Asymptotics of Learning with Deep Structured (Random) Features

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

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

February 22, 2024

Abstract

For a large class of feature maps we provide a tight asymptotic characterisation of the test error associated with learning the readout layer, in the high-dimensional limit where the input dimension, hidden layer widths, and number of training samples are proportionally large. This characterization is formulated in terms of the population covariance of the features. Our work is partially motivated by the problem of learning with Gaussian rainbow neural networks, namely deep non-linear fully-connected networks with random but structured weights, whose row-wise covariances are further allowed to depend on the weights of previous layers. For such networks we also derive a closed-form formula for the feature covariance in terms of the weight matrices. We further find that in some cases our results can capture feature maps learned by deep, finite-width neural networks trained under gradient descent.

1 Introduction

Deep neural networks are the backbone of most successful machine learning algorithms in the past decade. Despite their ubiquity, a firm theoretical understanding of the very basic mechanism behind their capacity to adapt to different types of data and generalise across different tasks remains, to a large extent, elusive. For instance, what is the relationship between the inductive bias introduced by the network architecture and the representations learned from the data, and how does it correlate with generalisation? Albeit the lack of a complete picture, insights can be found in recent empirical and theoretical works.

On the theoretical side, a substantial fraction of the literature has focused on the study of deep networks at initialisation, motivated by the lazy training regime of large-width networks with standard scaling. Besides the mathematical convenience, the study of random networks at initialisation have proven to be a valuable theoretical testbed – allowing in particular to capture some empirically observed behaviour, such as the double-decent [1] and benign overfitting [2] phenomena. As such, proxys for networks at initialisation, such as the Random Features (RF) model [3] have thus been the object of considerable theoretical attention, with their learning being asymptotically characterized in the two-layer case [4–10] and the deep case [11–14]. With the exception of [6] (limited to two-layer networks) and [14] (limited to linear networks), all the analyses for non-linear deep RFs assume unstructured random weights. In sharp contrast, the weights of trained neural networks are fundamentally structured - restricting the scope of these results to networks at initialization.

Indeed, an active research direction consists of empirically investigating how the statistics of the weights in trained neural networks encode the learned information, and how this translates to properties of the predictor, such as inductive biases [15, 16]. Of particular relevance to our work is a recent observation by [17] that a random (but structured) network with the weights sampled from an ensemble with matching statistics can retain a comparable performance to the original trained neural networks. In particular, for some tasks it was shown that second order statistics suffices – defining a Gaussian *rainbow network* ensemble.

Our goal in this manuscript is to provide an exact asymptotic characterization of the properties of *Gaussian rainbow networks*, i.e. deep, non-linear networks with structured random weights. Our **main contributions** are:

- We derive a tight asymptotic characterization of the test error achieved by performing ridge regression with Lipschitz-continuous feature maps, in the high-dimensional limit where the dimension of the features and the number of samples grow at proportional rate. This class of feature maps encompasses as a particular case Gaussian rainbow network features.
- The asymptotic characterization is formulated in terms of the population covariance of the features. For Gaussian rainbow networks, we explicit a closed-form expression of this covariance, formulated as in the unstructured case [12] as a simple linear recursion depending on the weight matrices of each layer. These formulae extend similar results of [12, 18] for independent and unstructured weights to the case of structured –and potentially correlated– weights.
- We empirically find that our theoretical characterization captures well the learning curves of some networks trained by gradient descent in the lazy regime.

Related works

Random features — Random features (Rfs) were introduced in [3] as a computationally efficient way of approximating large kernel matrices. In the shallow case, the asymptotic spectral density of the conjugate kernel was derived in [19–21]. The test error was on the other hand characterized in [9, 10] for ridge regression, and extended to generic convex losses by [4, 6, 8], and in [22–24] for other penalties. RFs have been studied as a model for networks in the lazy regime, see e.g. [25–28];

Deep RFs – Recent work have addressed the problem of extending these results to deeper architectures. In the case of linear networks, a sharp characterization of the test error is provided in [11] for the case of unstructured weights and [14] in the case of structured weights. For non-linear RFs, [12] provides deterministic equivalents for the sample covariance matrices, and [12, 13] provide a tight characterization of the test error. Deep random networks have been also studied in the context of Gaussian processes by [29, 30], Bayesian neural networks in [11, 31–35] and inference in [36–40]. The recent work of [17] provides empirical evidence that for a given trained neural network, a resampled network from an ensemble with matching statistics (*rainbow networks*) might achieve comparable generalization performance, thereby partly bridging the gap between random networks and trained networks.

2 Setting

Consider a supervised learning task with training data $(\mathsf{x}_i, \mathsf{y}_i)_{i\in[n]}$. In this manuscript, we are interested in studying the statistics of linear predictors $f_\theta(\mathsf{x}) = \frac{1}{\sqrt{p}}\theta^\top\varphi(\mathsf{x})$ for a class of fixed feature maps $\varphi : \mathbf{R}^d \to \mathbf{R}^p$ and weights $\theta \in \mathbf{R}^p$ trained via empirical risk minimization:

$$\hat{\theta}_\lambda = \min_{\theta\in\mathbf{R}^p} \sum_{i\in[n]} (\mathsf{y}_i - f_\theta(\mathsf{x}_i))^2 + \lambda||\theta||^2. \tag{1}$$

Of particular interest is the generalization error:

$$\mathcal{E}_{\text{gen}}(\hat{\theta}_\lambda) = \mathbf{E}\left(\mathsf{y} - f_{\hat{\theta}_\lambda}(\mathsf{x})\right)^2 \tag{2}$$

where the expectation is over a fresh sample from the same distribution as the training data. More precisely, our results will hold under the following assumptions.

Assumption 2.1 (Labels)**.** We assume that the labels y_i are generated by another feature map $\varphi_* : \mathbf{R}^d \to \mathbf{R}^k$ as

$$\mathsf{y}_i = \frac{1}{\sqrt{k}}\theta_*^\top\varphi_*(\mathsf{x}_i) + \varepsilon_i, \tag{3}$$

where $\varepsilon \in \mathbf{R}^n$ is an additive noise vector (independent of the covariates x_i) of zero mean and covariance $\Sigma := \mathbf{E}\,\varepsilon\varepsilon^\top$, and $\theta_* \in \mathbf{R}^k$ is a deterministic weight vector.

Assumption 2.2 (Data & Features)**.** We assume that the covariates x_i are independent and come from a distribution such that

(i) the feature maps φ, φ_* are centered in the sense $\mathbf{E}\,\varphi(\mathsf{x}_i) = 0$, $\mathbf{E}\,\varphi_*(\mathsf{x}_i) = 0$,

(ii) the feature covariances

$$\Omega := \mathbf{E}\,\varphi(\mathsf{x}_i)\varphi(\mathsf{x}_i)^\top \in \mathbf{R}^{p\times p}, \qquad \Psi := \mathbf{E}\,\varphi_*(\mathsf{x}_i)\varphi_*(\mathsf{x}_i)^\top \in \mathbf{R}^{k\times k}, \qquad \Phi := \mathbf{E}\,\varphi(\mathsf{x}_i)\varphi_*(\mathsf{x}_i)^\top \in \mathbf{R}^{p\times k}, \tag{4}$$

have uniformly bounded spectral norm.

(iii) scalar Lipschitz functions of the feature matrices

$$X := (\varphi(\mathsf{x}_1), \ldots, \varphi(\mathsf{x}_n)) \in \mathbf{R}^{p\times n}, \qquad Z := (\varphi_*(\mathsf{x}_1), \ldots, \varphi_*(\mathsf{x}_n)) \in \mathbf{R}^{k\times n} \tag{5}$$

are uniformly sub-Gaussian.

