\mathbb{E}_x\left[h_L^\star(x)h_L^{\star\top}(x)\right]a_\star + \tfrac{1}{k_L}\left\langle a^\top\mathbb{E}_x\left[h_L(x)h_L'(x)^\top\right]a'\right\rangle_{a,\{W_\ell\},a',\{W_\ell'\}^{i.i.d}_{\sim\mathbb{P}}} - 2\tfrac{1}{k_L}\left\langle a_\star^\top\mathbb{E}_x\left[h_L^\star(x)h_L^\top(x)\right]a\right\rangle\right] + \Delta \\
&= \frac{1}{k_L}\mathbb{E}_{\mathcal{D},a_\star,\{W_\ell^\star\}_\ell}\left[\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 a_\star^\top\left(\prod_{\ell=1}^{L}\frac{W_\ell^{\star\top}}{\sqrt{k_{\ell-1}}}\right)^\top \Sigma \left(\prod_{\ell=1}^{L}\frac{W_\ell^\star}{\sqrt{k_{\ell-1}}}\right)a_\star\right. \\
&\qquad + \sum_{\ell_0=1}^{L-1}\left(\kappa_*^{(\ell_0)}\right)^2 \prod_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2 a_\star\left(\prod_{\ell=\ell_0+1}^{L}\frac{W_\ell^{\star\top}}{\sqrt{k_{\ell-1}}}\right)^\top\left(\prod_{\ell=\ell_0+1}^{L}\frac{W_\ell^\star}{\sqrt{k_{\ell-1}}}\right)a_\star\bigg] \\
&\quad + \frac{1}{k_L}\mathbb{E}_{\mathcal{D},a_\star,\{W_\ell^\star\}_\ell}\left[\left\langle\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 a^\top\left(\prod_{\ell=1}^{L}\frac{W_\ell^\top}{\sqrt{k_{\ell-1}}}\right)^\top \Sigma\left(\prod_{\ell=1}^{L}\frac{W_\ell'^\top}{\sqrt{k_{\ell-1}}}\right)a'\right\rangle_{a,\{W_\ell\},a',\{W_\ell'\}^{i.i.d}_{\sim\mathbb{P}}}\right] \\
&\quad - 2\frac{1}{k_L}\mathbb{E}_{\mathcal{D},a_\star,\{W_\ell^\star\}_\ell}\left[\left\langle\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 a_\star^\top\left(\prod_{\ell=1}^{L}\frac{W_\ell^{\star\top}}{\sqrt{k_{\ell-1}}}\right)^\top \Sigma\left(\prod_{\ell=1}^{L}\frac{W_\ell}{\sqrt{k_{\ell-1}}}\right)a\right\rangle\right] + \Delta \\
&= \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2(\rho_1^L - q) + \sum_{\ell_0=1}^{L-1}\left(\kappa_*^{(\ell_0)}\right)^2\prod_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2\rho_{\ell_0+1}^L + \left(\kappa_*^L\right)^2\rho_a
\end{aligned}
\tag{157}
$$

Replacing the ρ_ℓ^L order parameters by their definitions, one finally reaches

$$
\epsilon_g^{\mathrm{Bayes}} - \Delta = \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\left(\Delta_a\left(\int z\,\mathrm{d}\mu(z)\right)\prod_{\ell=1}^{L}\Delta_\ell - q\right) + \sum_{\ell_0=1}^{L-1}\left(\kappa_*^{(\ell_0)}\right)^2\Delta_a\prod_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell + \left(\kappa_*^{(L)}\right)^2\Delta_a
$$

which is equation (12).

D. Equivalent shallow network

In this appendix, we show that the shallow network (16)

$$
y^{\mathrm{lin}}(x) = f_\star\left(\frac{\sqrt{\rho}\,\theta^\top x}{\sqrt{d}} + \sqrt{\epsilon_r}\,\xi\right),
\tag{158}
$$

possesses the same Bayes free energy as the deep network (1), implying in particular that they are characterized by the same Bayes error. Again, let us defined the notations

$$
P_w(w) \propto e^{-\frac{1}{2}\|w\|^2}
\tag{159}
$$

and the posterior

$$P'_{\text{out}}(y|z) \propto \int \frac{d\xi}{\sqrt{2\pi\epsilon_r}} e^{-\frac{1}{2\epsilon_r}\xi^2} \delta(y - f_\star(z + \xi)) \tag{160}$$

. We will again resort to the replica trick. While this exact replica computation has not, to the authors' knowledge, been detailed in previous work, it is nevertheless very standard and close to existing studies, e.g. (Barbier et al., 2017; Loureiro et al., 2021). For these reasons, we shall frequently point to related computations and be concise.

The replicated partition function reads

$$Z^s = \int \prod_{a=1}^s dw_a P_w(w_a) d\theta P_w(\theta) \prod_{\mu=1}^n \mathbb{E}_{y^\mu, x^\mu} \int dy^\mu P'_{\text{out}}\left(y^\mu \Big| \frac{\theta^\top x^\mu}{\sqrt{d}}\right) \prod_{a=1}^s P'_{\text{out}}\left(y^\mu \Big| \frac{w_a^\top x^\mu}{\sqrt{d}}\right) \tag{161}$$

Define the overlap matrix

$$q_{ab} = \frac{w_a^\top \Sigma w_b}{d} \qquad\qquad m_a = \frac{\theta^\top \Sigma w_a}{d} \tag{162}$$

and the associated Dirac conjugates $\{\hat{q}_{ab}\}, \{\hat{m}_a\}$, the partition function becomes

$$Z^s = \int \prod_{a\leq b} dq_{ab} d\hat{q}_{ab} \prod_{a=1}^s dm_a d\hat{m}_a e^{-\sum_{a\leq b} dq_{ab}\hat{q}_{ab} - \sum_a dm_a \hat{m}_a} e^{ds\Psi_w + \alpha ds\Psi_y} \tag{163}$$

where we introduced the entropic potential Ψ_w

$$e^{ds\Psi_w} = \int \prod_{a=1}^s dw_a P_w(w_a) d\theta P_w(\theta) e^{\sum_{a\leq b} \hat{q}_{ab} w_a^\top \Sigma w_b + \sum_a \hat{m}_a \theta^\top \Sigma w_a} \tag{164}$$

and the loss potential Ψ_y

$$e^{\alpha ds\Psi_y} = \mathbb{E}_{\{\lambda_a, \mu\}} \int dy P'_{\text{out}}(y|\mu) \prod_{a=1}^s P'_{\text{out}}(y|\lambda_a) \tag{165}$$

where the local fields $\{\lambda_a\}, \mu$ are Gaussian with statistics

$$\langle \mu^2 \rangle = 1, \qquad\qquad \langle \mu\lambda_a \rangle = m_a, \qquad\qquad \langle \lambda_a \lambda_b \rangle = q_{ab}. \tag{166}$$

We assume the replica symmetric ansatz with the Nishimori identities

$$q_{ab} = (\rho - q)\delta_{ab} + q \tag{167}$$
$$m_a = q \tag{168}$$
$$\hat{m}_a = \hat{q} \tag{169}$$
$$\hat{q}_{ab} = \hat{q}(1 - \delta_{ab}) \tag{170}$$

The computation of Ψ_y is standard and leads to (see e.g. (Loureiro et al., 2021))

$$\Psi_y = \int D\xi \int dy \mathcal{Z}'\left(y, \sqrt{q}\xi, \rho \int z d\mu(z) - q\right) \ln \mathcal{Z}'\left(y, \sqrt{q}\xi, \rho \int z d\mu(z) - q\right) \tag{171}$$

with

$$\mathcal{Z}'(y, \omega, V) = \mathbb{E}_x^{\mathcal{N}(\omega, V)} P'_{\text{out}}(y|x) = \int Dx D\xi \, \delta(y - f_\star(\sqrt{\epsilon_r}\xi + \sqrt{V}x + \omega))$$

$$= \int Dx D\xi \, \delta(y - f_\star(\sqrt{\Delta}\xi + \sqrt{V + \epsilon_r - \Delta}x + \omega))$$

$$= \mathcal{Z}_\ell(y, \omega, V + \epsilon_r - \Delta) \tag{172}$$

35

$\mathcal{Z}_\ell$ (109) has been introduced in Appendix C. As for the entropic potential, introducing the Hubbard Stratonovitch field η, and using the notation D for integrands with normal distributions,

$$
\begin{aligned}
e^{sd\Psi_w} &= \int D\theta D\eta \left[\int dw \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1+\rho\hat{q}\Sigma}{2} w^\top w + \rho\hat{q}\theta^\top \Sigma w + \sqrt{\rho\hat{q}} w^\top \Sigma^{\frac{1}{2}} \eta} \right]^s \\
&= \int D\theta D\eta e^{-\frac{s}{2}\ln\det(1+\hat{q}\Sigma)+\frac{s}{2}(\sqrt{\rho\hat{q}}\Sigma^{\frac{1}{2}}\eta+\rho\hat{q}\Sigma\theta)^\top (1+\rho\hat{q}\Sigma)^{-1}(\sqrt{\rho\hat{q}}\Sigma^{\frac{1}{2}}\eta+\rho\hat{q}\Sigma\theta)} \\
&= e^{-\frac{s}{2}\ln\det(1+\rho\hat{q}\Sigma)} \int \frac{dz}{\sqrt{\det 2\pi(\rho\hat{q}\Sigma + \rho^2\hat{q}^2\Sigma^2)}} e^{-\frac{1}{2}z^\top(\rho\hat{q}\Sigma+\rho^2\hat{q}^2\Sigma^2)^{-1}z+\frac{s}{2}z^\top(1+\rho\hat{q}\Sigma)^{-1}z} \\
&= e^{-\frac{s}{2}\ln\det(1+\rho\hat{q}\Sigma)-\frac{1}{2}\ln\det[\mathbb{I}-s(\rho\hat{q}\Sigma+\rho^2\hat{q}^2\Sigma^2)(1+\rho\hat{q}\Sigma)^{-1}]}.
\end{aligned}
\tag{173}
$$

Finally

$$
\Psi_w = \frac{1}{2}\ln\det(1+\rho\hat{q}\Sigma) - \frac{1}{2}\operatorname{Tr}\left[(\rho\hat{q}\Sigma + \rho^2\hat{q}^2\Sigma^2)(1+\rho\hat{q}\Sigma)^{-1}\right]
\tag{174}
$$

Putting everything together

$$
\begin{aligned}
f &= \operatorname*{extr}_{q,\hat{q}} -\frac{1}{2}\hat{q}q + \frac{1}{2d}\ln\det\left[1 + \Delta_a \prod_\ell \Delta_\ell \prod_\ell \left(\kappa_1^{(\ell)}\right)^2 \hat{q}\Sigma\right] - \frac{1}{2d}\operatorname{Tr}\left[\frac{\Delta_a \prod_\ell \Delta_\ell \prod_\ell \left(\kappa_1^{(\ell)}\right)^2 \hat{q}\Sigma + \left(\Delta_a \prod_\ell \Delta_\ell \prod_\ell \left(\kappa_1^{(\ell)}\right)^2\right)^2 \hat{q}^2\Sigma^2}{1 + \Delta_a \prod_\ell \Delta_\ell \prod_\ell \left(\kappa_1^{(\ell)}\right)^2 \hat{q}\Sigma}\right] \\
&\quad + \alpha \int D\xi \int dy \mathcal{Z}\left(y, \sqrt{q}\xi, \rho\int z d\mu(z) - q\right) \ln\mathcal{Z}\left(y, \sqrt{q}\xi, \rho\int z d\mu(z) - q\right) \\
&= \operatorname*{extr}_{q,\hat{q}} -\frac{1}{2}\hat{q}q + \frac{1}{2d}\ln\det\left[1 + \Delta_a \prod_\ell \Delta_\ell \hat{q}\Sigma\right] - \frac{1}{2d}\operatorname{Tr}\left[\frac{\hat{q}\Sigma + \Delta_a\prod_\ell\Delta_\ell\hat{q}^2\Sigma^2}{\frac{1}{\prod_\ell\left(\kappa_1^{(\ell)}\right)^2} + \hat{q}\Sigma}\right] \\
&\quad + \alpha \int D\xi \int dy \mathcal{Z}\left(y, \sqrt{\prod_\ell\left(\kappa_1^{(\ell)}\right)^2 q}\xi, \prod_\ell\left(\kappa_1^{(\ell)}\right)^2(r_1^L - q) + \epsilon_r - \Delta\right) \ln\mathcal{Z}_\ell\left(y, \sqrt{\prod_\ell\left(\kappa_1^{(\ell)}\right)^2 q}\xi, \prod_\ell\left(\kappa_1^{(\ell)}\right)^2(r_1^L - q) + \epsilon_r - \Delta\right) \\
&= \operatorname*{extr}_{q,\hat{q}} -\frac{1}{2}\hat{q}q + \frac{1}{2d}\ln\det\left[1 + \Delta_a \prod_\ell \Delta_\ell \hat{q}\Sigma\right] - \frac{1}{2d}\operatorname{Tr}\left[\frac{\hat{q}\Sigma + \Delta_a\prod_\ell\Delta_\ell\hat{q}^2\Sigma^2}{\frac{1}{\prod_\ell\left(\kappa_1^{(\ell)}\right)^2} + \hat{q}\Sigma}\right] \\
&\quad + \alpha \int D\xi \int dy \mathcal{Z}\left(y, \sqrt{\prod_\ell\left(\kappa_1^{(\ell)}\right)^2 q}\xi, \prod_\ell\left(\kappa_1^{(\ell)}\right)^2(r_1^L - q) + \sum_{\ell_0=1}^{L-1}\left(\kappa_*^{(\ell_0)}\right)^2\Delta_a \prod_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell + \left(\kappa_*^{(L)}\right)^2\Delta_a\right) \\
&\quad \ln\mathcal{Z}_\ell\left(y, \sqrt{\prod_\ell\left(\kappa_1^{(\ell)}\right)^2 q}\xi, \prod_\ell\left(\kappa_1^{(\ell)}\right)^2(r_1^L - q) + \sum_{\ell_0=1}^{L-1}\left(\kappa_*^{(\ell_0)}\right)^2\Delta_a \prod_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell + \left(\kappa_*^{(L)}\right)^2\Delta_a\right)
\end{aligned}
\tag{175}
$$

which is exactly the Bayes free energy (154) for the multi-layer target. The expressions of the Bayes MSE and classification errors for the single layer network are known (Loureiro et al., 2021) and exactly recover the expressions (12) and (5).

