Deterministic equivalent and error universality of deep random features learning

Dominik Schröder, Hugo Cui, Daniil Dmitriev, Bruno Loureiro

Deterministic equivalent and error universality of deep random features learning

Dominik Schröder[1*], Hugo Cui[2*], Daniil Dmitriev[3], and Bruno Loureiro[4]

[1]Department of Mathematics, ETH Zurich, 8006 Zürich, Switzerland
[2]Statistical Physics Of Computation lab., Institute of Physics, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
[3]Department of Mathematics, ETH Zurich and ETH AI Center, 8092 Zürich, Switzerland
[4]Département d'Informatique, École Normale Supérieure (ENS) - PSL & CNRS, F-75230 Paris cedex 05, France
dschroeder@ethz.ch, hugo.cui@epfl.ch, daniil.dmitriev@ai.ethz.ch, bruno.loureiro@di.ens.fr
[*]*Main contributions*

February 2, 2023

Abstract

This manuscript considers the problem of learning a random Gaussian network function using a fully connected network with frozen intermediate layers and trainable readout layer. This problem can be seen as a natural generalization of the widely studied random features model to deeper architectures. First, we prove Gaussian universality of the test error in a ridge regression setting where the learner and target networks share the same intermediate layers, and provide a sharp asymptotic formula for it. Establishing this result requires proving a deterministic equivalent for traces of the deep random features sample covariance matrices which can be of independent interest. Second, we conjecture the asymptotic Gaussian universality of the test error in the more general setting of arbitrary convex losses and generic learner/target architectures. We provide extensive numerical evidence for this conjecture. In light of our results, we investigate the interplay between architecture design and implicit regularization.

1 Introduction

Despite the incredible practical progress in the applications of deep neural networks to almost all fields of knowledge, our current theoretical understanding thereof is still to a large extent incomplete. Recent progress on the theoretical front stemmed from the investigation of simplified settings, which despite their limitations are often able to capture some of the key properties of "real life" neural networks. A notable example is the recent stream of works on random features (RFs), originally introduced by [1] as a computationally efficient approximation technique for kernel methods, but more recently studied as a surrogate model for two-layers neural networks in the lazy regime [2–5]. RFs are a particular instance of random neural networks, whose statistical properties have been investigated in a sizeable body of works [6–10]. The problem of training the readout layer of such networks has been addressed in the shallow (one hidden layer) case by [4, 5], who provide sharp asymptotic characterizations for the test error. A similar study in the generic deep case is, however, still missing. In this manuscript, we bridge this gap by considering the problem of learning the last layer of a deep, fully-connected random neural network, hereafter referred to as the *deep random features* (dRF) model. More precisely, our **main contributions** in this manuscript are:

- In Section 3, we state Theorem 3.3, which proves an asymptotic deterministic equivalent for the traces of the product of deterministic matrices with both conjugate kernel and sample covariance matrix of the layer-wise post-activations.
- As a consequence of Thm. 3.3, in Section 4 we derive a sharp asymptotic formula for the test error of the dRF model in the particular case where the target and learner networks share the same intermediate layers, and when the readout layer is trained with the squared loss. This result establishes the Gaussian equivalence of the test error for ridge regression in this setting.
- Finally, we conjecture (and provide strong numerical evidence for) the Gaussian universality of the dRF model for general convex losses, and generic target/learner network architectures. More specifically, we provide exact

asymptotic formulas for the test error that leverage recent progress in high-dimensional statistics [11] and a closed-form formula for the population covariance of network activations appearing in [12]. These formulas show that in terms of second-order statistics, the dRF is equivalent to a linear network with noisy layers. We discuss how this effective noise translates into a depth-induced implicit regularization in Section 5.

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

Related work

Random features were first introduced by [1]. The asymptotic spectral density of the single-layer conjugate kernel was characterized in [3, 13, 14]. Sharp asymptotics for the test error of the RF model appeared in [4, 15] for ridge regression, [5, 16] for general convex losses and [17, 18] for other penalties. The implicit regularization of RFs was discussed in [19]. The RFs model has been studied in many different contexts as a proxy for understanding overparametrisation, e.g. in uncertainty quantification [20], ensembling [21, 22], the training dynamics [23, 24], but also to highlight the limitations of lazy training [25–28];

Deep random networks were shown to converge to Gaussian processes in [6, 7]. They were also studied in the context of inference in [29, 30], and as generative priors to inverse problems in [31–33]. The distribution of outputs of deep random nets was characterized in [9, 10]. Close to our work is [8], which provide exact formulas for the asymptotic spectral density and Stieltjes transform of the NTK and conjugate kernel in the proportional limit. Our formulas for the sample and population covariance are complementary to theirs. The test error of linear-width deep networks has been recently studied in [34, 35] through the lens of Bayesian learning;

Gaussian universality of the test error for the RFs model was shown in [4], conjectured to hold for general losses in [5] and was proven in [36, 37]. Gaussian universality has also been shown to hold for other classes of features, such as two-layer NTK [38], kernel features [19, 24, 39, 40]. [11] provided numerical evidence for Gaussian universality of more general feature maps, including pre-trained deep features.

Deterministic equivalents of sample covariance matrices have first been established in [41, 42] for separable covariances, generalizing the seminal work [43] on the free convolution of spectra in an anisotropic sense. More recently these results have been extended to non-separable covariances, first in tracial [44], and then also in anisotropic sense [45, 46].

2 Setting & preliminaries

Let $(\mathrm{x}^\mu, y^\mu) \in \mathbb{R}^d \times \mathcal{Y}$, $\mu \in [n] \coloneqq \{1, \cdots, n\}$, denote some training data, with $\mathrm{x}^\mu \sim \mathcal{N}(0_d, \Omega_0)$ independently and $y^\mu = f_\star(\mathrm{x}^\mu)$ a (potentially random) target function. This work is concerned with characterising the learning performance of generalised linear estimation:

$$\hat{y} = \sigma\left(\frac{\theta^\top \varphi(\mathrm{x})}{\sqrt{k}}\right), \tag{1}$$

with *deep random features* (dRF):

$$\varphi(\mathrm{x}) \coloneqq \underbrace{(\varphi_L \circ \varphi_{L-1} \circ \cdots \circ \varphi_2 \circ \varphi_1)}_{L}(\mathrm{x}), \tag{2}$$

where the post-activations are given by:

$$\varphi_\ell(h) = \sigma_\ell\left(\frac{1}{\sqrt{k_{\ell-1}}} W_\ell \cdot h\right), \quad \ell \in [L]. \tag{3}$$

The weights $\{W_\ell \in \mathbb{R}^{k_\ell \times k_{\ell-1}}\}_{\ell \in [L]}$ are assumed to be independently drawn Gaussian matrices with i.i.d. entries $(W_\ell)_{ij} \sim \mathcal{N}(0, \Delta_\ell)$ $\forall 1 \leq i \leq k_\ell$, $1 \leq j \leq k_{\ell-1}$. To alleviate notation, sometimes it will be convenient to denote $k_L = k$. Only the readout weights $\theta \in \mathbb{R}^k$ in (1) are trained according to the usual regularized *empirical risk minimization* procedure:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^k}{\operatorname{argmin}} \left[\sum_{\mu=1}^{n} \ell(y^\mu, \theta^\top \varphi(\mathrm{x}^\mu)) + \frac{\lambda}{2}||\theta||^2\right], \tag{4}$$

where $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ is a loss function, which we assume convex, and $\lambda > 0$ sets the regularization strength.

To assess the training and test performances of the empirical risk minimizer (4), we let $g : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ be any performance metric (e.g. the loss function itself or, in the case of classification, the probability of misclassifying), and define the test error:

$$\epsilon_g(\hat{\theta}) \coloneqq \mathbf{E}\left[g(y, \hat{\theta}^\top \varphi(\mathrm{x}))\right] \tag{5}$$

Our main goal in this work is to provide a sharp characterization of (5) in the proportional asymptotic regime $n, d, k_\ell \to \infty$ at fixed $\mathcal{O}(1)$ ratios $\alpha \coloneqq n/d$ and $\gamma_\ell \coloneqq k_\ell/d$, for all layer index $\ell \in [L]$. This requires a precise characterization of the *sample and population covariances* and the *Gram* matrices of the post-activations.

2.1 Background on sample covariance matrices

Marchenko-Pastur and free probability: We briefly introduce basic nomenclature on sample covariance matrices. For a random vector $x \in \mathbb{R}^d$ with mean zero $\mathbf{E}\, x = 0$ and covariance $\Sigma \coloneqq \mathbf{E}\, xx^\top \in \mathbb{R}^{d\times d}$, we call the matrix $\widehat{\Sigma} \coloneqq \mathcal{X}\mathcal{X}^\top/n \in \mathbb{R}^{d\times d}$ obtained from n independent copies $x_1, \dots, x_n$ of x written in matrix form as $\mathcal{X} \coloneqq (x_1, \dots, x_n)$ the *sample covariance matrix* corresponding to the *population covariance matrix* Σ. The *Gram matrix* $\check{\Sigma} \coloneqq \mathcal{X}^\top\mathcal{X}/n \in \mathbb{R}^{n\times n}$ has the same non-zero eigenvalues as the sample covariance matrix but unrelated eigenvectors. The systematic mathematical study of sample covariance and Gram matrices has a long history dating back to [47]. While in the "classical" statistical limit $n \to \infty$ with d being fixed the sample covariance matrix converges to the population covariance matrix $\widehat{\Sigma} \to \Sigma$, in the proportional regime $d \sim n \gg 1$ the non-trivial asymptotic relationship between the spectra of $\widehat{\Sigma}$ and Σ has first been obtained in the seminal paper [43]: the empirical spectral density $\mu(\widehat{\Sigma}) \coloneqq d^{-1}\sum_{\lambda\in\mathrm{Spec}(\widehat{\Sigma})}\delta_\lambda$ of $\widehat{\Sigma}$ is approximately equal to the *free multiplicative convolution* of $\mu(\Sigma)$ and a Marchenko-Pastur distribution μ_{MP}^c of aspect ratio $c = d/n$,

$$\mu(\widehat{\Sigma}) \approx \mu(\Sigma) \boxtimes \mu_{\mathrm{MP}}^{d/n}. \tag{6}$$

Here the free multiplicative convolution $\mu \boxtimes \mu_{\mathrm{MP}}^c$ may be defined as the unique distribution ν whose Stieltjes transform $m = m_\nu(z) \coloneqq \int (x-z)^{-1}\,\mathrm{d}\nu(x)$ satisfies the scalar *self-consistent equation*

$$zm = \frac{z}{1-c-czm} m_\mu\left(\frac{z}{1-c-czm}\right). \tag{7}$$

The spectral asymptotics (6) originally were obtained in the case of Gaussian $\mathcal{X}$ or, more generally, for separable correlations $\mathcal{X} = \sqrt{\Sigma}Y$ for some i.i.d. matrix $Y \in \mathbb{R}^{d\times n}$. These results were later extended [44] to the general case under essentially optimal assumptions on concentrations of quadratic forms $x^\top Ax$ around their expectation $\mathrm{Tr}\, A\Sigma$.

Deterministic equivalents: It has only been recognised much later [41, 42] that the relationship (6) between the asymptotic spectra of Σ and $\widehat{\Sigma}, \check{\Sigma}$ actually extends to eigenvectors as well, and that the resolvents $\widehat{G}(z) \coloneqq (\widehat{\Sigma} - z)^{-1}$, $\check{G}(z) \coloneqq (\check{\Sigma} - z)^{-1}$ are asymptotically equal to *deterministic equivalents*

$$\widehat{M}(z) \coloneqq -\frac{(\Sigma\check{m}(z) + I_d)^{-1}}{z}, \quad \check{M}(z) \coloneqq \check{m}(z) I_n, \tag{8}$$

also in an *anisotropic* rather than just a tracial sense, highlighting that despite the simple relationship between their averaged traces

$$\widehat{m}(z) \coloneqq m_{\mu(\Sigma)\boxtimes\mu_{\mathrm{MP}}^c}(z), \quad \check{m}(z) = \frac{c-1}{z} + c\widehat{m}(z),$$

the sample covariance and Gram matrices carry rather different non-spectral information. The anisoptric concentration of resolvents (or in physics terminology, the self-averaging) has again first been obtained in the Gaussian or separable cases [41, 42]. The extension to general sample covariance matrices was only achieved much more recently [45, 46] under Lipschitz concentration assumptions. In this work we specifically use the deterministic equivalent for sample covariance matrices with general covariance from [46] and extend it to cover Gram matrices.

Application to the deep random features model: In this work we apply the general theory of anisotropic deterministic equivalents to the deep random features model. As discussed in Section 4, to prove error universality even for the simple ridge regression case, it is not enough to only consider the spectral convergence of the matrices, and a stronger result is warranted. The application of non-linear activation functions makes the model neither Gaussian nor separable, hence our analysis relies on the deterministic equivalents from [46] and our extension to Gram matrices, which appear naturally in the explicit error derivations.

2.2 Notation

We will adopt the following notation:

- For $A \in \mathbb{R}^{n\times n}$ we denote $\langle A\rangle := {}^1\!/_n \operatorname{tr} A$.
- For matrices $A \in \mathbb{R}^{n\times m}$ we denote the operator norm (with respect to the ℓ^2-vector norm) by $\|A\|$, the max-norm by $\|A\|_{\max} := \max_{ij}|A_{ij}|$, and the Frobenius norm by $\|A\|_{\mathrm{F}}^2 := \sum_{ij}|A_{ij}|^2$.
- For any distribution μ we denote the push-forward under the map $\lambda \mapsto a\lambda + b$ by $a \otimes \mu \oplus b$ in order to avoid confusion with e.g. the convex combination $a\mu_1 + (1-a)\mu_2$ of measures μ_1, μ_2.
- We say that a sequence of random variables $(X_n)_n$ is *stochastically dominated* by another sequence $(Y_n)_n$ if for all small $\epsilon > 0$ and large $D < \infty$ it holds that $P(X_n > n^\epsilon Y_n) \le n^{-D}$ for large enough n, and in this case write $X_n \prec Y_n$.

3 Deterministic equivalents

Consider the sequence of variances defined by the recursion

$$r_{\ell+1} = \Delta_{\ell+1}\, \mathbf{E}_{\xi\sim\mathcal{N}(0,r_\ell)}\left[\sigma_\ell(\xi)^2\right] \tag{9}$$

with initial condition $r_1 := \Delta_1 {}^{\langle\Omega_0\rangle}\!/_d$ and coefficients

$$\begin{aligned}\kappa_1^\ell &= \frac{1}{r_\ell}\,\mathbf{E}_{\xi\sim\mathcal{N}(0,r_\ell)}\left[\xi\sigma_\ell(\xi)\right],\\ \kappa_*^\ell &= \sqrt{\mathbf{E}_{\xi\sim\mathcal{N}(0,r_\ell)}\left[\sigma_\ell(\xi)^2\right] - r_\ell\left(\kappa_1^\ell\right)^2}.\end{aligned} \tag{10}$$

3.1 Rigorous results on the multi-layer sample covariance and Gram matrices

Our main result on the anisotropic deterministic equivalent of dRFs follows from iterating the following proposition. We consider a data matrix $X_0 \in \mathbb{R}^{d\times n}$ whose Gram matrix concentrates as

$$\left\|\frac{X_0^\top X_0}{d} - r_1 I\right\|_{\max} \prec \frac{1}{\sqrt{n}}, \quad \left\|\frac{X_0}{\sqrt{d}}\right\| \prec 1 \tag{11}$$

for some positive constant r_1. The Assumption (11) for instance is satisfied if the columns x of X_0 are independent with mean $\mathbf{E}\,\mathrm{x} = 0$ and covariance $\mathbf{E}\,\mathrm{x}\,\mathrm{x}^\top = \Omega_0 \in \mathbb{R}^{d\times d}$ (together with some mild assumptions on the fourth moments), in which case $r_1 = \langle\Omega_0\rangle$ is the normalised trace of the covariance. We then consider $X_1 := \sigma_1(W_1X_0/\sqrt{d})$ assuming the entries of $W_1 \in \mathbb{R}^{k_1\times d}$ are iid. $\mathcal{N}(0,1)$ elements, and σ_1 satisfies $\mathbf{E}_{\xi\sim\mathcal{N}(0,1)}\,\sigma_1(\sqrt{r_1}\xi) = 0$ in the proportional $n \sim d \sim k_1$ regime. Upon changing σ_1 there is no loss in generality in assuming $\Delta_1 = 1$ which we do for notational convenience.

Proposition 3.1 (Deterministic equivalent for RF)**.** *For any deterministic A and Lipschitz-continuous activation function σ_1, under the assumptions above, we have that, for any $z \in \mathbf{C}\setminus\mathbb{R}_+$*

$$\left|\left\langle A\Big[\Big(\frac{X_1^\top X_1}{k_1} - z\Big)^{-1} - \widehat{M}(z)\Big]\right\rangle\right| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}},$$

and

$$\left|\left\langle A\Big(\frac{X_1X_1^\top}{k_1} - z\Big)^{-1}\right\rangle - \langle A\rangle\breve{m}(z)\right| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}},$$

where $\delta := \operatorname{dist}(z,\mathbb{R}_+)$,

$$\begin{aligned}-z\widehat{M}(z) &:= \Big(\breve{m}(z)\Sigma_{\mathrm{lin}} + I\Big)^{-1},\\ \Sigma_{\mathrm{lin}} &:= (\kappa_1^1)^2\frac{X_0^\top X_0}{d} + (\kappa_*^1)^2 I,\end{aligned} \tag{12}$$

and

$$\widehat{m}(z) := m_{\mu(\Sigma_{\text{lin}})\boxtimes\mu_{\text{MP}}^{n/k_1}}(z), \quad \breve{m}(z) = \frac{n-k_1}{nz} + \frac{n}{k_1}\widehat{m}(z).$$

Furthermore, Assumption (11) *holds true with* X_0, r_1 *replaced by* X_1, r_2*, respectively, and we have that* $\operatorname{dist}(-1/\breve{m}(z), \mathbb{R}_+) \geq \operatorname{dist}(z, \mathbb{R}_+)$.

Remark 3.2. The tracial version of Proposition 3.1 has appeared multiple times in the literature, e.g. [44]. It implies that the spectrum $\widehat{\mu}_1$ of $X_1^\top X_1/k_1$ is approximately given by the free multiplicative convolution

$$\begin{aligned}\widehat{\mu}_1 &\approx \mu\Big((\kappa_1^1)^2\frac{X_0^\top X_0}{d} + (\kappa_*^1)^2 I\Big) \boxtimes \mu_{\text{MP}}^{n/k_1} \\ &= \Big(\mu\Big((\kappa_1^1)^2\frac{X_0^\top X_0}{d}\Big) \boxplus \delta_{(\kappa_*^1)^2}\Big) \boxtimes \mu_{\text{MP}}^{n/k_1}.\end{aligned} \tag{13}$$

In case $c \leq 1$, i.e. when μ_{MP}^c has no atom at 0, it was shown in [48] that

$$\sqrt{\mu \boxtimes \mu_{\text{MP}}^c} \boxplus_c \sqrt{\mu' \boxtimes \mu_{\text{MP}}^c} = \sqrt{(\mu \boxplus \mu') \boxtimes \mu_{\text{MP}}^c} \tag{14}$$

which allows to simplify (13). Here $\boxplus_c$ is the *rectangular free convolution* which models the distribution of singular values of the addition of two free rectangular random matrices, and the square-root is to be understood as the push-forward of the square-root map. Applying (14) to (13) yields

$$\sqrt{\widehat{\mu}_1} \approx \Big(\kappa_1^1 \otimes \sqrt{\widehat{\mu}_0 \boxtimes \mu_{\text{MP}}^{n/k_1}}\Big) \boxplus_{n/k_1} \kappa_*^1 \otimes \sqrt{\mu_{\text{MP}}^{n/k_1}}, \tag{15}$$

suggesting that the non-zero singular values of $X_1/\sqrt{k}$ can be modeled by the non-zero singular values of the *Gaussian equivalent model*:

$$c'W'X_0 + c''W'' \tag{16}$$

for some suitably chosen constants c', c'' and independent Gaussian matrices W, W'.