Assumption 2.3 (Proportional regime)**.** The number of samples n and the feature dimensions p, k are all large and comparable, see Theorem 3.1 later.

Remark 2.4. *We formulated Assumption 2.2 as a joint assumption on the covariates distribution and the feature maps. A conceptually simpler but less general condition would be to assume that*

(ii') the covariates x_i are Gaussian with bounded covariance $\Omega_0 := \mathbf{E}\,\mathsf{x}_i\mathsf{x}_i^\top$

(iii') the feature maps φ, φ_* are Lipschitz-continuous

instead of Assumptions (ii) and (iii).

The setting above defines a quite broad class of problems, and the results that follow in Section 3 will hold under these generic assumptions. The main class of feature maps we are interested in are *deep structured feature models.*

Definition 2.5 (Deep structured feature model)**.** For any $L \in \mathbb{N}$ and dimensions $d, p_1, \ldots, p_L = p$, let $\varphi_1, \ldots, \varphi_L : \mathbf{R} \to \mathbf{R}$ be Lipschitz-continuous *activation functions* $|\varphi_l(a) - \varphi_l(b)| \lesssim |a - b|$ applied entrywise, and let $W_1 \in \mathbf{R}^{p_1\times d}, W_2 \in \mathbf{R}^{p_2\times p_1}, \ldots$ be deterministic *weight matrices* with uniformly bounded spectral norms, $\|W_l\| \lesssim 1$. We then call

$$\varphi(\mathsf{x}) := \varphi_L\left(W_L\varphi_{L-1}\left(\cdots W_2\varphi_1\left(W_1\mathsf{x}\right)\right)\right). \tag{6}$$

a *deep structured feature* model.

Note that eq. (6) defines a Lipschitz-continuous map[1] $\varphi\colon \mathbf{R}^d \to \mathbf{R}^p, \varphi_*\colon \mathbf{R}^d \to \mathbf{R}^k$ and therefore if both φ, φ_* are deep structured feature models (with distinct parameters in general), then Assumption 2.2 is satisfied whenever the feature maps φ, φ_* are centered[2] with respect to Gaussian covariates x_i. As hinted in the introduction we will be particularly interested in one sub-class of Definition 2.5 known as *Gaussian rainbow networks.*

Definition 2.6 (Gaussian rainbow ensemble)**.** Borrowing the terminology of [17], we define a fully-connected, L-layer *Gaussian rainbow network* as a random variant of Definition 2.5 where for each ℓ the hidden-layer weights $W_\ell = Z_\ell C_\ell^{1/2}$ are random matrices with $Z_\ell \in \mathbf{R}^{p_{\ell+1}\times p_\ell}$ having zero mean and i.i.d. variance $1/p_\ell$ Gaussian entries and $C_\ell \in \mathbf{R}^{p_\ell\times p_\ell}$ being uniformly bounded covariance matrices, which we allow to depend on previous layer weights $Z_1, \ldots, Z_{l-1}$.

Note that Gaussian rainbow networks above can be seen as a generalization of the deep random features model studied in [12, 13, 41], with the crucial difference that the weights are structured.

Notations

For square matrices $A \in \mathbf{R}^{n\times n}$ we denote the averaged trace by $\langle A\rangle := n^{-1}\operatorname{Tr} A$, and for rectangular matrices $A \in \mathbf{R}^{n\times m}$ we denote the Frobenius norm by $\|A\|_F^2 := \sum_{ij}|a_{ij}|^2$, and the operator norm by $\|A\|$. For families of non-negative random variables $X(n), Y(n)$ we say that X is *stochastically dominated* by Y, and write $X \prec Y$, if for all ϵ, D it holds that $P(X(n) \geq n^\epsilon Y(n)) \leq n^{-D}$ for n sufficiently large.

[1] $\|\varphi(W\mathsf{x}) - \varphi(W\mathsf{x}')\|^2 = \sum_i|\varphi(w_i^\top\mathsf{x}) - \varphi(w_i^\top\mathsf{x}')|^2 \lesssim \sum_i|w_i^\top(\mathsf{x}-\mathsf{x}')|^2 = \|W(\mathsf{x}-\mathsf{x}')\|^2 \lesssim \|\mathsf{x}-\mathsf{x}'\|^2$

[2] It is sufficent that e.g. ϕ_l is odd, and x_i is centered.

3 Test error of Lipschitz feature models

Under Assumptions 2.1 and 2.2 the generalization error from Eq. (2) is given by

$$\mathcal{E}_{\text{gen}}(\lambda) = \frac{\theta_*^\top \Psi \theta_*}{k} + \frac{\theta_*^\top Z X^\top G \Omega G X Z^\top \theta_*}{kp^2} + \frac{n}{p}\left\langle \frac{X^\top G \Omega G X \Sigma}{p} \right\rangle - 2\frac{\theta_*^\top \Phi^\top G X Z^\top \theta_*}{kp}, \tag{7}$$

in terms of the *resolvent* $G = G(\lambda) := (XX^\top/p + \lambda)^{-1}$.

Our main result is a rigorous asymptotic expression for Eq. (7). To that end define, $m(\lambda)$ to be the unique solution to the equation

$$\frac{1}{m(\lambda)} = \lambda + \left\langle \Omega\Big(I + \frac{n}{p} m(\lambda)\Omega\Big)^{-1}\right\rangle, \tag{8}$$

and define

$$M(\lambda) = \Big(\lambda + \frac{n}{p}\lambda m(\lambda)\Omega\Big)^{-1} \tag{9}$$

which is the *deterministic equivalent* of the resolvent, $M(\lambda) \approx G(\lambda)$, see Theorem 3.3 later. The fact that eq. (8) admits a unique solution $m(\lambda) > 0$ which is continuous in λ follows directly from continuity and monotonicity. Moreover, from

$$0 \le \left\langle \Omega\Big(I + \frac{n}{p} m\Omega\Big)^{-1}\right\rangle \le \min\left\{\langle\Omega\rangle, \frac{\operatorname{rank}\Omega}{n}\frac{1}{m}\right\}$$

we obtain the bounds

$$\max\left\{\frac{1}{\lambda + \langle\Omega\rangle}, \frac{1 - \frac{\operatorname{rank}\Omega}{n}}{\lambda}\right\} \le m(\lambda) \le \frac{1}{\lambda}. \tag{10}$$

We also remark that $m(\lambda)$ depends on Ω only through its eigenvalues $\omega_1, \ldots, \omega_p$, while $M(\lambda)$ depends on the eigenvectors. The asymptotic expression Eq. (12) for the generalization error derived below depends on the eigenvalues of Ω, the overlap of the eigenvectors of Ω with the eigenvectors of Φ, and the overlap of the eigenvectors of Ψ, Φ with θ_*.

Theorem 3.1. *Under Assumption 2.1, Assumption 2.2 and Assumption 2.3 for fixed $\lambda > 0$ we have the asymptotics*

$$\mathcal{E}_{\text{gen}}(\lambda) = \mathcal{E}_{\text{gen}}^{\text{rmt}}(\lambda) + O\Big(\frac{1}{\sqrt{n}}\Big), \tag{11}$$

in the proportional $n \sim k \sim p$ regime, where

$$\mathcal{E}_{\text{gen}}^{\text{rmt}}(\lambda) := \frac{1}{k}\theta_*^\top \frac{\Psi - \frac{n}{p}m\lambda\Phi(M + \lambda M^2)\Phi^\top}{1 - \frac{n}{p}(\lambda m)^2\langle\Omega M \Omega M\rangle}\theta_* + \langle\Sigma\rangle \frac{(\lambda m)^2\frac{n}{p}\langle M\Omega M\Omega\rangle}{1 - \frac{n}{p}(\lambda m)^2\langle\Omega M\Omega M\rangle}. \tag{12}$$

In the general case of comparable parameters we have the asymptotics with a worse error of

$$\frac{1}{\sqrt{\min\{n,p,k\}}}\Big(1 + \frac{\max\{n,p,k\}}{\min\{n,p,k\}}\Big).$$

Remark 3.2 (Relation to previous results). *We focus on the misspecified case as this presents the main novelty of the present work. In the wellspecified case $Z = X$ our model essentially reduces to linear regression with data distribution $x = \varphi(\mathsf{x})$. There has been extensive research on the generalization error of linear regression, see e.g. in [42–45] and the references therein.*

(a) We confirm Conjecture 1 of [23] under assumption 2.2. The expression for the error term in Theorem 3.1 matches the expression obtained in [23] for a Gaussian covariates teacher-student model.