E. ERM : the shallow case

In this appendix, we contrast the Bayes optimal MSE for regression (12) obtained in Appendix B and C to the test error achieved by linear ERM methods. Are successively described ridge regression, and random features regression (Rahimi & Recht, 2007), alongside its infinite width limit, kernel regression. Sharp asymptotics for the test error of those ERM methods can be accessed by combining the multi-layer linearization (10) (see also Appendix A) with the results of (Loureiro et al., 2021).

We start by reviewing the results of (Loureiro et al., 2021). They consider learning from the dataset $\{v^\mu, y^\mu\}_{\mu=1}^n$, where $v^\mu \in \mathbb{R}^p$ represent the student features, assuming that the labels are generated from a generalized linear model as

$$
y^\mu = f_\star\left(\frac{1}{k_L} a_\star^\top h_L^{\star\mu}\right)
$$

with $h_L^{\star\mu} \in \mathbb{R}^{k_L}$. Finally, suppose that $h_L^{\star\mu}, v^\mu$ are jointly Gaussian

$$\begin{pmatrix} h_L^{\star\mu} \\ v^\mu \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{bmatrix}\right).$$

The trains set can be learnt resorting to ERM with loss function $g(\cdot)$

$$\hat{w} = \underset{w}{\operatorname{argmin}} \sum_{\mu=1}^{n} g\left(y^\mu, \frac{w^\top v^\mu}{\sqrt{p}}\right) + \frac{\lambda}{2}||w||^2$$

with the test error being measured as

$$\epsilon_g = \mathbb{E}_{\mathcal{D}}\mathbb{E}_{v,h_L^\star}\left[\hat{g}\left(\hat{f}\left(\frac{\hat{w}^\top v^\mu}{\sqrt{p}}\right), f_\star\left(\frac{1}{k_L}a_\star^\top h_L^{\star\mu}\right)\right)\right]$$

for some metric $\hat{g}$ and estimator function $\hat{f}$. Finally, the proportional limit $p, k_L, n \to \infty$ with finite ratios $\alpha := {}^n/_p, \gamma := {}^{k_L}/_p$. Then the theorem of (Loureiro et al., 2021) states that

$$\epsilon_g \to \mathbb{E}_{(\nu,\lambda)}\left[\hat{g}(\hat{f}(\lambda), f_\star(\nu)\right]$$

where

$$\begin{pmatrix} \nu \\ \lambda \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \frac{a_\star^\top \Psi a_\star}{k_L} & m \\ m & q \end{bmatrix}\right).$$

The *self-overlap* q and the *magnetization* m are two scalar values characterized as the solutions of a system of scalar *saddle-point* equations, depending only on Ω, Φ, the loss function g and the output channel $f_\star$. (Loureiro et al., 2021) provide detailed examples of these saddle-point equations for standard loss functions, which we do not replicate here for conciseness.

This provide a very versatile framework to evaluate the sharp asymptotics of the test error of linear ERM methods, provided one crucially assumes that the student features v and the last layer post-activations $h_L^\star$ are jointly Gaussian. Such Gaussian equivalence results are known to hold rigorously for a variety of settings (Mei & Montanari, 2019; Hu & Lu, 2022a); (Loureiro et al., 2021; Cui et al., 2021; 2022) have also provided heuristic and numerical evidence of this equivalence holding for more intricate set-ups. The results of the present manuscript fit in this last line of works, and establishes heuristically and numerically that the equivalence holds for the presently considered multilayer network targets, and linear ERM methods. A rigorous proof thereof falls beyond the scope of the present paper.

Assuming the Gaussian equivalence, the test error of ERM methods on the target (1) can be accessed by evaluating the covariances Ψ, Ω, Φ using the result (10) (see also Appendix (A) for a detailed discussion and derivation), and inject these expressions in the saddle-point equations of (Loureiro et al., 2021). In the following, we sequentially detail the final saddle point equations for ridge regression, random features regression, and kernel regression.

Linear regression The sharp asymptotics for the test error in (Loureiro et al., 2021) can be applied, leveraging the characterization (10) for the covariances. This straightforwardly yields, in the notations of (Loureiro et al., 2021):

$$\begin{cases} \rho = \kappa_1^{\star 2}\rho_w + \kappa_\star^{\star 2}\rho_a + \Delta, \\ \Omega = \Sigma \\ \Phi = \sqrt{\Delta_a \Delta_w}\kappa_1^\star \Sigma, \end{cases}$$

with the corresponding saddle-point equations reading

$$\begin{cases} \hat{V} = \frac{\alpha}{1+V} \\ \hat{q} = \alpha\frac{\kappa_1^{\star 2}\rho_w + \kappa_\star^{\star 2}\rho_a + q - 2\kappa_1^\star m + \Delta}{(1+V)^2} \\ \hat{m} = \frac{\kappa_1^\star \alpha}{1+V} \end{cases} \qquad \begin{cases} V = \int d\mu(z)\frac{z}{\lambda I_d + \hat{V}z} \\ q = \int d\mu(z)\frac{\Delta_a\Delta_w\hat{m}^2z^3 + \hat{q}\Sigma^2}{(\lambda I_d + \hat{V}z)^2} \\ m = \Delta_a\Delta_w\hat{m}\int d\mu(z)\frac{z^2}{\lambda I_d + \hat{V}z} \end{cases} \qquad (176)$$

The generalization error is

$$\epsilon_g = \kappa_1^{\star 2}\rho_w + \kappa_\star^{\star 2}\rho_a + q - 2\kappa_1^\star m \qquad (177)$$

Random Features Denoting $F \in \mathbb{R}^{k \times d}$ the RF matrix, which we will take $\mathcal{N}(0, \Delta_f)$, and denoting

$$\kappa_1 = \frac{1}{\Delta_f \frac{\operatorname{Tr}\Sigma}{d}} \mathbb{E}_z^{\mathcal{N}(0, \Delta_f \frac{\operatorname{Tr}\Sigma}{d})}[z\sigma(z)] \qquad \kappa_* = \sqrt{\mathbb{E}_z^{\mathcal{N}(0, \Delta_f \frac{\operatorname{Tr}\Sigma}{d})}[\sigma(z)^2] - \kappa_1^2 \Delta_f \frac{\operatorname{Tr}\Sigma}{d}} \tag{178}$$

the GET coefficients associated to the RF non linearity $\sigma(\cdot)$ (9), the covariances read (10) (see Appendix A)

$$\begin{cases} \rho = \kappa_1^{\star 2}\rho_w + \kappa_*^{\star 2}\rho_a + \Delta, \\ \Omega = \kappa_1^2 \frac{F\Sigma F^\top}{d} + \kappa_*^2 I_k \\ \Phi = \sqrt{\Delta_a \Delta_w} \kappa_1 \kappa_1^\star \Sigma \frac{F^\top}{\sqrt{d}}. \end{cases}$$

From which it follows, leveraging the asymptotic results of (Loureiro et al., 2021),

$$\begin{cases} \hat{V} = \frac{\frac{\alpha}{\gamma}}{1+V} \\ \hat{q} = \frac{\alpha}{\gamma} \frac{\kappa_1^{\star 2}\rho_w + \kappa_*^{\star 2}\rho_a + q - 2\kappa_1^\star m + \Delta}{(1+V)^2} \\ \hat{m} = \sqrt{\gamma} \frac{\kappa_1^\star \frac{\alpha}{\gamma}}{1+V} \end{cases} \qquad \begin{cases} V = \frac{1}{\gamma d}\operatorname{Tr}\left[\frac{\kappa_1^2 \frac{F\Sigma F^\top}{d} + \kappa_*^2 I_k}{\lambda I_k + \hat{V}(\kappa_1^2 \frac{F\Sigma F^\top}{d} + \kappa_*^2 I_k)}\right] \\ q = \frac{1}{\gamma d}\operatorname{Tr}\left[\frac{\Delta_a \Delta_w \hat{m}^2 \kappa_1^2 \frac{F\Sigma^2 F^\top}{d}(\kappa_1^2 \frac{F\Sigma F^\top}{d} + \kappa_*^2 I_k) + \hat{q}(\kappa_1^2 F\Sigma F^\top + \kappa_*^2 I_k)^2}{(\lambda I_k + \hat{V}(\kappa_1^2 \frac{F\Sigma F^\top}{d} + \kappa_*^2 I_k))^2}\right] \\ m = \sqrt{\gamma}\frac{\Delta_a \Delta_w \hat{m}}{\gamma d}\operatorname{Tr}\left[\frac{\kappa_1^2 \frac{F\Sigma^2 F^\top}{d}}{\lambda I_k + \hat{V}(\kappa_1^2 \frac{F\Sigma F^\top}{d} + \kappa_*^2 I_k)}\right] \end{cases} \tag{179}$$

The error is the same as for ridge regression (177). In the case where $\Sigma = I_d$ (which we assume from now on), $F\Sigma F^\top$ and $F\Sigma^2 F^\top$ a jointly diagonalizable and introducing the spectral density $\rho(s)$ of $FF^\top/d$:

$$\begin{cases} \hat{V} = \frac{\frac{\alpha}{\gamma}}{1+V} \\ \hat{q} = \frac{\alpha}{\gamma} \frac{\kappa_1^{\star 2}\rho_w + \kappa_*^{\star 2}\rho_a + q - 2\kappa_1^\star m + \Delta}{(1+V)^2} \\ \hat{m} = \sqrt{\gamma} \frac{\kappa_1^\star \frac{\alpha}{\gamma}}{1+V} \end{cases} \qquad \begin{cases} V = \int d\rho(s) \frac{\kappa_1^2 s + \kappa_*^2}{\lambda + \hat{V}(\kappa_1^2 s + \kappa_*^2)} \\ q = \int d\rho(s) \frac{\Delta_a \Delta_w \hat{m}^2 \kappa_1^2 s(\kappa_1^2 s + \kappa_*^2) + \hat{q}(\kappa_1^2 s + \kappa_*^2)^2}{(\lambda + \hat{V}(\kappa_1^2 s + \kappa_*^2))^2} \\ m = \sqrt{\gamma}\Delta_a \Delta_w \hat{m} \int d\rho(s) \frac{\kappa_1^2 s}{\lambda + \hat{V}(\kappa_1^2 s + \kappa_*^2)} \end{cases}. \tag{180}$$

Introducing the Stieljes transform

$$g(z) = \int d\rho(s)\frac{1}{s-z}, \tag{181}$$

Then the saddle point equations can be expressed as

$$\begin{cases} \hat{V} = \frac{\frac{\alpha}{\gamma}}{1+V} \\ \hat{q} = \frac{\alpha}{\gamma} \frac{\kappa_1^{\star 2}\rho_w + \kappa_*^{\star 2}\rho_a + q - 2\kappa_1^\star m\sqrt{\Delta_a \Delta_w} + \Delta}{(1+V)^2} \\ \hat{m} = \sqrt{\Delta_a \Delta_w}\sqrt{\gamma} \frac{\kappa_1^\star \frac{\alpha}{\gamma}}{1+V} \end{cases} \qquad \begin{cases} V = \frac{1}{\hat{V}} - \frac{\lambda}{\hat{V}^2 \kappa_1^2}g\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right) \\ q = \frac{\hat{m}^2 + \hat{q}}{\hat{V}^2} - \frac{1}{\kappa_1^2 \hat{V}^2}\left(\frac{2\lambda(\hat{m}^2 + \hat{q})}{\hat{V}} + \hat{m}^2 \kappa_*^2\right)g\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right) \\ \quad + \frac{\lambda}{\kappa_1^4 \hat{V}^3}\left(\frac{\lambda(\hat{m}^2 + \hat{q})}{\hat{V}} + \hat{m}^2 \kappa_*^2\right)g'\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right) \\ m = \sqrt{\gamma}\frac{\hat{m}}{\hat{V}}\left[1 - \frac{1}{\kappa_1^2}\left(\frac{\lambda}{\hat{V}} + \kappa_*^2\right)g\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right)\right]. \end{cases}$$

Kernels We know consider the infinite width limit $\gamma \to \infty$ limit of (180), corresponding to GP kernel regression. For Gaussian projection, $\rho(\cdot)$ is Marcenko-Pastur, and reads

$$\rho(s) = \left(1 - \frac{1}{\gamma}\right)\delta(s) + \nu(s) \tag{182}$$

where

$$\nu(s) = \frac{1}{\gamma 2\pi\Delta_f}\frac{\sqrt{(s - \Delta_f(\sqrt{\gamma}-1)^2)(\Delta_f(\sqrt{\gamma}+1)^2 - s)}}{s}\mathbb{1}_{\Delta_f(\sqrt{\gamma}-1)^2 \le s \le \Delta_f(\sqrt{\gamma}+1)^2}. \tag{183}$$

The measure $\nu(\cdot)$ simplifies in the $\gamma \to \infty$ limit. For any test function $h(\cdot)$:

$$\int d\nu(x)h(x) = \int_{\sigma^2(\gamma - 2\sqrt{\gamma}+1)}^{\sigma^2(\gamma + 2\sqrt{\gamma}+1)} \frac{dx}{2\pi\gamma\Delta_f}\frac{\sqrt{(x - \Delta_f(\sqrt{\gamma}-1)^2)(\Delta_f(\sqrt{\gamma}+1)^2 - x)}}{x}h(x)$$

$$= \int_{-1}^{1} \frac{2dx}{\pi}\sqrt{(1-x)(1+x)}\frac{h(\gamma\Delta_f + 2\sqrt{\gamma}\Delta_f x)}{\gamma + 2\sqrt{\gamma}x} = \frac{h(\Delta_f\gamma)}{\gamma} + \mathcal{O}\left(\frac{1}{\sqrt{\gamma}}\right) \tag{184}$$

meaning

$$\rho(s) \approx \left(1 - \frac{1}{\gamma}\right)\delta(s) + \frac{1}{\gamma}\delta(s - \gamma\Delta_f). \tag{185}$$