The last assertion of Proposition 3.1 allows to iterate over an arbitrary (but finite) number of layers. Indeed, after one layer we have

$$\begin{aligned}\Big(\frac{X_1^\top X_1}{k_1} - z_1\Big)^{-1} &\approx \Big(-\breve{m}(z_1)z_1\Sigma_{\text{lin}} - z_1\Big)^{-1} \\ &= c_1\Big(\frac{X_0^\top X_0}{k_0} - z_0\Big)^{-1},\end{aligned} \tag{17}$$

using the definitions from Theorem 3.3 for c_1, z_0 below.

Theorem 3.3 (Deterministic equivalent for dRF)**.** *For any deterministic* A *and Lipschitz-continious activation functions* $\sigma_1, \dots, \sigma_\ell$ *satisfying* $\mathbf{E}_{\xi\sim\mathcal{N}(0,1)}\,\sigma_m(\sqrt{r_m}\xi) = 0$*, under the Assumption* (11) *above, we have that for any* $z_\ell \in \mathbf{C} \setminus \mathbb{R}_+$

$$\Big|\Big\langle A\Big(\frac{X_\ell^\top X_\ell}{k_\ell} - z_\ell\Big)^{-1}\Big\rangle - c_1\cdots c_\ell \breve{m}_0\langle A\rangle\Big| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}}$$

and that

$$\Big|\Big\langle A\Big(\frac{X_\ell X_\ell^\top}{k_\ell} - z_\ell\Big)^{-1}\Big\rangle - \breve{m}_\ell\langle A\rangle\Big| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}},$$

where $\delta := \operatorname{dist}(z_\ell, \mathbb{R}_+)$*, and we recursively define*

$$\begin{aligned}\Sigma_{\text{lin}}^{\ell-1} &:= (\kappa_1^\ell)^2\frac{X_{\ell-1}^\top X_{\ell-1}}{k_{\ell-1}} + (\kappa_*^\ell)^2 I, \\ \breve{m}_\ell &:= \frac{n-k_\ell}{nz_\ell} + \frac{n}{k_\ell} m_{\mu(\Sigma_{\text{lin}}^{\ell-1})\boxtimes\mu_{\text{MP}}^{n/k_\ell}}(z_\ell) \\ -\frac{1}{c_\ell} &:= \breve{m}_\ell z_\ell(\kappa_1^\ell)^2, \quad z_{\ell-1} := c_\ell z_\ell - \Big(\frac{\kappa_*^\ell}{\kappa_1^\ell}\Big)^2\end{aligned} \tag{18}$$

for $\ell \geq 1$ *and finally*

$$\breve{m}_0 := \frac{d-n}{nz_0} + \frac{d^2}{n^2} m_{\mu(\Omega_0)\boxtimes\mu_{\text{MP}}^{d/n}}\Big(\frac{d}{n}z_0\Big). \tag{19}$$

Proofs of Proposition 3.1 and Theorem 3.3 are given in App. A.

Remark 3.4. The same iteration argument has appeared before in [8]. The main difference to our present work is the anisotropic nature of our estimate which allows to test both sample covariance, as well as Gram resolvent against arbitrary deterministic matrices. As we will discuss in the next section, this is crucial in order to provide closed-form asymptotics for the test error of the deep random features model.

3.2 Closed-formed formula for the population covariance

In Proposition 3.1 and Theorem 3.3 we iteratively considered $X_\ell^\top X_\ell/k_\ell$ as a sample-covariance matrix with population covariance

$$\mathbf{E}_{W_\ell} \frac{X_\ell^\top X_\ell}{k_\ell} = \mathbf{E}_w \, \sigma_\ell\Big(\frac{X_{\ell-1}^\top w}{\sqrt{k_{\ell-1}}}\Big)\sigma_\ell\Big(\frac{w^\top X_{\ell-1}}{\sqrt{k_{\ell-1}}}\Big) \approx \Sigma_{\rm lin}^\ell$$

and from this obtained formulas for the deterministic equivalents for both $X_\ell^\top X_\ell$ and $X_\ell X_\ell^\top$. A more natural approach would be to consider $X_\ell X_\ell^\top/n$ as a sample covariance matrix with population covariance

$$\Omega_\ell := \mathbf{E}_{X_0} \frac{X_\ell X_\ell^\top}{n}, \tag{20}$$

noting that the matrix X_ℓ conditioned on $W_1, \dots, W_\ell$ has independent columns. Theorem A.3 and Proposition A.4 apply also in this setting, but lacking a rigorous expression for Ω_ℓ the resulting deterministic equivalent is less descriptive than the one from Theorem 3.3. A heuristic closed-form formula for the population covariance which is conjectured to be exact was recently derived in [12]. We now discuss this result, and for the sake of completeness provide a derivation in Appendix App. B. Consider the sequence of matrices $\{\Omega_\ell^{\rm lin}\}_\ell$ defined by the recursion

$$\Omega_{\ell+1}^{\rm lin} = \kappa_1^{(\ell+1)2} \frac{W_{\ell+1}\Omega_\ell^{\rm lin} W_{\ell+1}^\top}{k_\ell} + \kappa_*^{(\ell+1)2} I_{k_{\ell+1}}. \tag{21}$$

with $\Omega_0^{\rm lin} := \Omega_0$. Informally, $\Omega_\ell^{\rm lin}$ provides an asymptotic approximation of Ω_ℓ in the sense that the normalized distance $||\Omega_\ell^{\rm lin} - \Omega_\ell||_F/\sqrt{d}$ is of order $\mathcal{O}(1/\sqrt{d})$. Besides, the recursion (21) implies that $\Omega_\ell^{\rm lin}$ can be expressed as a sum of products of Gaussian matrices (and transposes thereof), and affords a straightforward way to derive an analytical expression its asymptotic spectral distribution. This derivation is presented in App. B.

It is an interesting question whether an approximate formula for the population covariance matrix like the one in Equation (21) can be obtained indirectly via Theorem 3.3. There is extensive literature on this *inverse problem*, i.e. how to infer spectral properties of the population covariance spectrum from the sample covariance spectrum, e.g. [49] but we leave this avenue to future work.

3.3 Consistency of Theorem 3.3 and the approximate population covariance

What we can note, however, is that Equation (21) is *consistent* with Theorem 3.3. We demonstrate this in case of equal dimensions $n = d = k_1 = \cdots = k_\ell$ to avoid unnecessary technicalities due to the zero eigenvalues. We define

$$\widehat{\mu}_\ell := \mu\Big(\frac{X_\ell^\top X_\ell}{k_\ell}\Big) = \breve{\mu}_\ell := \mu\Big(\frac{X_\ell X_\ell^\top}{n}\Big) \tag{22}$$

and recall that Proposition 3.1 implies that

$$\widehat{\mu}_\ell \approx ((\kappa_1^l)^2 \otimes \widehat{\mu}_{l-1} \oplus (\kappa_*^l)^2) \boxtimes \mu_{\rm MP}. \tag{23}$$

On the other hand (6) applied to the sample covariance matrix $X_\ell X_\ell^\top/n$ with population covariance $\Omega_\ell \approx \Omega_\ell^{\rm lin}$ implies that

$$\begin{aligned}
\breve{\mu}_\ell &\approx \mu(\Omega_\ell^{\rm lin}) \boxtimes \mu_{\rm MP} \\
&= \mu\Big((\kappa_1^\ell)^2 \frac{W_\ell \Omega_{\ell-1}^{\rm lin} W_\ell^\top}{k_{\ell-1}} + (\kappa_*^\ell)^2 I_{k_\ell}\Big) \boxtimes \mu_{\rm MP} \\
&\approx \Big((\kappa_1^\ell) \otimes \mu(\Omega_{\ell-1}^{\rm lin}) \boxtimes \mu_{\rm MP} \oplus (\kappa_*^\ell)^2\Big) \boxtimes \mu_{\rm MP} \\
&\approx \Big((\kappa_1^\ell) \otimes \breve{\mu}_{\ell-1} \oplus (\kappa_*^\ell)^2\Big) \boxtimes \mu_{\rm MP},
\end{aligned} \tag{24}$$

demonstrating that both approaches lead to the same recursion. Here in the third step we applied (6) to the sample covariance matrix $\sqrt{\Omega_{\ell-1}^{\rm lin}} W_\ell^\top$, and in the fourth step used the first approximation for ℓ replaced by $\ell - 1$.

4 Gaussian universality of the test error

In the second part of this work, we discuss how the results on the asymptotic spectrum of the empirical and population covariances of the features can be used to provide sharp expressions for the test and training errors (5) when the labels are generated by a deep random neural network:

$$f_\star(\mathrm{x}^\mu) = \sigma^\star\left(\frac{\theta_\star^\top \varphi^\star(\mathrm{x}^\mu)}{\sqrt{k^\star}}\right). \tag{25}$$

The feature map $\varphi^\star$ denotes the composition $\varphi^\star_{L^\star} \circ \ldots \circ \varphi^\star_1$ of the $L^\star + 1$ layers:

$$\varphi^\star_\ell(\mathrm{x}) = \sigma^*_\ell\left(\frac{1}{\sqrt{k^\star_{\ell-1}}} W^\star_\ell \cdot \mathrm{x}\right),$$

and $\theta_\star \in \mathbb{R}^{k^\star}$ is the last layer weights. To alleviate notations, we denote $k^\star := k^\star_L$. The weight matrices $\{W^\star_\ell\}_{\ell\in[L^\star]}$ have i.i.d Gaussian entries sampled from $\mathcal{N}(0, \Delta^\star_\ell)$. Note that we do not require the sequence of activations $\{\sigma^\star_\ell\}_\ell$ and widths $\{\gamma_\ell := {}^{k^\star_\ell}/_d\}_\ell$ to match with those of the learner dRF (2). We address in succession

- The well-specified case where the target and learner networks share the same intermediate layers (i.e. same architecture, activations and weights) $\varphi^\star_\ell = \varphi_\ell$, $\ell \in [L]$ with $L^\star = L$, and the readout of the dRF is trained using ridge regression. This is equivalent to the interesting setting of ridge regression on a linear target, with features drawn from a non-Gaussian distribution, resulting from the propagation of Gaussian data through several non-linear layers.
- The general case where the target and learner possess generically distinct architectures, activations and weights, and a generic convex loss.

In both cases, we provide a sharp asymptotic characterization of the test error. Furthermore, we establish the equality of the latter with the test error of an equivalent learning problem on *Gaussian samples* with matching population covariance, thereby showing the Gaussian universality of the test error. In the well-specified case, our results are rigorous, and make use of the deterministic equivalent provided by Theorem 3.3. In the fully generic case, we formulate a conjecture, which we strongly support with finite-size numerical experiments.

4.1 Well-specified case

We first establish the Gaussian universality of the test error of dRFs in the matched setting $\varphi = \varphi^\star$, for a readout layer trained using a square loss. This corresponds to $\mathcal{Y} = \mathbb{R}$, $\ell(y, \hat{y}) = {}^1/_2(y - \hat{y})^2$. This case is particularly simple since the empirical risk minimization problem (4) admits the following closed form solution:

$$\hat{\theta} = {}^1/\sqrt{k}(\lambda I_k + {}^1/_k X_L X_L^\top)^{-1} X_L y \tag{26}$$

where we recall the reader $X_L \in \mathbb{R}^{k\times n}$ is the matrix obtained by stacking the last layer features column-wise and $y \in \mathbb{R}^n$ is the vector of labels. For a given target function, computing the test error boils down to a random matrix theory problem depending on variations of the trace of deterministic matrices times the resolvent of the features sample covariance matrices (c.f. App. C for a derivation):

$$\begin{aligned}\epsilon_g(\hat{\theta}) = \Delta\left(\left\langle \Omega_L \left(\lambda I_k + {}^1/_k X_L X_L\right)^{-1}\right\rangle + 1\right)\\ - \lambda(\lambda - \Delta)\partial_\lambda\left\langle \Omega_L \left(\lambda I_k + {}^1/_k X_L X_L\right)^{-1}\right\rangle\end{aligned} \tag{27}$$

Applying Theorem 3.3 yields the following corollary:

Corollary 4.1 (Ridge universality of matched target)**.** *Let $\lambda > 0$. In the asymptotic limit $n, d, k_\ell \to \infty$ with fixed $\mathcal{O}(1)$ ratios $\alpha = {}^n/_d$, $\gamma_\ell := {}^{k_\ell}/_d$ and under the assumptions of Theorem 3.3, the asymptotic test error of the ridge estimator (26) on the target (25) with $L = L^*$ and $\varphi^*_\ell = \varphi_\ell$ and additive Gaussian noise with variance $\Delta > 0$ is given by:*

$$\begin{aligned}\epsilon_g(\hat{\theta}) \xrightarrow{k\to\infty} \epsilon^\star_g = \Delta\left(\langle\Omega_L\rangle \breve{m}_L(-\lambda) + 1\right)\\ - \lambda(\lambda - \Delta)\langle\Omega_L\rangle\partial_\lambda \breve{m}_L(-\lambda)\end{aligned} \tag{28}$$

where $\breve{m}_L$ can be recursively computed from (18) respectively. In particular, this implies Gaussian universality of the asymptotic mean-squared error in this model, since (28) exactly agrees with the asymptotic test error of ridge regression on Gaussian data $\mathrm{x} \sim \mathcal{N}(0_d, \Omega_L)$ derived in [50].

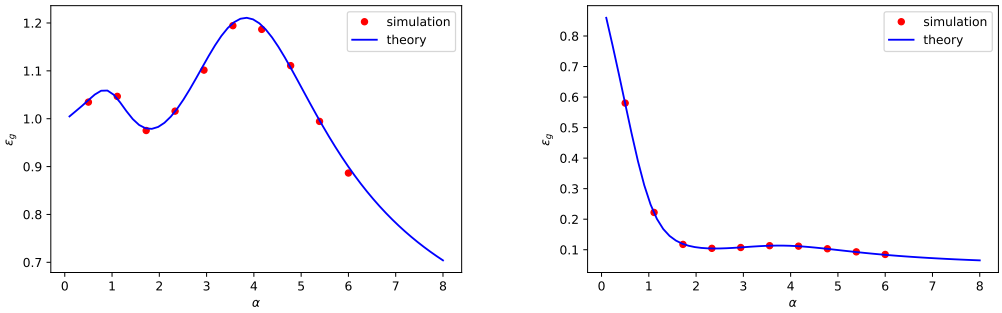

Figure 1: Learning curves $\epsilon_g(\alpha)$ for ridge regression ($\sigma_\star = id$, $\ell(y,z) = 1/2(y-z)^2$, and $g(y,\hat{y}) = (y-\hat{y})^2$) . Red dots correspond to numerical simulations on the learning model (2) (25), averaged over 20 runs. The solid line correspond to sharp asymptotic characterization provided by conjecture 4.3, and detailed in App. D. (left) 2-layers target ($L^\star = 1$,$\sigma_1^\star = \text{sign}$), (right) single-layer target ($L^\star = 0$). Both are learnt with a $2-$hidden layers RF (2) with $\sigma_{1,2}(x) = \tanh(2x)$ activation and regularization $\lambda = 0.001$.

A detailed derivation of (27) and Corollary 4.1 is given in App. C, together with a discussion of possible extensions to deterministic last-layer weights and general targets. Note that, while it is not needed to establish the Gaussian equivalence of ridge dRF regression in the well-specified case, the trace of the population covariance $\langle \Omega_L \rangle$ can be explicitly computed from the closed-form formula (21).

4.2 General case

Despite the major progress stemming from the application of the random matrix theory toolbox to learning problems, the application of the latter has been mostly limited to quadratic problems where a closed-form expression of the estimators, such as (26), are available. Proving universality results akin to Corollary 4.1 beyond quadratic problems is a challenging task, which has recently been the subject of intense investigation. In the context of generalized linear estimation (4), universality of the test error for the $L = 1$ random features model under a generic convex loss function was heuristically studied in [5], where the authors have shown that the asymptotic formula for the test error obtained under the Gaussian design assumption perfectly agreed with finite-size simulations with the true features. This Gaussian universality of the test error was later proven by [37] by combining a Lindeberg interpolation scheme with a generalized central limit theorem. Our goal in the following is to provide an analogous contribution as [5] to the case of multi-layer random features. This result builds on a rigorous, closed-form formula for the asymptotic test error of misspecified generalized linear estimation in the high-dimensional limit considered here, which was derived in [11].

We show that in the high-dimensional limit the asymptotic test error for the model introduced in Section 2 is in the *Gaussian universality class*. More precisely, the test error of this model is asymptotically equivalent to the test error of an equivalent Gaussian covariate model (GCM) consisting of doing generalized linear estimation on a dataset $\check{\mathcal{D}} = \{v^\mu, \check{y}^\mu\}_{\mu\in[n]}$ with labels $\check{y}^\mu = f_\star(1/\sqrt{k^\star}\theta_\star^\top u^\mu)$ and jointly Gaussian covariates:

$$(u,v) \sim \mathcal{N}\begin{pmatrix} \Psi_{L^\star} & \Phi_{L^\star L} \\ \Phi_{L^\star L}^\top & \Omega_L \end{pmatrix} \tag{29}$$

where we recall Ω_L is the variance of the model features (20) and $\Phi \in \mathbb{R}^{k^\star \times k}$ and $\Psi \in \mathbb{R}^{k^\star \times k^\star}$ are the covariances between the model and target features and the target variance respectively:

$$\Phi_{L^\star L} := \mathbf{E}\left[\varphi^\star(\mathrm{x})\varphi(\mathrm{x})^\top\right], \ \Psi_{L^\star} := \mathbf{E}\left[\varphi^\star(\mathrm{x})\varphi^\star(\mathrm{x})^\top\right] \tag{30}$$

This result adds to a stream of recent universality results in high-dimensional linear estimation [11, 38, 51], and generalizes the random features universality of [15, 36, 37] to $L > 1$. It can be summarized in the following conjecture:

Conjecture 4.2. *In the high-dimensional limit $n, d, k_\ell \to \infty$ at fixed $\mathcal{O}(1)$ ratios $\alpha := n/d$ and $\gamma_\ell := k_\ell/d$, the test error of the empirical risk minimizer (4) trained on $\mathcal{D} = \{(\mathrm{x}^\mu, y^\mu)\}_{\mu\in[n]}$ with covariates $\mathrm{x}^\mu \sim \mathcal{N}(0_d, \Omega_0)$ and labels from (25) is equal to the one of a Gaussian covariate model (29) with matching second moments Ψ, Φ, Ω as defined in (20) and (30).*

We go a step further and provide a sharp asymptotic expression for the test error. Construct recursively the sequence

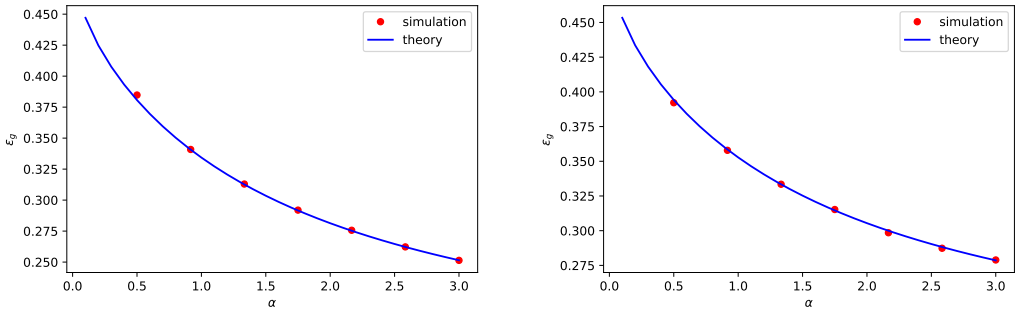

Figure 2: Learning curves $\epsilon_g(\alpha)$ for logistic regression ($\sigma_\star = \mathrm{sign}$, $\ell(y,z) = \ln(1+e^{-yz})$ and metric $g(y,\hat{y}) = 1-\Theta(y\hat{y})$). Red dots correspond to numerical simulations on the learning model (2) (25), averaged over 20 runs. The solid line correspond to sharp asymptotic characterization provided by conjecture 4.3, and detailed in App. D. (left) single-layer target ($L^\star = 0$), (right) two-layer target ($L^\star = 1$, $\sigma_1^\star = \mathrm{erf}$) (25) hidden sign layer. Both are learnt with a depth $L = 2$ dRF (2) with activation $\sigma_{1,2}(x) = \tanh(2x)$ and regularization $\lambda = 0.05$ (top) and $\sigma_{1,2}(x) = \mathrm{erf}(x)$ and $\lambda = 0.1$ (bottom).

of matrices

$$\Psi_{\ell+1}^{\mathrm{lin}} = \left(\kappa_1^{\star(\ell+1)}\right)^2 \frac{W_{\ell+1}^\star \Psi_\ell^{\mathrm{lin}} W_{\ell+1}^{\star\top}}{k_\ell^\star} + \left(\kappa_*^{\star(\ell+1)}\right)^2 I_{k_{\ell+1}^\star} \tag{31}$$

with the initial condition $\Omega_0^{\mathrm{lin}} = \Psi_0^{\mathrm{lin}} := \Omega_0$. Further define

$$\Phi_{L^\star L}^{\mathrm{lin}} = \left(\prod_{\ell=L^\star}^{1} \frac{\kappa_1^{\star\ell} W_\ell^\star}{\sqrt{k_\ell^\star}}\right) \cdot \Omega_0 \cdot \left(\prod_{\ell=1}^{L} \frac{\kappa_1^\ell W_\ell^\top}{\sqrt{k_\ell}}\right). \tag{32}$$

The sequence $\{\kappa_1^{\star\ell}\kappa_*^{\star\ell}\}_{\ell=1}^{L^\star}$ is define by (10) with $\sigma_\ell^\star, \Delta_\ell^\star$. In the special case $L^\star = 0$, which correspond to a single-index target function, the first product in $\Phi_{L^\star L}^{\mathrm{lin}}$ should be replaced by I_d. This particular target architecture is also known, in the case $L = 1$, as the *hidden manifold model* [5, 52] and affords a stylized model for structured data. The present paper generalizes these studies to arbitrary depths L. One is then equipped to formulate the following, stronger, conjecture:

Conjecture 4.3. *In the same limit as in Conjecture 4.2, the test error of the empirical risk minimizer (4) trained on $\mathcal{D} = \{(\mathrm{x}^\mu, y^\mu)\}_{\mu\in[n]}$ with covariates $\mathrm{x}^\mu \sim \mathcal{N}(0_d, \Omega_0)$ and labels from (25) is equal to the one of a Gaussian covariate model (29) with the matrices $\Psi_{L^\star}^{\mathrm{lin}}, \Omega_L^{\mathrm{lin}}, \Phi_{L^\star L}^{\mathrm{lin}}$ (21),(32).*

Conjecture 4.3 allows to give a fully analytical sharp asymptotic characterization of the test error, which we detail in App. D. Importantly, observe that it also affords compact closed-form formulae for the population covariances $\Omega_L, \Phi_{L^\star L}, \Psi_{L^\star}$. In particular the spectrum of $\Psi_{L^\star}^{\mathrm{lin}}, \Omega_L^{\mathrm{lin}}$ can be analytically computed and compares excellently with empirical numerical simulations. We report those results in detail in App. B. Figs. 1 and 2 present the resulting theoretical curve and contrasts them to numerical simulations in dimensions $d = 1000$, revealing an excellent agreement.