(b) Independently and concurrently to the current work [46] (partially confirming a conjecture made in [47]) obtained similar results under different assumptions. Most importantly [46] considers one-layer unstructured random feature models and computes the empirical generalization error *for a deterministic data set, while we consider general Lipschitz features of random data, and compute the generalization error.*

(c) In the unstructured random feature model [10, 48] obtained an expression for the generalization error under the assumption that the target model is linear or rotationally invariant.

The novelty of Theorem 3.1 compared to many of the previous works is, besides the level of generality, two-fold:

(i) We obtain a deterministic equivalent for the generalization error involving the population covariance Φ and the sample covariance $XZ^\top$ in the general misspecified setting.

(ii) Our deterministic equivalent is *anisotropic*, allowing to evaluate Eq. (7) for *fixed* targets θ_* and structured noise covariance $\Sigma \neq I$.

Some of the previous rigorous results on the generalization error of ridge regression have been limited to the well-specified case, $X = Z$, since in this particular case the second term of Eq. (7) can be simplified to

$$\frac{XX^\top}{p} G\Omega G \frac{XX^\top}{p} = (1 - \lambda G)\Omega(1 - \lambda G). \tag{13}$$

When computing deterministic equivalents for terms as $G\Omega G$, some previous results have relied on the "trick" of differentiating a generalized resolvent matrix $\widetilde{G}(\lambda, \lambda') := (XX^\top/p + \lambda'\Omega + \lambda)^{-1}$ with respect to λ'. Our approach is more robust and not limited to expressions which can be written as certain derivatives.

To illustrate Item (ii), the conventional approach in the literature to approximating e.g. the third term on the right hand side of Eq. (7) in the case $\Sigma = I$ would be to use the cyclicity of the trace to obtain

$$\begin{aligned}\frac{1}{p^2} \operatorname{Tr} X^\top G\Omega G X &= \frac{1}{p} \operatorname{Tr} G \frac{XX^\top}{p} G\Omega \\ &= \langle G\Omega \rangle - \lambda \langle G^2 \Omega \rangle.\end{aligned} \tag{14}$$

Then upon using Eq. (8) and $\langle G\Omega \rangle \approx \langle M\Omega \rangle$, the first term of Eq. (14) can be approximated by $1/(\lambda m(\lambda)) - 1$, while for the second term it can be argued that this approximation also holds in derivative sense to obtain

$$\langle G^2 \Omega \rangle = -\frac{\mathrm{d}}{\mathrm{d}\,\lambda} \langle G\Omega \rangle \approx -\frac{\mathrm{d}}{\mathrm{d}\,\lambda} \frac{1}{\lambda m(\lambda)} = \frac{\lambda m'(\lambda) + m(\lambda)}{(\lambda m(\lambda))^2}$$

By differentiating Eq. (8), solving for m' and simplifying, it can be checked that this result agrees with the second term of Eq. (12) in the special case $\Sigma = I$. However, it is clear that any approach which only relies on *scalar* deterministic equivalents is inherently limited in the type of expressions which can be evaluated. Instead, our approach involving *anisotropic deterministic equivalents* has no inherent limitation on the structure of the expressions to be evaluated.

An alternative to evaluating rational expressions of X, Z, commonly used in similar contexts, is the technique of *linear pencils* [46, 48]. The idea here is to represent rational functions of X, Z as blocks of inverses of larger random matrices which depend linearly X, Z. The downside of linear pencils is that even for simple rational expressions the linearizations become complicated, sometimes even requiring the use of computer algebra software for the analysis[3] In comparison we believe that our approach is more direct and flexible.

3.1 Proof of Theorem 3.1

We present the proof of Theorem 3.1 in details in Appendix A. The main steps and ingredients for the proof of Theorem 3.1 consist of the following:

Concentration:

As a first step we establish *concentration estimates* for Lipschitz functions of X, Z and its columns. A key aspect is the concentration of quadratic forms in the columns $x_i := \varphi(\mathsf{x}_i)$ of X:

$$|x_i^\top A x_i - \mathbf{E}\, x_i^\top A x_i| = |x_i^\top A x_i - \operatorname{Tr} \Omega A| \prec \|A\|_F$$

which follows from the Hanson-Wright inequality [49]. The concentration step is very similar to analagous considerations in previous works [50, 51] but we present it for completeness. The main property used extensively in the subsequent analysis is that traces of resolvents with deterministic observables concentrate as

$$|\langle A[G(\lambda) - \mathbf{E}\, G(\lambda)] \rangle| \prec \frac{\langle |A|^2 \rangle^{1/2}}{n \lambda^{3/2}}. \tag{15}$$

Anisotropic Marchenko-Pastur Law:

As a second step we prove an anisotropic Marchenko-Pastur law for the resolvent G, of the form:

[3] For instance [48] used block matrices with up to 16×16 blocks in order to evaluate the asymptotic test error.

Theorem 3.3. *For arbitrary deterministic matrices A we have the high-probability bound*

$$|\langle (G(\lambda) - M(\lambda)A\rangle| \prec \frac{\langle |A|^2\rangle}{n\lambda^3}, \tag{16}$$

in the proportional $n \sim p$ regime[4].

Remark 3.4. *Tracial Marchenko-Pastur laws (case $A = I$ above) have a long history, going back to [52] in the isotropic case $\Omega = I$, [53] in the general case with separable covariance $x = \sqrt{\Omega} z$ and [54] under quadratic form concentration assumption. Anisotropic Marchenko-Pastur laws under various conditions and with varying precision have been proven e.g. in [47, 50, 55, 56].*

For the proof of Theorem 3.3 the resolvent $\check{G} := (X^\top X/p + \lambda)^{-1} \in \mathbf{R}^{n\times n}$ of the *Gram matrix* $X^\top X$ plays a key role. The main tool used in this step are the commonly used *leave-one-out identities*, e.g.

$$Gx_i = \lambda \check{G}_{ii} G_{-i} x_i, \quad G_{-i} := \Big(\sum_{j\neq i} \frac{x_j x_j^\top}{p} + \lambda\Big)^{-1} \tag{17}$$

which allow to decouple the randomness due the i-th column from the remaining randomness. Such identities are used repeatedly to derive the approximation

$$\mathbf{E}\, G \approx \Big(\frac{n}{p}\lambda \langle \mathbf{E}\, \check{G}\rangle \Omega + \lambda\Big)^{-1} \tag{18}$$

in Frobenius norm, which, together with the relation $1 - \lambda\langle \check{G}\rangle = \frac{p}{n}\big(1 - \lambda\langle G\rangle\big)$ between the traces of G and $\check{G}$, yields a self-consistent equation for $\langle \check{G}\rangle$. This self-consistent equation is an approximate version of Eq. (8), justifying the definition of m. The *stability* of the self-consistent equation then implies the averaged asymptotic equivalent

$$|m - \langle \mathbf{E}\, \check{G}\rangle| \lesssim \frac{1}{n\lambda^2}. \tag{19}$$

and therefore by Eq. (18) finally

$$\|M - \mathbf{E}\, G\|_F \lesssim \frac{1}{n^{1/2}\lambda^3}, \tag{20}$$

which together with Eq. (15) implies Theorem 3.3.