The saddle point equation become approximately

$$
\begin{cases}
\hat{V} = \frac{\frac{\alpha}{\gamma}}{1+V} \\
\hat{q} = \frac{\alpha}{\gamma}\frac{\kappa_1^{\star 2}\rho_w + \kappa_*^{\star 2}\rho_a + q - 2\kappa_1^\star m + \Delta}{(1+V)^2} \\
\hat{m} = \sqrt{\gamma}\frac{\kappa_1^\star \frac{\alpha}{\gamma}}{1+V}
\end{cases}
\qquad
\begin{cases}
V &= \left(1 - \frac{1}{\gamma}\right)\frac{\kappa_*^2}{\lambda + \hat{V}(\kappa_*^2)} + \frac{1}{\gamma}\frac{\kappa_1^2\gamma\Delta_f + \kappa_*^2}{\lambda + \hat{V}(\kappa_1^2\gamma\Delta_f + \kappa_*^2)} \\
q &= \left(1 - \frac{1}{\gamma}\right)\frac{\hat{q}\kappa_*^4}{(\lambda + \hat{V}\kappa_*^2)^2} + \frac{1}{\gamma}\frac{\Delta_a\Delta_w\hat{m}^2\kappa_1^2\gamma\Delta_f(\kappa_1^2\gamma\Delta_f + \kappa_*^2) + \hat{q}(\kappa_1^2\gamma\Delta_f + \kappa_*^2)^2}{(\lambda + \hat{V}(\kappa_1^2\gamma\Delta_f + \kappa_*^2))^2} \\
m &= \sqrt{\gamma}\Delta_a\Delta_w\hat{m}\frac{1}{\gamma}\frac{\kappa_1^2\gamma\Delta_f}{\lambda + \hat{V}(\kappa_1^2\gamma\Delta_f + \kappa_*^2)}.
\end{cases}
\tag{186}
$$

Rescaling

$$\hat{V}, \hat{q} \leftarrow \gamma\hat{V}, \gamma\hat{q} \qquad\qquad \hat{m} \leftarrow \sqrt{\gamma}\hat{m} \tag{187}$$

The equations (186) reduce to

$$
\begin{cases}
\hat{V} = \frac{\alpha}{1+V} \\
\hat{q} = \alpha\frac{\kappa_1^{\star 2}\rho_w + \kappa_*^{\star 2}\rho_a + q - 2\kappa_1^\star m + \Delta}{(1+V)^2} \\
\hat{m} = \alpha\frac{\kappa_1^\star}{1+V}
\end{cases}
\qquad
\begin{cases}
V &= \frac{\kappa_*^2}{\lambda} + \frac{\kappa_1^2\Delta_f}{\lambda + \hat{V}\kappa_1^2\Delta_f} \\
q &= \frac{\Delta_a\Delta_w\hat{m}^2\kappa_1^4\Delta_f^2 + \hat{q}\kappa_1^4\Delta_f^2}{(\lambda + \hat{V}\kappa_1^2\Delta_f)^2} \\
m &= \Delta_a\Delta_w\hat{m}\frac{\kappa_1^2\Delta_f}{\lambda + \hat{V}\kappa_1^2\Delta_f}
\end{cases}
. \tag{188}
$$

F. Optimality of kernel ERM on shallow targets

In this appendix, we demonstrate that optimally regularized ridge regression and kernel regression achieve the Bayes optimal MSE, and provide a derivation of the explicit formulas for the oprimal regularizations (21) and (25). Note that this extends the very closely related results of (Sahraee-Ardakan et al., 2022), which were derived for a GP target, corresponding to the infinite width limit of (1). In fact, (21) and (25) can be heuristically derived by evaluating the formulas of (Sahraee-Ardakan et al., 2022) on the equivalent linear teacher (16).

F.1. Ridge ERM achieves Bayes optimal error

We first show that optimally regularized ridge regression achieves the Bayes optimal MSE (12). As a side result, the formula (21) is obtained.

Optimal regularization for ridge   Combining (177) and (176), one can derive a self-consistent equation for ϵ_g:

$$
\epsilon_g = \kappa_*^2\rho_a + \int d\mu(z)\frac{\kappa_1^2\Delta_w\Delta_a\hat{V}^2 z^3 + \hat{q}z^2 - 2\kappa_1^2\Delta_w\Delta_a\lambda\hat{V}z^2 - 2\kappa_1^2\Delta_w\Delta_a\hat{V}^2 z^3 + \kappa_1^2\Delta_w\Delta_a\lambda^2 z + \kappa_1^2\Delta_w\Delta_a\hat{V}^2 z^3 + 2\kappa_1^2\Delta_w\Delta_a\lambda\hat{V}z^2}{(\lambda I_d + \hat{V}z)^2}
$$

$$
\overset{\tilde{\lambda} \equiv \frac{\lambda}{\hat{V}}}{=} \kappa_*^2\rho_a + \int d\mu(z)\frac{\kappa_1^2\Delta_a\Delta_w\tilde{\lambda}^2 z + \frac{\epsilon_g}{\alpha}z^2}{(\tilde{\lambda}I_d + z)^2} \tag{189}
$$

Requiring ϵ_g to be minimal with respect to λ (equivalently, $\tilde{\lambda}$):

$$
0 = \int d\mu(z)\frac{2\kappa_1^2\Delta_w\Delta_a\tilde{\lambda}z}{(\tilde{\lambda}I_d + z)^2} - 2\frac{2\kappa_1^2\Delta_w\Delta_a\tilde{\lambda}^2 z + \frac{\epsilon_g}{\alpha}z^2}{(\tilde{\lambda}I_d + z)^3} \tag{190}
$$

$$
= \int d\mu(z)\frac{z^2}{(\tilde{\lambda}I_d + z)^3} \times 2\kappa_1^2\Delta_w\Delta_a\left(\tilde{\lambda} - \frac{\epsilon_g}{\alpha\kappa_1^2\Delta_w\Delta_a}\right) \tag{191}
$$

It follows the relation

$$
\lambda = \frac{\hat{V}\epsilon_g}{\alpha\kappa_1^2\Delta_w\Delta_a} = \frac{\epsilon_g}{\kappa_1^2\Delta_w\Delta_a(1+V)}. \tag{192}
$$

Explicit characterization of the optimal regularization   We now proceed to show (21). Rewrite (192) as

$$\epsilon_g = \alpha \kappa_1^2 \Delta_w \Delta_a \tilde{\lambda}, \tag{193}$$

where we remind the auxiliary variable $\tilde{\lambda} \equiv \lambda/\hat{V}$. Replacing ϵ_g in (189) for $\tilde{\lambda}$, one reaches

$$\tilde{\lambda} = \frac{1}{\alpha}\left(\frac{\kappa_*^2 \Delta_a + \Delta}{\Delta_w \Delta_a \kappa_1^2} + \int \mathrm{d}\mu(z) \frac{\tilde{\lambda}z}{\tilde{\lambda}I_d + z} \right) \tag{194}$$

On the other hand, the saddle point equations for $\hat{V}, V$ (176) can be rewritten as

$$\frac{1}{\hat{V}} = \frac{1}{\alpha}\left(1 + \frac{1}{\hat{V}} \int \mathrm{d}\mu(z) \frac{z}{\tilde{\lambda}I_d + z} \right) \tag{195}$$

Multiplying by λ on both sides:

$$\tilde{\lambda} = \frac{1}{\alpha}\left(\lambda + \int \mathrm{d}\mu(z) \frac{\tilde{\lambda}z}{\tilde{\lambda}I_d + z} \right) \tag{196}$$

Identifying (194) and (196) yield

$$\lambda = \frac{\kappa_*^2 \Delta_a + \Delta}{\Delta_w \Delta_a \kappa_1^2}, \tag{197}$$

which is (21) for $L = 1$ hidden layer.

Connection to (Sahraee-Ardakan et al., 2022)   In the two layer case, this result interestingly corresponds to the optimal regularization given by Theorem 4.1 in (Sahraee-Ardakan et al., 2022), provided it is applied to the GP corresponding to the *infinite-width* limit of the target. Following (Sahraee-Ardakan et al., 2022), the optimal regularization for ridge regression over such a GP reads

$$\lambda = \frac{c_0 + \sigma^2}{c_2},$$

which translating into our notations with

$$c_2 = \Delta_w \Delta_a \kappa_1^2 \qquad\qquad c_0 = \Delta_a \kappa_*^2 \qquad\qquad \sigma^2 = \Delta \tag{198}$$

leads back to

$$\lambda = \frac{\kappa_*^2 \Delta_a + \Delta}{\Delta_w \Delta_a \kappa_1^2}. \tag{199}$$

Equivalence Ridge/Bayes   Finally, we show that ridge regression, regularized with (21), achieves the Bayes optimal MSE (12). We will show to this end that $-\kappa_1^2 q_w = q - 2\kappa_1 m$, taking inspiration from (116) to define $\hat{q}_w = \alpha\kappa_1^2/\epsilon_g = (\hat{V}\Delta_w \Delta_a)/\lambda$. In the following, λ is understood as the ridge optimal regularization (21).

$$
\begin{aligned}
-\frac{q - 2\kappa_1 m}{\kappa_1^2} &= \frac{1}{\kappa_1^2} \int \mathrm{d}\mu(z) \frac{\Delta_w \Delta_a \kappa_1^2 \hat{V}^2 z^3 + \hat{q}z^2 - 2\lambda\Delta_w \Delta_a \kappa_1^2 \hat{V}z^2 - 2\Delta_w \Delta_a \kappa_1^2 \hat{V}^2 z^3}{(\lambda I_d + \hat{V}z)^2} \\
&= \frac{1}{\kappa_1^2} \int \mathrm{d}\mu(z) \frac{\frac{\hat{V}^2}{\alpha}\epsilon_g z^2 + \Delta_w \Delta_a \kappa_1^2 \hat{V}^2 z^3}{(\lambda I_d + \hat{V}z)^2} \\
&= \frac{1}{\kappa_1^2} \int \mathrm{d}\mu(z) \frac{\hat{V}\lambda\Delta_w \Delta_a \kappa_1^2 + \Delta_w \Delta_a \kappa_1^2 \hat{V}^2 z^3}{(\lambda I_d + \hat{V}z)^2} \\
&= \frac{1}{\kappa_1^2} \int \mathrm{d}\mu(z) \frac{\hat{V}\lambda\Delta_w \Delta_a \kappa_1^2 z^2 + \Delta_w \Delta_a \kappa_1^2 \hat{V}^2 z^3}{(\lambda I_d + \hat{V}z)^2} \\
&= \int \mathrm{d}\mu(z) \frac{\hat{q}_w z^2}{I_d + \Delta_w \Delta_a \hat{q}_w z} \\
&= \int \mathrm{d}\mu(z) \frac{\alpha\kappa_1^2 z^2}{\epsilon_g + \Delta_w \Delta_a \alpha\kappa_1^2 z}
\end{aligned}
\tag{200}
$$

which corresponds to the equation (116). Renaming $q_w := -q - 2\kappa_1 m/\kappa_1^2$ allows to recover the two-layer Bayes MSE.

F.2. Kernels achieve Bayes optimal error

This subsection similarly addresses the case of kernels.

Optimal regularization for RF kernel ERM   Combining (177) and (188), one can as in the ridge case derive a self-consistent equation for the generalization error:

$$\epsilon_g - \Delta = \kappa_*^{\star 2}\rho_a + \frac{\Delta_w\Delta_a\kappa_1^{\star 2}\lambda^2 + \Delta_w\Delta_a\kappa_1^{\star 2}\kappa_1^4\Delta_f^2\hat{V} + 2\Delta_w\Delta_a\kappa_1^{\star 2}\lambda\hat{V}\kappa_1\Delta_f - \Delta_w\Delta_a\kappa_1^{\star 2}\kappa_1^4\Delta_f^2\hat{V} + \frac{\hat{V}^2}{\alpha}\kappa_1^4\Delta_f^2\epsilon_g - 2\Delta_w\Delta_a\kappa_1^{\star 2}\lambda\hat{V}\kappa_1\Delta_f}{(\lambda + \hat{V}\kappa_1^2\Delta_f)^2}$$

$$\overset{\tilde{\lambda}\equiv\frac{\lambda}{\hat{V}}}{=} \kappa_*^{\star 2}\rho_a + \frac{\Delta_w\Delta_a\kappa_1^{\star 2}\tilde{\lambda}^2 + \frac{\kappa_1^4\Delta_f^2\epsilon_g}{\alpha}}{(\tilde{\lambda} + \kappa_1^2\Delta_f)^2} \tag{201}$$

Requiring $\frac{\partial \epsilon_g}{\partial \tilde{\lambda}} = 0$ leads to

$$0 = \frac{2\Delta_w\Delta_a\kappa_1^{\star 2}\tilde{\lambda}^2 + 2\Delta_w\Delta_a\kappa_1^{\star 2}\kappa_1^2\Delta_f\tilde{\lambda} - 2\Delta_w\Delta_a\kappa_1^{\star 2}\tilde{\lambda}^2 - 2\frac{\kappa_1^4\Delta_f^2\epsilon_g}{\alpha}}{(\tilde{\lambda} + \kappa_1^2\Delta_f)^3} = 2\kappa_1^2\Delta_f\frac{\Delta_w\Delta_a\kappa_1^{\star 2}\tilde{\lambda} - \frac{\epsilon_g\kappa_1^2\Delta_f}{\alpha}}{(\tilde{\lambda} + \kappa_1^2\Delta_f)^3} \tag{202}$$

The optimal regularization thus reads

$$\lambda = \frac{\hat{V}\epsilon_g\kappa_1^2\Delta_f}{\Delta_w\Delta_a\alpha\kappa_1^{\star 2}} = \frac{\epsilon_g\kappa_1^2\Delta_f}{\Delta_w\Delta_a\kappa_1^{\star 2}(1+V)}. \tag{203}$$