5 Depth-induced implicit regularization

An informal yet extremely insightful takeaway from Conjecture 4.3, and in particular the closed-form expressions (21), is that the activations in a deep non-linear dRF (2) share the same population statistics as the activations in a deep *noisy* linear network, with layers

$$\varphi_\ell^{\mathrm{lin}}(\mathrm{x}) = \kappa_1^\ell \frac{W_\ell^\top \mathrm{x}}{\sqrt{k_{\ell-1}}} + \kappa_*^\ell \xi_\ell, \tag{33}$$

where $\xi_\ell \sim \mathcal{N}(0_{k_\ell}, I_{k_\ell})$ is a Gaussian noise term. It is immediate to see that (33) lead to the same recursion as (21). This observation, which was made in the concomitant work [12], essentially allows to equivalently think of the problem of learning using a dRF (2) as one of learning with linear noisy network. Indeed, Conjecture 4.3 essentially suggests that the asymptotic test error depends on the second-order statistics of the last layer acrivations, shared between the dRF and the equivalent linear network. Finally, it is worthy to stress that, while the learner dRF is deterministic conditional

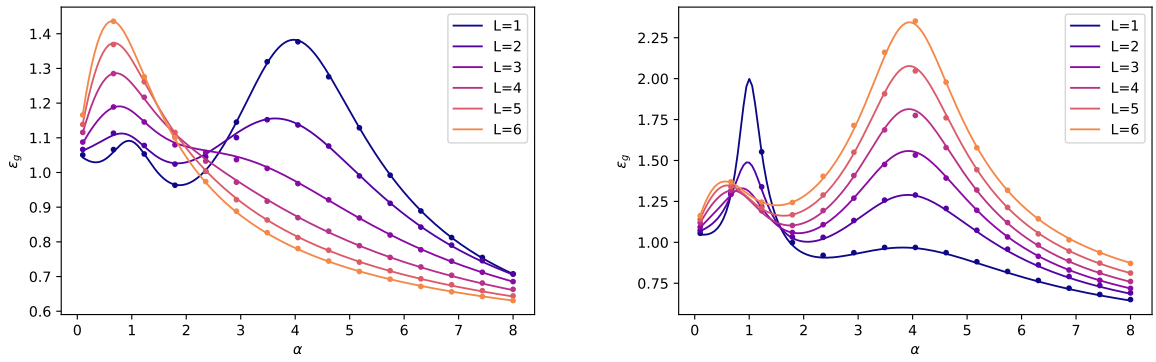

Figure 3: Learning curves for ridge regression on a 1-hidden layer target function ($\gamma_1^\star = 2$, $\sigma_1^\star = \text{sign}$) using a $L-$hidden layers learner with widths $\gamma_1 = ... = \gamma_L = 4$ and $\sigma_{1,...,L} = \tanh$ activation (left) or $\sigma_{1,...,L}(x) = 1.1 \times \text{sign}(x) \times \min(2, |x|)$ clipped linear activation (right), for depths $1 \leq L \leq 6$. The regularization is $\lambda = 0.001$. Solid lines represent theoretical curves evaluated from the sharp characterization of conjecture 4.3, while numerical simulations, averaged over 50 runs, are indicated by dots. The linear peak can be observed at $\alpha = 1$, while the non-linear peak occurs for $\alpha = \gamma = 4$ [53]. Despite sharing the same architecture, the use of different activations induces different implicit regularizations, leading to the linear (resp. non-linear) peak being further suppressed as the depth increases for the clipped linear activation (resp. tanh activation).

on the weights $\{W_\ell\}$, the equivalent linear network (33) is intrinsically stochastic in nature due to the effective noise injection ξ_ℓ at each layer. Statistical common sense dictates that this effective noise injection has a regularizing effect, by introducing some randomness in the learning, and helps mitigating overfitting. Since the effective noise is a product of the propagation through a non-linear layer, this suggest that *adding random non linear layers induces an implicit regularization*. We explore this intuition in this last section.

Observe first that the equivalent noisy linear network (33) reduces to a simple shallow noisy linear model

$$\hat{y}_\theta^{\text{lin}}(\mathrm{x}) = \sigma\left(\frac{1}{\sqrt{k}}\theta^\top\left(A_L \cdot \mathrm{x} + \xi_L\right)\right) \tag{34}$$

where the effective weight matrix A is

$$A_L := \prod_{\ell=1}^{L}\left(\kappa_1^\ell \frac{W_\ell}{\sqrt{k_{\ell-1}}}\right)$$

and the effective noise ξ_L is Gaussian with covariance C_ξ^L

$$C_\xi^L = \sum_{\ell_0=1}^{L-1}(\kappa_*^{\ell_0})^2\left(\prod_{\ell=\ell_0+1}^{L}\frac{\kappa_1^\ell W_\ell^\top}{\sqrt{k_{\ell-1}}}\right)^\top\left(\prod_{\ell=\ell_0+1}^{L}\frac{\kappa_1^\ell W_\ell^\top}{\sqrt{k_{\ell-1}}}\right) + (\kappa_*^L)^2 I_k.$$

The signal-plus-noise structure of the equivalent linear features (34) has profound consequences on the level of the learning curves of the model (2):

- When $\alpha = 1$, there are as many training samples as the dimension of the data $d-$ dimensional submanifold $A_L\,\mathrm{x}$, resulting in a standard interpolation peak. The noise part ξ_L induces an implicit regularization which helps mitigate the overfitting.
- As $\alpha = \gamma_L$, the number of training samples matches the dimension k_L of the noise, and the *noise* part is used to interpolate the training samples, resulting in another peak. This second peak is referred to as the non-linear peak by [53].

Therefore, there exists an interplay between the two peaks, with higher noise ξ_L both helping to mitigate the linear peak, and aggravating the non-linear peak. The depth of the network plays a role in that it modulates the amplitudes of the signal part and the noise part, depending on the activation through the recursions (10).

We give two illustrations of the regularization effect of depth in Fig. 3. Two activations are considered : $\sigma_a = \tanh$ (for which the noise level, as measure by $\text{tr}\, C_\xi^L$ decreases with depth) , and a very weakly non-linear activation

$\sigma_b(x) = 1.1 \times \text{sign}(x) \times \min(2, |x|)$, corresponding to a linear function clipped between -2.2 and 2.2 (for which $\text{tr}\, C_\xi^L$ increases with depth). Note that, because for σ_a the effective noise decreases with depth, the linear peak is aggravated for deeper networks, while the non-linear peak is simultaneously suppressed. Conversely, for σ_b, additional layers introduce more noise and cause a higher non-linear peak, while the induced implicit regularization mitigates the linear peak. Further discussion about the effect of architecture design on the generalization ability of dRFs (2) is provided in App. E.

6 Conclusion

We study the problem of learning a deep random network target function by training the readout layer of a deep network, with frozen random hidden layers (deep Random Features). We first prove an asymptotic deterministic equivalent for the conjugate kernel and sample covariance of the activations in a deep Gaussian random networks. This result is leveraged to establish a sharp asymptotic characterization of the test error in the specific case where the learner and teacher networks share the same intermediate layers, and the readout is learnt using a ridge loss. This proves the Gaussian universality of the test error of ridge regression on non-linear features corresponding to the last layer activations. In the fully generic case, we conjecture a sharp asymptotic formula for the test error, for fully general target/learner architectures and convex loss. The formulas suggest that the dRF behaves like a linear noisy network, characterized by an implicit regularization. We explore the consequences of this equivalence on the interplay between the architecture of the dRF and its generalization ability.

Acknowledgements

We thank Gabriele Sicuro for discussion during the course of this project. BL acknowledges support from the *Choose France - CNRS AI Rising Talents* program. DD is supported by ETH AI Center doctoral fellowship. DS is supported by SNSF Ambizione Grant PZ00P2_209089. HC acknowledges support from the ERC under the European Union's Horizon 2020 Research and Innovation Program Grant Agreement 714608-SMiLe.

References

[1] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[2] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[3] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. *J. Stat. Mech. Theory Exp.*, 2019(12), 2019.

[4] Song Mei and Andrea Montanari. The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve. *Commun. Pure Appl. Math.*, 75(4):667–766, 2022.

[5] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Generalisation error in learning with random features and the hidden manifold model. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3452–3462. PMLR, 13–18 Jul 2020.

[6] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[7] Alexander G. De G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

[8] Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[9] Jacob Zavatone-Veth and Cengiz Pehlevan. Exact marginal prior distributions of finite bayesian neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3364–3375. Curran Associates, Inc., 2021.

[10] Lorenzo Noci, Gregor Bachmann, Kevin Roth, Sebastian Nowozin, and Thomas Hofmann. Precise characterization of the prior predictive distribution of deep relu networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 20851–20862. Curran Associates, Inc., 2021.

[11] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *J. Stat. Mech. Theory Exp.*, 2022(11):Paper No. 114001, 78, 2022.

[12] Hugo Cui, Lenka Zdeborová, and Florent Krzakala. Optimal learning of deep random networks of extensive-width, 2023. *Private communication.*

[13] Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3063–3071. PMLR, 10–15 Jul 2018.

[14] Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electron. J. Probab.*, 26:Paper No. 150, 37, 2021.

[15] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Appl. Comput. Harmon. Anal.*, 59:3–84, 2022.

[16] Oussama Dhifallah and Yue M. Lu. A precise performance analysis of learning with random features. *arXiv:2008.11904*, 2020.

[17] Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and minimum-ℓ1-norm interpolated classifiers. *Ann. Statist.*, 50(3):1669–1695, 2022.

[18] David Bosch, Ashkan Panahi, Ayca Özcelikkale, and Devdatt Dubhash. Double descent in random feature models: Precise asymptotic analysis for general convex regularization. *arXiv:2204.02678*, 2022.

[19] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clement Hongler, and Franck Gabriel. Implicit regularization of random feature models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4631–4640. PMLR, 13–18 Jul 2020.

[20] Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. A study of uncertainty quantification in overparametrized high-dimensional models. *arXiv:2210.12760*, 2022.

[21] Stéphane D'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance(s) in the lazy regime. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2280–2290. PMLR, 13–18 Jul 2020.

[22] Bruno Loureiro, Cedric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala. Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14283–14314. PMLR, 17–23 Jul 2022.

[23] Antoine Bodin and Nicolas Macris. Model, sample, and epoch-wise descents: exact solution of gradient flow in the random feature model. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21605–21617. Curran Associates, Inc., 2021.

[24] Blake Bordelon and Cengiz Pehlevan. Learning curves for SGD on structured features. In *International Conference on Learning Representations*, 2022.

[25] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[26] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *J. Stat. Mech. Theory Exp.*, 2021(12):Paper No. 124009, 110, 2021.

[27] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[28] Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborova. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8936–8947. PMLR, 18–24 Jul 2021.

[29] Andre Manoel, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Multi-layer generalized linear estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2098–2102, 2017.

[30] Marylou Gabrié, Andre Manoel, Clément Luneau, jean barbier, Nicolas Macris, Florent Krzakala, and Lenka Zdeborová. Entropy and mutual information in models of deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[31] Benjamin Aubin, Bruno Loureiro, Antoine Maillard, Florent Krzakala, and Lenka Zdeborová. The spiked matrix model with generative priors. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[32] Paul Hand, Oscar Leong, and Vlad Voroninski. Phase retrieval under a generative prior. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[33] Benjamin Aubin, Bruno Loureiro, Antoine Baker, Florent Krzakala, and Lenka Zdeborová. Exact asymptotics for phase retrieval and compressed sensing with random generative priors. In Jianfeng Lu and Rachel Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 55–73. PMLR, 20–24 Jul 2020.

[34] Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Phys. Rev. X*, 11:031059, 2021.

[35] S. Ariosto, R. Pacelli, M. Pastore, F. Ginelli, M. Gherardi, and P. Rotondo. Statistical mechanics of deep learning beyond the infinite-width limit. *arXiv:2209.04882*, 2022.

[36] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. The Gaussian equivalence of generative models for learning with shallow neural networks. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, Proceedings of Machine Learning Research. 145, pages 426–471, 2021.

[37] Hong Hu and Yue M. Lu. Universality Laws for High-Dimensional Learning with Random Features. *IEEE Trans. Inf. Theory*, 2022.

[38] Andrea Montanari and Basil N. Saeed. Universality of empirical risk minimization. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4310–4312. PMLR, 02–05 Jul 2022.

[39] Hugo Cui, Luca Saglietti, and Lenka Zdeborov'a. Large deviations for the perceptron model and consequences for active learning. *Mach. Learn. Sci. Technol.*, 2:45001, 2019.

[40] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Error rates for kernel classification under source and capacity conditions. *ArXiv*, abs/2201.12655, 2022.

[41] Z. Burda, A. Görlich, A. Jarosz, and J. Jurkiewicz. Signal and noise in correlation matrix. *Phys. A*, 343(1-4):295–310, 2004.

[42] Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probab. Theory Related Fields*, 169(1-2):257–352, 2017.

[43] V A Marčenko and L A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

[44] Zhidong Bai and Wang Zhou. Large sample covariance matrices without independence structures in columns. *Statist. Sinica*, 18(2):425–442, 2008.

[45] Cosme Louart and Romain Couillet. Concentration of measure and large random matrices with an application to sample covariance matrices. *arXiv:1805.08295*, 2018.

[46] Clément Chouard. Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure. *arXiv:2211.13044*, 2022.

[47] John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1-2):32–52, 1928.

[48] Florent Benaych-Georges. On a surprising relation between the Marchenko-Pastur law, rectangular and square free convolutions. *Ann. Inst. Henri Poincaré Probab. Stat.*, 46(3):644–652, 2010.

[49] Noureddine El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.*, 36(6):2757–2790, 2008.

[50] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Stat.*, 46(1):247–279, 2018.

[51] Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová. Gaussian universality of linear classifiers with random labels in high-dimension, 2022.

[52] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Phys. Rev. X*, 10:041044, 2020.

[53] Stéphane D'Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: where and why do they appear? *J. Stat. Mech. Theory Exp.*, 2021(12):Paper No. 124002, 21, 2021.

[54] Radosław Adamczak. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electron. Commun. Prob.*, 20, 2015.

[55] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.

[56] Stéphane D'Ascoli, Marylou Gabrié, Levent Sagun, and Giulio Biroli. On the interplay between data structure and loss function in classification problems. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8506–8517. Curran Associates, Inc., 2021.

[57] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10131–10143. Curran Associates, Inc., 2021.

[58] Theodor Misiakiewicz. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv:2204.10425*, 2022.

[59] Hong Hu and Yue M. Lu. Sharp asymptotics of kernel ridge regression beyond the linear regime. *arXiv:2205.06798*, 2022.

[60] Noureddine El Karoui, Derek Bean, Peter J. Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[61] Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10112–10123. Curran Associates, Inc., 2020.

[62] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Stat.*, 50(2):949–986, 2022.

[63] Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23549–23588. PMLR, 17–23 Jul 2022.

A Anisotropic deterministic equivalent

A.1 Sample covariance matrices

Consider a random vector $x \in \mathbb{R}^d$ with $\mathbf{E}\, x = 0$ and $\mathbf{E}\, xx^\top = \Sigma$ and for $n \in \mathbf{N}$ construct $\mathcal{X} = (x_1, \dots, x_n) \in \mathbb{R}^{d\times n}$ using n independent copies $x_1, \dots, x_n$ of x. We are interested in the sample covariance and Gram matrices

$$\widehat{\Sigma} := \frac{\mathcal{X}\mathcal{X}^\top}{n} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{d\times d} \qquad \text{and} \qquad \check{\Sigma} := \frac{\mathcal{X}^\top\mathcal{X}}{n} = \Big(\frac{x_i^\top x_j}{n}\Big)_{i,j=1}^n \in \mathbb{R}^{n\times n} \tag{35}$$

and their resolvents

$$\widehat{G}(z) := (\widehat{\Sigma} - z)^{-1} \in \mathbb{C}^{d\times d} \qquad \text{and} \qquad \check{G}(z) := (\check{\Sigma} - z)^{-1} \in \mathbb{C}^{n\times n}. \tag{36}$$

The expectations of the sample covariance and Gram matrices are

$$\mathbf{E}\,\widehat{\Sigma} = \Sigma, \quad \mathbf{E}\,\check{\Sigma} = \frac{d}{n}\langle\Sigma\rangle I_n, \tag{37}$$

where we introduced the averaged trace $\langle A\rangle := m^{-1}\operatorname{Tr} A$ for $A \in \mathbb{R}^{m\times m}$.

Note that while the two resolvents behave differently as matrices, their traces are related due to the fact that the non-zero eigenvalues of $\widehat{\Sigma}$ and $\check{\Sigma}$ agree, whence

$$\langle\widehat{G}(z)\rangle = \frac{n}{d}\langle\check{G}(z)\rangle + \frac{n-d}{pz}. \tag{38}$$

The classical result on normalised traces of sample covariance and Gram resolvents is the following variance estimate under essentially optimal conditions.