Compared to most previous anisotropic deterministic equivalents as in [56] we measure the error of the approximation Eq. (16) with respect to the Frobenius of the observable A. As in the case of unified local laws for Wigner matrices [57] this idea renders the separate handling of quadratic form bound unnecessary, considerably streamlining the proof. To illustrate the difference note that specializing A to be rank-one $A = xy^\top$ in

$$|y^\top (G - M)x| = |\mathrm{Tr}(G - M)A| \prec \begin{cases} \|A\| \\ \langle |A|^2\rangle^{1/2} \end{cases}$$

results in a trivial estimate $\|x\|\|y\|$ in the case of the spectral norm, and in the optimal estimate $\|x\|\|y\|/\sqrt{p}$ in the case of the Frobenius norm.

Anisotropic Multi-Resolvent Equivalents

The main novelty of the current work lies in Proposition A.4 which asymptotically evaluates the expressions on the right-hand-side of Eq. (7). A key property of the deterministic equivalents is that the approximation is *not* invariant under multiplication. E.g. for the last term in Eq. (7) we have the approximations $G \approx M$ and $\frac{1}{n}XZ^\top = \frac{1}{n}\sum x_i z_i^\top \approx \Phi$, while for the product the correct deterministic equivalent is

$$G\frac{XZ^\top}{n} \approx \lambda m M \Phi, \tag{21}$$

i.e. the is an additional factor of $m\lambda$. In this case the additional factor can be obtained from a direct application of the leave-one-out identity Eq. (17) to the product $G\frac{XZ^\top}{n}$, but the derivation of the multi-resolvent equivalents requires

[4] See the precise statement in the comparable regime in Eq. (51) later

more involved arguments. When expanding the multi-resolvent expression $\langle GAGB\rangle$ we obtain an approximative self-consistent equation of the form

$$\langle GAGB\rangle \approx \langle MAMB\rangle + \frac{n}{p}(m\lambda)^2\langle MBM\Omega\rangle\langle GAG\Omega\rangle.$$

Using a stability analysis this yields a deterministic equivalent for the special form $\langle GAG\Omega\rangle$ which then can be used for the general case. The second term of Eq. (7) requires the most carefuly analysis due to the interplay of the multi-resolvent expression and the dependency among Z, X.

4 Population covariance for rainbow networks

Theorem 3.1 characterizes the test error for learning using Lipschitz feature maps as a function of the three features population (cross-)covariances Ω, Φ, Ψ. For the particular case where both the target and learner feature maps are drawn from the Gaussian rainbow ensemble from Definition 2.6, these population covariances can be expressed in closed-form in terms of combinations of products of the weights matrices. Consider two rainbow networks

$$\begin{aligned} \varphi(\mathsf{x}) &= \varphi_L(W_L\varphi_{L-1}(\dots\varphi_1(W_1\mathsf{x}))) \\ \varphi_*(\mathsf{x}) &= \psi_{\widetilde{L}}(V_{\widetilde{L}}\psi_{\widetilde{L}-1}(\dots\psi_1(V_1\mathsf{x}))) \end{aligned} \tag{22}$$

with depths $L, \widetilde{L}$. The approach we introduce here is in theory capable of obtaining linear or polynomial approximations to Ω, Φ, Ψ under very general assumptions. However, for definiteness we focus on a class of correlated rainbow networks in which, for all $k \neq j$, the k-th row of W_ℓ is independent from the j-th row of W_ℓ, V_ℓ as this allows for particularly simple expressions for the linearized covariances[5]. Note that we explicitly allow for weights to be correlated across layers.

Assumption 4.1 (Correlated rainbow networks)**.** By symmetry we assume without loss of generality $L \leq \widetilde{L}$. Furthermore, we assume that

(a) for $\ell \leq L \leq \widetilde{L}$ all the internal widths p_ℓ of W_ℓ, V_ℓ agree,

(b) for all $\ell \leq \widetilde{L}$, the dimensions scale proportionally, i.e. $n \sim d \sim p_\ell$,

(c) for $\ell \leq L \leq \widetilde{L}$ the rows w_ℓ, v_ℓ of W_ℓ, V_ℓ are mean-zero and i.i.d. with

$$C_\ell := p_\ell \,\mathbf{E}\, w_\ell w_\ell^\top,\ \widetilde{C}_\ell := p_\ell \,\mathbf{E}\, v_\ell v_\ell^\top,\ \check{C}_\ell := p_\ell \,\mathbf{E}\, w_\ell v_\ell^\top,$$

(d) for two (possibly identical) rows u, z, and for any matrix A, quadratic forms admit concentration, w.h.p.[6]

$$u^\top A z - \mathrm{Tr}(A\,\mathbf{E}\, z u^\top) \lesssim n^{-1/2}, \tag{23}$$

(e) for all $\ell \leq \widetilde{L}$, operator norms of (cross-)covariance matrices admit uniform bounds

$$\|C_\ell\| + \|\widetilde{C}_\ell\| + \|\check{C}_\ell\| \lesssim 1. \tag{24}$$

Under Assumption 4.1 the *linearized population covariances* can be defined recursively as follows:

Definition 4.2 (Linearized population covariances)**.** Define the sequence of matrices $\Omega_\ell^{\mathrm{lin}}, \Phi_\ell^{\mathrm{lin}}, \Psi_\ell^{\mathrm{lin}}$ by the recursions

$$\begin{aligned} \Omega_\ell^{\mathrm{lin}} &= (\kappa_\ell^1)^2 W_\ell \Omega_{\ell-1}^{\mathrm{lin}} W_\ell^\top + (\kappa_\ell^*)^2 I \\ \Psi_\ell^{\mathrm{lin}} &= (\tilde{\kappa}_\ell^1)^2 V_\ell \Psi_{\ell-1}^{\mathrm{lin}} V_\ell^\top + (\tilde{\kappa}_\ell^*)^2 I \\ \Phi_\ell^{\mathrm{lin}} &= \kappa_\ell^1 \tilde{\kappa}_\ell^1 W_\ell \Phi_{\ell-1}^{\mathrm{lin}} V_\ell^\top + (\check{\kappa}_\ell^*)^2 I, \end{aligned} \tag{25}$$

with $\Omega_0^{\mathrm{lin}} = \Psi_0^{\mathrm{lin}} = \Phi_0^{\mathrm{lin}} = \Omega_0$ the input covariance. The coefficients $\{\kappa_\ell^1, \tilde{\kappa}_\ell^1, \kappa_\ell^*, \tilde{\kappa}_\ell^*\}$ are defined by the recursion

$$\kappa_\ell^1 := \mathbf{E}\,\varphi_\ell'(N_\ell), \quad \tilde{\kappa}_\ell^1 := \mathbf{E}\,\psi_\ell'(\widetilde{N}_\ell) \tag{26}$$

[5] The identity matrices in Eq. (25) are a direct consequence of this assumption. In case of weight matrices with varying row-norms or covariances across rows the resulting expression would be considerably more complicated.