Explicit characterization for the optimal regularization   One is now in position to derive (25). Combining (201) and (203), the optimal λ for kernel regression is the solution on the λ variable of the system of equations

$$\begin{cases} \lambda = \frac{\hat{V}\epsilon_g\kappa_1^2\Delta_f}{\Delta_w\Delta_a\alpha\kappa_1^{\star 2}} \\ \epsilon_g = \Delta + \kappa_*^{\star 2}\rho_a + \frac{\Delta_w\Delta_a\kappa_1^{\star 2}\tilde{\lambda}^2 + \frac{\kappa_1^4\Delta_f^2\epsilon_g}{\alpha}}{(\tilde{\lambda}+\kappa_1^2\Delta_f)^2} \\ \frac{1}{\hat{V}} = \frac{1}{\alpha}\left(1 + \frac{\kappa_*^2}{\lambda} + \frac{\kappa_1^2\Delta_f}{\lambda+\hat{V}\kappa_1^2\Delta_f}\right) \end{cases} . \tag{204}$$

Introducing the variable $\tilde{\lambda} \equiv \lambda/\hat{V}$, and plugging the first line into the second, this simplifies to

$$\begin{cases} \epsilon_g = \alpha\frac{\Delta_w\Delta_a}{\Delta_f}\frac{\kappa_1^{\star 2}}{\kappa_1^2}\tilde{\lambda} \\ \tilde{\lambda} = \frac{\Delta_f\kappa_1^2}{\Delta_w\kappa_1^{\star 2}}\frac{1}{\alpha}\left(\kappa_*^{\star 2} + \frac{\Delta}{\Delta_a} + \frac{\Delta\kappa_1^{\star 2}\tilde{\lambda}}{\tilde{\lambda}+\kappa_1^2\Delta_f}\right) \\ \tilde{\lambda} = \frac{1}{\alpha}\left(\lambda + \kappa_*^2 + \frac{\kappa_1^2\Delta_f\tilde{\lambda}}{\tilde{\lambda}+\kappa_1^2\Delta_f}\right) \end{cases} . \tag{205}$$

Substracting the third line from the second yields an expression for λ:

$$\frac{\Delta_f\kappa_1^2}{\Delta_w\kappa_1^{\star 2}}\left(\kappa_*^{\star 2} + \frac{\Delta}{\Delta_a}\right) = \lambda + \kappa_*^2. \tag{206}$$

In other words, the optimal regularization is explicitly given by

$$\lambda = \kappa_1^2\Delta_f\left(\frac{\Delta_a\kappa_*^{\star 2} + \Delta}{\kappa_1^{\star 2}\Delta_w\Delta_a} - \frac{\kappa_*^2}{\kappa_1^2\Delta_f}\right) \tag{207}$$

Alternative argument using (Sahraee-Ardakan et al., 2022)   Theorem 2.1 of (Sahraee-Ardakan et al., 2022) shows that regularizing a kernel method with λ_{GP} yields the same test error as ridge regression with regularization λ_r, provided these two regularization are related as

$$\lambda_{\mathrm{GP}} = c_2 \times \lambda_r - c_0, \tag{208}$$

i.e. in our notations

$$\lambda_{\mathrm{GP}} = \Delta_f\kappa_1^2 \times \lambda_r - \kappa_*^2. \tag{209}$$

Taking λ_r to be the optimal regularizer for ridge regression derived in previous subsection, this leads to

$$\lambda_{\mathrm{GP}} = \Delta_f\kappa_1^2\left(\frac{(\kappa_*^\star)^2\Delta_a + \Delta}{\Delta_w\Delta_a(\kappa_1^\star)^2}\right) - \kappa_*^2 = \kappa_1^2\Delta_f\left(\frac{\Delta_a\kappa_*^{\star 2} + \Delta}{\kappa_1^{\star 2}\Delta_w\Delta_a} - \frac{\kappa_*^2}{\kappa_1^2\Delta_f}\right) \tag{210}$$

Equivalence kernel/BO We now show the optimally regularized error achieves the Bayes optimal MSE (12). More precisely, we show $q_w = -(q - 2\kappa_1^\star m)/\kappa_1^{\star 2}$:

$$
-\frac{q - 2\kappa_1^\star m}{\kappa_1^{\star 2}} = -\frac{1}{\kappa_1^{\star 2}} \frac{\Delta_w \Delta_a \kappa_1^{\star 2} \kappa_1^4 \Delta_f^2 \hat{V}^2 + \kappa_1^3 \Delta_f^2 \frac{\hat{V}^2}{\alpha} \epsilon_g - 2\Delta_w \Delta_a \kappa_1^{\star 2} \kappa_1^2 \Delta_f \hat{V} \lambda - 2\Delta_w \Delta_a \kappa_1^{\star 2} \kappa_1^4 \Delta_f^2 \hat{V}^2}{(\lambda + \hat{V}\kappa_1 \Delta_f)^2}
$$

$$
= \frac{\Delta_w \Delta_a (\Delta_f^2 \kappa_1^4 \hat{V}^2 + 2\kappa_1^2 \Delta_f \lambda \hat{V}) - \frac{\kappa_1^4 \Delta_f^2}{\kappa_1^{\star 2}} \frac{\hat{V}}{\alpha} \epsilon_g}{(\lambda + \hat{V}\kappa_1 \Delta_f)^2}
$$

$$
= \frac{\Delta_w \Delta_a \frac{\alpha \kappa_1^{\star 2}}{\epsilon_g}}{1 + \Delta_w \Delta_a \frac{\alpha \kappa_1^{\star 2}}{\epsilon_g}} \overset{\hat{q}_w = \alpha \kappa_1^{\star 2}/\epsilon_g}{\equiv} \frac{\Delta_w \Delta_a \hat{q}_w}{1 + \Delta_w \Delta_a \hat{q}_w} \tag{211}
$$

which is exactly the update for q_w (116).

F.3. A short RMT argument for finite width RF

We provide in this subsection a short argument, at a physics level of rigor, that RF cannot achieve Bayes optimality away from the infinite width (kernel) limit, for simplicity in the isotropic $\Sigma = I_d$ case. Remember that, in the framework of (Loureiro et al., 2021), the covariance between the student features is $\Omega = \kappa_1^2 FF^\top/d + \kappa_*^2 I_p$, while the average effective cross-covariance between teacher and student features can be taken as $\Phi = \kappa_1 \kappa_1^\star F^\top/\sqrt{d}$. The labels are generated from the target features (last layer post-activation), and not from the student features, rendering the setting slightly intricate. However, under the assumption that the student features and teacher post-activations are jointly Gaussian, (Clarté et al., 2022) provide an equivalent model where the labels are generated directly from the student features. Introduce the effective teacher

$$
\check{\theta} \equiv \Omega^{-\frac{1}{2}} \Phi^\top a^\star \tag{212}
$$

and the effective noise variance

$$
\check{\Delta} = \frac{1}{d} \text{Tr}\left[\Psi - \Phi \Omega^{-1} \Phi^\top \right] = (\kappa_1^\star)^2 \left[1 - \gamma \int d\rho(s) \frac{\kappa_*^2 s}{\kappa_1^2 s + \kappa_*^2} \right] = (\kappa_1^\star)^2 \left[1 - \gamma + \gamma \frac{\kappa_*^2}{\kappa_1^2} g\left(-\frac{\kappa_*^2}{\kappa_1^2} \right) \right], \tag{213}
$$

where $g(\cdot)$ is the Stieljes transform of the Marcenko Pastur law. Note that $\check{\Delta} = \mathcal{O}(1)$ even in the $\gamma \to \infty$ limit. Then random feature regression on a two-layer target is equivalent to the ERM problem

$$
y \sim \frac{1}{\sqrt{p}} \check{\theta}^\top u + \sqrt{\check{\Delta}} \mathcal{N}(0,1),
$$

$$
\check{\mathcal{R}}(w) = \sum_{\mu=1}^{n} \frac{1}{2} \left(y^\mu - \frac{1}{\sqrt{p}} w^\top u^\mu \right)^2 + \frac{\lambda}{2} w^\top w, \qquad u \sim \mathcal{N}(0, \Omega). \tag{214}
$$

Absorbing the Ω dependence into the weights, this problem is further equivalent to

$$
y \sim \frac{1}{\sqrt{d}} a_\star^\top \Phi u + \sqrt{\check{\Delta}} \mathcal{N}(0,1),
$$

$$
\check{\mathcal{R}}(w) = \frac{1}{2\check{\Delta}} \sum_{\mu=1}^{n} \left(y^\mu - \frac{1}{\sqrt{d}} w^\top u^\mu \right)^2 + \frac{\check{\Delta}\lambda\gamma}{2} w^\top \Omega^{-1} w, \qquad u \sim \mathcal{N}(0, I_p) \tag{215}
$$

Note that when $a_\star \sim \mathcal{N}(0, I_d)$, then $\Phi^\top a_\star \sim \mathcal{N}(0, \Phi^\top \Phi/\gamma)$. As discussed in the main text, ERM using a linear model and a square loss, as implied by (215), is equivalent to conducting Bayesian inference, with prior $\check{\Delta}$ over the additive noise and prior $1/\lambda\gamma\check{\Delta}\Omega$ over the weights. Note that this prior matches the ground truth covariance if and only if $\exists \lambda$, $\Phi^\top \Phi/\gamma = 1/\lambda\gamma\check{\Delta}\Omega$. This requires at least the spectra to match, i.e

$$
\frac{(\kappa_1^\star)^2}{\gamma} \text{spec}\left(\kappa_1^2 FF^\top/d \right) \propto \frac{1}{\gamma\lambda} \text{spec}\left(\kappa_1^2 FF^\top/d \right) + \frac{\kappa_*^2}{\gamma\lambda}, \tag{216}
$$

which can only be verified asymptotically as $\gamma \to \infty$, when the spectrum has a finite fraction $1/\gamma$ of $\mathcal{O}(\gamma)$ eigenvalues and $(\gamma - 1)/\gamma$ of zero eigenvalues, i.e. in the infinite width (kernel) limit. This implies that doing *finite width* RF regression is equivalent to doing Bayes inference with *mismatched* prior over the effective weights, and that therefore the test error achieved by RF regression cannot be Bayes optimal. This is numerically obersved in Fig. 1, where the test error of RF is bounded away from the Bayes MSE baseline (12).

G. ERM : the deep case

In this section, we extend the discussion of Appendix E and F to the general multi-layer case.

G.1. Ridge regression

Similarly to the two-layer case, these results leverage the asymptotics of (Loureiro et al., 2021), using, in their notations,

$$
\begin{cases}
\rho = \prod\limits_{\ell=1}^{L^\star} \left(\kappa_1^{\star(\ell)}\right)^2 \rho_1^{L^\star} + \sum\limits_{\ell_0=1}^{L-1} \kappa_*^{\star(\ell_0)} \prod\limits_{\ell=\ell_0+1}^{L^\star} \left(\kappa_1^{(\ell)}\right)^2 \rho_{\ell_0+1}^{L^\star} + \left(\kappa_*^{L^\star}\right)^2 \rho_a + \Delta, \\
\Omega = \Sigma \\
\Phi = \Delta_a^{\frac{1}{2}} \prod\limits_{\ell=1}^{L^\star} \Delta_\ell^{\frac{1}{2}} \kappa_1^{\star(\ell)} \Sigma
\end{cases}
$$

The replica saddle-point equations for a ridge student then read:

$$
\begin{cases}
\hat{V} = \frac{\alpha}{1+V} \\
\hat{q} = \alpha \frac{\prod\limits_{\ell=1}^{L^\star} \left(\kappa_1^{\star(\ell)}\right)^2 \rho_1^{L^\star} + \sum\limits_{\ell_0=1}^{L-1} \kappa_*^{\star(\ell_0)} \prod\limits_{\ell=\ell_0+1}^{L^\star} \left(\kappa_1^{(\ell)}\right)^2 \rho_{\ell_0+1}^{L^\star} + \left(\kappa_*^{L^\star}\right)^2 \rho_a + q - 2 \prod\limits_{\ell=1}^{L^\star} \kappa_1^{\star(\ell)} m + \Delta}{(1+V)^2} \\
\hat{m} = \frac{\prod\limits_{\ell=1}^{L^\star} \kappa_1^{\star(\ell)} \alpha}{1+V}
\end{cases}
\qquad
\begin{cases}
V = \int \mathrm{d}\mu(z) \frac{z}{\lambda I_d + \hat{V} z} \\
q = \int \mathrm{d}\mu(z) \frac{\Delta_a \prod\limits_{\ell=1}^{L^\star} \Delta_\ell \hat{m}^2 z^3 + \hat{q} z^2}{(\lambda I_d + \hat{V} z)^2} \\
m = \Delta_a \prod\limits_{\ell=1}^{L^\star} \Delta_\ell \hat{m} \int \mathrm{d}\mu(z) \frac{z^2}{\lambda I_d + \hat{V} z}.
\end{cases}
\tag{217}
$$

The generalization error is in the multilayer case

$$
\epsilon_g = \prod_{\ell=1}^{L^\star} \left(\kappa_1^{\star(\ell)}\right)^2 \rho_1^{L^\star} + \sum_{\ell_0=1}^{L^\star-1} \kappa_*^{\star(\ell_0)} \prod_{\ell=\ell_0+1}^{L^\star} \left(\kappa_1^{\star(\ell)}\right)^2 \rho_{\ell_0+1}^{L^\star} + \left(\kappa_*^{L^\star}\right)^2 \rho_a + q - 2 \prod_{\ell=1}^{L^\star} \kappa_1^{\star(\ell)} m + \Delta.
\tag{218}
$$

Optimal regularization for ridge   Combining (20) and (19), a self-consistent equation for ϵ_g can be reached:

$$
\epsilon_g \overset{\tilde{\lambda} \equiv \frac{\lambda}{\hat{V}}}{=} \epsilon_r + \int \mathrm{d}\mu(z) \frac{\prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell \tilde{\lambda}^2 z + \frac{\epsilon_g}{\alpha} z^2}{(\tilde{\lambda} I_d + z)^2}
\tag{219}
$$

Requiring ϵ_g to be minimal with respect to λ (equivalently, $\tilde{\lambda}$):

$$
0 = \int \mathrm{d}\mu(z) \frac{2 \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell \tilde{\lambda} z}{(\tilde{\lambda} I_d + z)^2} - 2 \frac{\prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell \tilde{\lambda}^2 z + \frac{\epsilon_g}{\alpha} z^2}{(\tilde{\lambda} I_d + z)^3}
\tag{220}
$$

$$
= \int \mathrm{d}\mu(z) \frac{z^2}{(\tilde{\lambda} I_d + z)^3} \times 2 \prod_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_a \prod_{\ell=1}^{L} \Delta_\ell \left(\tilde{\lambda} - \frac{\epsilon_g}{\alpha \Delta_a \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell} \right)
\tag{221}
$$

It follows the relation

$$
\lambda = \frac{\hat{V} \epsilon_g}{\alpha \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell} = \frac{\epsilon_g}{(1+V) \Delta_a \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell}.
\tag{222}
$$

One is now in position to derive the optimal regularization (21).