Theorem A.1 (Tracial convergence of sample covariance matrices with general population [44]). *Assume that* $\|\Sigma\| \lesssim 1$, $d/n \sim 1$ *and that*

$$\mathbf{E}\left|\frac{x^\top Ax}{d} - \mathbf{E}\,\frac{x^\top Ax}{d}\right|^2 = \mathbf{E}\left|\frac{x^\top Ax}{d} - \langle\Sigma A\rangle\right|^2 = o(\|A\|) \tag{39}$$

for all deterministic matrices A. *Then it holds that*

$$\mathbf{E}\left|\langle(\widehat{\Sigma} - z)^{-1}\rangle - \widehat{m}(z)\right|^2 = o(1), \quad \mathbf{E}\left|\langle(\check{\Sigma} - z)^{-1}\rangle - \check{m}(z)\right|^2 = o(1), \quad \text{as} \quad n, d \to \infty, \tag{40}$$

for all fixed $z \in \mathbb{C}\setminus\mathbb{R}_+$, *where* $\check{m} = \check{m}(z)$ *is the unique solution to the scalar equation*

$$1 - \frac{d}{n} + z\check{m} = -\frac{d}{n}\langle(\Sigma\check{m} + 1)^{-1}\rangle. \tag{41}$$

and

$$\widehat{m}(z) := \frac{n}{d}\check{m}(z) + \frac{d-n}{d}\frac{1}{-z} \tag{42}$$

Here $\widehat{m}$ is the solution to the Marchenko-Pastur equation (6) and the correspodning measure is the free multiplicative convolution of the empirical spectral measure $\mu(\Sigma) := d^{-1}\sum_{\lambda\in\operatorname{Spec}(\Sigma)}\delta_\lambda$ of Σ and a Marchenko-Pastur distribution μ_{MP}^c of aspect ratio $c = d/n$. Thus, by Stieltjes inversion the result of Theorem A.1 can be phrased as

$$\mu\Big(\frac{\mathcal{X}^\top\mathcal{X}}{n}\Big) = \mu(\check{\Sigma}) \approx \frac{d}{n}\mu(\Sigma)\boxtimes\mu_{\mathrm{MP}}^{d/n} + \frac{n-d}{n}\delta_0, \quad \mu\Big(\frac{\mathcal{X}\mathcal{X}^\top}{n}\Big) = \mu(\widehat{\Sigma}) \approx \mu(\Sigma)\boxtimes\mu_{\mathrm{MP}}^{d/n} \tag{43}$$

in a weak and global sense. Note that we have the limits

$$\lim_{c\to\infty}\mu(\Sigma)\boxtimes\mu_{\mathrm{MP}}^c = \delta_0, \quad \lim_{c\to 0}\mu(\Sigma)\boxtimes\mu_{\mathrm{MP}}^c = \mu(\Sigma) \tag{44}$$

which are precisely the expected behaviour since for large $c = d/n$ the rank n of $\mathcal{X}\mathcal{X}^\top$ grows much smaller than d and therefore the empirical measure $\mu(\widehat{\Sigma})$ is concentrated on the origin, while for small $c = d/n$ by the law of large numbers $\mathcal{X}\mathcal{X}^\top/n \approx \mathbf{E}\,\mathcal{X}\mathcal{X}^\top/n = \Sigma$.

A.2 Anisotropic deterministic equivalents

The tracial result from Theorem A.1 only allows to control the eigenvalues of $\widehat{\Sigma}, \check{\Sigma}$ but not the eigenvectors. There has been extensive work on non-tracial deterministic equivalents of $\widehat{\Sigma}, \check{\Sigma}$, either in the form of entrywise asymptotics $\check{G}_{ij} \approx \cdots$, isotropic asymptotics $x^\top \widehat{G} y \approx \cdots$ for deterministic vectors x, y or functional tracial asymptotics $\langle A\widehat{G}\rangle \approx \cdots$ for deterministic matrices A. Any of these results contain non-trivial information on how $\widehat{G}, \check{G}$ behave as matrices in the asymptotic limit and can be used to infer information on eigenvectors.

For separable correlations an optimal local law in isotropic and tracial form has been obtained in [42]:

Theorem A.2 ([42], Theorem 3.6). *If $\mathcal{X} = \Sigma^{1/2} X$ for some matrix X with independent identically distributed entries[1] with mean 0 and variance 1, and the spectral density $\mu(\Sigma) \boxtimes \mu_{\mathrm{MP}}^{d/n}$ is* regular[2]*, then it holds that*

$$\left|\langle(\widehat{\Sigma} - z)^{-1}\rangle - \widehat{m}(z)\right| + \left|\langle(\check{\Sigma} - z)^{-1}\rangle - \check{m}(z)\right| \prec \frac{1}{n \operatorname{Im} z}, \tag{45}$$

in tracial sense, and for any deterministic vectors x, y

$$\left|x^\top\left[(\widehat{\Sigma} - z)^{-1} - (-\Sigma\check{m}(z)z - z)^{-1}\right]y\right| + \left|x^\top(\check{\Sigma} - z)^{-1}y - \check{m}(z)x^\top y\right| \prec \frac{\|x\|\|y\|}{\sqrt{n \operatorname{Im} z}} \tag{46}$$

in isotropic sense.

Note that in particular, matrix $\widehat{G}(z)$ asymptotically is equal to a resolvent

$$\widehat{M}(z) := \left(-\Sigma\check{m}(z)z - z\right)^{-1} \tag{47}$$

of the population covariance Σ, while $\check{G}$ asymptotically is a scalar multiple of the identity.

More recently a functional tracial local law (albeit with very much suboptimal dependence on the spectral parameter) for $\widehat{G}$ has been obtained in [46]:

Theorem A.3 ([46], Proposition 2.4). *If $\|\Sigma\| \le C$ and $\mathcal{X}$ satisfies, for some positive constants c, C, σ*

$$P(|f(\mathcal{X}) - \mathbf{E}\, f(\mathcal{X})| \ge t) \le Ce^{-c(t/\sigma)^2} \quad \forall \text{ 1-Lipschitz } f : (\mathbb{R}^{d\times n}, \|\cdot\|_F) \to (\mathbb{R}, |\cdot|), \tag{48}$$

we have that for all deterministic matrices A and $|z| \lesssim 1$ with high probability[3].

$$\left|\langle A(\widehat{\Sigma} - z)^{-1} - A(-\check{m}(z)z\Sigma - z)^{-1}\rangle\right| \le \frac{\sqrt{\langle AA^*\rangle \log n}}{n \operatorname{dist}(z, \mathbb{R}_+)^9}, \tag{49}$$

where $\check{m} = \check{m}(z)$ is the unique solution to the scalar equation

$$1 - \frac{d}{n} + z\check{m} = -\frac{d}{n}\langle(\Sigma\check{m} + 1)^{-1}\rangle. \tag{50}$$

Note that the functional tracial formulation with convergence rate $1/n$ and error in terms of the Frobenius norm of A automatically includes an isotropic local law as a special case. Indeed, for $A = xy^\top$ it follows that

$$y^\top\left((\hat{\Sigma} - z)^{-1} - (-\check{m}z\Sigma - z)^{-1}\right)x \prec \frac{\|x\|\|y\|}{\sqrt{n}\delta^9}, \tag{51}$$

where we denote here and in the future $\delta \equiv \delta(z) := \operatorname{dist}(z, \mathbb{R}_+)$. In this work we extend the functional tracial local law from [46] to the case of $\check{G}$ and obtain the following result:

Proposition A.4 (Functional local law for Gram matrices). *Under the assumptions of Theorem A.3 we have that*

$$\left|\langle A(\check{\Sigma} - z)^{-1}\rangle - \check{m}(z)\langle A\rangle\right| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}}. \tag{52}$$

[1] with finite moments of all orders

[2] See Definition 2.7 in [42]

[3] The statement in [46] literally gives $\operatorname{Im} z$ rather than $\operatorname{dist}(z, \mathbb{R}_+)$ but the proof verbatim gives the stronger bound since $\operatorname{Im} z$ is merely used as a lower bound on the smallest singular value of a matrix of the type $AA^* - z$

Note that the bound in Proposition A.4 is weaker than the bound in Theorem A.3, and both results are very much weaker than Theorem A.2 in the dependence on the spectral parameter. In light of related results it is natural to conjecture the following:

Conjecture A.5. *Assume that quadratic forms of x concentrate as*

$$\left|\frac{x^\top Ax}{d} - \langle \Sigma A\rangle\right| \prec \frac{\langle AA^*\rangle^{1/2}}{\sqrt{d}} \tag{53}$$

for any deterministic matrix A, and that $\|\Sigma\| \lesssim 1$. Then we have the functional tracial estimates

$$\begin{aligned}\left|\langle zA(\widehat{\Sigma}-z)^{-1}\rangle - \langle A(-\breve{m}(z)\Sigma - I)^{-1}\right| &\prec \frac{\langle AA^*\rangle^{1/2}}{n\delta}\\ \left|\langle A(\breve{\Sigma}-z)^{-1}\rangle - \breve{m}(z)\langle A\rangle\right| &\prec \frac{\langle AA^*\rangle^{1/2}}{n\delta}.\end{aligned} \tag{54}$$

Note that the Lipschitz concentration required in Theorem A.3 is much stronger than the quadratic form concentration of Conjecture A.5 because it implies that the column vectors x of $\mathcal{X}$ satisfy

$$P(|f(x) - \mathbf{E}\, f(x)| \geq t) \leq C\exp\Big(-\frac{t^2}{C\lambda_f^2}\Big) \tag{55}$$

for all λ_f-Lipschitz $f\colon \mathbb{R}^d \to \mathbb{R}$. Therefore by Hanson-Wright ([54], Thm. 2.4)

$$P\Big(\left|\frac{x^\top Ax}{d} - \langle \Sigma A\rangle\right| \geq \frac{t\langle AA^*\rangle^{1/2}}{\sqrt{d}} + \frac{t\|A\|}{d}\Big) \leq Ce^{-\min\{t^2,t\}/C} \tag{56}$$

and, since also $\|A\| \leq \sqrt{d}\langle AA^*\rangle^{1/2}$, we have that with high probability

$$\left|\frac{x^\top Ax}{d} - \langle \Sigma A\rangle\right| \leq \log d\Big(\frac{\langle AA^*\rangle^{1/2}}{\sqrt{d}} + \frac{\|A\|}{d}\Big) \lesssim \log d\frac{\langle AA^*\rangle^{1/2}}{\sqrt{d}}. \tag{57}$$

Let us now turn to the proof of Proposition A.4. We will need the following result of Lipschitzness of the resolvent function, see e.g. [46]

Lemma A.6. *The map $\breve{G}\colon \mathcal{X} \to \big(\mathcal{X}^\top\mathcal{X}/n - z\big)^{-1}$ is $(3\delta^{-2}|z|^{1/2}n^{-1/2})$-Lipschitz with respect to Frobenius norm.*

Proof of Proposition A.4. Denote $\breve{m} \equiv \breve{m}(z)$. By the Schur complement formula we have

$$\breve{G}_{ii} = -\Big(z + z\frac{x_i^\top \hat{G}^{(i)}x_i}{n}\Big)^{-1} = -\Big(z + zc\langle \Sigma\hat{G}^{(i)}\rangle\Big)^{-1} + O\Big(\frac{1}{\sqrt{n}\delta^9}\Big) = \breve{m} + O\Big(\frac{1}{\sqrt{n}\delta^9}\Big), \tag{58}$$

using

$$\langle \Sigma\hat{G}^{(i)}\rangle = \langle \Sigma\breve{G}\rangle + \frac{1}{n}\left\langle \Sigma\frac{\hat{G}^{(i)}x_ix_i^\top\hat{G}^{(i)}}{1 + x_i\hat{G}^{(i)}x_i/n}\right\rangle = -\frac{1}{z}\langle \Sigma(\breve{m}\Sigma + I)^{-1}\rangle + O\Big(\frac{1}{n\delta^9}\Big) \tag{59}$$

and

$$z - c\langle \Sigma(\breve{m}\Sigma + I)^{-1}\rangle = z - \frac{c}{\breve{m}} + \frac{c}{\breve{m}^2}\langle(\breve{m}\Sigma + I)^{-1}\rangle = z - \frac{c}{\breve{m}} - \frac{1}{\breve{m}}\Big(1 - c + z\breve{m}\Big) = -\frac{1}{\breve{m}} \tag{60}$$

in the last step. Next, for off-diagonal elements we have, again by Schur-complement, that

$$\breve{G}_{ij} = z\breve{G}_{ii}\breve{G}^{(i)}_{jj}\frac{x_i^\top\hat{G}^{(ij)}x_j}{n} = z\breve{G}_{ii}\Big(\breve{G}_{jj} - \frac{\breve{G}_{ij}\breve{G}_{ji}}{\breve{G}_{jj}}\Big)\frac{x_i^\top\hat{G}^{(ij)}x_j}{n}. \tag{61}$$

Here from the first equality already a bound size $n^{-1/2}\delta^{-4}$ follows. Thus, together with Equation (58) it follows that

$$\breve{G}_{ij} = \breve{m}^2 z\frac{x_i^\top\hat{G}^{(ij)}x_j}{n} + O\Big(\frac{1}{n\delta^8}\Big), \tag{62}$$

and therefore by mean-zero assumption that $\mathbf{E}\,\breve{G}_{ij} = O(1/n\delta^8)$. This together with Equation (58) implies that

$$\Big\|\mathbf{E}\,\breve{G} - \breve{m}(z)I\Big\|_{\mathrm{F}} = O\Big(\frac{1}{\delta^9}\Big). \tag{63}$$

We write

$$\left|\langle A\check{G}\rangle - \check{m}(z)\langle A\rangle\right| \le \left|\langle A\check{G}\rangle - \mathbf{E}\langle A\check{G}\rangle\right| + \left|\mathbf{E}\langle A\check{G}\rangle - \check{m}(z)\langle A\rangle\right|. \tag{64}$$

Note that from Lemma A.6 and Cauchy-Schwarz inequality,

$$\text{the map } \mathcal{X} \to \left\langle A\left(\frac{\mathcal{X}^\top\mathcal{X}}{n} - z\right)\right\rangle \quad \text{is} \quad \frac{3|z|^{1/2}\langle AA^*\rangle^{1/2}}{n\delta^2}\text{-Lipschitz}, \tag{65}$$

therefore,

$$\left|\langle A\check{G}\rangle - \mathbf{E}\langle A\check{G}\rangle\right| \prec \frac{|z|^{1/2}\langle AA^*\rangle^{1/2}}{n\delta^2} \tag{66}$$

Also, from (63), we have

$$\left|\mathbf{E}\langle A\check{G}\rangle - \check{m}(z)\langle A\rangle\right| \le \frac{1}{\sqrt{n}}\langle AA^*\rangle^{1/2}\left\|\mathbf{E}\,\check{G} - \check{m}(z)I\right\|_F \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}}. \tag{67}$$

The statement of the Proposition follows from (64), (66) and (67). □

A.3 Random feature model

We consider a one-layer random feature model, with a scalar function $\sigma_1(x)$ applied entrywise.

$$\sigma_1\Big(\frac{W_1X_0}{\sqrt{d}}\Big), \quad X_0 \in \mathbb{R}^{d\times n}, \quad W_1 \in \mathbb{R}^{k_1\times d}. \tag{68}$$

We require the following assumptions.

Assumption A.7 (Gaussian weight). *Entries of W_1 are iid. $\mathcal{N}(0,1)$ elements.*

Assumption A.8 (Orthogonal and bounded data). *For a positive constant r_1, X_0 satisfies*

$$\left\|\frac{X_0^\top X_0}{d} - r_1 I\right\|_{\max} \prec \frac{1}{\sqrt{n}}, \quad \left\|\frac{X_0}{\sqrt{d}}\right\|_{op} \prec 1. \tag{69}$$

Assumption A.9 (Nonlinearity). *The scalar function σ_1 is λ_σ-Lipschitz and satisfies $\langle\sigma_1\rangle_{\mathcal{N}(r_1)} = 0$, where*

$$\langle f\rangle_{\mathcal{N}(\sigma^2)} := \frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}} f(x)\exp\Big(-\frac{x^2}{2\sigma^2}\Big)\,\mathrm{d}x. \tag{70}$$

Assumption A.10 (Proportional regime). *For some constants c_1, c_2,*

$$c_1 n \le \min\{d, k_1\} \le \max\{d, k_1\} \le c_2 n, \quad 0 < c_1 < c_2 < \infty. \tag{71}$$

For simplicity, we set the variance of the weight matrix to be equal to 1, although the results can be easily extended to arbitrary variance Δ, by scaling the function σ_1.

Let $\tilde{w}_i$ denote the ith row of W_1. We define

$$X_1 := \sigma_1\Big(\frac{W_1X_0}{\sqrt{d}}\Big) = \left(\sigma_1\Big(\frac{X^\top\tilde{w}_1}{\sqrt{d}}\Big)\cdots\sigma_1\Big(\frac{X^\top\tilde{w}_{k_1}}{\sqrt{d}}\Big)\right)^\top \in \mathbb{R}^{k_1\times n} \tag{72}$$

as a matrix with independent identically distributed rows and corresponding sample covariance matrix

$$\widehat{\Sigma} := \frac{X_1^\top X_1}{k_1} = \frac{1}{k_1}\sigma_1\Big(\frac{X_0^\top W_1^\top}{\sqrt{d}}\Big)\sigma_1\Big(\frac{W_1X_0}{\sqrt{d}}\Big). \tag{73}$$

We have $(X_1)_{ij} = \sigma_1(\xi_{ij})$, for $\xi_{ij} := \tilde{w}_i^\top x_j \sim \mathcal{N}\left(0, \|x_j\|^2/d\right)$, where x_j is the jth column of X_0. In order to analyze functions of Gaussian variables, we use the following decomposition.

Lemma A.11 (Hermite decomposition). *For any Lipschitz-continuous f and any $\sigma > 0$ we have the σ-Hermite expansion*[4]*,*

$$f(x) = \sum_{k\geq 0} \frac{\sigma^k}{k!} \operatorname{He}_k\Big(\frac{x}{\sigma}\Big) \langle f^{(k)} \rangle_{\mathcal{N}(\sigma^2)} \tag{74}$$

where

$$\operatorname{He}_k(x) := (-1)^k \exp\Big(\frac{x^2}{2}\Big) \frac{\mathrm{d}^k}{\mathrm{d}x^k} \exp\Big(-\frac{x^2}{2}\Big) \tag{75}$$

with $\operatorname{He}_k(x)$ being the standard Hermite polynomials $\operatorname{He}_0(x) = 1$, $\operatorname{He}_1(x) = x$, $\operatorname{He}_2(x) = x^2 - 1$, etc.

Note that the Hermite polynomials are pairwise orthogonal with respect to the Gaussian density. More precisely,

$$\mathbf{E} \operatorname{He}_k(N_1) \operatorname{He}_j(N_2) = \delta_{jk} k! \operatorname{Cov}(N_1, N_2)^k \tag{76}$$

for jointly Gaussian N_1, N_2 with $\mathbf{E}\, N_1 = \mathbf{E}\, N_2 = 0$ and $\mathbf{E}\, N_1^2 = \mathbf{E}\, N_2^2 = 1$. By applying (74) twice and using (76) we obtain the Parseval identity

$$\langle f^2 \rangle_{\mathcal{N}(\sigma)} = \sum_{k\geq 0} \frac{\sigma^{2k}}{k!} \langle f^{(k)} \rangle^2_{\mathcal{N}(\sigma)}. \tag{77}$$

In the proof of the deterministic equivalent for the deep random features model, we rely on techniques developed in [45, 46] which use concentration of measure theory to analyze random matrices. This approach works particularly well with common neural network architectures, where one can view transformations from layer to layer as Lipschitz mappings. The following Lemma establishes Lipschitzness of required functions.

Lemma A.12. *Let $f(x)$ be a λ-Lipschitz function. Let $x, y, w \in \mathbb{R}^d$, $W \in \mathbb{R}^{k\times d}$ and $X \in \mathbb{R}^{d\times n}$. The following maps are Lipschitz, assuming $f(x)$ is applied entrywise:*

$$w \to f\left(\frac{x^\top w}{\sqrt{d}}\right) \qquad \text{and} \qquad W \to f\left(\frac{WX}{\sqrt{d}}\right), \tag{78}$$

with Lipschitz constants $\lambda\left\|x/\sqrt{d}\right\|$ and $\lambda\left\|X/\sqrt{d}\right\|$ respectively. Furthermore, under the event $Q := \{|f(x^\top w/\sqrt{d})| \lesssim 1 \wedge |f(y^\top v/\sqrt{d})| \lesssim 1|\}$, the map

$$w \to f\left(\frac{x^\top w}{\sqrt{d}}\right) f\left(\frac{y^\top w}{\sqrt{d}}\right) \tag{79}$$

is also Lipschitz with corresponding constant $\alpha \lesssim \lambda\left(\left\|x/\sqrt{d}\right\| + \left\|y/\sqrt{d}\right\|\right)$.