[6] This concentration holds in particular when rows u, z are Lipschitz concentrated with constant $O(n^{-1/2})$, see Theorem A.3.

and

$$\begin{aligned}\kappa_\ell^* &= \sqrt{\mathbf{E}[\varphi_\ell(N_\ell)^2] - r_\ell(\kappa_\ell^1)^2} \\ \tilde{\kappa}_\ell^* &= \sqrt{\mathbf{E}[\psi_\ell(\widetilde{N}_\ell)^2] - \tilde{r}_\ell(\tilde{\kappa}_\ell^1)^2} \\ \breve{\kappa}_\ell^* &= \sqrt{\mathbf{E}[\varphi_\ell(N_\ell)\psi_\ell(\widetilde{N}_\ell)] - \breve{r}_\ell\kappa_\ell^1\tilde{\kappa}_\ell^1},\end{aligned} \tag{27}$$

where $N_\ell, \widetilde{N}_\ell$ are jointly mean-zero Gaussian with $\mathbf{E}\, N_\ell^2 = r_\ell$, $\mathbf{E}\, \widetilde{N}_\ell^2 = \tilde{r}_\ell$, $\mathbf{E}\, N_\ell \widetilde{N}_\ell = \breve{r}_\ell$, with

$$r_\ell = \mathrm{Tr}[C_\ell \Omega_{\ell-1}^{\mathrm{lin}}],\ \tilde{r}_\ell = \mathrm{Tr}[\widetilde{C}_\ell \Psi_{\ell-1}^{\mathrm{lin}}],\ \breve{r}_\ell = \mathrm{Tr}[\breve{C}_\ell^\top \Phi_{\ell-1}^{\mathrm{lin}}].$$

Finally, for $\tilde{L} \geq \ell \geq L+1$, define

$$\Phi_\ell^{\mathrm{lin}} = \breve{\kappa}_\ell^1 \Phi_{\ell-1}^{\mathrm{lin}} \widetilde{W}_\ell^\top, \tag{28}$$

with still $\breve{\kappa}_\ell^1, \breve{\kappa}_\ell^*$ just as before, and Ψ_ℓ^{lin} with the same recursion (25).

Conjecture 4.3. *The populations covariances Ω, Φ, Ψ involved in Theorem 3.1 can be asymptotically approximated with the last iterates of the linear recursions of Definition 4.2, i.e.*

$$\|\Omega - \Omega_L^{\mathrm{lin}}\|_F + \|\Psi - \Psi_{\tilde{L}}^{\mathrm{lin}}\|_F + \|\Phi - \Phi_{\tilde{L}}^{\mathrm{lin}}\|_F \lesssim 1 \tag{29}$$

Note that the linearization from Definition 4.2 also provides good approximation to the population covariances $\Omega_\ell, \Phi_\ell, \Psi_\ell$ of the post-activations at intermediate layers ℓ. The method we use to rigorously derive the linearizations is in theory applicable to any depths, however the estimates quickly become tedious. To keep the present work at a manageable length we provide a rigorous proof of concept only for the simplest multi-layer case.

Theorem 4.4. *Under Assumption 4.1 with $L = 1, \widetilde{L} = 2$ we have*

$$\begin{aligned}\|\Omega_1 - \Omega_1^{\mathrm{lin}}\|_F + \|\Psi_1 - \Psi_1^{\mathrm{lin}}\|_F + \|\Phi_1 - \Phi_1^{\mathrm{lin}}\|_F \lesssim 1 \\ \|\Psi_2 - \Psi_2^{\mathrm{lin}}\|_F + \|\Phi_2 - \Phi_2^{\mathrm{lin}}\|_F \lesssim 1\end{aligned}$$

with high probability.

Remark 4.5 (Comparison)**.** *The approach we take here is somewhat different from previous works [12, 41, 58] on (multi-layer) random feature models. In these previous results, the deterministic equivalent for the resolvent was obtained using primarily the randomness of the weights, resulting in relatively stringent assumptions (Gaussianity and independence between layers). This layer-by-layer recursive approach resulted in a deterministic equivalent for the resolvent which is* consistent *with a sample covariance matrix with linearized population covariance. Here we take the direct approach of considering feature models with arbitrary structured features, and then linearize the population covariances in a separate step for random features.*

4.1 Proof of Theorem 4.4

We sketch the main tools used in the argument and we refer the reader to Proposition B.11 and Theorem B.12 for the formal proof. In the proof, we crucially rely on the theory of Wiener chaos expansion and Stein's method (see [59]). Gaussian Wiener chaos is a generalization of Hermite polynomial expansions, which previously have been used for approximate linearization [12, 41] in similar contexts. The basic idea is to decompose random variables $F = F(\mathsf{x})$ which are functions of the Gaussian random vector x, into pairwise uncorrelated components

$$F = \mathbf{E}\, F + \sum_{p\geq 1} I_p\Big(\frac{\mathbf{E}\, D^p F}{p!}\Big), \tag{30}$$

where I_p is a so called *multiple integral* (generalizing Hermite polynomials) and D^p is the p-th Malliavin derivative. By applying this to the one-layer quantities $\varphi_1(w^\top \mathsf{x}), \psi_1(u^\top \mathsf{x})$ we obtain, for instance

$$\begin{aligned}&\mathbf{E}\, \varphi_1(w^\top \mathsf{x})\psi_1(v^\top \mathsf{x}) \\ &= \sum_{p\geq 1} \frac{1}{p!}\, \mathbf{E}\, \varphi_1^{(p)}(w^\top \mathsf{x})\, \mathbf{E}\, \psi_1^{(p)}(u^\top \mathsf{x}) \langle w, v\rangle^p,\end{aligned} \tag{31}$$

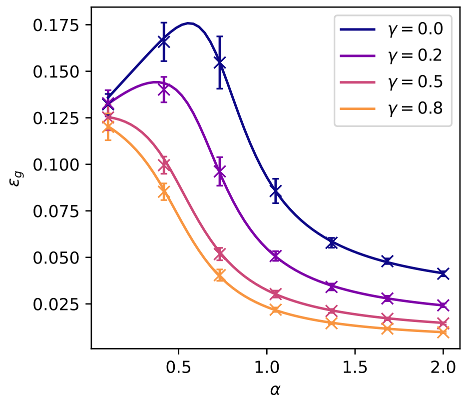

Figure 1: Test error for a target $\theta_*^\top \tanh(W_* x)$, when learning with a four-layer Gaussian rainbow network with feature map $\varphi(x) = \tanh(W_3 \tanh(W_2 \tanh(W_1 x)))$. All width were taken equal to the input dimension d, and the regularization employed is $\lambda = 10^{-4}$. The student weights are correlated across layers, with $W_1 = W_2$, and the covariance C_3 of W_3 depending on W_1 as $C_3 = (W_1 W_1^\top + {}^1\!/\!{}_2 I)^{-1}$. Target/student correlations are also present, with $\check{C}_1 = {}^1\!/\!{}_2 I$. The covariances $C_1, C_2, \tilde{C}_1$ were finally taken to have a spectrum with power-law decay, parametrized by γ. All details are provided in App. C. Solid lines: theoretical prediction of Theorem 3.1, in conjunction with the closed-form expression for the features population covariance of Definition 4.2. Crosses : numerical simulations in $d = 1000$. All experimental points were averaged over 20 instances, with error bars representing one standard deviation. Different colors represent different values for the parameter γ, with small (large) values indicating slow (fast) covariance eigenvalue decay.

which for independent w, v we can truncate after $p = 1$, giving rise to the linearization.

For the multi-layer case we combine the chaos expansion with Stein's method in order to prove *quantitative central limit theorems* of the type

$$d_W(F, N) \lesssim \mathbf{E}|\mathbf{E}\, F^2 - \langle DF, -DL^{-1}F \rangle| \tag{32}$$

for the Wasserstein distance d_W, where

$$F := w^\top \varphi_1(W\mathsf{x}), \quad N \sim \mathcal{N}(0, \mathbf{E}\, F^2), \tag{33}$$

and L^{-1} is the pseudo-inverse of the *generator of the Ornstein–Uhlenbeck semigroup*.