Rewrite (222) as

$$
\epsilon_g = \alpha \tilde{\lambda} \Delta_a \prod_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell,
$$

where we remind the auxilary variable $\tilde{\lambda} \equiv \lambda/\hat{V}$. Replacing ϵ_g in (219) for $\tilde{\lambda}$, one reaches

$$\tilde{\lambda} = \frac{1}{\alpha} \left(\frac{\epsilon_r}{\Delta_a \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)} \right)^2 \Delta_\ell} + \int \mathrm{d}\mu(z) \frac{\tilde{\lambda} z}{\tilde{\lambda} I_d + z} \right). \tag{223}$$

On the other hand, the saddle point equations for $\hat{V}, V$ (176) can be rewritten as

$$\frac{1}{\hat{V}} = \frac{1}{\alpha} \left(1 + \frac{1}{\hat{V}} \int \mathrm{d}\mu(z) \frac{z}{\tilde{\lambda} I_d + z} \right) \tag{224}$$

Multiplying by λ on both sides:

$$\tilde{\lambda} = \frac{1}{\alpha} \left(\lambda + \int \mathrm{d}\mu(z) \frac{\tilde{\lambda} z}{\tilde{\lambda} I_d + z} \right). \tag{225}$$

Identifying (223) and (225) yield

$$\lambda = \frac{\epsilon_r}{\Delta_a \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)} \right)^2 \Delta_\ell} \tag{226}$$

which is (21).

Equivalence Ridge/BO We now show that under the optimal regularization (21), the saddle point equations for ridge regression (20) reduce to (13). It then straightforwardly follow that ridge regression achieves the Bayes optimal MSE. In the following, λ is understood as the ridge optimal regularization (21). Define

$$\hat{q} = \frac{\alpha \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)} \right)^2}{\epsilon_g} = \frac{\hat{V} \Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell}{\lambda}$$

to connect with (13). Then

$$-\frac{q - 2\prod_{\ell=1}^{L}\kappa_1^{(\ell)}m}{\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2} = \frac{1}{\kappa_1^2}\int d\mu(z)\frac{\hat{q}z^2 + \Delta_a \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell\left(\hat{V}^2 z^3 + -2\lambda\hat{V}z^2 - 2\hat{V}^2 z^3\right)}{(\lambda I_d + \hat{V}z)^2}$$

$$= \frac{1}{\kappa_1^2}\int d\mu(z)\frac{\frac{\hat{V}^2}{\alpha}\epsilon_g z^2 + \Delta_a \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell \hat{V}^2 z^3}{(\lambda I_d + \hat{V}z)^2}$$

$$= \Delta_a \prod_{\ell=1}^{L}\Delta_\ell \int d\mu(z)\frac{\hat{V}\lambda + \hat{V}^2 z^3}{(\lambda I_d + \hat{V}z)^2}$$

$$= \Delta_a \prod_{\ell=1}^{L}\Delta_\ell \int d\mu(z)\frac{\hat{V}\lambda z^2 + \hat{V}^2 z^3}{(\lambda I_d + \hat{V}z)^2}$$

$$= \int d\mu(z)\frac{\hat{q}z^2}{I_d + \Delta_a \prod_{\ell=1}^{L}\Delta_\ell \hat{q}z}$$

$$= \int d\mu(z)\frac{\alpha \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 z^2}{\epsilon_g I_d + \Delta_a \prod_{\ell=1}^{L}\Delta_\ell \alpha \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 z}$$

$$(227)$$

which is exactly the Bayes Optimal equation (13).

G.2. Random nets / features

The linearization (10), further detailed in Appendix A, translate in the framework of (Loureiro et al., 2021) into

$$\begin{cases} \rho = \prod_{\ell=1}^{L^\star}\left(\kappa_1^{\star(\ell)}\right)^2 \rho_1^{L^\star} + \sum_{\ell_0=1}^{L-1}\kappa_*^{\star(\ell_0)}\prod_{\ell=\ell_0+1}^{L^\star}\left(\kappa_1^{(\ell)}\right)^2 \rho_{\ell_0+1}^{L^\star} + \left(\kappa_*^{L^\star}\right)^2\rho_a + \Delta, \\ \Omega = \kappa_1^2 \frac{F\Sigma F^\top}{d} + \kappa_*^2 I_k \\ \Phi = \sqrt{\Delta_a \prod_{\ell=1}^{L}\Delta_\ell}\kappa_1 \prod_{\ell=1}^{L}\kappa_1^{\star(\ell)}\Sigma\frac{F^\top}{\sqrt{d}} \end{cases}$$

yielding the saddle point equations

$$\begin{cases} \hat{V} = \frac{\frac{\alpha}{\gamma}}{1+V} \\ \hat{q} = \frac{\alpha}{\gamma}\frac{\prod_{\ell=1}^{L^\star}\left(\kappa_1^{\star(\ell)}\right)^2\rho_1^{L^\star} + \sum_{\ell_0=1}^{L-1}\kappa_*^{\star(\ell_0)}\prod_{\ell=\ell_0+1}^{L^\star}\left(\kappa_1^{(\ell)}\right)^2\rho_{\ell_0+1}^{L^\star} + \left(\kappa_*^{L^\star}\right)^2\rho_a + q - 2\prod_{\ell=1}^{L^\star}\kappa_1^{\star(\ell)}m + \Delta}{(1+V)^2} \\ \hat{m} = \sqrt{\gamma}\frac{\prod_{\ell=1}^{L^\star}\kappa_1^{\star(\ell)}\frac{\alpha}{\gamma}}{1+V} \end{cases}$$

$$(228)$$

$$\begin{cases} V = \int d\rho(s)\frac{\kappa_1^2 s + \kappa_*^2}{\lambda + \hat{V}(\kappa_1^2 s + \kappa_*^2)} \\ q = \int d\rho(s)\frac{\Delta_a \prod_{\ell=1}^{L^\star}\Delta_\ell \hat{m}^2 \kappa_1^2 s(\kappa_1^2 s + \kappa_*^2) + \hat{q}(\kappa_1^2 s + \kappa_*^2)^2}{(\lambda + \hat{V}(\kappa_1^2 s + \kappa_*^2))^2} \\ m = \sqrt{\gamma}\Delta_a \prod_{\ell=1}^{L^\star}\Delta_\ell \hat{m}\int d\rho(s)\frac{\kappa_1^2 s}{\lambda + \hat{V}(\kappa_1^2 s + \kappa_*^2)} \end{cases}$$

$$(229)$$

G.3. GP kernels

By generalizing the 2-layer equations to the multi-layer case,

$$
\begin{cases}
\hat{V} = \frac{\alpha}{1+V} \\
\hat{q} = \alpha \frac{\prod\limits_{\ell=1}^{L^\star}\left(\kappa_1^{\star(\ell)}\right)^2 \rho_1^{L^\star} + \sum\limits_{\ell_0=1}^{L-1}\kappa_*^{\star(\ell_0)}\prod\limits_{\ell=\ell_0+1}^{L^\star}\left(\kappa_1^{(\ell)}\right)^2\rho_{\ell_0+1}^{L^\star} + \left(\kappa_*^{L^\star}\right)^2\rho_a + q - 2\prod\limits_{\ell=1}^{L^\star}\kappa_1^{\star(\ell)}m + \Delta}{(1+V)^2} \\
\hat{m} = \alpha \frac{\prod\limits_{\ell=1}^{L^\star}\kappa_1^{\star(\ell)}}{1+V}
\end{cases}
\qquad
\begin{cases}
V = \frac{\kappa_*^2}{\lambda} + \frac{\kappa_1^2\Delta_f}{\lambda + \hat{V}\kappa_1^2\Delta_f} \\
q = \frac{\Delta_a\prod\limits_{\ell=1}^{L^\star}\Delta_\ell\hat{m}^2\kappa_1^4\Delta_f^2 + \hat{q}\kappa_1^4\Delta_f^2}{(\lambda + \hat{V}\kappa_1^2\Delta_f)^2} \\
m = \Delta_a\prod\limits_{\ell=1}^{L^\star}\Delta_\ell\hat{m}\frac{\kappa_1^2\Delta_f}{\lambda + \hat{V}\kappa_1^2\Delta_f}
\end{cases}
\tag{230}
$$

Optimal regularization for RF kernel ERM   Combining (177) and (188), one can as in the ridge case derive a self-consistent equation for the generalization error:

$$
\epsilon_g - \Delta = \epsilon_r + \frac{\Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell\lambda^2 + \Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell\kappa_1^4\Delta_f^2\hat{V} + 2\Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell\lambda\hat{V}\kappa_1\Delta_f - \Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell\kappa_1^4\Delta_f^2\hat{V} + \frac{\hat{V}^2}{\alpha}\kappa_1^4\Delta_f^2\epsilon_g - 2\Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell\lambda\hat{V}\kappa_1\Delta_f}{(\lambda + \hat{V}\kappa_1^2\Delta_f)^2}
$$

$$
\overset{\tilde{\lambda}\equiv\frac{\lambda}{\hat{V}}}{=} \epsilon_r + \frac{\Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell\tilde{\lambda}^2 + \frac{\kappa_1^4\Delta_f^2\epsilon_g}{\alpha}}{(\tilde{\lambda}+\kappa_1^2\Delta_f)^2}
\tag{231}
$$

Requiring $\frac{\partial\epsilon_g}{\partial\tilde{\lambda}} = 0$ leads to

$$
0 = \frac{2\Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell\left(\tilde{\lambda}^2 + 2\kappa_1^2\Delta_f\tilde{\lambda} - 2\tilde{\lambda}^2\right) - 2\frac{\kappa_1^4\Delta_f^2\epsilon_g}{\alpha}}{(\tilde{\lambda}+\kappa_1^2\Delta_f)^3} = 2\kappa_1^2\Delta_f\frac{\Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell\tilde{\lambda} - \frac{\epsilon_g\kappa_1^2\Delta_f}{\alpha}}{(\tilde{\lambda}+\kappa_1^2\Delta_f)^3}.
\tag{232}
$$

The optimal regularization thus reads

$$
\lambda = \frac{\hat{V}\epsilon_g\kappa_1^2\Delta_f}{\alpha\Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell} = \frac{\epsilon_g\kappa_1^2\Delta_f}{\Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell(1+V)}.
\tag{233}
$$

The optimal λ is therefore the solution on the λ variable of the system of equations

$$
\begin{cases}
\lambda = \frac{\hat{V}\epsilon_g\kappa_1^2\Delta_f}{\alpha\Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell} \\
\epsilon_g = \epsilon_r + \frac{\Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell\tilde{\lambda}^2 + \frac{\kappa_1^4\Delta_f^2\epsilon_g}{\alpha}}{(\tilde{\lambda}+\kappa_1^2\Delta_f)^2} \\
\frac{1}{\hat{V}} = \frac{1}{\alpha}\left(1 + \frac{\kappa_*^2}{\lambda} + \frac{\kappa_1^2\Delta_f}{\lambda+\hat{V}\kappa_1^2\Delta_f}\right)
\end{cases}.
\tag{234}
$$

Introducing the variable $\tilde{\lambda}\equiv\lambda/\hat{V}$, and plugging the first line into the second, this simplifies to

$$
\begin{cases}
\epsilon_g = \alpha\frac{\Delta_a\prod\limits_{\ell=1}^{L}\Delta_\ell\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2}{\Delta_f}\frac{\tilde{\lambda}}{\kappa_1^2} \\
\tilde{\lambda} = \frac{\Delta_f\kappa_1^2}{\Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell}\frac{1}{\alpha}\left(\epsilon_r + \frac{\Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell\tilde{\lambda}}{\tilde{\lambda}+\kappa_1^2\Delta_f}\right) \\
\tilde{\lambda} = \frac{1}{\alpha}\left(\lambda + \kappa_*^2 + \frac{\kappa_1^2\Delta_f\tilde{\lambda}}{\tilde{\lambda}+\kappa_1^2\Delta_f}\right)
\end{cases}.
\tag{235}
$$

Substracting the third line from the second yields an expression for λ:

$$
\frac{\Delta_f\kappa_1^2}{\Delta_a\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell}\epsilon_r = \lambda + \kappa_*^2.
\tag{236}
$$

In other words the optimal regularization is explicitly given by

$$\lambda = \kappa_1^2 \Delta_f \left(\frac{\epsilon_r}{\Delta_a \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell} - \frac{\kappa_*^2}{\kappa_1^2 \Delta_f} \right) \tag{237}$$

corresponding to (25).

Equivalence GP kernel/BO   Under optimal regularization (25) the saddle-point equations for the Bayes-optimal MSE can be recovered:

$$-\frac{q - 2\prod\limits_{\ell=1}^{L} \kappa_1^{(\ell)} m}{\prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2} = -\frac{1}{\prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2} \frac{\Delta_a \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell \kappa_1^4 \Delta_f^2 \hat{V}^2 + \kappa_1^3 \Delta_f^2 \frac{\hat{V}^2}{\alpha} \epsilon_g - 2\Delta_a \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell \kappa_1^2 \Delta_f \hat{V} \lambda - 2\Delta_a \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell \kappa_1^4 \Delta_f^2 \hat{V}^2}{(\lambda + \hat{V}\kappa_1 \Delta_f)^2}$$

$$= \frac{\Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell (\Delta_f^2 \kappa_1^4 \hat{V}^2 + 2\kappa_1^2 \Delta_f \lambda \hat{V}) - \frac{\kappa_1^4 \Delta_f^2}{\prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2} \frac{\hat{V}}{\alpha} \epsilon_g}{(\lambda + \hat{V}\kappa_1 \Delta_f)^2}$$

$$= \frac{\Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell \frac{\alpha \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2}{\epsilon_g}}{1 + \Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell \frac{\alpha \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2}{\epsilon_g}} = \frac{\Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell \alpha \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2}{\epsilon_g + \Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell \alpha \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2} \tag{238}$$

which recovers (116).