Proof. Lipschitz property of the first and second map follows directly from Cauchy-Schwarz inequality. For the third map, since the product of Lipschitz functions is not necessarily Lipschitz, one needs to condition on the "good" event Q. For simplicity, denote $f(a, b) := f\left(a^\top b/\sqrt{d}\right)$. Under Q we can write, for some vectors $u, v \in \mathbb{R}^d$,

$$\begin{aligned}
&|f(x,w)f(y,w) - f(x,v)f(y,v)| \\
&\quad \leq |f(x,w)f(y,w) - f(x,w)f(y,v)| + |f(x,w)f(y,v) - f(x,v)f(y,v)| \\
&\quad = |f(x,w)||f(y,w) - f(y,v)| + |f(y,v)||f(x,w) - f(x,v)| \\
&\quad \lesssim \lambda\left(\left\|x/\sqrt{d}\right\| + \left\|y/\sqrt{d}\right\|\right) \|w - v\|.
\end{aligned} \tag{80}$$

□

Recall the notations

$$\begin{aligned}
r_2 &:= \langle \sigma_1^2 \rangle_{\mathcal{N}(r_1)} \\
\kappa_1^1 &:= \langle \sigma_1' \rangle_{\mathcal{N}(r_1)} \\
\kappa_*^1 &:= \sqrt{\langle \sigma_1^2 \rangle_{\mathcal{N}(r_1)} - r_1 (\kappa_1^1)^2}
\end{aligned} \tag{81}$$

for the proof of Proposition 3.1. We state technical Lemmas.

[4]Note that despite the appearance of the derivative smoothness is not required as by integration by parts the derivative can be transferred to the smooth Gaussian weight.

Lemma A.13. *For $w \sim \mathcal{N}(0, I)$, λ_σ-Lipschitz function $\sigma(x)$ and $\left\|x/\sqrt{d}\right\| \lesssim 1$, with high probability*

$$\left|\sigma\left(\frac{x^\top w}{\sqrt{d}}\right)\right| \lesssim 1. \tag{82}$$

Proof. Since the map $w \to \sigma\left(\frac{x^\top w}{\sqrt{d}}\right)$ is $\left\|x/\sqrt{d}\right\|\lambda_\sigma$-Lipschitz, we have by Gaussian concentration theorem (see e.g. Theorem 5.2.2 in [55]) that

$$P\left(\left|\sigma\left(\frac{x^\top w}{\sqrt{d}}\right) - \mathbf{E}_w\,\sigma\left(\frac{x^\top w}{\sqrt{d}}\right)\right| \geq t\right) \leq e^{-\frac{t^2}{2\|x/\sqrt{d}\|^2\lambda_\sigma^2}}. \tag{83}$$

Next, by Equation (90), for each $i \in [n]$,

$$\mathbf{E}_w\,\sigma\left(\frac{x^\top w}{\sqrt{d}}\right) = \langle\sigma\rangle_{\mathcal{N}(\|x\|^2/d)} = O(1/\sqrt{n}), \tag{84}$$

which implies that, with high probability,

$$\left|\sigma\left(\frac{x^\top w}{\sqrt{d}}\right)\right| \lesssim 1. \tag{85}$$

□

Lemma A.14. *For $w \sim \mathcal{N}(0, I)$, the random variable $\sigma\left(\frac{x^\top w}{\sqrt{d}}\right)\sigma\left(\frac{w^\top y}{\sqrt{d}}\right)$ is subgaussian with high probability. Its subgaussian norm is $O(\lambda_\sigma(\left\|x/\sqrt{d}\right\| + \left\|y/\sqrt{d}\right\|))$*

Proof. Follows from Lemma A.13, Lemma A.12 and the Gaussian concentration theorem. □

Lemma A.15. *For matrices $A, B \in \mathbb{R}^{n\times n}$, we have*

1. $\|AB\|_F \leq \|A\|\|B\|_F$,
2. $\mathrm{Tr}(AB) \leq \|A\|_F\|B\|_F$,
3. $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, *if A and B are invertible.*

Lemma A.16. *For any positive semi-definite matrix Y and for any $z \in \mathbb{C} \setminus \mathbb{R}_+$, we have*

$$\left\|(Y - z)^{-1}\right\| \leq \mathrm{dist}(z, \mathbb{R}_+)^{-1}. \tag{86}$$

Proof of Proposition 3.1. Define the population covariance matrix

$$\Sigma_X := \mathbf{E}_w\,\sigma\Big(\frac{X_0^\top w}{\sqrt{d}}\Big)\sigma\Big(\frac{w^\top X_0}{\sqrt{d}}\Big) \in \mathbb{R}^{n\times n}, \quad w \sim \mathcal{N}(0, I). \tag{87}$$

Using Hermite series expansion (74) and (76), for fixed X_0, we can write an explicit form

$$\Sigma_X = \sum_{a\geq 0}\frac{1}{a!}D_X^{(a)}\Big(\frac{X_0^\top X_0}{d}\Big)^{\odot a}D_X^{(a)}, \tag{88}$$

where we defined the diagonal matrix

$$D_X^{(a)} := \mathrm{diag}\Big(\langle\sigma^{(a)}\rangle_{\mathcal{N}(\|x_1\|^2/d)}, \ldots, \langle\sigma^{(a)}\rangle_{\mathcal{N}(\|x_n\|^2/d)}\Big). \tag{89}$$

From Assumption A.8 and standard perturbation analysis it follows that

$$\langle\sigma\rangle_{\mathcal{N}(\|x_i\|^2/d)} = \langle\sigma\rangle_{\mathcal{N}(r_1)} + O\Big(\frac{1}{\sqrt{n}}\Big) = O\Big(\frac{1}{\sqrt{n}}\Big) \tag{90}$$

and

$$\langle\sigma'\rangle_{\mathcal{N}(\|x_i\|^2/d)} = \langle\sigma'\rangle_{\mathcal{N}(r_1)} + O\Big(\frac{1}{\sqrt{n}}\Big). \tag{91}$$

Therefore, we can conclude that, for off-diagonal $i \neq j$,

$$\begin{aligned}(\Sigma_X)_{ij} &= \sum_{a\geq 0} \frac{\langle \sigma^{(a)} \rangle_{\mathcal{N}(\|x_i\|^2/d)} \langle \sigma^{(a)} \rangle_{\mathcal{N}(\|x_j\|^2/d)}}{a!} \left(\frac{x_i^\top x_j}{d}\right)^a \\ &= \langle \sigma' \rangle^2_{\mathcal{N}(r_1)} \frac{x_i^\top x_j}{d} + O\left(\frac{1}{n}\right) = (\Sigma_{\text{lin}})_{ij} + O\left(\frac{1}{n}\right)\end{aligned} \tag{92}$$

and for diagonal entries we can write directly from (87),

$$\begin{aligned}(\Sigma_X)_{ii} = \langle \sigma^2 \rangle_{\mathcal{N}(\|x_i\|^2/d)} &= \langle \sigma' \rangle^2_{\mathcal{N}(r_1)} \frac{\|x_i\|^2}{d} + \left(\langle \sigma^2 \rangle_{\mathcal{N}(r_1)} - r_1 \langle \sigma' \rangle^2_{\mathcal{N}(r_1)}\right) + O\left(\frac{1}{\sqrt{n}}\right) \\ &= (\Sigma_{\text{lin}})_{ii} + O\left(\frac{1}{\sqrt{n}}\right).\end{aligned} \tag{93}$$

Summing over all indices i, j we get that

$$\|\Sigma_X - \Sigma_{\text{lin}}\|_{\text{F}} = O(1). \tag{94}$$

Let us define $\breve{m}(\Sigma, z)$ as the solution to the following equation:

$$\breve{m} = \frac{d-n}{nz} - \frac{d}{zn} \langle (\Sigma \breve{m} + 1)^{-1} \rangle, \tag{95}$$

and $\breve{m}_X \coloneqq \breve{m}(\Sigma_X, z), \breve{m}_{\text{lin}} \coloneqq \breve{m}(\Sigma_{\text{lin}}, z)$. Consider the sequence of approximations (in a functional tracial sense):

$$\left(\frac{X_1^\top X_1}{k_1} - z\right)^{-1} \approx \left(-\breve{m}_X z \Sigma_X - z\right)^{-1} \approx \left(-\breve{m}_{\text{lin}} z \Sigma_X - z\right)^{-1} \approx \left(-\breve{m}_{\text{lin}} z \Sigma_{\text{lin}} - z\right)^{-1}. \tag{96}$$

The first approximation follows from Theorem A.3 applied to the matrix $\mathcal{X} = X_1^\top$. The matrix $\mathcal{X}$ is concentrated due to Lemma A.12 and Gaussian concentration theorem.

The second approximation requires proving a stability property of the function $\breve{m}(\Sigma, z)$. In particular, we write

$$\begin{aligned}&\left|\left\langle A\left[\left(-\breve{m}_X z \Sigma_X - z\right)^{-1} - \left(-\breve{m}_{\text{lin}} z \Sigma_X - z\right)^{-1}\right]\right\rangle\right| \\ &= \left|\left\langle A\left[\left(-\breve{m}_X z \Sigma_X - z\right)^{-1} (z(\breve{m}_X - \breve{m}_{\text{lin}})\Sigma_X) \left(-\breve{m}_{\text{lin}} z \Sigma_X - z\right)^{-1}\right]\right\rangle\right| \\ &\leq \frac{|\breve{m}_X - \breve{m}_{\text{lin}}|}{|z|^2\sqrt{n}} \langle AA^* \rangle^{1/2} \left\|\left(\breve{m}_X \Sigma_X + I\right)^{-1}\right\| \left\|\left(\breve{m}_{\text{lin}} \Sigma_X + I\right)^{-1}\right\| \|\Sigma_X\|_F \\ &\leq |z|^{-2} |\breve{m}_X - \breve{m}_{\text{lin}}| \langle AA^* \rangle^{1/2}.\end{aligned} \tag{97}$$

Now, we analyze the difference between $\breve{m}_X$ and $\breve{m}_{\text{lin}}$. According to (95), we can write

$$\begin{aligned}\Delta := |\breve{m}_X - \breve{m}_{\text{lin}}| &= \frac{d}{|z|n^2} \operatorname{Tr}\left[(\breve{m}_X \Sigma_X + I)^{-1} - (\breve{m}_{\text{lin}} \Sigma_{\text{lin}} + I)^{-1}\right] \\ &\lesssim \frac{1}{|z|n} \operatorname{Tr}\left[(\breve{m}_X \Sigma_X + I)^{-1} (\breve{m}_{\text{lin}} \Sigma_{\text{lin}} - \breve{m}_X \Sigma_X)(\breve{m}_X \Sigma_{\text{lin}} + I)^{-1}\right] \\ &\leq \frac{1}{|z|n} \left\|(\breve{m}_X \Sigma_X + I)^{-1}\right\|_F \|\breve{m}_{\text{lin}} \Sigma_{\text{lin}} - \breve{m}_X \Sigma_X\|_F \left\|(\breve{m}_{\text{lin}} \Sigma_{\text{lin}} + I)^{-1}\right\| \\ &\leq \frac{1}{|z|\sqrt{n}} \|\breve{m}_{\text{lin}} \Sigma_{\text{lin}} - \breve{m}_X \Sigma_X\|_F \\ &\leq \frac{1}{|z|\sqrt{n}} \|\breve{m}_{\text{lin}} \Sigma_{\text{lin}} - \breve{m}_{\text{lin}} \Sigma_X\|_F + \frac{d}{|z|n^{3/2}} \|\breve{m}_{\text{lin}} \Sigma_X - \breve{m}_X \Sigma_X\|_F \\ &= \frac{|\breve{m}_{\text{lin}}|}{|z|\sqrt{n}} \|\Sigma_{\text{lin}} - \Sigma_X\|_F + \frac{\|\Sigma_X\|}{|z|\sqrt{n}} \Delta.\end{aligned} \tag{98}$$

Since $\|\Sigma_X\||z|^{-1}n^{-1/2} \ll 1$, we obtain using (94) that $|\breve{m}_X - \breve{m}_{\text{lin}}| \lesssim |z|^{-1}n^{-1/2}$, and thus, for the second approximation,

$$\left|\left\langle A\left[\left(-\breve{m}_X z \Sigma_X - z\right)^{-1} - \left(-\breve{m}_{\text{lin}} z \Sigma_X - z\right)^{-1}\right]\right\rangle\right| \lesssim \frac{1}{\delta^3\sqrt{n}} \langle AA^* \rangle^{1/2}. \tag{99}$$

For the third approximation, we can write

$$\begin{aligned}
&\left|\left\langle A\Big[\Big(-\breve{m}_{\rm lin} z\Sigma_X - z\Big)^{-1} - \Big(-\breve{m}_{\rm lin} z\Sigma_{\rm lin} - z\Big)^{-1}\Big]\right\rangle\right| = \frac{1}{|z||\breve{m}_{\rm lin}|} B, \\
\text{where}\quad B &:= \left|\langle A(\Sigma_X + 1/\breve{m})^{-1} - A(\Sigma_{\rm lin} + 1/\breve{m})^{-1}\rangle\right| \\
&\le \frac{1}{\sqrt{n}}\langle AA^*\rangle^{1/2}\left\|(\Sigma_X + 1/\breve{m})^{-1} - (\Sigma_{\rm lin} + 1/\breve{m})^{-1}\right\|_{\rm F} \\
&= \frac{1}{\sqrt{n}}\langle AA^*\rangle^{1/2}\left\|(\Sigma_X + 1/\breve{m})^{-1}(\Sigma_{\rm lin} - \Sigma_X)(\Sigma_{\rm lin} + 1/\breve{m})^{-1}\right\|_{\rm F} \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^2\sqrt{n}},
\end{aligned} \tag{100}$$

where in the last inequality we used (94).

Combining all the approximations together, we have proved that

$$\left|\left\langle A\Big[\Big(\frac{X_1^\top X_1}{k_1} - z\Big)^{-1} - \Big(-\breve{m}_{\rm lin} z\Sigma_{lin} - z\Big)^{-1}\Big]\right\rangle\right| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}}. \tag{101}$$

Next, we will verify that Assumption A.8 holds true when we replace matrix X_0 by X_1 and r_1 by r_2. In particular, we want to show that, with high probability,

$$\left\|\frac{X_1^\top X_1}{k_1} - r_2 I\right\|_{\max} = O\left(\frac{1}{\sqrt{n}}\right). \tag{102}$$

Note that Equations (92, 93) show that

$$\|\Sigma_X - r_2 I\|_{\max} = O\left(\frac{1}{\sqrt{n}}\right). \tag{103}$$

We have that

$$\left(\frac{X_1^\top X_1}{k_1}\right)_{ij} = \frac{1}{k_1}\sum_{l=1}^{k_1}\sigma\left(\frac{x_i^\top \tilde{w}_l}{\sqrt{d}}\right)\sigma\left(\frac{\tilde{w}_l^\top x_j}{\sqrt{d}}\right) = \frac{1}{k_1}\sum_{l=1}^{k_1} Y_l, \quad \text{where } Y_l := \sigma\left(\frac{x_i^\top \tilde{w}_l}{\sqrt{d}}\right)\sigma\left(\frac{\tilde{w}_l^\top x_j}{\sqrt{d}}\right). \tag{104}$$

Note that Y_l are independent random variables and from Lemma A.14 it follows that the subgaussian norm of Y_l is $O(\lambda_\sigma\left\|X/\sqrt{d}\right\|)$. Therefore, from Hoeffding inequality, we have that

$$P\left(\left|\left(\frac{X_1^\top X_1}{k_1}\right)_{ij} - (\Sigma_X)_{ij}\right| \ge t\right) \le 2e^{-\frac{ct^2 k_1}{\lambda_\sigma\|X/\sqrt{d}\|}}, \tag{105}$$

from which, applying union bound, we can deduce that

$$\left\|\frac{X_1^\top X_1}{k_1} - \Sigma_X\right\|_{\max} = O\left(\frac{1}{\sqrt{n}}\right). \tag{106}$$

Combining Equations (103) and (106) we get the required maximum norm bound. Next, with a standard ε-net argument (see, e.g. [46], Proposition 3.4) we can show that

$$\frac{1}{\sqrt{d}}\|X_1 - \mathbf{E}\,X_1\| \prec 1. \tag{107}$$

Since $\sqrt{n}\|\mathbf{E}\,X_1\|_{\max} \lesssim 1$ it follows that

$$\left\|\frac{\mathbf{E}\,X_1}{\sqrt{d}}\right\| \le \sqrt{\frac{nk_1}{d}}\|\mathbf{E}\,X_1\|_{\max} \lesssim \sqrt{\frac{k_1}{d}} \lesssim 1. \tag{108}$$

Finally, the claim that $\operatorname{dist}(-1/\breve{m}(z), \mathbb{R}_+) \ge \operatorname{dist}(z, \mathbb{R}_+)$ follows elementarily from the fixed point equation, see e.g. Proposition 6.2 in [46]. □

Proof of Theorem 3.3. This follows directly from iteratively applying Proposition 3.1 until we reach

$$\Big(\frac{X_\ell^\top X_\ell}{k_\ell} - z_\ell\Big)^{-1} \approx c_1 \cdots c_\ell\Big(\frac{X_0^\top X_0}{d} - z_0\Big)^{-1} \tag{109}$$

in the last layer, where "≈" is to be understood in the sense of Proposition 3.1. Now, using that $X_0 X_0^\top/n$ is a sample covariance matrix with population covariance matrix Ω_0, it follows that

$$\Big(\frac{X_0^\top X_0}{d} - z_0\Big)^{-1} = \frac{d}{n}\Big(\frac{X_0^\top X_0}{n} - \frac{d}{n} z_0\Big)^{-1} \approx \frac{d}{n}\Big(\frac{d}{n} m_{\mu(\Omega_0)\boxtimes \mu_{\mathrm{MP}}^{d/n}}\Big(\frac{d}{n} z_0\Big) + \frac{d-n}{d z_0}\Big), \tag{110}$$

where we used Proposition A.4 once more in the final step. □

B Closed-form formulae for population covariances

B.1 Multi-Layer linearization

In this Appendix, we provide a (heuristic) derivation of closed-form expressions for the population covariances:

$$\Omega_L \coloneqq \mathbf{E}\left[\varphi(\mathrm{x})\varphi(\mathrm{x})^\top\right], \qquad \Phi_{L^\star L} \coloneqq \mathbf{E}\left[\varphi^\star(\mathrm{x})\varphi(\mathrm{x})^\top\right], \qquad \Psi_{L^\star} \coloneqq \mathbf{E}\left[\varphi^\star(\mathrm{x})\varphi^\star(\mathrm{x})^\top\right]. \tag{111}$$

This derivation has appeared in [12], and we include it here for the sake of completeness.

Reminder of the results Consider the dRF (2) and target (25), with data $x \sim \mathcal{N}(0, \Omega_0)$. Ω_0 is assumed to possess extensive Frobenius norm and trace, i.e. there exists constant c, c' so that asymptotically (noting $k_0 = d$)

$$c < \frac{1}{d}\operatorname{tr}\Omega_0^2 = \frac{1}{d}||\Omega_0||_F^2 < c' < \infty, \qquad c < \frac{1}{d}\operatorname{tr}\Omega_0 < c' < \infty. \tag{112}$$

In terms of the limiting spectral density μ, these assumptions imply that the first and second moments are finite and non zero. Consider the sequence of variances defined by the recurrence

$$r_{\ell+1}^{(\star)} = \Delta_{\ell+1}^{(\star)} \mathbb{E}_z^{\mathcal{N}(0,r_\ell)}\left[\sigma_\ell^{(\star)}(z)^2\right] \tag{113}$$

with the initial condition

$$r_1^{(\star)} = \Delta_1^{(\star)} \frac{1}{d}\operatorname{tr}\Omega_0 \tag{114}$$

and the GET [5, 36, 52] coefficients

$$\kappa_1^{\ell(\star)} = \frac{1}{r_\ell^{(\star)}} \mathbb{E}_z^{\mathcal{N}(0,r_\ell)}\left[z\sigma_\ell^{(\star)}(z)\right] \qquad \kappa_*^{\ell(\star)} = \sqrt{\mathbb{E}_z^{\mathcal{N}(0,r_\ell)}\left[\sigma_\ell^{(\star)}(z)^2\right] - r_\ell^{(\star)}\left(\kappa_1^{\ell(\star)}\right)^2}. \tag{115}$$

Define the sequence of matrices

$$\Omega_{\ell+1}^{\mathrm{lin}} = \kappa_1^{\ell 2} \frac{W_{\ell+1}\Omega_\ell^{\mathrm{lin}} W_{\ell+1}^\top}{k_\ell} + \kappa_*^{\ell 2} I_{k_{\ell+1}} \tag{116}$$

$$\Psi_{\ell+1}^{\mathrm{lin}} = \kappa_1^{\star\ell 2} \frac{W_{\ell+1}^\star \Psi_\ell^{\mathrm{lin}} W_{\ell+1}^{\star\top}}{k_\ell^\star} + \kappa_*^{\star\ell 2} I_{k_{\ell+1}^\star} \tag{117}$$

with initialization

$$\Omega_0^{\mathrm{lin}} \coloneqq \Psi_0^{\mathrm{lin}} = \Omega_0, \tag{118}$$

and the matrix

$$\Phi_{\ell^\star \ell}^{\mathrm{lin}} = \prod_{r=1}^{\ell}\prod_{s=1}^{\ell^\star} \kappa_1^r \kappa_1^{s\star} \times \frac{W_{\ell^\star} \cdot \ldots \cdot W_1^\star \cdot \Sigma \cdot W_1^\top \cdot \ldots \cdot W_\ell^\top}{\prod_{r=0}^{\ell-1}\prod_{s=0}^{\ell^\star-1}\sqrt{k_r k_s^\star}}. \tag{119}$$

Then $\Omega_L \approx \Omega_L^{\mathrm{lin}}, \Psi_{L^\star} \approx \Psi_{L^\star}$ and $\Phi_{\ell^\star \ell} \approx \Phi_{\ell^\star \ell}^{\mathrm{lin}}$. $A \approx B$ is understood as $||A-B||_F^2/d = \mathcal{O}(1/d)$.