4.2 Discussion of Theorem 4.4

The population covariances thus admit simple approximate closed-form expressions as linear combinations of products of relevant weight matrices. These expressions generalize similar linearizations introduced in [12, 13, 18, 41, 58] for the case of weights which are both unstructured and independent, and iteratively build upon earlier results for the two-layer case developed in [4, 6, 7, 9]. In fact, the expressions leveraged in these works can be recovered as a special case for $C_\ell = \tilde{C}_\ell = I$ (isotropic weights) and $\check{C}_\ell = 0$ (independence). Importantly, note that possible correlation between weights across different layers do not enter in the reported expressions. In practice, we have observed in all probed settings the test error predicted by Theorem 3.1, in conjunction with the linearization formulae for the features covariance, to match well numerical experiments.

Figure 1 illustrates a setting where many types of weights correlations are present. It represents the learning curves of a four-layer Gaussian rainbow network with feature map $\tanh(W_3 \tanh(W_2 \tanh(W_1 \mathsf{x})))$, learning from a two-layer target $\theta_*^\top \tanh(\widetilde{W} \mathsf{x})$. To illustrate our result, we consider both target/student correlations $\check{C}_1 = {}^1\!/\!{}_2 I$, and inter-layer correlations $W_1 = W_2$. We furthermore took the covariance of the third layer to depend on the weights of the first layer, $C_3 = (W_1 W_1^\top + {}^1\!/\!{}_2 I)^{-1}$. In order to have structured weights, the covariances $\tilde{C}_1, C_1, C_2$ were chosen to have a power-law spectrum. All details on the experimental details and parameters are exhaustively provided in Appendix C. Note that despite the presence of such non-trivial correlations, the theoretical prediction of Theorem 3.1 using the linearized closed-form formulae of Def. 4.2 for the features covariances (solid lines) captures compellingly the test error evaluated in numerical experiment (crosses).

Finally, we note that akin to [12], as a consequence of the simple linear recursions, it follows that the Gaussian rainbow network feature map φ shares the same second moments, and thus by Theorem 3.1 the same test error, as an

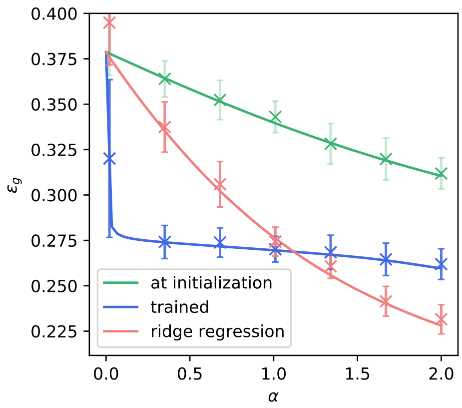

Figure 2: Crosses : Test error when training the readout layer only of a tanh-activated three-layer neural network at initialization (green) and after training (blue), using the `Pytorch` implementation of the full-batch Adam [61] optimizer, over 3000 epochs with leraning rate 10^{-4} and $n_0 = 1400$ samples, in dimension $d = 1000$. (red): ridge regression. The data is sampled from an isotropic Gaussian distribution. In all training procedures, an ℓ_2 optimization was employed, and the strength thereof optimized over using cross-validation. Solid lines represent the theoretical prediction of Theorem 3.1, using the linearized formulae of Definition 4.2 for the features population covariances Ω, Ψ, Φ. Crosses represent numerical experiments. Each simulation point is averaged over 10 instances, with error bars representing one standard deviation.

equivalent *linear stochastic* network $\varphi^{\text{lin}} = \psi_L \circ \cdots \circ \psi_1$, with

$$\psi_\ell(x) = \kappa_\ell^1 W_\ell x + \kappa_\ell^* \xi_\ell \tag{34}$$

where $\xi_\ell \sim \mathcal{N}(0, I)$ a stochastic noise. This equivalent viewpoint has proven fruitful in yielding insights on the implicit bias of RFs [12, 60] and on the fundamental limitations of deep networks in the proportional regime [18]. In the Section 5 we push this perspective further, by heuristically finding that the linearization and Theorem 3.1 can also describe deterministic networks trained with gradient descent in the lazy regime.

5 Linearizing trained neural networks

The previous discussion addressed feature maps associated to random Gaussian networks. However, note that the linearization itself only involves products of the weights matrices, and coefficient depending on weight covariances which can straightforwardly be estimated therefrom. The linearization 4.2 can thus be readily heuristically evaluated for feature maps associated to deterministic *trained* finite-width neural networks. As we discuss later in this section, the resulting prediction for the test error captures well the learning curves when re-training the readout weights of the network in a number settings. Naturally, such settings correspond to lazy learning regimes [60], where the network feature map is effectively *linear*, thus little expressive. However, these trained feature map, albeit linear, can still encode some inductive bias, as shown by [62] for one gradient step in the shallow case. In this section, we briefly explore these questions for fully trained deep networks, through the lens of our theoretical results.

Fig. 2 contrasts the test error achieved by linear regression (red), and regression on the feature map associated to a three-layer student at initialization (green) and after 3000 epochs of end-to-end training using full-batch Adam [61] at learning rate 10^{-4} and weight decay 10^{-3} over $n_0 = 1400$ training samples (blue). For all curves, the readout weights were trained using ridge regression, with regularization strength optimized over using cross-validation. Solid curves indicate the theoretical predictions of Thm. 3.1 leveraging the closed-form linearized formulae 4.2 for the features covariance. Interestingly, even for the deterministic trained network features, the formula captures the learning curve well. This observation temptingly suggests to interpret the feature map $\varphi(x)$ as the stochastic linear map

$$\varphi^g(x) = W_{\text{eff.}} x + C_{\text{eff.}}^{1/2} \xi \tag{35}$$

where $W_{\text{eff.}} \in \mathbf{R}^{p \times d}$ is proportional to the product of all the weight matrices and $\xi \sim \mathcal{N}(0, I)$ is a stochastic noise colored by the covariance

$$C_{\text{eff.}} \equiv \sum_{\ell=1}^{L-1} \left(\kappa_\ell^* \prod_{s=\ell+1}^{L} \kappa_s^1 \right)^2 \hat{W}_L \dots \hat{W}_{\ell+1} \hat{W}_{\ell+1}^\top \dots \hat{W}_L^\top + (\kappa_L^*)^2 I. \tag{36}$$

We denoted $\{\hat{W}_\ell\}_{1\le\ell\le L}$ the trained weights. Note that the effective linear network (35) simply corresponds to the composition of the equivalent stochastic linear layers (34). A very similar expression for the covariance of the effective structured noise (36) appeared in [12] for the random case with unstructured and untrained random weights. The effective linear model (35) affords a concise viewpoint on a deep finite-width non-linear network trained in the lazy regime. On an intuitive level, during training, the network effectively tunes the two matrices $W_{\text{eff.}}, C_{\text{eff.}}$ which parametrize the effective model (35). The effective weights $W_{\text{eff.}}$ controls the (linear) representation of the data, while the colored noise $C_{\text{eff.}}^{1/2}\xi$ in (35) can be loosely interpreted as inducing an effective regularization.

In fact, despite the fact that all three feature maps represented in Fig. 2 are effectively just linear feature maps, they can still encode very different biases, yielding different phenomenology. In particular, remark that the trained feature map (blue) is outperformed by mere ridge regression (red) at large sample complexities, despite the former having been priorly trained on n_0 additional samples – suggesting the trained weights $W_{\text{eff.}}, C_{\text{eff.}}$ learned some form of inductive bias which is helpful at small and moderate sample complexities, but ultimately harmful for large sample complexities.