H. Classification

H.1. Bayes-optimal error for classification

In this section, we treat the case of classification, corresponding to a sign readout $f_\star(x) = \mathrm{sign}(x)$ in (1). Since the finite temperature free energy of Appendix C is derived under no assumption over the output channel, it is amenable to being readily specialized to the present setting. The output partition function reads

$$\mathcal{Z}(y, \omega, V) = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{\omega}{\sqrt{2(V+\Delta)}}\right)\right]. \tag{239}$$

The saddle point equations can be massaged into the compact form

$$\begin{cases} q = \int \frac{\hat{q}\Delta_a^2 \prod\limits_{\ell=1}^{L} \Delta_\ell^2 z^2}{\hat{q}z\Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell + 1} \mathrm{d}\mu(z) \\[2em] \hat{q} = \frac{2\alpha \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2}{\Delta_a \int z\mathrm{d}\mu(z) \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell + \epsilon_r - \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 q} \\[2em] \int \frac{d\xi}{(2\pi)^{\frac{3}{2}}} \frac{2e^{-\frac{1}{2}\frac{\Delta_a \int z\mathrm{d}\mu(z) \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell + \epsilon_r + \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 q}{\Delta_a \int z\mathrm{d}\mu(z) \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell + \epsilon_r - \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 q} \xi^2}}{1 - \mathrm{erf}\left(\frac{\prod\limits_{\ell=1}^{L} \kappa_1^{(\ell)} \sqrt{q}\xi}{\sqrt{2\left(\Delta_a \int z\mathrm{d}\mu(z) \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell + \epsilon_r - \prod\limits_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 q\right)}}\right)} \end{cases}$$

The interested reader is referred to (Cui et al., 2019) for a full derivation in a closely related case.

H.2. Bayes optimal classification error

We detail here the derivation of the Bayes misclassification error (15). The Bayes optimal estimator for a test input x is (Opper & Haussler, 1991)

$$\text{sign}\left[\left\langle \text{sign}\left(a^\top h_L(x)\right)\right\rangle_{a,\{W_\ell\}_\ell \sim \mathbb{P}}\right] \tag{240}$$

where $h_L(x)$ is the post-activation of the student network and $\mathbb{P}$ is the Bayes posterior

$$\mathbb{P}(a, \{W_\ell\}) = \frac{1}{Z} e^{-\frac{||a||^2}{2\Delta_a} - \sum_\ell \frac{||W_\ell||_F^2}{2\Delta_\ell}} \prod_{\mu=1}^n \mathbb{E}_{\xi \sim \mathcal{N}(0,\Delta)}\left[\delta\left(y^\mu - \text{sign}(\hat{y}(x^\mu) + \xi)\right)\right].$$

Then

$$
\begin{aligned}
1 - \epsilon_g &= \mathbb{E}_{\mathcal{D},a_\star,\{W_\ell^\star\}_\ell}\mathbb{E}_x \Theta\left[\frac{1}{\sqrt{k_L}}a_\star^\top h_L^\star(x) \times \left\langle \text{sign}\left(\frac{1}{\sqrt{k_L}}a^\top h_L(x)\right)\right\rangle\right] \\
&= \mathbb{E}_{\mathcal{D},a_\star,\{W_\ell^\star\}_\ell}\mathbb{E}_x \int dv d\hat{v} e^{-iv\hat{v}} \Theta\left[\frac{1}{\sqrt{k_L}}a_\star^\top h_L^\star(x) \times v\right] e^{i\hat{v}\left\langle \text{sign}\left(\frac{1}{\sqrt{k_L}}a^\top h_L(x)\right)\right\rangle} \\
&= \mathbb{E}_{\mathcal{D},a_\star,\{W_\ell^\star\}_\ell}\mathbb{E}_x \int dv d\hat{v} e^{-iv\hat{v}} \Theta\left[\frac{1}{\sqrt{k_L}}a_\star^\top h_L^\star(x) \times v\right] \sum_{k=0}^\infty \frac{(i\hat{v})^k}{k!}\left\langle \text{sign}\left(\frac{1}{\sqrt{k_L}}a^\top h_L(x)\right)\right\rangle^k \\
&= \int dv d\hat{v} e^{-iv\hat{v}} \sum_{k=0}^\infty \frac{(i\hat{v})^k}{k!}\mathbb{E}_x \mathbb{E}_{\mathcal{D},a_\star,\{W_\ell^\star\}_\ell}\Theta\left[\frac{1}{\sqrt{k_L}}a_\star^\top h_L^\star(x) \times v\right]\left\langle \text{sign}\left(\frac{1}{\sqrt{k_L}}a^\top h_L(x)\right)\right\rangle^k.
\end{aligned} \tag{241}
$$

The computation of the bracketed average can be carried out using the replica trick, observing

$$
\begin{aligned}
\left\langle \text{sign}\left(\frac{1}{\sqrt{k_L}}a^\top h_L(x)\right)\right\rangle &= \mathbb{E}_{\mathcal{D},a_\star,\{W_\ell^\star\}_\ell}\lim_{s\to 0} Z^{s-1}\int da P_a(a)\prod_\ell dW_\ell P_w(W_\ell)\prod_{\mu=1}^n P_{\text{out}}(y^\mu|\hat{y}(x^\mu))\text{sign}\left(\hat{y}(x)\right) \\
&= \mathbb{E}_{\mathcal{D},a_\star,\{W_\ell^\star\}_\ell}\int \prod_{a=1}^s da^a P_a(a^a)\prod_{a=1}^s\prod_\ell dW_\ell^a P_w(W_\ell^a)\prod_{\mu=1}^n\prod_{a=1}^s P_{\text{out}}(y^\mu|\hat{y}^a(x^\mu))\text{sign}\left(\hat{y}^1(x)\right)
\end{aligned} \tag{242}
$$

where $\hat{y}^1$ is the first replica of the student (2), with weights $a^a, \{W_\ell^a\}$. It follows that

$$
\begin{aligned}
\left\langle \text{sign}\left(\frac{1}{\sqrt{k_L}}a^\top h_L(x)\right)\right\rangle^k &= \mathbb{E}_{\mathcal{D},a_\star,\{W_\ell^\star\}_\ell}\int \prod_{a=1}^s\prod_{\alpha=1}^k da^{a\alpha} P_a(a^{a\alpha})\prod_{a=1}^s\prod_{\alpha=1}^k\prod_\ell dW_\ell^{a\alpha} P_w(W_\ell^{a\alpha}) \\
&\qquad \prod_{\mu=1}^n\prod_{\alpha=1}^k \text{sign}\left(\hat{y}^{1\alpha}(x)\right)\prod_{a=1}^s P_{\text{out}}(y^\mu|\hat{y}^{a\alpha}(x^\mu)),
\end{aligned} \tag{243}
$$

with the power k effectively introducing a second level of replication.

$$
\begin{aligned}
\mathbb{E}_x \mathbb{E}_{a_\star,\{W_\ell^\star\}_\ell}\Theta\left[\frac{1}{\sqrt{k_L}}a_\star^\top h_L^\star(x) \times v\right]&\left\langle \text{sign}\left(\frac{1}{\sqrt{k_L}}a^\top h_L(x)\right)\right\rangle^k \\
&= \mathbb{E}_{\mathcal{D},a_\star,\{W_\ell^\star\}_\ell}\int \prod_{a=1}^s\prod_{\alpha=1}^k da^{a\alpha} P_a(a^{a\alpha})\prod_{a=1}^s\prod_{\alpha=1}^k\prod_\ell dW_\ell^{a\alpha} P_w(W_\ell^{a\alpha}) \\
&\quad \mathbb{E}_{\{x^\mu\}_{\mu=1}^n}\prod_{\mu=1}^n\prod_{a=1}^s\prod_{\alpha=1}^k P_{\text{out}}(y^\mu|\hat{y}^{a\alpha}(x^\mu))\mathbb{E}_x \underbrace{\prod_{\alpha=1}^k \text{sign}\left(\hat{y}^{1\alpha}(x)\right)\Theta(v \times y(x))}_{(a)}.
\end{aligned} \tag{244}
$$

Following the discussion in Appendix A (see also Section 2 in the main text), and under the assumption that $\hat{y}^{1\alpha}(x), y(x)$ may be as a first approximation treated as Gaussian variables, it follows from the linearization (10) that $\hat{y}^{1\alpha}(x), y(x)$ can be replaced inside the expectation $\mathbb{E}_x$ in (a) by their equivalent linear models,

$$
\hat{y}_{\mathrm{lin}}^{1\alpha}(x) := \frac{1}{\sqrt{k_L}} a_{1\alpha}^\top \left[\kappa_*^{(L)} \xi_L^\alpha + \sum_{\ell=2}^{L-1} \kappa_*^{(\ell-1)} \left(\prod_{\ell'=L}^{\ell} \kappa_1^{(\ell')} \frac{W_{\ell'}^{1\alpha}}{\sqrt{k_{\ell'-1}}} \right) \xi_{\ell-1}^\alpha + \left(\prod_{\ell=L}^{1} \kappa_1^{(\ell)} \frac{W_\ell^{1\alpha}}{\sqrt{k_{\ell-1}}} \right) x \right]
$$

$$
y_{\mathrm{lin}}(x) := \frac{1}{\sqrt{k_L}} a_\star^\top \left[\kappa_*^{(L)} \xi_L^\star + \sum_{\ell=2}^{L-1} \kappa_*^{(\ell-1)} \left(\prod_{\ell'=L}^{\ell} \kappa_1^{(\ell')} \frac{W_{\ell'}^\star}{\sqrt{k_{\ell'-1}}} \right) \xi_{\ell-1}^\star + \left(\prod_{\ell=L}^{1} \kappa_1^{(\ell)} \frac{W_\ell^\star}{\sqrt{k_{\ell-1}}} \right) x \right] + \sqrt{\Delta} \xi \tag{245}
$$

where $\xi_\ell^{\star/\alpha} \sim \mathcal{N}(0, I_{k_\ell})$ are independently sampled Gaussian noise vectors. Note that the replication procedure leading to the introduction of the second replication level, indexed by α, means that the ξ^α are also mutually independent. Finally, one can write

$$
(a) = \mathbb{E}_{x, \{\xi_\ell^\alpha\}_{\ell,\alpha}, \{\xi_\ell^\star\}_\ell} \prod_{\alpha=1}^{k} \mathrm{sign}\left(\hat{y}_{\mathrm{lin}}^{1\alpha}(x) \right) \Theta(v \times y_{\mathrm{lin}}(x)). \tag{246}
$$

The full equation for the test error is then

$$
1 - \epsilon_g = \mathbb{E}_{x, \{\xi_\ell^\alpha\}_{\ell,\alpha}, \{\xi_\ell^\star\}_\ell, \xi} \int \mathrm{d}v \mathrm{d}\hat{v} e^{-iv\hat{v}} \sum_{k=0}^{\infty} \frac{(i\hat{v})^k}{k!} \Theta\left[y_{\mathrm{lin}}(x) \times v \right] \underbrace{\mathbb{E}_{\mathcal{D}, a_\star, \{W_\ell^\star\}_\ell} \left\langle \mathrm{sign}\left(\hat{y}_{\mathrm{lin}}(x) \right) \right\rangle^k}_{(b)}. \tag{247}
$$

We henceforth focus on computing (b). Resuming the replica computation,

$$
(b) = \int \prod_{a=1}^{s} \prod_{\alpha=1}^{k} \mathrm{d}a^{a\alpha} P_a(a^{a\alpha}) \prod_{a=1}^{s} \prod_{\alpha=1}^{k} \prod_{\ell}^{L} \mathrm{d}W_\ell^{a\alpha} P_w(W_\ell^{a\alpha})
$$

$$
\prod_{\mu=1}^{n} \prod_{\alpha=1}^{k} \mathrm{sign}\left(\hat{y}_{\mathrm{lin}}^{1\alpha}(x) \right) \mathbb{E}_{\mathcal{D}, a_\star, \{W_\ell^\star\}_\ell} \prod_{a=1}^{s} \prod_{\alpha=1}^{k} P_{\mathrm{out}}(y^\mu | \hat{y}_{\mathrm{lin}}^{a\alpha}(x^\mu)). \tag{248}
$$

Define the local fields

$$
z_a^\alpha := \frac{1}{\sqrt{k_L}} a_{1\alpha}^\top \xi_L^\alpha,
$$

$$
z_\ell^\alpha := \frac{1}{\sqrt{k_L}} a_{1\alpha}^\top \left(\prod_{\ell'=L}^{\ell} \frac{W_{\ell'}^{1\alpha}}{\sqrt{k_{\ell'-1}}} \right) \xi_{\ell-1}^\alpha, \qquad \forall 2 \le \ell \le L
$$

$$
z_1^\alpha := \frac{1}{\sqrt{k_L}} a_{1\alpha}^\top \left(\prod_{\ell=L}^{1} \frac{W_\ell^{1\alpha}}{\sqrt{k_{\ell-1}}} \right) x, \tag{249}
$$

so that

$$
\hat{y}_{\mathrm{lin}}^{1\alpha} = \kappa_*^{(L)} z_a^\alpha + \sum_{\ell=2}^{L-1} \kappa_*^{(\ell-1)} \prod_{\ell'=\ell}^{L} \kappa_1^{(\ell)} z_\ell^\alpha + \prod_{\ell=1}^{L} \kappa_1^{(\ell)} z_1^\alpha. \tag{250}
$$