Example for $L = 2$ We give for concreteness an example for $L^\star = 1, L = 2$ (RF teacher, 2-layer DRN student). The recursions (116)(119) for the student reads for $L = 2$

$$\Omega_2 = (\kappa_1^1)^2(\kappa_1^2)^2 \frac{W_2 W_1 \Sigma W_1^\top W_2^\top}{k_1 d} + (\kappa_1^2)^2(\kappa_*^1)^2 \frac{W_2 W_2^\top}{k_1} + (\kappa_*^2)^2 I_{k_1} \tag{120}$$

$$\Psi_1 = (\kappa_1^{1\star})^2 \frac{W_1^\star \Sigma W_1^{\star\top}}{d} + (\kappa_*^{1\star})^2 I_{k_1^\star} \tag{121}$$

$$\Phi_{1,2} = \kappa_1^1 \kappa_1^2 \kappa_1^{1\star} \frac{W_1^\star \Sigma W_2^\top W_1^\top}{d\sqrt{k_1}} \tag{122}$$

Equivalent Linear Net Note that the linearization means one can think of the ℓ-th layer as a noisy linear layer,

$$\varphi_\ell(x)^{\text{lin}} \approx \kappa_1^\ell \frac{1}{\sqrt{k_{\ell-1}}} W_\ell \cdot x + \kappa_*^\ell \xi_\ell \tag{123}$$

with $\xi_\ell \in \mathbb{R}^{k_\ell}$ an i.i.d Gaussian noise indepent layer from layer, and also independent between the teacher and student *provided the teacher and student weights are drawn independently*. Similarly for the teacher:

$$\varphi_\ell^\star(x) \approx \kappa_1^{\star\ell} \frac{1}{\sqrt{k_{\ell-1}^\star}} W_\ell^\star \cdot x + \kappa_*^{\star\ell} \xi_\ell^\star \tag{124}$$

This provides a simple way to rederive the relations (116) and (119).

B.2 Derivation sketch for Ω_L

We first derive a relation between the covariance of the post-activations at two successive layers, and then iterate. Remark that since the computation for $\Psi_{L^\star}$ is identical *mutatis mutandis*, we only address here Ω_L.

Propagation through a single layer Consider the auxiliary single-layer problem

$$h(x) = \sigma\left(\frac{1}{\sqrt{d}} W \cdot x\right) \tag{125}$$

with $x \sim \mathcal{N}(0, \Sigma)$. Suppose recursively that Σ statisfies the properties (112). The population covariance of the post-activations h reads

$$\Omega_{ij} = \langle h_i(x) h_j(x) \rangle_x = \int \frac{e^{-\frac{1}{2}\begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} \frac{w_i^\top \Sigma w_i}{d} & \frac{w_i^\top \Sigma w_j}{d} \\ \frac{w_i^\top \Sigma w_j}{d} & \frac{w_j^\top \Sigma w_j}{d} \end{pmatrix}^{-1} \begin{pmatrix} u \\ v \end{pmatrix}}}{\sqrt{\det 2\pi \begin{pmatrix} \frac{w_i^\top \Sigma w_i}{d} & \frac{w_i^\top \Sigma w_j}{d} \\ \frac{w_i^\top \Sigma w_j}{d} & \frac{w_j^\top \Sigma w_j}{d} \end{pmatrix}}} \sigma(u)\sigma(v). \tag{126}$$

Note that have

$$\mathbb{E}_w \frac{w^\top \Sigma w}{d} = \frac{\Delta}{d} \operatorname{tr} \Sigma \equiv r, \tag{127}$$

which by assumption is of order 1. Diagonalizing $\Sigma = U \Lambda U^\top$ and noting that $U^\top w$ is still Gaussian with independent entries,

$$\mathbb{V}_w \left[\frac{w^\top \Sigma w}{d}\right] = \frac{1}{d^2} \sum_{i=1}^d \lambda_i^2 \mathbb{V}_w \left[(U^\top w)_i^2\right] = \frac{2\Delta}{d^2} \operatorname{tr} \Sigma^2 = \frac{2\Delta}{d} \frac{||\Sigma||_F^2}{d} = \mathcal{O}\left(\frac{1}{d}\right) \tag{128}$$

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 $i \neq j$:

$$\begin{aligned} \Omega_{ij} &= \int \frac{e^{-\frac{1}{2}\frac{1}{r^2 - \mathcal{O}\left(\frac{1}{d}\right)}(ru^2 + rv^2)} e^{\frac{1}{r^2 - \mathcal{O}\left(\frac{1}{d}\right)} \frac{w_i^\top \Sigma w_j}{d} uv}}{2\pi \sqrt{r^2 - \mathcal{O}\left(\frac{1}{d}\right)}} \sigma(u)\sigma(v) \\ &= \left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}} \sigma(z)\right)^2 + \frac{1}{r} \frac{w_i^\top \Sigma w_j}{d} \left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}} z\sigma(z)\right)^2 + \mathcal{O}\left(\frac{1}{d}\right) \\ &= \kappa_1^2 \times \frac{w_i^\top \Sigma w_j}{d}. \end{aligned} \tag{129}$$

on the diagonal ($i = j$), this becomes

$$\Omega_{ii} = \int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}} \sigma(z)^2 = \kappa_*^2 + r\kappa_1^2 \tag{130}$$

yielding

$$\Omega = \kappa_1^2 \frac{W\Sigma W^\top}{d} + \kappa_*^2 I_k \tag{131}$$

with

$$\kappa_1 = \frac{1}{r}\mathbb{E}_z^{\mathcal{N}(0,r)}\left[z\sigma(z)\right] \qquad \kappa_*^2 = \mathbb{E}_z^{\mathcal{N}(0,r)}\left[\sigma(z)^2\right] - r \times \kappa_1^2 \tag{132}$$

This extends the GET [5] generalization used in [56] to arbitrary input covariances.

Iterating layer to layer (115) and (116) follow by straightforward recursion from the single-layer results (132) and (131). One just need to connect (113) to the single-layer variance r (127).

$$\begin{aligned} r_{\ell+1} &= \Delta_{\ell+1}\frac{1}{k_\ell}\operatorname{tr}\Omega_\ell \\ &= \Delta_{\ell+1}\left(\frac{1}{k_\ell}\left(\kappa_1^\ell\right)^2\operatorname{tr}\left[\frac{W_\ell\Omega_{\ell-1}W_\ell^\top}{k_{\ell-1}}\right] + \left(\kappa_*^\ell\right)^2\right) \\ &= \Delta_{\ell+1}\left(\left(\kappa_1^\ell\right)^2 r_\ell + \left(\kappa_*^\ell\right)^2\right) \\ &= \Delta_{\ell+1}\mathbb{E}_z^{\mathcal{N}(0,r_\ell)}\left[\sigma_\ell(z)^2\right] \end{aligned} \tag{133}$$

We used

$$\begin{aligned} \frac{1}{k_\ell}\operatorname{tr}\left[\frac{W_\ell\Omega_{\ell-1}W_\ell^\top}{k_{\ell-1}}\right] &= \frac{1}{k_{\ell-1}}\sum_{i=1}^{k_{\ell-1}}\lambda_i^{\ell-1}\frac{1}{k_\ell}\left(U^\top W_\ell^\top W_\ell U\right)_{ii} \\ &= \frac{1}{k_{\ell-1}}\sum_{i=1}^{k_{\ell-1}}\lambda_i^{\ell-1}\Delta_\ell \\ &= \Delta_\ell\frac{1}{k_{\ell-1}}\operatorname{tr}\Omega_{\ell-1} = r_\ell \end{aligned} \tag{134}$$

We used that $W_\ell U$ is also an i.i.d Gaussian matrix. 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

$$\frac{1}{k}||\kappa_1^2\frac{W\Sigma W^\top}{d} + \kappa_*^2 I_k||_F^2 \geq \kappa_*^2 > 0. \tag{135}$$

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

$$\begin{aligned} \frac{1}{k}||\kappa_1^2\frac{W\Sigma W^\top}{d} + \kappa_*^2 I_k||_F^2 &\leq \frac{1}{k}||\kappa_1^2\frac{W\Sigma W^\top}{d}||_F^2 + \kappa_*^2 \\ &\leq \kappa_*^2 + \frac{||W||_F^4}{d^2}\frac{||\Sigma||_F^2}{k} \\ &\leq \kappa_*^2 + c' < \infty. \end{aligned} \tag{136}$$

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

$$\frac{1}{k}\operatorname{Tr}\left[\kappa_1^2\frac{W\Sigma W^\top}{d} + \kappa_*^2 I_k\right] = \kappa_*^2 + \frac{\kappa_1^2}{kd}\operatorname{Tr}\left[\Sigma W^\top W\right]. \tag{137}$$

Bounding

$$0 \leq \frac{\kappa_1^2}{kd}\operatorname{Tr}\left[\Sigma W^\top W\right] = \frac{\kappa_1^2}{kd}\sum_{i=1}^{k} w_i^\top\Sigma w_i = \kappa_1^2\frac{1}{d}\operatorname{Tr}\{\Sigma\} \leq \kappa_1^2 c', \tag{138}$$

where the last bound holds asymptotically almost surely.

B.3 Derivation sketch for $\Phi_{L^\star L}$

We now turn to the cross-covariance $\Phi_{L^\star L}$ between the post-activations of two random networks with independent weights. Again, we first establish a preliminary result, addressing the statistics of two correlated Gaussians propagating through non-linear layers with independently drawn weights.

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

$$(u, v) \sim \mathcal{N}\begin{pmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{pmatrix} \tag{139}$$

each independently propagated through a non-linear layer

$$h^\star(u) = \sigma_\star\left(\frac{1}{\sqrt{d_\star}} W_\star \cdot u\right), \qquad h(v) = \sigma\left(\frac{1}{\sqrt{d}} W \cdot v\right). \tag{140}$$

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

$$\Phi^h_{ij} = \langle h^\star_i(u) h_j(v) \rangle_{u,v} = \int \frac{e^{-\frac{1}{2}\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} \frac{w_i^{\star\top}\Sigma w_i^\star}{d_\star} & \frac{w_i^{\star\top}\Sigma w_j}{\sqrt{d_\star d}} \\ \frac{w_i^{\star\top}\Sigma w_j}{\sqrt{d_\star d}} & \frac{w_j^\top \Sigma w_j}{d} \end{pmatrix}^{-1}\begin{pmatrix} x \\ y \end{pmatrix}}}{\sqrt{\det 2\pi \begin{pmatrix} \frac{w_i^{\star\top}\Sigma w_i^\star}{d_\star} & \frac{w_i^{\star\top}\Sigma w_j}{\sqrt{d_\star d}} \\ \frac{w_i^{\star\top}\Sigma w_j}{\sqrt{d_\star d}} & \frac{w_j^\top \Sigma w_j}{d} \end{pmatrix}}} \sigma_\star(x)\sigma(y) \tag{141}$$

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

$$r_\star \equiv \frac{\Delta_\star}{d_\star} \operatorname{tr} \Psi \qquad\qquad r \equiv \frac{\Delta}{d} \operatorname{tr} \Omega \tag{142}$$

Plugging these definitions into the above:

$$\begin{aligned}
\Phi^h_{ij} &= \int \frac{e^{-\frac{1}{2}\frac{1}{r_\star r - \mathcal{O}\left(\frac{1}{d}\right)}(r x^2 + r_\star y^2)} e^{\frac{1}{r_\star r - \mathcal{O}\left(\frac{1}{d}\right)}\frac{w_i^{\star\top}\Phi w_j}{\sqrt{d_\star d}} xy}}{2\pi\sqrt{r_\star r - \mathcal{O}\left(\frac{1}{d}\right)}} \sigma_\star(x)\sigma(y) \\
&= \left(\int \frac{e^{-\frac{1}{2r_\star}z^2}}{\sqrt{2\pi r_\star}} \sigma_\star(z)\right)\left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}} \sigma(z)\right) + \frac{1}{r_\star r}\frac{w_i^{\star\top}\Phi w_j}{\sqrt{d_\star d}}\left(\int \frac{e^{-\frac{1}{2r_\star}z^2}}{\sqrt{2\pi r_\star}} z\sigma_\star(z)\right)\left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}} z\sigma(z)\right) + \mathcal{O}\left(\frac{1}{d}\right) \\
&:= \kappa_1 \kappa_1^\star \times \frac{w_i^{\star\top}\Phi w_j}{\sqrt{d_\star d}}
\end{aligned} \tag{143}$$

yielding

$$\Phi^h = \kappa_1 \kappa_1^\star \frac{W_\star \Phi W^\top}{\sqrt{d_\star d}} \tag{144}$$

with

$$\kappa_1 = \frac{1}{r}\mathbb{E}_z^{\mathcal{N}(0,r)}\left[z\sigma(z)\right] \qquad\qquad \kappa_1 = \frac{1}{r_\star}\mathbb{E}_z^{\mathcal{N}(0,r_\star)}\left[z\sigma_\star(z)\right] \tag{145}$$

One Gaussian propagating through one layer We will need another result, addressing again two correlated Gaussians, with only one propagating through a non-linear layer. Consider two jointly Gaussian variables $u \in \mathbb{R}^{d_\star},\ v \in \mathbb{R}^d$

$$(u, v) \sim \mathcal{N}\begin{pmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{pmatrix} \tag{146}$$

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

$$h(v) = \sigma\left(\frac{1}{\sqrt{k}} W \cdot v\right). \tag{147}$$

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

$$\Phi^h_{ij} = \langle u_i h_j(v)\rangle_{u,v} = \int \frac{e^{-\frac{1}{2}\begin{pmatrix} x & y\end{pmatrix}\begin{pmatrix} \Psi_{ii} & \frac{\Phi_i w_j}{\sqrt{k}} \\ \frac{\Phi_i w_j}{\sqrt{k}} & \frac{w_j^\top \Sigma w_j}{k}\end{pmatrix}^{-1}\begin{pmatrix} x \\ y\end{pmatrix}}}{\sqrt{\det 2\pi \begin{pmatrix} \Psi_{ii} & \frac{\Phi_i w_j}{\sqrt{k}} \\ \frac{\Phi_i w_j}{\sqrt{k}} & \frac{w_j^\top \Sigma w_j}{k}\end{pmatrix}}} x\sigma(y) \tag{148}$$

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

$$r \equiv \frac{\Delta}{k}\operatorname{tr}\Omega \tag{149}$$

Plugging this definition into the above:

$$\begin{aligned}\Phi^h_{ij} &= \int \frac{e^{-\frac{1}{2}\frac{1}{\Psi_{ii}r-\mathcal{O}\left(\frac{1}{d}\right)}(rx^2+\Psi_{ii}y^2)} e^{\frac{1}{\Psi_{ii}r-\mathcal{O}\left(\frac{1}{d}\right)}\frac{\Phi_i w_j}{\sqrt{k}}xy}}{2\pi\sqrt{\Psi_{ii}r-\mathcal{O}\left(\frac{1}{d}\right)}} x\sigma(y) \\ &= \left(\int \frac{e^{-\frac{1}{2\Psi_{ii}}z^2}}{\sqrt{2\pi\Psi_{ii}}}z\right)\left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}}\sigma(z)\right) + \frac{1}{\Psi_{ii}r}\frac{\Phi_i w_j}{\sqrt{k}}\left(\int \frac{e^{-\frac{1}{2\Psi_{ii}}z^2}}{\sqrt{2\pi\Psi_{ii}}}z^2\right)\left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}}z\sigma(z)\right) + \mathcal{O}\left(\frac{1}{d}\right) \\ &:= \kappa_1 \times \frac{\Phi_i w_j}{\sqrt{k}}\end{aligned} \tag{150}$$

yielding

$$\Phi^h = \kappa_1 \frac{\Phi W^\top}{\sqrt{k}} \tag{151}$$

with

$$\kappa_1 = \frac{1}{r}\mathbb{E}_z^{\mathcal{N}(0,r)}\left[z\sigma(z)\right]. \tag{152}$$

Iterating To establish (32), we iterate (144) $\min(L, L_\star)$ times, and followed by $\max(L, L_\star) - \min(L, L_\star)$ iterations of the single layer relation (151), so as to finish propagating the data through the deeper (teacher (25) or student (2)) network.

B.4 Spectrum of the covariances

In this section, we derive the spectrum of the linearized covariance (21), which is a result of indenpendent interest.