Acknowledgements

We thank Florentin Guth, Nathanaël Cuvelle-Magar, Stéphane Mallat, Lenka Zdeborová for fruitful discussions during the course of this project. We thank the Institut d'Études Scientifiques de Cargèse for the hospitality, where discussions for this project were held throughout the 2023 *"Statistical physics & machine learning back together again"* program organised by Vittorio Erba, Damien Barbier, Florent Krzakala, BL and Lenka Zdeborová. BL acknowledges support from the *Choose France - CNRS AI Rising Talents* program. DD is supported by ETH AI Center doctoral fellowship and ETH Foundations of Data Science initiative. 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] Mikhail Belkin, Siyuan Ma, Soumik Mandal, Mikhail Belkin, and Daniel Hsu. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. U. S. A.*, 116(32):15849–15854, 2019.

[2] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. USA*, 117(48):30063–30070, 2020.

[3] 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.

[4] 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.

[5] 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(4), 2020.

[6] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. 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.

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

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

[9] Song Mei and Andrea Montanari. The generalization error of random features regression: precise asymptotics and the double descent curve. *Comm. Pure Appl. Math.*, 75(4):667–766, 2022.

[10] 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.

[11] Jacob A. Zavatone-Veth, Cengiz Pehlevan, and William L. Tong. Contrasting random and learned features in deep Bayesian linear regression. *Phys. Rev. E*, 105(6), 2022.

[12] Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 30285–30320. PMLR, 23–29 Jul 2023.

[13] David Bosch, Ashkan Panahi, and Babak Hassibi. Precise asymptotic analysis of deep random feature models, 2023.

[14] Jacob A Zavatone-Veth and Cengiz Pehlevan. Learning curves for deep structured gaussian feature models. *arXiv preprint arXiv:2303.00564*, 2023.

[15] Matthias Thamm, Max Staats, and Bernd Rosenow. Random matrix analysis of deep neural network weight matrices. *Phys. Rev. E*, 106(5):Paper No. 054124, 15, 2022.

[16] Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

[17] Florentin Guth, Brice Ménard, Gaspar Rochette, and Stéphane Mallat. A rainbow in deep network black boxes. *arXiv preprint arXiv:2305.18512*, 2023.

[18] Hugo Cui, Florent Krzakala, and Lenka Zdeborova. Bayes-optimal learning of deep random networks of extensive-width. In *International Conference on Machine Learning*, pages 6468–6521. PMLR, 2023.

[19] 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.

[20] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. *J. Stat. Mech. Theory Exp.*, (12):124005, 14, 2019.

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

[22] Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and minimum-$\ell 1$-norm interpolated classifiers. *Ann. Statist.*, 50(3):1669–1695, 2022.

[23] 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.

[24] 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.

[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 Zdeborová. 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] 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.

[30] 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.

[31] Omry Cohen, Or Malka, and Zohar Ringel. Learning curves for overparametrized deep neural networks: A field theory perspective. *Phys. Rev. Res.*, 3:023034, Apr 2021.

[32] Gadi Naveh, Oded Ben David, Haim Sompolinsky, and Zohar Ringel. Predicting the outputs of finite deep neural networks trained with noisy gradients. *Phys. Rev. E*, 104:064301, Dec 2021.

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

[34] Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. *ArXiv*, abs/1909.05989, 2019.

[35] R. Pacelli, S. Ariosto, M. Pastore, F. Ginelli, M. Gherardi, and P. Rotondo. A statistical mechanics framework for bayesian deep neural networks beyond the infinite-width limit. *Nature Machine Intelligence*, 5(12):1497–1507, Dec 2023.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. Preprint, 2023.

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

[44] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. Preprint, 2021.

[45] Chen Cheng and Andrea Montanari. Dimension free ridge regression. Preprint, 2022.

[46] Hugo Latourelle-Vigeant and Elliot Paquette. Matrix Dyson equation for correlated linearizations and test error of random features regression. Preprint, 2023.

[47] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *Ann. Appl. Probab.*, 28(2):1190–1248, 2018.

[48] Ben Adlam and Jeffrey Pennington. The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization. Preprint, 2020.

[49] Radosław Adamczak. A note on the hanson-wright inequality for random vectors with dependencies. *Electron. Commun. Prob.*, 20, 2015.

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

[51] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *Ann. Appl. Probab.*, 28(2):1190–1248, 2018.

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

[53] J.W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995.

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

[55] Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Statistics & Probability Letters*, 81(5):592–602, 2011.

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

[57] Giorgio Cipolloni, László Erdős, and Dominik Schröder. Rank-uniform local law for Wigner matrices. *Forum Math., Sigma*, 10, 2022.

[58] Clément Chouard. Deterministic equivalent of the Conjugate Kernel matrix associated to Artificial Neural Networks. Preprint, 2023.

[59] Ivan Nourdin and Giovanni Peccati. *Normal approximations with Malliavin calculus: from Stein's method to universality*, volume 192. Cambridge University Press, 2012.

[60] 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.

[61] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[62] Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation, 2022.

[63] Radosł aw Adamczak. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electron. Commun. Probab.*, 20:no. 72, 13, 2015.

[64] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. Preprint, 2010.

[65] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

A Anisotropic asymptotic equivalents

Recall from Assumption 2.2 that we assume that the feature matrices X, Z are Lipschitz-concentrated in the following sense (considering the vectors space of rectangular matrices equipped with the Frobenius norm):

Definition A.1 (Lipschitz concentration)**.** We say that a random vector x in a normed vector space $\mathcal{X}$ is Lipschitz-concentrated with constant μ if there exists a constant C such that for all 1-Lipschitz functions $f\colon \mathcal{X}\to\mathbf{R}$ it holds that

$$\mathbb{P}(|f(x)-\mathbf{E}\,f(x)|\ge t)\le C\exp\Big(-\frac{t^2}{C\mu^2}\Big). \tag{37}$$

A sufficent condition for Lipschitz concentration is that that the columns $x_i=\varphi(\mathrm{x}_i)$ are Lipschitz functions of Gaussian random vectors x_i of bounded covariance $\Omega_0:=\mathbf{E}\,\mathrm{x}_i\mathrm{x}_i^\top$, c.f. Remark 2.4. Indeed, let $\widetilde\varphi(\mathrm{g}):=\varphi(\sqrt{\Omega_0}\mathrm{g})$ and consider standard Gaussian vectors $\mathrm{g}_i,\dots,\mathrm{g}_n$. We recall that standard Gaussian random vectors are Lipschitz-concentrated with a constant which is independent of the dimension:

Theorem A.2 (Gaussian concentration)**.** *Let* g *be a random vector with independent standard Gaussian entries. Then* g *is Lipschitz-concentrated with constant* $\mu=1$.

Therefore we can stack the Gaussian vectors $\mathrm{g}_1,\dots,\mathrm{g}_n$ into $\mathrm{g}\in\mathbf{R}^{np}$ and write $X=X(\mathrm{g})=(\widetilde\varphi(\mathrm{g}_1),\dots,\widetilde\varphi(\mathrm{g}_n))$. Then X is Lipschitz-concentrated with dimension-independent constant by Theorem A.2 since for any Lipschitz $f\colon\mathbf{R}^{p\times n}\to\mathbf{R}$ it holds that $\mathrm{g}\mapsto f(X(\mathrm{g}))$ is Lipschitz due to

$$|f(X(\mathrm{g}))-f(X(\mathrm{g}'))|^2\le\|X(\mathrm{g})-X(\mathrm{g})'\|_F^2=\sum_i\|\widetilde\varphi(\mathrm{g}_i)-\widetilde\varphi(\mathrm{g}_i')\|^2\lesssim\sum_i\|\mathrm{g}_i-\mathrm{g}_i'\|^2=\|\mathrm{g}-\mathrm{g}'\|^2. \tag{38}$$