Enforcing the definitions of the local fields using the Fourier representation of Dirac peaks, (b) becomes

$$
(b) = \int \prod_{\alpha=1}^{k} \mathrm{d}z_1^\alpha \mathrm{d}\hat{z}_1^\alpha e^{-z_1^\alpha \hat{z}_1^\alpha} \left(\prod_{\ell=2}^{L} \mathrm{d}z_\ell^\alpha \mathrm{d}\hat{z}_\ell^\alpha e^{-z_\ell^\alpha \hat{z}_\ell^\alpha} \right) \mathrm{d}z_a^\alpha \mathrm{d}\hat{z}_a^\alpha e^{-z_a^\alpha \hat{z}_a^\alpha} \int \prod_{a=1}^{s} \prod_{\alpha=1}^{k} \mathrm{d}a^{a\alpha} P_a(a^{a\alpha}) \prod_{a=1}^{s} \prod_{\alpha=1}^{k} \prod_{\ell}^{L} \mathrm{d}W_\ell^{a\alpha} P_w(W_\ell^{a\alpha})
$$

$$
\prod_{\alpha=1}^{k} e^{\hat{z}_a^\alpha \frac{1}{\sqrt{k_L}} a_{1\alpha}^\top \xi_L^\alpha + \sum_{\ell=2}^{L} \hat{z}_\ell^\alpha \frac{1}{\sqrt{k_L}} a_{1\alpha}^\top \left(\prod_{\ell'=L}^{\ell} \frac{W_{\ell'}^{1\alpha}}{\sqrt{k_{\ell'-1}}} \right) \xi_{\ell-1}^\alpha + \hat{z}_1^\alpha \frac{1}{\sqrt{k_L}} a_{1\alpha}^\top \left(\prod_{\ell=L}^{1} \frac{W_\ell^{1\alpha}}{\sqrt{k_{\ell-1}}} \right) x}
$$

$$
\prod_{\mu=1}^{n} \prod_{\alpha=1}^{k} \mathrm{sign}\left(\hat{y}_{\mathrm{lin}}^{1\alpha}(x) \right) \mathbb{E}_{\mathcal{D}, a_\star, \{W_\ell^\star\}_\ell} \prod_{a=1}^{s} \prod_{\alpha=1}^{k} P_{\mathrm{out}}(y^\mu | \hat{y}_{\mathrm{lin}}^{a\alpha}(x^\mu)). \tag{251}
$$

49

The computation for the trace and energy potentials $\Psi_{t,y}$ follows the same steps as detailed in Appendix C, leading to the same contributions to the multi-layer free energy. On the other hand, the entropy potential Ψ_w is changed, as it now includes the second line. As in Appendix C, we decompose Ψ_w in input, hidden, and readout layer contributions and detail the derivation for each.

Input layer . We detail the computation of the entropy contribution at the input layer $\tilde{\Psi}_1$, where the tilde distinguishes the $\hat{z}$ dependent potential from the one computed in Appendix C. Note crucially that in particular $\tilde{\Psi}_1$ has a part which is not proportional to the first replica index s, and therefore contributes non-vanishingly in the $s \to 0$ limit.

$$
e^{\tilde{\Psi}_1} = \prod_{j=1}^{d} \int \prod_{a=1}^{s} \prod_{\alpha=1}^{k} dw_{a\alpha}^1 P_w(w_{a\alpha}^1) e^{\sum_{a\alpha \leq b\beta} \sigma_j(\hat{q}_1^L)_{a\alpha,b\beta} \frac{\left(w_{a\alpha}^{1\top} \prod_{\ell=2}^{L} W_{a\alpha}^\ell a_{a\alpha}\right) \times \left(w_{b\beta}^{1\top} \prod_{\ell=2}^{L} W_{b\beta}^\ell a_{b\beta}\right)}{\prod_{\ell=1}^{L} k_\ell}}
$$

$$
\times e^{+\sum_{a\alpha} \sigma_j \hat{m} \frac{\left(w_{a\alpha}^{1\top} \prod_{\ell=2}^{L} W_{a\alpha}^\ell a_{a\alpha}\right) \times \left((w_\star^1)_j^\top \prod_{\ell=2}^{L^\star} W_\star^\ell a_\star\right)}{\prod_{\ell=1}^{L} \sqrt{k_\ell} \prod_{\ell=1}^{L^\star} \sqrt{k_\ell^\star}} + \sum_\alpha \hat{z}_1^\alpha \frac{1}{\sqrt{k_L}} a_{1\alpha}^\top \left(\prod_{\ell=L}^{2} \frac{W_\ell^{1\alpha}}{\sqrt{k_{\ell-1}}}\right) w_{1\alpha}^1 \frac{1}{\sqrt{d}} x_j}. \tag{252}
$$

Again, we assume replica symmetry with respect to the double replica, which together with the Nishimori conditions imply

$$
(\hat{q}_1^L)_{a\alpha,b\beta} = \delta_{ab}\delta_{\alpha\beta}\hat{q}, \tag{253}
$$

together with $\hat{m} = \hat{q}$. As in Appendix C, introduce the local fields

$$
\eta_{a\alpha} \equiv \frac{w_{a\alpha}^{1\top} \prod_{\ell=2}^{L} W_{a\alpha}^\ell a_{a\alpha}}{\prod_{\ell=1}^{L} \sqrt{k_\ell}}, \tag{254}
$$

with statistics

$$
\langle \eta_{a\alpha}\eta_{b\beta}\rangle = \delta_{ab}\delta_{\alpha\beta}\Delta_1\rho_2^L. \tag{255}
$$

Therefore

$$
e^{\tilde{\Psi}_1} = \prod_{j=1}^{d} \int \prod_{a\alpha} \frac{d\eta_{a\alpha}}{\sqrt{2\pi\Delta_1\rho_2^L}} e^{-\frac{1}{2\Delta_1\rho_2^L}\eta_{a\alpha}^2} e^{\sum_{a\alpha \leq b\beta}\sigma_j(\hat{q}_1^L)_{a\alpha,b\beta}\eta_{a\alpha}\eta_{b\beta} + \sum_{a\alpha}\sigma_j\hat{m}\eta_\alpha\eta_j^* + \sum_\alpha \hat{z}_1^\alpha \eta_{1\alpha}\frac{1}{\sqrt{d}}x_j}
$$

$$
= e^{sd\Psi_1} \times e^{\frac{1}{2}\sum_{j=1}^{d}\frac{x_j^2}{d}\frac{\Delta_1\rho_2^L+\hat{q}\sigma_j}{((\Delta_1\rho_2^L)^{-1}+\sigma_j\hat{q})^2}\sum_{\alpha=1}^{k}(\hat{z}_1^\alpha)^2 + \frac{1}{2}\sum_{j=1}^{d}\frac{x_j^2}{d}\frac{\hat{q}\sigma_j}{((\Delta_1\rho_2^L)^{-1}+\sigma_j\hat{q})^2}\sum_{\alpha,\beta}\hat{z}_1^\alpha\hat{z}_1^\beta + \sum_{j=1}^{d}\sigma_j\hat{q}\eta_j^*\frac{x_j}{\sqrt{d}}\frac{1}{\Delta_1\rho_2^L)^{-1}+\sigma_j\hat{q}}\sum_{\alpha=1}^{k}\hat{z}_1^\alpha}
$$

$$
= \int \frac{d\eta}{\sqrt{2\pi}} e^{-\frac{\eta^2}{2}} \prod_{\alpha=1}^{k} \left[e^{\frac{1}{2}\sum_{j=1}^{d}\frac{x_j^2}{d}\frac{\Delta_1\rho_2^L+\hat{q}\sigma_j}{((\Delta_1\rho_2^L)^{-1}+\sigma_j\hat{q})^2}(\hat{z}_1^\alpha)^2 + \sqrt{\sum_{j=1}^{d}\frac{x_j^2}{d}\frac{\hat{q}\sigma_j}{((\Delta_1\rho_2^L)^{-1}+\sigma_j\hat{q})^2}}\eta\hat{z}_1^\alpha + \sum_{j=1}^{d}\sigma_j\hat{q}\eta_j^*\frac{x_j}{\sqrt{d}}\frac{1}{\Delta_1\rho_2^L)^{-1}+\sigma_j\hat{q}}\hat{z}_1^\alpha} \right], \tag{256}
$$

where we introduced a Hubbard-Stratonovitch field η in the last line and took the $s \to 0$ limit.

Hidden layers In this paragraph, we turn to the middle layers $\ell \geq 2$.

$$
e^{\tilde{\Psi}_{\ell_0}} = \prod_{j=1}^{d} \left[\int \prod_{a\alpha} dw_{a\alpha} e^{-\frac{\Delta_{\ell_0}^{-1}}{2}\sum_{a\alpha}w_{a\alpha}^\top w_{a\alpha} - \frac{1}{2}\hat{r}_{\ell_0}^L \sum_{a\alpha}\left(\frac{w_{a\alpha}^\top \prod_{\ell=\ell_0+1}^{L} W_{a\alpha}^\ell a_{a\alpha}}{\prod_{\ell=\ell_0}^{L}\sqrt{k_\ell}}\right)^2 + \sum_{a\alpha}\hat{z}_{\ell_0}^\alpha \frac{w_{1\alpha}^\top \prod_{\ell=\ell_0+1}^{L} W_{1\alpha}^\ell a_{1\alpha}}{\prod_{\ell=\ell_0}^{L}\sqrt{k_\ell}}\frac{(\xi_{\ell_0-1}^\alpha)_j}{\sqrt{k_{\ell_0-1}}}} \right]. \tag{257}
$$

Introducing the local field

$$\eta_{a\alpha} := \frac{w_{a\alpha}^\top \prod\limits_{\ell=\ell_0+1}^{L} W_{a\alpha}^\ell a_{a\alpha}}{\prod\limits_{\ell=\ell_0}^{L} \sqrt{k_\ell}}, \tag{258}$$

with the statistics

$$\langle \eta_{a\alpha} \eta_{b\beta} \rangle = \delta_{ab} \delta_{\alpha\beta} \Delta_{\ell_0} \rho_{\ell_0+1}^L, \tag{259}$$

this becomes

$$e^{\tilde{\Psi}_{\ell_0}} = \prod_{j=1}^{d} \left[\int \prod_{a\alpha} \frac{d\eta_{a\alpha}}{\sqrt{2\pi \Delta_{\ell_0} \rho_{\ell_0+1}^L}} e^{-\frac{1}{2}\hat{r}_{\ell_0}^L \sum\limits_{a\alpha} \eta_{a\alpha}^2 + \sum\limits_{a\alpha} \hat{z}_{\ell_0}^\alpha \eta_{a\alpha} \frac{(\xi_{\ell_0}^\alpha - 1)_j}{\sqrt{k_{\ell_0-1}}}} \right]$$

$$= e^{sd\Psi_{\ell_0}} \times e^{\frac{1}{2} \sum\limits_{j=1}^{d} \Delta_{\ell_0} \rho_{\ell_0+1}^L \sum\limits_{\alpha=1}^{k} (\hat{z}_{\ell_0}^\alpha)^2 \frac{(\xi_{\ell_0}^\alpha - 1)_j^2}{k_{\ell_0-1}}}. \tag{260}$$

Readout layer By the same token, the new readout layer entropy potential reads

$$e^{\tilde{\Psi}_a} = e^{\frac{1}{2} \sum\limits_{j=1}^{d} \Delta_L \sum\limits_{\alpha=1}^{k} (\hat{z}_a^\alpha)^2 \frac{(\xi_L^\alpha)_j^2}{k_L}} \tag{261}$$

End of the computation Finally, putting all the pieces together, the computation of (b) follows as

$$(b) = \int \frac{d\eta}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta^2} \left[\int \prod_{\ell=1}^{L} d\hat{z}_\ell dz_\ell d\hat{z}_a dz_a e^{-\sum\limits_{\ell=1}^{L} z_\ell \hat{z}_\ell - \hat{z}_a z_a + \frac{1}{2}\Delta_a \frac{\sum\limits_{j=1}^{d}(\xi_L)_j^2}{k_L} \hat{z}_a^2 + \frac{1}{2} \sum\limits_{\ell=2}^{L} \Delta_\ell \rho_{\ell+1}^L \frac{\sum\limits_{j=1}^{d}(\xi_\ell)_j^2}{k_{\ell-1}} \hat{z}_\ell^2 + \frac{1}{2}\tilde{\Delta}_1(\hat{z}_1)^2 + \mu \hat{z}_1} \right.$$

$$\left. \times \operatorname{sign}\left(\kappa_*^{(L)} z_a + \sum_{\ell=2}^{L-1} \kappa_*^{(\ell-1)} \prod_{\ell'=\ell}^{L} \kappa_1^{(\ell)} z_\ell + \prod_{\ell=1}^{L} \kappa_1^{(\ell)} z_1 \right) \right]^k. \tag{262}$$

We introduced the shorthands

$$\begin{cases} \mu = \sqrt{\sum\limits_{j=1}^{d} \frac{x_j^2}{d} \frac{\hat{q}\sigma_j}{((\Delta_1 \rho_2^L)^{-1} + \sigma_j \hat{q})^2}} \eta + \sum\limits_{j=1}^{d} \sigma_j \hat{q} \eta_j^* \frac{x_j}{\sqrt{d}} \frac{1}{(\Delta_1 \rho_2^L)^{-1} + \sigma_j \hat{q}} \\ \tilde{\Delta}_1 = \sum\limits_{j=1}^{d} \frac{x_j^2}{d} \frac{\Delta_1 \rho_2^L + \hat{q}\sigma_j}{((\Delta_1 \rho_2^L)^{-1} + \sigma_j \hat{q})^2} \end{cases} \tag{263}$$

Note that the quantities $\sum\limits_{j=1}^{d} (\xi_L)_j^2 / k_L$ and $\sum\limits_{j=1}^{d} (\xi_\ell)_j^2 / k_{\ell-1}$ concentrate around their unit mean. Carrying out the $\hat{z}$ integrals:

$$(b) = \int \frac{d\eta}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta^2} \left[\int \frac{dz_a \prod\limits_{\ell} dz_\ell}{\sqrt{(2\pi)^{L+1} \Delta_a \tilde{\Delta}_1 \prod\limits_{\ell=2}^{L} \Delta_\ell \rho_{\ell+1}^L}} e^{-\frac{1}{2\Delta_a} z_a^2 - \frac{1}{2} \sum\limits_{\ell=2}^{L} \frac{z_\ell^2}{\Delta_\ell \rho_{\ell+1}^L} - \frac{1}{2\tilde{\Delta}_1}(z_1 - \mu)^2} \right.$$

$$\left. \times \operatorname{sign}\left(\kappa_*^{(L)} z_a + \sum_{\ell=2}^{L-1} \kappa_*^{(\ell-1)} \prod_{\ell'=\ell}^{L} \kappa_1^{(\ell)} z_\ell + \prod_{\ell=1}^{L} \kappa_1^{(\ell)} z_1 \right) \right]^k$$

$$= \int \frac{d\eta}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta^2} \left[\int \frac{dz}{\sqrt{2\pi\Delta_z}} e^{-\frac{1}{2\Delta_z}\left(z - \mu \prod\limits_{\ell=1}^{L} \kappa_1^{(\ell)} \right)^2} \times \operatorname{sign}(z) \right]^k$$

$$= \int \frac{d\eta}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta^2} \operatorname{erf}\left(\frac{\mu \prod\limits_{\ell=1}^{L} \kappa_1^{(\ell)}}{\sqrt{2\Delta_z}} \right)^k \tag{264}$$