Useful identities We remind first some useful facts. For $W \in \mathbb{R}^{k_\ell \times k_{\ell-1}}$ with i.i.d Gaussian entries and $\Sigma \in \mathbb{R}^{k_{\ell-1}\times k_{\ell-1}}$ a deterministic matrix admitting a limiting spectral density $\mu_{\ell-1}$ as $k_{\ell-1} \to \infty$, we have, from the fact that $XX^\top$ and $X^\top X$ share the same spectrum up to a zero eigenvalues,

$$\mu_{\frac{1}{k_{\ell-1}}W\Sigma W^\top} = \frac{k_{\ell-1}}{k_\ell} \times \left[\frac{k_\ell}{k_{\ell-1}} \otimes \mu_{\frac{1}{k_\ell}\Sigma^{\frac{1}{2}}W^\top W\Sigma^{\frac{1}{2}}}\right] + \frac{k_\ell - k_{\ell-1}}{k_\ell}\delta \tag{153}$$

The spectrum of $\frac{1}{k_\ell}\Sigma^{\frac{1}{2}}W^\top W\Sigma^{\frac{1}{2}}$ is given by

$$\mu_{\mathrm{MP}}^{\frac{k_{\ell-1}}{k_\ell}} \boxtimes \mu_{\ell-1} \tag{154}$$

where $\mu_{\mathrm{MP}}^{\frac{k_{\ell-1}}{k_\ell}}$ is the Marcenko-Pastur distribution with aspect ratio $k_{\ell-1}/k_\ell$. In terms of Stieltjes transforms:

$$m_{\frac{1}{k_{\ell-1}}W\Sigma W^\top}(z) = \left(\frac{k_{\ell-1}}{k_\ell}\right)^2 \times m_{\frac{1}{k_\ell}\Sigma^{\frac{1}{2}}W^\top W\Sigma^{\frac{1}{2}}}\left(\frac{k_{\ell-1}}{k_\ell}z\right) + \left(\frac{k_{\ell-1}}{k_\ell} - 1\right)\frac{1}{z} \tag{155}$$

Using the Marcenko-Pastur map and using the shorthand $\gamma_\ell = \frac{k_{\ell-1}}{k_\ell}$, we reach that the Stieltjes transform for $\frac{1}{k_{\ell-1}} W\Sigma W^\top$ is the solution of

$$m(z) = \int \frac{(\gamma_\ell - 1)xm(z) - \gamma_\ell}{zxm(z) + \gamma_\ell z} d\mu_{\ell-1}(x) = \frac{\gamma_\ell - 1}{z} - \frac{\gamma_\ell^2}{z} \int \frac{1}{xm(z) + \gamma_\ell} d\mu_{\ell-1}(x) \tag{156}$$

Spectrum Ω_ℓ^{lin} The spectral distribution μ_ℓ of Ω_ℓ^{lin} is then given by the recursion relation

$$\mu_\ell = \left(\kappa_1^\ell\right)^2 \otimes \left[\frac{k_{\ell-1}}{k_\ell} \times \left[\frac{k_\ell}{k_{\ell-1}} \otimes \mu_{\text{MP}}^{\frac{k_{\ell-1}}{k_\ell}} \boxtimes \mu_{\ell-1} \right] + \frac{k_\ell - k_{\ell-1}}{k_\ell} \delta \right] \oplus \left(\kappa_*^\ell\right)^2 \tag{157}$$

with initial condition $\mu_0 = \mu_{\Omega_0}$. This translates to

$$m_\ell(z) = \frac{\gamma_\ell - 1}{z - \left(\kappa_*^\ell\right)^2} - \frac{\gamma_\ell^2 \left(\kappa_1^\ell\right)^2}{(z - \left(\kappa_*^\ell\right)^2) m_\ell(z)} m_{\ell-1}\left(- \frac{\gamma_\ell}{\left(\kappa_1^\ell\right)^2 m_\ell(z)} \right). \tag{158}$$

Numerical scheme We now discuss a numerical scheme to solve (158). Note that each $m_{\ell-1}$ is only evaluated at

$$z_{\ell-1} \equiv - \frac{\gamma_\ell}{\left(\kappa_1^\ell\right)^2 m_\ell(z)} \tag{159}$$

To solve this numerically we keep two arrays $(m_0, ..., m_L)$ and $(z_0, ...z_L)$, with $m_\ell \equiv m_\ell(z_\ell)$. For simplicity consider the case where the input covariance is identity, meaning

$$m_0(z_0) = \frac{1}{1 - z_0} \tag{160}$$

Then until convergence we iterate

$$\forall 0 \le i \le \ell - 1, \qquad z_i \leftarrow - \frac{\gamma_{i+1}}{\left(\kappa_1^i\right)^2 m_{i+1}} \tag{161}$$

and keep

$$z_L = \lambda + i\eta \tag{162}$$

with $\eta = 0^+$ and λ the value at which we wish to evaluate the density $\mu_L(\lambda)$. Then we update

$$\forall 1 \le i \le L, \qquad m_i \leftarrow \frac{\gamma_i - 1 - \sqrt{(\gamma_i - 1)^2 - 4 \frac{m_{i-1}\gamma_i^2}{\left(\kappa_1^i\right)^2} \left(z_i - \left(\kappa_*^i\right)^2\right)}}{2\left(z_i - \left(\kappa_*^i\right)^2\right)} \tag{163}$$

where we solved the update (158) directly, which is empirically yielding better convergence than directly iterating (158).

Fig. 4 shows the theoretical asymptotic distribution (157) for $3-$layer RFs with sigmoid and sign activations, which is found to display excellent agreement with numerical estimations of the population covariance estimated with 10^5 independent samples. Fig. 5 shows the asymptotic distribution across $L = 5$ layers for a rectangular $\tanh$ network. In alignment to the observations of [8] for the conjugate kernel in similar models, the support of the distribution increases with depth, alongside an increase in the density of small eigenvalues. Note that the presence of small eigenvalues has been linked in a variety of settings [15, 57–59] to an effective additional implicit ℓ_2 regularization when using (2) to perform regression. This intuition is further discussed in Section 5.

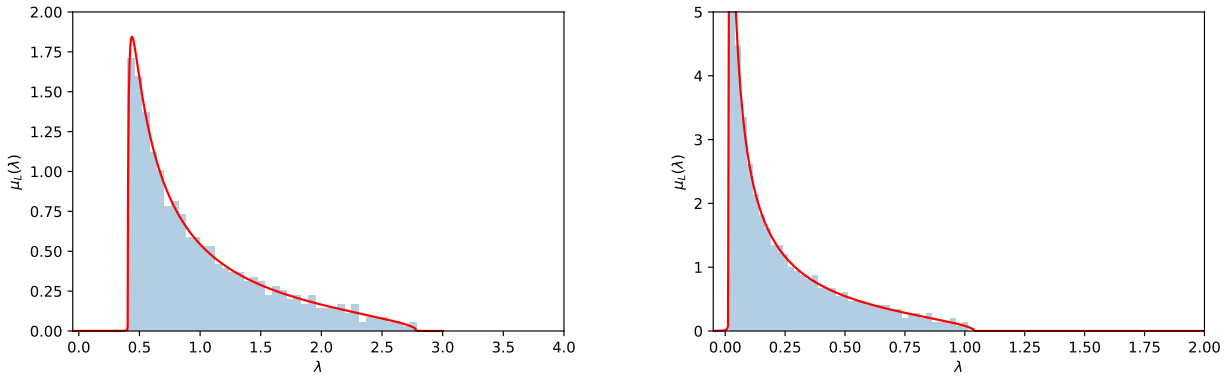

Figure 4: Limiting spectral distributions for the post-activation covariance Ω_2 (21) of a $2-$ hidden layers network (2), with architectures $\gamma_1 = 6/5, \gamma_2 = 3/5$ and activation $\sigma_1 = \sigma 2 = \tanh(2\cdot)$ (top), and $\gamma_1 = 7/10, \gamma_2 = 6/5$ and activation $\sigma_1 = \sigma 2 = \text{sign}$ (bottom) (red) Theoretical asymptotic spectral distribution obtained from solving the recursion (157) (see Appendix B for further details on the numerical scheme) (blue) Empirical distribution, estimated from the sample covariance of 10^5 samples, in dimension $d = 1000$.

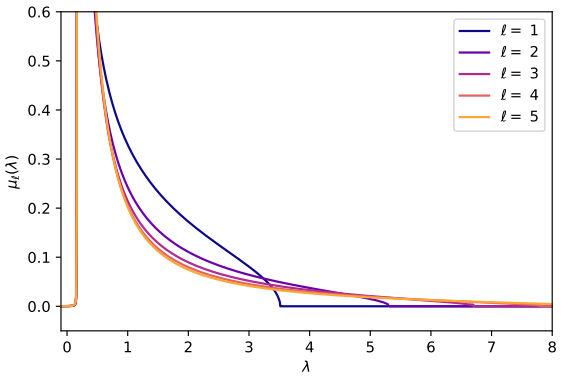

Figure 5: Evolution of the asymptotic spectral distribution μ_ℓ of the post-activations $h_\ell(\mathrm{x})$, for $1 \leq \ell \leq L = 5$, for a network with architecture $\gamma_1 = ... = \gamma_5 = 1$ and $\sigma_1 = ...\sigma_5 = \tanh$ activation, and isotropic data $\Omega_0 = I_d$. Propagation through non-linear layers tends to extend the support of the distribution, and also increase the density of small eigenvalues.

C Error universality of ridge regression

In this Appendix we provide a detailed derivation of (27) and Corollary 4.1. First, we start by recapping the setting for this Corollary. Here, we are interested in characterizing the asymptotic mean-squared test error:

$$\mathcal{E}_{\text{gen.}}(\hat{\theta}) = \mathbb{E}\left(y - \frac{\hat{\theta}^\top \varphi(\mathrm{x})}{\sqrt{k}}\right)^2 \tag{164}$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}^k$ are the L-layers random features defined in (2) and $\hat{\theta} \in \mathbb{R}^k$ is the ridge estimator:

$$\begin{aligned}\hat{\theta} &= \text{argmin}\left[\sum_{\mu=1}^{n}\left(y^\mu - \frac{\hat{\theta}^\top \varphi(\mathrm{x}^\mu)}{\sqrt{k}}\right)^2 + \frac{\lambda}{2}||\theta||_2^2\right] \\ &= \frac{1}{\sqrt{k}}\left(\lambda I_k + \frac{1}{k} X_L X_L^\top\right)^{-1} X_L y\end{aligned} \tag{165}$$

where, following the notation in the main, we have defined the features matrix $X_L \in \mathbb{R}^{k\times n}$ by stacking together $\varphi(\mathrm{x}^\mu)$ column-wise and the label vector $y \in \mathbb{R}^n$. In particular, in Corollary 4.1 we focus in the case where the labels are generated, up to additive Gaussian noise, by a L-layers random features target with the same architecture. Explicitly,

this can be written as:

$$y^\mu = \frac{\theta_\star^\top \varphi(\mathrm{x}^\mu)}{\sqrt{k}} + z^\mu \tag{166}$$

where $\theta_\star \sim \mathcal{N}(0_k, I_k)$ and $z^\mu \sim \mathcal{N}(0, \Delta)$ independently. Note that, for the purposes of the discussion here we do not need to assume the inputs $\mathrm{x}^\mu \in \mathbb{R}^d$ are Gaussian, but only that the data matrix $X_0 \in \mathbb{R}^{d\times n}$ satisfies the concentration condition (11). In particular, this implies that the results in this Appendix hold for the test error (164) conditionally on the training inputs X_0.

From here, the computation is standard, and closely follows other works deriving closed-form asymptotics for ridge regression under different assumptions, e.g. [50, 60–62]. First, note we can rewrite:

$$\begin{aligned}
\mathcal{E}_{\text{gen.}}(\hat{\theta}) &= \mathbb{E}_{\mathbf{z},\theta_\star,z,\mathrm{x}} \left(\frac{\theta_\star^\top \varphi(\mathrm{x})}{\sqrt{k}} + z - \frac{\hat{\theta}^\top \varphi(\mathrm{x})}{\sqrt{k}} \right)^2 \\
&\overset{(a)}{=} \frac{1}{k} \mathbb{E}_{,\mathbf{z},\theta_\star} (\hat{\theta} - \theta_\star)^\top \mathbb{E}_{\mathrm{x}} \left[\varphi(\mathrm{x}) \varphi(\mathrm{x})^\top \right] (\hat{\theta} - \theta_\star) + \Delta \\
&\overset{(b)}{=} \frac{1}{k} \mathbb{E}_{\mathbf{z},\theta_\star} \left[(\hat{\theta} - \theta_\star)^\top \Omega_L (\hat{\theta} - \theta_\star) \right]
\end{aligned} \tag{167}$$

where in (a) we used the independence of the test sample took the z average explicitly, and in (b) we have used the definition (87). Focusing on:

$$\begin{aligned}
\hat{\theta} - \theta_\star &= {}^1\!/\!\sqrt{k} \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} X_L^\top ({}^1\!/\!\sqrt{k} X_L^\top \theta_\star + \mathbf{z}) - \theta_\star \\
&= \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} \left({}^1\!/\!_k X_L X_L^\top - I_k \right) \theta_\star + {}^1\!/\!\sqrt{k} \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} X \, \mathbf{z} \qquad (168) \\
&\overset{(c)}{=} -\lambda \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} \theta_\star + {}^1\!/\!\sqrt{k} \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} X \, \mathbf{z} \qquad (169)
\end{aligned}$$

where in (c) we have used the following version of the Woodbury identity:

$$\lambda \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} = I_k - {}^1\!/\!_k \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} X_L X_L^\top \tag{170}$$

Inserting the above in (167):

$$\begin{aligned}
\mathcal{E}_{\text{gen.}}(\hat{\theta}) &= {}^{\lambda^2}\!/\!_k \, \mathbb{E}_{\theta_\star} \left[\theta_\star^\top \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} \Omega_L \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} \theta_\star \right] + \\
&\quad + {}^1\!/\!_{k^2} \, \mathbb{E}_{\mathbf{z}} \left[\mathbf{z}^\top X_L^\top \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} \Omega_L \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} X_L \, \mathbf{z} \right] + \Delta \\
&\overset{(d)}{=} \lambda^2 \left\langle \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} \Omega_L \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} \right\rangle + \\
&\quad + \Delta \left\langle {}^1\!/\!_k X_L X_L^\top \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} \Omega_L \left(\lambda I_k + {}^1\!/\!_k X_L X_L^\top \right)^{-1} \right\rangle + \Delta
\end{aligned} \tag{171}$$

where in (d) we took the expectations over the noise and target weights and used the definition $\langle \cdot \rangle \equiv {}^1\!/\!_k \operatorname{tr}(\cdot)$ with the cyclicity of the trace. We can put the expression above in a shape in which Theorem 3.3 apply by adding and subtracting λI_k to ${}^1\!/\!_k X_L X_L^\top$ the second trace term. This leads to the expression (27) quoted in the main text:

$$\begin{aligned}
\mathcal{E}_{\text{gen.}}(\hat{\theta}) &= \Delta \left(\left\langle \Omega_L \left(\lambda I_k + {}^1\!/\!_k X_L X_L \right)^{-1} \right\rangle + 1 \right) + \lambda(\lambda - \Delta) \left\langle \Omega_L \left(\lambda I_k + {}^1\!/\!_k X_L X_L \right)^{-2} \right\rangle \\
&= \Delta \left(\left\langle \Omega_L \left(\lambda I_k + {}^1\!/\!_k X_L X_L \right)^{-1} \right\rangle + 1 \right) - \lambda(\lambda - \Delta) \partial_\lambda \left\langle \Omega_L \left(\lambda I_k + {}^1\!/\!_k X_L X_L \right)^{-1} \right\rangle
\end{aligned} \tag{172}$$

Note that the last expression requires applying Theorem 3.3 to the derivative of the resolvent. In general, this can be justified by writing a squared resolvent $G(z)^2 = (H - z)^{-2}$ of some non-negative matrix $H \geq 0$ in terms of a Cauchy-integral

$$G(z)^2 = G'(z) = \frac{1}{2\pi \mathrm{i}} \oint_\gamma \frac{1}{(w - z)^2} G(w) \, \mathrm{d}w, \tag{173}$$

where γ is any contour around z not crossing $\mathbb{R}_+$. In this way some local law of the type $|\langle A(G - M) \rangle| \prec \epsilon \operatorname{dist}(z, \mathbb{R}_+)^{-k}$ can be transferred to the derivative as

$$\langle A(G'(z) - M'(z)) \rangle = \frac{1}{2\pi \mathrm{i}} \oint_\gamma \frac{1}{(w - z)^2} \langle A(G(w) - M(w)) \rangle \, \mathrm{d}w = O_\prec \Big(\frac{\epsilon}{\operatorname{dist}(z, \mathbb{R}_+)^{k+1}} \Big), \tag{174}$$

by choosing γ to be a small circle of radius $\mathrm{dist}(z, \mathbb{R}_+)/2$ around z, using that the deterministic equivalent is also holomorphic away from $\mathbb{R}_+$.

Therefore, in the high-dimensional limit where $n, k_\ell, d \to \infty$ at fixed ratios $\alpha = {}^{n}/_{d}$ and $\gamma_\ell = {}^{k_\ell}/_{d}$, under the assumptions of Theorem 3.3 for the input data $X_0 \in \mathbb{R}^{d\times n}$ (11) and the architecture of the deep random features and for $\lambda > 0$[5] we can apply Theorem 3.3 to write the asymptotic limit of the test error:

$$\lim_{k\to\infty} \mathcal{E}_{\text{gen.}}(\hat{\theta}) = \mathcal{E}^{\star}_{\text{gen.}}(\lambda, \Delta, \alpha, \gamma_\ell, \kappa_1^\ell, \kappa_\star^\ell) \equiv \Delta\left(\langle\Omega_L\rangle \breve{m}_L(-\lambda) + 1\right) - \lambda(\lambda - \Delta)\langle\Omega_L\rangle \partial_\lambda \breve{m}_L(-\lambda) \tag{175}$$

where $\breve{m}_L(z)$ can be computed recursively from (18) for a given regularization strength $\lambda > 0$, noise level $\Delta > 0$, sample complexity $\alpha > 0$ and features architecture $(\gamma_\ell, \sigma_\ell)_{\ell\in[L]}$. On the other hand, it follows from the recursion (21) that the trace of the last-layer covariance $\langle\Omega_L\rangle$ admits the compact expression

$$\langle\Omega_L\rangle = \sum_{\ell=1}^{L-1} \left(\kappa_*^\ell\right)^2 \prod_{\ell'=\ell+1}^{L} \left(\kappa_1^{\ell'}\right)^2 \Delta_{\ell'} + \left(\kappa_*^L\right)^2 + \frac{1}{d}\langle\Omega_0\rangle \prod_{\ell=1}^{L} \left(\kappa_1^\ell\right)^2 \Delta_\ell \tag{176}$$

in terms only of the coefficients (10).

Note that (175) agrees exactly with the formula for the asymptotic test error of ridge regression on a equivalent Gaussian dataset $\mathcal{D} = \{(v^\mu, y^\mu)\}_{\mu\in[n]}$:

$$y^\mu = \frac{\theta_\star^\top v^\mu}{\sqrt{k}} + z^\mu, \qquad\qquad v^\mu \sim \mathcal{N}(0_k, \Omega_L). \tag{177}$$

which, to our best knowledge, was first derived in [50]. This establishes the Gaussian universality of the asymptotic test error for this model.

C.1 Possible extensions

We now discuss some possible extensions of the universality result above. They require, however, a more involved analysis, which we leave for future work. Our goal here is simply to highlight other possible applications of our deterministic equivalent in Thm. 3.3.

Deterministic last-layer weights: The first extension is to generalize the result above to deterministic last layer weights $\theta_\star$. Indeed, [63] shows that for ridge regression on a deterministic target $y^\mu = {}^{1}/_{\sqrt{k}}\, \theta_\star^\top \varphi(\mathrm{x}^\mu)$[6], the test error can be asymptotically estimated from the *generalized cross-validation* (GCV) estimator, defined as:

$$\mathrm{GCV}_\lambda = \lambda \left\langle (\lambda I_k + \hat{\Omega}_L)^{-1} \right\rangle \mathcal{E}_{\text{train.}}(\hat{\theta}) \tag{178}$$

where $\hat{\Omega}_L = {}^{1}/_{n} X_L X_L^\top$ is the *sample covariance matrix* of the features and $\mathcal{E}_{\text{train.}}(\hat{\theta})$ is the training error associated to the ridge estimator:

$$\mathcal{E}_{\text{train.}}(\hat{\theta}) = \frac{1}{n}\sum_{\mu=1}^{n} \left(y^\mu - \frac{\hat{\theta}^\top \varphi(\mathrm{x}^\mu)}{\sqrt{k}}\right)^2 \tag{179}$$

In particular, it is shown that:

Theorem C.1 (Thm. 8 of [63]). *Assume that*

$$\left|\langle\left(\frac{X_L^\top X_L}{n} - z\right)^{-1}\rangle - \breve{m}(z)\right| + \left|v^\top\left[\left(\frac{X_L X_L^\top}{zn} - I\right)^{-1} + \left(\Omega_L \breve{m}(z) + 1\right)^{-1}\right] v\right| \prec \frac{\sqrt{\mathrm{Im}\,\breve{m}(z)}}{\sqrt{n\,\mathrm{Im}\,z}}, \tag{180}$$

for all deterministic vectors v with $v^\top \Omega_L v \le 1$, where

$$\breve{m}(z) = \frac{k_L - n}{nz} + \frac{k_L}{n} m_{\mu(\Omega_L)\boxtimes\mu_{\mathrm{MP}}^{k_L/n}}(z). \tag{181}$$

Then for all $\lambda > 0$ it holds that

$$\left|GCV_\lambda - \mathcal{E}_{\text{gen.}}(\hat{\theta})\right| \lesssim n^{-1/2+o(1)} \theta_\star^\top \Omega_L \theta_\star \left[\frac{||\Omega_L||_{\text{op.}}}{\lambda} + \left(\frac{\mathrm{tr}\,\Omega_L}{\lambda n}\right)^{3/2}\right] \tag{182}$$

[5] Technically, we don't need to assume the regularization is bounded away from here. It suffices to take it decaying slower than $n^{-1/18}$ for Thm. 3.3 to apply.

[6] For simplicity, we discuss the noiseless $\Delta = 0$ case here. See Appendix B of [63] for a discussion of noisy targets

Applying Theorem A.3 for fixed weights $W_1, \dots, W_L$ shows[7] that assumption (180) is satisfied in the proportional regime $k_L \sim n$, up to a worse z-dependence of the error $\mathrm{dist}(z, \mathbb{R}_+)^{-9}$ rather than $(\operatorname{Im} \breve{m}(z)/\operatorname{Im} z)^{1/2}$, and only for bounded vectors $\|v\| \lesssim 1$.

Instead, our result Theorem 3.3 proves a preliminary version of (180) with an explicit deterministic equivalent only depending on the input population covariance Ω_0 rather than the output population covariance Ω_L, at the price of having an error which is larger by a factor of $\sqrt{n}$. It is an interesting question whether our error rates can be improved to imply Equation (180) which is left for future work.

General case: As discussed in the introduction, in the general case we are interested in a target:

$$f_\star(\mathrm{x}^\mu) = \frac{1}{\sqrt{k_\star}} \theta_\star^\top \varphi_\star(\mathrm{x}^\mu), \qquad \theta_\star \sim \mathcal{N}(0_{k_\star}, I_{k_\star}), \tag{183}$$

where the $L_\star$ multi-layer random features $\varphi_\star : \mathbb{R}^d \to \mathbb{R}^{k_\star}$ are not necessarily the same as the L multi-layer random features $\varphi : \mathbb{R}^d \to \mathbb{R}^k$. As discussed in the introduction, this contains as a special case the *hidden-manifold model* (HMM), introduced in [52] as a model for structured high-dimensional data where the labels depend only on the coordinates of a lower-dimensional "latent space". While in Section 4 we provide an exact but heuristic formula to compute the error in this case (valid for arbitrary convex losses), the challenge in proving it with random matrix theory methods in the case of ridge regression comes from the fact that this is a mismatched model. Indeed, naively writing the expression for the test error in this case:

$$\begin{aligned} \mathcal{E}_{\text{gen.}}(\hat{\theta}) &= \mathbb{E}_{\theta_\star, \mathrm{x}} \left(\frac{\theta_\star^\top \varphi_\star(\mathrm{x})}{\sqrt{k_\star}} + z - \frac{\hat{\theta}^\top \varphi(\mathrm{x})}{\sqrt{k}} \right)^2 \\ &= \langle \Psi_{L_\star} \rangle + \frac{2}{\sqrt{k_\star k}} \mathbb{E}_{\theta_\star} \left[\theta_\star^\top \Phi_{L_\star L} \hat{\theta} \right] + \frac{1}{k} \mathbb{E}_{\theta_\star} \left[\hat{\theta}^\top \Omega_L \hat{\theta} \right] \end{aligned} \tag{184}$$

where we recall the reader of the definitions:

$$\Psi_\ell = \mathbb{E}\left[h_\ell(\mathrm{x}) h_\ell(\mathrm{x})^\top \right], \qquad \Phi_{\ell\ell'} = \mathbb{E}\left[h_\ell(\mathrm{x}) h_{\ell'}(\mathrm{x})^\top \right]. \tag{185}$$

Indeed, applying Thm. 3.3 to the expression above is not as straightforward as above. To see this, focus on the second term:

$$\mathbb{E}_{\theta_\star} \left[\theta_\star^\top \Phi_{L_\star L} \hat{\theta} \right] = \operatorname{tr}\left[\Phi_{L_\star L} (\lambda I_k + {}^1\!/_k X_L X_L^\top)^{-1} X_L^\top X_{L_\star} \right] \tag{186}$$

where we defined the target feature matrix $X_{L_\star} \in \mathbb{R}^{k_\star \times n}$ with columns given by $\varphi(\mathrm{x}^\mu) \in \mathbb{R}^{k_\star}$. This would, naively, require a more refined deterministic equivalent than Thm. 3.3 provides. Possible alternative approaches would be to rewrite the misspecification as an effective additive noise (e.g. as in Appendix B of [20]) and derive a local-law akin to Assumption 180 with a control over the noise (see Appendix B of [63] for a discussion) or to use the linear pencil method as in [4]. This provides an interesting avenue for future work.

[7] Technically this requires some argument that with high probability the deep RF model with quenched weights satisfies Lipschitz concentration with respect to X_0

D Exact asymptotics for the general case

In this appendix, we detail the sharp asymptotic characterization for the test error of the dRF (2) on a deep random network target (25), for regression (Fig. 1) and classification (Fig. 2).

The backbone of the derivation is the theorem of [11], which fully characterizes the test error of the GCM (29) in terms of the covariance matrices $\Psi_{L_\star}, \Omega_L$ and $\Phi_{L_\star L}$. In the original work of [11], these matrices for the dRF model had to be estimated numerically through a Monte-Carlo algorithm. In the present work however, the closed-form expressions afforded by (21), (31) and (32), which we remind in the next subsection, now afford a way to access fully analytical formulas. We successively detail these characterizations for ridge regression and logistic regression readouts.

D.1 Reminder of second-order statistics of network activations

Before providing detailed asymptotic characterizations for the test error of ridge and logistic regression, we first provide a reminder for the expressions of the linearized matrices $\Psi^{\rm lin}_{L_\star}, \Omega^{\rm lin}_L$ and $\Phi^{\rm lin}_{L_\star L}$ (21,31,30). Using conjecture 4.3, these matrices can then be used in the formulas of [11] to access fully analytical formulas for the test errors, in terms only of the target network weights (25) and the coefficients (10). The following expressions follow from expliciting the solution of the recursions (21,31,30)

$$\Omega^{\rm lin}_L = \left(\prod_{\ell'=1}^{L} \frac{\kappa_1^{\ell'} W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right)^\top \Omega_0 \left(\prod_{\ell'=1}^{L} \frac{\kappa_1^{\ell'} W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right) + \sum_{\ell=1}^{L-1} \left(\kappa_*^\ell\right)^2 \left(\prod_{\ell'=\ell+1}^{L} \frac{\kappa_1^{\ell'} W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right)^\top \Omega_0 \left(\prod_{\ell'=\ell+1}^{L} \frac{\kappa_1^{\ell'} W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right) + \left(\kappa_*^L\right)^2 I_{k_L} \tag{187}$$

$$\Psi^{\rm lin}_{L_\star} = \left(\prod_{\ell'=1}^{L_\star} \frac{\kappa_1^{\ell'\star} W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right)^\top \Omega_0 \left(\prod_{\ell'=1}^{L_\star} \frac{\kappa_1^{\ell'\star} W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right) + \sum_{\ell=1}^{L_\star-1} \left(\kappa_*^{\ell\star}\right)^2 \left(\prod_{\ell'=\ell+1}^{L_\star} \frac{\kappa_1^{\ell'\star} W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right)^\top \Omega_0 \left(\prod_{\ell'=\ell+1}^{L_\star} \frac{\kappa_1^{\ell'\star} W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right) + \left(\kappa_*^{L_\star\star}\right)^2 I_{k_{L_\star}} \tag{188}$$

$$\Phi^{\rm lin}_{LL_\star} = \left(\prod_{\ell=L^\star}^{1} \frac{\kappa_1^{\ell\star} W_\ell^\star}{\sqrt{k_\ell^\star}}\right) \cdot \Omega_0 \cdot \left(\prod_{\ell=1}^{L} \frac{\kappa_1^{\ell} W_\ell^\top}{\sqrt{k_\ell}}\right) \tag{189}$$

In the special case where the teacher has depth $L_\star = 0$ (i.e. possesses an architecture with no hidden layer), the above expression reduce to

$$\Psi^{\rm lin}_0 = \Omega_0 \tag{190}$$

$$\Phi^{\rm lin}_{LL_\star} = \Omega_0 \cdot \left(\prod_{\ell=1}^{L} \frac{\kappa_1^{\ell} W_\ell^\top}{\sqrt{k_\ell}}\right). \tag{191}$$

The $L_\star = 0, L = 1$ case has been studied in the literature [5, 52] as the *Hidden Manifold Model*. The present work encompasses the analysis of its generalization to deep learners with $L > 1$ hidden layers.

D.2 Ridge regression

We consider the supervised learning problem of training the readout weights θ of the dRF (2) on a dataset $\mathcal{D} = \{x^\mu, y^\mu\}_{\mu=1}^n$, with $x^\mu \sim \mathcal{N}(0_d, \Omega_0)$ independently. The labels are given by a deep random network

$$y^\mu = \frac{\theta_\star^\top \varphi_\star(x^\mu)}{\sqrt{k_{L_\star}}} + z^\mu, \tag{192}$$

where $z^\mu \sim \mathcal{N}(0, \Delta)$ is a Gaussian additive noise and the teacher feature map is

$$\varphi_\star(x) = \varphi^\star_{L_\star} \circ \ldots \circ \varphi^\star_1(x). \tag{193}$$

Note that compared to (25), we have adopted the notation $\theta_\star := W^\star_{L_\star+1} \in \mathbb{R}^{k_{L_\star}}$ for the sake of clarity. Defining

$$\rho := \frac{\theta_\star^\top \Psi_{L_\star} \theta_\star}{k_{L_\star}}, \tag{194}$$

We consider the problem training the last layer θ of the learner dRF (2) with ridge regression, by minimizing the risk

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \left\{ \sum_{\mu=1}^{n} \left(y^\mu - \frac{\theta^\top \varphi(x^\mu)}{\sqrt{k_L}} \right)^2 + \frac{\lambda}{2} ||\theta||^2 \right\}. \tag{195}$$

Building on the theorem of [11] and conjecture 4.3, the mean squared error achieved by this ERM algorithm is given by

$$\epsilon_g := \mathbb{E}_{\mathcal{D}} \mathbb{E}_{x \sim \mathcal{N}(0_d, \Omega_0)} \left(f_\star(x) - \frac{\hat{\theta}^\top \varphi(x)}{\sqrt{k_L}} \right)^2 = \rho + q - 2m, \tag{196}$$

with q, m the solutions of the system of equations

$$\begin{cases} \hat{V} = \frac{1}{\gamma_L} \frac{\alpha}{1+V} \\ \hat{q} = \frac{1}{\gamma_L} \alpha \frac{\rho + q - 2m}{(1+V)^2} \\ \hat{m} = \frac{1}{\sqrt{\gamma_L \gamma^\star_{L_\star}}} \frac{\alpha}{1+V} \end{cases}, \qquad \begin{cases} V = \frac{1}{k_L} \operatorname{tr} \left(\lambda I_d + \hat{V} \Omega_L^{\text{lin}} \right)^{-1} \Omega_L^{\text{lin}} \\ q = \frac{1}{k_L} \operatorname{tr} \left[\left(\hat{q} \Omega_L^{\text{lin}} + \hat{m}^2 \Phi_{LL_\star}^{\text{lin}\top} \theta_\star \theta_\star^\top \Phi_{LL_\star}^{\text{lin}} \right) \Omega_L^{\text{lin}} \left(\lambda I_d + \hat{V} \Omega_L^{\text{lin}} \right)^{-2} \right] \\ m = \sqrt{\frac{\gamma_L}{\gamma^\star_{L_\star}}} \frac{\hat{m}}{k_L} \operatorname{tr} \Phi_{LL_\star}^{\text{lin}\top} \theta_\star \theta_\star^\top \Phi_{LL_\star}^{\text{lin}} \left(\lambda I_d + \hat{V} \Omega_L^{\text{lin}} \right)^{-1} \end{cases}. \tag{197}$$

D.3 Logistic regression

We now turn to the classification setting, when the labels are given by a deep random network with sign readout

$$y^\mu = \text{sign} \left(\frac{\theta_\star^\top \varphi_\star(x^\mu)}{\sqrt{k_{L_\star}}} \right). \tag{198}$$

Note that this corresponds to $\sigma_{L_\star+1} = \text{sign}$. and the dRF readout weights θ are trained with logistic regression, using the ERM

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \left\{ \sum_{\mu=1}^{n} \ln \left(1 + e^{-y^\mu \frac{\theta^\top \varphi(x^\mu)}{\sqrt{k_L}}} \right) + \frac{\lambda}{2} ||\theta||^2 \right\}. \tag{199}$$

By the same token, introducing following [11] 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

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

It follows from [11] and Conjecture 4.3 that the associated test error reads

$$\epsilon_g := \mathbb{E}_{\mathcal{D}} \mathbb{P}_{x \sim \mathcal{N}(0_d, \Omega_0)} \left(f_\star(x) \neq \text{sign} \left(\frac{\hat{\theta}^\top \varphi(x)}{\sqrt{k_L}} \right) \right) = \frac{1}{\pi} \arccos \frac{m}{\sqrt{\rho q}}, \tag{200}$$

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

$$\begin{cases} \hat{V} = -\frac{\alpha}{\gamma_L} \int \frac{d\xi e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \left[\sum_{y=\pm 1} Z \left(y, \frac{m}{\sqrt{q}} \xi, \rho - \frac{m^2}{q} \right) \partial_\omega f(y, \sqrt{q}\xi, V) \right] \\ \hat{q} = \frac{\alpha}{\gamma_L} \int \frac{d\xi e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \left[\sum_{y=\pm 1} Z \left(y, \frac{m}{\sqrt{q}} \xi, \rho - \frac{m^2}{q} \right) f(y, \sqrt{q}\xi, V)^2 \right] \\ \hat{m} = \frac{\alpha}{\sqrt{\gamma_L \gamma^\star_{L_\star}}} \int \frac{d\xi e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \left[\sum_{y=\pm 1} \partial_\omega Z \left(y, \frac{m}{\sqrt{q}} \xi, \rho - \frac{m^2}{q} \right) f(y, \sqrt{q}\xi, V) \right] \end{cases}$$

$$\begin{cases} V = \frac{1}{k_L} \operatorname{tr} \left(\lambda I_d + \hat{V} \Omega_L^{\text{lin}} \right)^{-1} \Omega_L^{\text{lin}} \\ q = \frac{1}{k_L} \operatorname{tr} \left[\left(\hat{q} \Omega_L^{\text{lin}} + \hat{m}^2 \Phi_{LL_\star}^{\text{lin}\top} \theta_\star \theta_\star^\top \Phi_{LL_\star}^{\text{lin}} \right) \Omega_L^{\text{lin}} \left(\lambda I_d + \hat{V} \Omega_L^{\text{lin}} \right)^{-2} \right] \\ m = \sqrt{\frac{\gamma_L}{\gamma^\star_{L_\star}}} \frac{\hat{m}}{k_L} \operatorname{tr} \Phi_{LL_\star}^{\text{lin}\top} \theta_\star \theta_\star^\top \Phi_{LL_\star}^{\text{lin}} \left(\lambda I_d + \hat{V} \Omega_L^{\text{lin}} \right)^{-1} \end{cases} \tag{201}$$

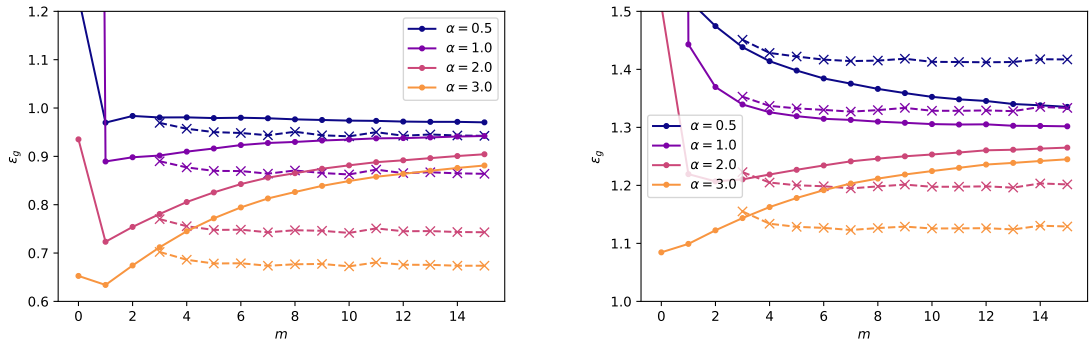

Figure 6: Test error for a regression task on a $L_\star = 1$ two-layer target with sign activation and width $\gamma_1^\star = 4$. Solid lines represent the test error of a dRF of depth m and widths $\gamma_1 = ... = \gamma_m = \gamma$, while dashed lines indicate the test errors of wide and shallow dRFs with architecture $\gamma_1 = \gamma_3 = \gamma$ and $\gamma_2 = (m-2) \times k$. All values were evaluated using the sharp asymptotic characterization of conjecture 4.3, see also App.D. The parameter m, which parameterizes the number of parameters in these two networks, is varied from 0 to 15. For $\gamma = 4$ (left), the deep architecture is consistently outperformed by the wide architecture. Closer to the interpolation peak, for $\gamma = 1$ (right), the implicit depth-induced regularization means that deeper architectures perform better than wider architectures.

E Architecture-induced implicit regularization

A seminal pursuit in machine learning research is the theoretical understanding of the interplay between the network architecture and it learning ability. While this is a challenging open question, the study of dRF (2), i.e. networks with intermediate layers frozen at initialization, allows to make some headway and gather some preliminary insight into these interrogations. It constitutes a highly stylized, but nonetheless versatile, playground for which some questions can be explored, and which hopefully pave the first preliminary steps in the understanding of networks trained end-to-end.

Section 5 in the main text discussed the regularization induced by depth in dRF architectures. In this section, we further explore, using conjecture 4.3 as a flexible toolbox to access asymptotic test errors, the role of other architectural features in the performance of dRF. Our purpose is mainly to complement the discussion of section E, and highlight some observations of interest. A more complete study falls out of the scope of the present manuscript and is left for future work. In this section, we briefly discuss two questions:

- For a fixed number of parameters, is it better to have a deep or wide architecture?
- What is the influence of a narrow (bottleneck) hidden layer on the test error?

E.1 Deeper or wider

For a given number of parameters $m\gamma d$, for $m \in [14]$ we explore the performance of

- A rectangular deep net of depth $L = m$ with width sequence $\gamma_1 = ... = \gamma_m = \gamma$.
- A wide net with widths $\gamma_1 = \gamma_3 = \gamma$ and $\gamma_2 = (m-2)\gamma$ of depth $L = 3$.

for $m \geq 3$, learning from a two-layers target with sign activation. The activation is taken to be $\tanh$ for all layers, in both networks. Note that in both architectures, the number of trainable parameter is also always the same, since the readout layer is in any case of width γd.

Fig. 6 compares the deep architecture (dashed lines) with the wide architecture (solid lines). In general, the wide architecture provides smaller test errors, in accordance with the intuition that additional layers introduce more effective noise and therefore generically prove detrimental to the learning ability of the dRF, see Fig. 6, right panel. However, as discussed in section 5 in the main text, the implicit regularization induced by this noise can help mitigate overfitting in some regimes. This is in particular the case in the vicinity of the interpolation peaks, for noisy targets. The left panel of Fig. 6 shows such a case, where deep architectures outperform wide architectures in small data regimes $\alpha = 0.5, 1$. If the explicit regularization λ is optimized over, this effect disappears.

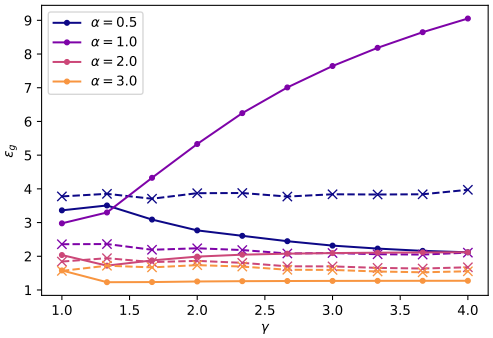

Figure 7: Regression problem over a $L_\star = 1$ target with sign activation and width $\gamma_1^\star$. Dashed lines represent the test error (evaluated using the sharp asymptotics of conjecture 4.3, see also App.D) of $L = 4$ dRFs, with $\gamma_1 = \gamma_2 = \gamma_4 = \gamma$ and a bottleneck third layer $\gamma_3 = 1/2$. Solid lines corresponds to a rectangular network with no bottleneck $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma$. Close to the interpolation peak ($\alpha = 1, 2$) the regularization induced by the bottleneck mitigates the overfitting and leads to smaller test errors.

E.2 Bottleneck hidden layer

Another question of interest is the effect of a very narrow hidden layer. Fig. 7 investigates the performance over a $L_\star = 1$ target with sign activation and width $\gamma_1^\star$, of $L = 4$ dRFs, with $\gamma_1 = \gamma_2 = \gamma_4 = \gamma$ and a bottleneck third layer $\gamma_3 = 1/2$. The parameter γ was varied between 1 and 4. As intuitively expected, the bottleneck, by forcing an intermediary low-dimensional representation, has a regularizing effect. While generically the bottleneck translates into a loss of information, it is beneficial in regimes where regularization is helpful, e.g. close to interpolation peaks or noisy settings. Such an instance is presented in Fig. 7. Again, if the explicit regularization λ is tuned, this effect disappears.