We introduced

$$\Delta_z = \left(\kappa_*^{(L)}\right)^2 \Delta_a + \sum_{\ell=2}^{L-1} \left(\kappa_*^{(\ell-1)} \prod_{\ell'=\ell}^{L} \kappa_1^{(\ell)}\right)^2 \Delta_\ell \rho_{\ell+1}^L + \prod_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \tilde{\Delta}_1. \tag{265}$$

This concludes the computation of (b). Returning to the computation of the classification error:

$$1 - \epsilon_g = \mathbb{E}_{x,\{\xi_\ell^\alpha\}_{\ell,\alpha},\{\xi_\ell^\star\}_{\ell},\xi} \int \frac{d\eta}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta^2} \int \mathrm{d}v \mathrm{d}\hat{v} e^{-iv\hat{v}} \sum_{k=0}^{\infty} \frac{(i\hat{v})^k}{k!} \Theta\left[y_{\mathrm{lin}}(x) \times v\right] \mathrm{erf}\left(\frac{\mu \prod_{\ell=1}^{L} \kappa_1^{(\ell)}}{\sqrt{2\Delta_z}}\right)^k$$

$$= \mathbb{E}_{x,\{\xi_\ell^\alpha\}_{\ell,\alpha},\{\xi_\ell^\star\}_{\ell},\xi} \int \frac{d\eta}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta^2} \Theta\left[y_{\mathrm{lin}}(x) \times \mathrm{erf}\left(\frac{\mu \prod_{\ell=1}^{L} \kappa_1^{(\ell)}}{\sqrt{2\Delta_z}}\right)\right]$$

$$= \mathbb{E}_{x,\{\xi_\ell^\alpha\}_{\ell,\alpha},\{\xi_\ell^\star\}_{\ell},\xi} \int \frac{d\eta}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta^2} \Theta\left[y_{\mathrm{lin}}(x) \times \mu\right]$$

$$= 1 - \frac{1}{\pi} \arccos\left(\frac{\mathbb{E}_{x,\{\xi_\ell^\alpha\}_{\ell,\alpha},\{\xi_\ell^\star\}_{\ell},\xi,\eta}[y_{\mathrm{lin}}(x) \times \mu]}{\sqrt{\mathbb{E}_{x,\{\xi_\ell^\alpha\}_{\ell,\alpha},\{\xi_\ell^\star\}_{\ell},\xi,\eta}[y_{\mathrm{lin}}(x)^2]} \times \sqrt{\mathbb{E}_{x,\{\xi_\ell^\alpha\}_{\ell,\alpha},\{\xi_\ell^\star\}_{\ell},\xi,\eta}[\mu^2]}}\right). \tag{266}$$

The average squared label can be computed as

$$\mathbb{E}_{x,\{\xi_\ell^\alpha\}_{\ell,\alpha},\{\xi_\ell^\star\}_{\ell},\xi,\eta}[y_{\mathrm{lin}}(x)^2]$$

$$= \frac{1}{k_L}\left(\kappa_*^{(L)}\right)^2 a_\star^\top a_\star + \sum_{\ell=2}^{L-1} \left(\kappa_*^{(\ell-1)}\right)^2 \frac{1}{k_L} a_\star^\top \left(\prod_{\ell'=L}^{\ell} \kappa_1^{(\ell')} \frac{W_{\ell'}^\star}{\sqrt{k_{\ell'-1}}}\right) \left(\prod_{\ell'=L}^{\ell} \kappa_1^{(\ell')} \frac{W_{\ell'}^\star}{\sqrt{k_{\ell'-1}}}\right)^\top a_\star$$

$$+ \frac{1}{k_L} a_\star^\top \left(\prod_{\ell=L}^{1} \kappa_1^{(\ell)} \frac{W_\ell^\star}{\sqrt{k_{\ell-1}}}\right) \Sigma \left(\prod_{\ell=L}^{1} \kappa_1^{(\ell)} \frac{W_\ell^\star}{\sqrt{k_{\ell-1}}}\right)^\top a_\star + \Delta$$

$$= \left(\kappa_*^{(L)}\right)^2 \Delta_a + \sum_{\ell=2}^{L-1} \left(\kappa_*^{(\ell-1)}\right)^2 \Delta_a \left(\prod_{\ell'=L}^{\ell} \left(\kappa_1^{(\ell')}\right)^2 \Delta_{\ell'}\right) + \Delta_a \left(\prod_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell\right) \frac{1}{d} \mathrm{Tr}\,\Sigma + \Delta$$

$$= \left(\kappa_*^{(L)}\right)^2 \Delta_a + \sum_{\ell=2}^{L-1} \left(\kappa_*^{(\ell-1)}\right)^2 \Delta_a \left(\prod_{\ell'=L}^{\ell} \left(\kappa_1^{(\ell')}\right)^2 \Delta_{\ell'}\right) + \Delta_a \left(\prod_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell\right) \int d\mu(z)z + \Delta, \tag{267}$$

where we used the fact that the scalar products in the first line all concentrate asymptotically, leveraging the assumption on Σ. The teacher/student correlation term can be computed as

$$\mathbb{E}_{x,\{\xi_\ell^\alpha\}_{\ell,\alpha},\{\xi_\ell^\star\}_{\ell},\xi,\eta}[y_{\mathrm{lin}}(x) \times \mu] = \prod_{\ell=1}^{L} \kappa_1^{(\ell)} \eta^{*\top} \frac{\hat{q}\Sigma}{(\Delta_1 \rho_2^L)^{-1} + \hat{q}\Sigma} \Sigma\eta^*$$

$$= \prod_{\ell=1}^{L} \kappa_1^{(\ell)} \mathrm{Tr}\left[\frac{\hat{q}\Sigma^2 \Delta_a^2 \prod_{\ell=1}^{L} \Delta_\ell^2}{1 + \hat{q}\Sigma\Delta_a \prod_{\ell=1}^{L} \Delta_\ell}\right]$$

$$= \prod_{\ell=1}^{L} \kappa_1^{(\ell)} \int d\mu(z) \frac{\hat{q}z^2 \Delta_a^2 \prod_{\ell=1}^{L} \Delta_\ell^2}{1 + \hat{q}z\Delta_a \prod_{\ell=1}^{L} \Delta_\ell}$$

$$= \prod_{\ell=1}^{L} \kappa_1^{(\ell)} \times q, \tag{268}$$

where we used the definition of η^* and the self-averaging of the scalar product in going from the first to the second line. Finally, the averaged squared student output is

$$
\mathbb{E}_{x,\{\xi_\ell^\alpha\}_{\ell,\alpha},\{\xi_\ell^\star\}_\ell,\xi,\eta}[\mu^2] = \hat{q}\,\mathrm{Tr}\left[\frac{\Sigma^2}{\left((\Delta_1\rho_2^L)^{-1}+\hat{q}\Sigma\right)^2}\right] + \hat{q}^2\,\mathrm{Tr}\left[\frac{\Sigma^3\Delta_a\prod\limits_{\ell=1}^{L}\Delta_\ell}{\left((\Delta_1\rho_2^L)^{-1}+\hat{q}\Sigma\right)^2}\right]
$$

$$
= \mathrm{Tr}\left[\frac{\hat{q}\Sigma^2\Delta_a^2\prod\limits_{\ell=1}^{L}\Delta_\ell^2}{1+\hat{q}\Delta_a\prod\limits_{\ell=1}^{L}\Delta_\ell\Sigma}\right] = \int d\mu(z)\frac{\hat{q}z^2\Delta_a^2\prod\limits_{\ell=1}^{L}\Delta_\ell^2}{1+\hat{q}z\Delta_a\prod\limits_{\ell=1}^{L}\Delta_\ell} = q. \tag{269}
$$

Putting these contributions together, one finally reaches

$$
\epsilon_g = \frac{1}{\pi}\arccos\left(\frac{\prod\limits_{\ell=1}^{L}\kappa_1^{(\ell)}\sqrt{q}}{\sqrt{\left(\kappa_*^{(L)}\right)^2\Delta_a + \sum\limits_{\ell=2}^{L-1}\left(\kappa_*^{(\ell-1)}\right)^2\Delta_a\left(\prod\limits_{\ell'=L}^{\ell}\left(\kappa_1^{(\ell')}\right)^2\Delta_{\ell'}\right) + \Delta_a\left(\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell\right)\int d\mu(z)z + \Delta}}\right)
\tag{270}
$$

which is (15).

H.3. ERM

In full similarity to the regression case discussed in Appendix E, sharp asymptotics of the classification error can be obtained by leveraging the covariance formulae (10) in the framework of (Loureiro et al., 2021), using in their notations

$$
\begin{cases}
\rho = \prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\rho_1^L + \sum\limits_{\ell_0=1}^{L-1}\kappa_*^{(\ell_0)}\prod\limits_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2\rho_{\ell_0+1}^L + \left(\kappa_*^L\right)^2\rho_a, \\
\Omega = \Sigma \\
\Phi = \Delta_a^{\frac{1}{2}}\prod\limits_{\ell=1}^{L}\Delta_\ell^{\frac{1}{2}}\kappa_1^{(\ell)}\Sigma
\end{cases}.
$$

We recall that we consider, for classification, the noiseless case $\Delta = 0$. The test error is given by

$$
\epsilon_g = \frac{1}{\pi}\arccos\left(\frac{m\prod\limits_{\ell=1}^{L}\kappa_1^{(\ell)}}{\sqrt{\left(\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\rho_1^L + \sum\limits_{\ell_0=1}^{L-1}\kappa_*^{(\ell_0)}\prod\limits_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2\rho_{\ell_0+1}^L + \left(\kappa_*^L\right)^2\rho_a\right)q}}\right), \tag{271}
$$

with ρ and q being given by a risk-dependent system of equations. We examine in succession ridge classification and logistic regression.

Ridge classification The saddle point equations which need to be solved in order to access the asymptotic limit of the test error for ridge classification then read

$$
\begin{cases}
\hat{V} = \frac{\alpha}{1+V} \\
\hat{q} = \alpha\frac{1+q-2\prod\limits_{\ell=1}^{L}\kappa_1^{(\ell)}\sqrt{\frac{2}{\pi\left(\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\rho_1^L+\sum\limits_{\ell_0=1}^{L-1}\kappa_*^{(\ell_0)}\prod\limits_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2\rho_{\ell_0+1}^L+\left(\kappa_*^L\right)^2\rho_a\right)}}m+\Delta}{(1+V)^2} \\
\hat{m} = \sqrt{\frac{2}{\pi\left(\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\rho_1^L+\sum\limits_{\ell_0=1}^{L-1}\kappa_*^{(\ell_0)}\prod\limits_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2\rho_{\ell_0+1}^L+\left(\kappa_*^L\right)^2\rho_a\right)}}\frac{\prod\limits_{\ell=1}^{L}\kappa_1^{(\ell)}\alpha}{1+V}
\end{cases}
\qquad
\begin{cases}
V = \int d\mu(z)\frac{z}{\lambda I_d+\hat{V}z} \\
q = \int d\mu(z)\frac{\Delta_a\prod\limits_{\ell=1}^{L}\Delta_\ell\hat{m}^2z^3+\hat{q}z^2}{(\lambda I_d+\hat{V}z)^2} \\
m = \Delta_a\prod\limits_{\ell=1}^{L}\Delta_\ell\hat{m}\int d\mu(z)\frac{z^2}{\lambda I_d+\hat{V}z}
\end{cases}.
\tag{272}
$$

Logistic regression By the same token, introducing following (Loureiro et al., 2021) the auxiliary functions

$$Z(y, \omega, V) := \frac{1}{2}\left(1 + \text{erf}\left(\frac{y\omega}{\sqrt{2V}}\right)\right)$$

and $f(y, \omega, V)$ defined as the solution of

$$f(y, \omega, V) = \frac{y}{1 + e^{y(Vf(y,\omega,V)+\omega)}}.$$

Using the abuse of notations

$$Z := Z\left(y, \frac{m\prod_{\ell=1}^{L}\kappa_1^{(\ell)}}{\sqrt{q}}\xi, \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\rho_1^L + \sum_{\ell_0=1}^{L-1}\kappa_*^{(\ell_0)}\prod_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2\rho_{\ell_0+1}^L + \left(\kappa_*^L\right)^2\rho_a - \frac{m^2\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2}{q}\right) \tag{273}$$

$$f := f(y, \sqrt{q}\xi, V) \tag{274}$$

the corresponding saddle point equations read

$$\begin{cases} \hat{V} = -\alpha\int\frac{d\xi e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}}\sum_{y=\pm 1}Z\partial_\omega f \\ \hat{q} = \alpha\int\frac{d\xi e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}}\sum_{y=\pm 1}Zf^2 \\ \hat{m} = \alpha\int\frac{d\xi e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}}\sum_{y=\pm 1}\partial_\omega Zf \end{cases} \qquad \begin{cases} V = \int d\mu(z)\frac{z}{\lambda I_d + \hat{V}z} \\ q = \int d\mu(z)\frac{\Delta_a\prod_{\ell=1}^{L}\Delta_\ell\hat{m}^2 z^3 + \hat{q}z^2}{(\lambda I_d + \hat{V}z)^2} \\ m = \Delta_a\prod_{\ell=1}^{L}\Delta_\ell\hat{m}\int d\mu(z)\frac{z^2}{\lambda I_d + \hat{V}z}. \end{cases} \tag{275}$$