Resolvent concentration

It will be useful to introduce also the resolvent of the associated Gram matrix $X^\top X/p$ which is given by

$$\check G=\Big(\frac{X^\top X}{p}+\lambda\Big)^{-1}. \tag{39}$$

The two resolvents are related by the identity

$$\frac{X^\top GX}{p}=\frac{1}{p}X^\top\Big(\frac{XX^\top}{p}+\lambda\Big)^{-1}X=\frac{X^\top X}{p}\Big(\frac{X^\top X}{p}+\lambda\Big)^{-1}=1-\lambda\check G. \tag{40}$$

Both resolvents $G,\check G$ are Lipschitz-continuous with respect to the Frobenius norm due to the resovlent identity

$$\Big(\frac{XX^\top}{p}+\lambda\Big)^{-1}-\Big(\frac{YY^\top}{p}+\lambda\Big)^{-1}=\Big(\frac{XX^\top}{p}+\lambda\Big)^{-1}\frac{(Y-X)Y^\top+X(Y-X)^\top}{p}\Big(\frac{YY^\top}{p}+\lambda\Big)^{-1} \tag{41}$$

and the bound

$$\|GX\|\le\sqrt{p\|G\|+p\lambda\|G^2\|}\le\sqrt{2p/\lambda}, \tag{42}$$

implying

$$\|G-G'\|_F\le 2\frac{\mu}{\lambda^{3/2}p^{1/2}}\|X-Y\|_F,\quad G:=\Big(\frac{XX^\top}{p}+\lambda\Big)^{-1},\quad G':=\Big(\frac{YY^\top}{p}+\lambda\Big)^{-1}. \tag{43}$$

Therefore we obtain that

$$|\langle A(G-\mathbf{E}\,G)\rangle|\lesssim\frac{\langle|A|^2\rangle^{1/2}}{\lambda^{3/2}p},\quad |\langle A(\check G-\mathbf{E}\,\check G)\rangle|\lesssim\frac{\langle|A|^2\rangle^{1/2}}{\lambda^{3/2}p^{1/2}n^{1/2}} \tag{44}$$

from Theorem A.2,

$$|\langle A(G-G')\rangle|\le\frac{1}{p}\|A\|_F\|G-G'\|_F\le\frac{2\langle|A|^2\rangle^{1/2}}{\lambda^{3/2}p}\|X-Y\|_F. \tag{45}$$

and the analogous estimate for $\check G-\check G'$. An important special case of eq. (44) is A being rank-one which yields

$$|x^\top Gy-\mathbf{E}\,x^\top Gy|\lesssim\frac{\|x\|\|y\|}{\lambda^{3/2}p^{1/2}},\quad |x^\top\check Gy-\mathbf{E}\,x^\top\check Gy|\lesssim\frac{\|x\|\|y\|}{\lambda^{3/2}p^{1/2}} \tag{46}$$

Quadratic form and norm concentration

The other important concentration result needed in the proof of Theorem 3.1 is the concentration of quadratic forms, see e.g. Theorem 2.3 in [63].

Theorem A.3. *If x is a random vector of mean zero satisfying Lipschitz concentration with constant μ, and A is a deterministic matrix, then*

$$|x^\top Ax - \mathbf{E}\, x^\top Ax| \lesssim \mu^2 \|A\|_F. \tag{47}$$

Finally we need some upper bound on the operator norm of $X/\sqrt{p}$ which can be obtained standard ϵ-net arguments,

$$\left\| \frac{XX^\top}{n} - \Omega \right\| \prec \frac{p}{n}, \tag{48}$$

see e.g. Remark 5.40 in [64].

Leave-one-out identities

Define the leave-one-out resolvent $G_{-i} = (\lambda + p^{-1}\sum_{j\neq i} x_j x_j^\top)^{-1}$ for which we have the identity

$$\begin{aligned} G &= G_{-i} - \frac{1}{p}\frac{G_{-i}x_i x_i^\top G_{-i}}{1 + x_i^\top G_{-i}x_i/p} = G_{-i} - \lambda \frac{G_{-i}x_i x_i^\top G_{-i}}{p}\check{G}_{ii} \\ Gx_i &= G_{-i}x_i\Big(1 - \frac{1}{p}\frac{x_i^\top G_{-i}x_i}{1 + x_i^\top G_{-i}x_i/p}\Big) = \frac{G_{-i}x_i}{1 + x_i^\top G_{-i}x_i/p} = \lambda \check{G}_{ii} G_{-i}x_i \end{aligned} \tag{49}$$

where the denominators can be simplified using

$$-\frac{1}{1 + x_i^\top G_{-i}x_i/p} = \frac{x_i^\top G_{-i}x_i}{1 + x_i^\top G_{-i}x_i/p} - 1 = \frac{x_i^\top G x_i}{p} - 1 = -\lambda(\check{G})_{ii} \tag{50}$$

due to (40).

Anisotropic Marchenko-Pastur Law

We are now ready to prove Theorem 3.3, the anisotropic Marchenko-Pastur Law. In the comparable regime from Theorem 3.1 we will show that

$$|\langle [G(\lambda) - M(\lambda)A\rangle| \prec \frac{\langle |A|^2\rangle^{1/2}}{p\lambda^3}\Big(1 + \frac{p}{n} + \frac{n}{p}\Big). \tag{51}$$

Proof of Theorem 3.3. For the resolvent G we obtain the equation

$$\begin{aligned} I &= \frac{\lambda}{p}\sum_i \Big((\mathbf{E}\,\check{G}_{ii})\,\mathbf{E}\,G_{-i}\Omega + \mathbf{E}(\check{G}_{ii} - \mathbf{E}\,\check{G}_{ii})G_{-i}x_i x_i^\top \Big) + \lambda\,\mathbf{E}\,G \\ &= \mathbf{E}\,G\Big(\lambda\frac{n}{p}\langle \mathbf{E}\,\check{G}\rangle\Omega + \lambda\Big) + \frac{\lambda}{p}\sum_i \Big(\langle \mathbf{E}\,\check{G}\rangle(\mathbf{E}\,G_{-i} - \mathbf{E}\,G)\Omega + \mathbf{E}(\check{G}_{ii} - \mathbf{E}\,\check{G}_{ii})G_{-i}x_i x_i^\top \Big) \end{aligned} \tag{52}$$

so that

$$\mathbf{E}\,G = \Big(\lambda\frac{n}{p}\langle \mathbf{E}\,\check{G}\rangle\Omega + \lambda\Big)^{-1} + \frac{\lambda}{p}\sum_i \Big(\langle \mathbf{E}\,\check{G}\rangle(\mathbf{E}\,G_{-i} - \mathbf{E}\,G)\Omega + \mathbf{E}(\check{G}_{ii} - \mathbf{E}\,\check{G}_{ii})G_{-i}x_i x_i^\top \Big)\Big(\lambda\frac{n}{p}\langle \mathbf{E}\,\check{G}\rangle\Omega + \lambda\Big)^{-1}. \tag{53}$$

Using the bounds

$$\|G_{-i}x_i x_i^\top - \mathbf{E}\,G_{-i}x_i x_i^\top\|_F \le \|G_{-i}x_i x_i^\top - G_{-i}\Omega\|_F + \|(G_{-i} - \mathbf{E}_{-i}\,G_{-i})\Omega\|_F \prec \frac{1}{\lambda} + \frac{1}{p^{1/2}\lambda^{3/2}}, \tag{54}$$

$$\|\mathbf{E}\,G_{-i} - \mathbf{E}\,G\|_F = \lambda|\check{G}_{ii}|\left\|\frac{G_{-i}x_i x_i^\top G_{-i}}{p}\right\|_F \prec \frac{1}{p}\Big(\|G_{-i}\Omega G_{-i}\|_F + \|G_{-i}(x_i x_i^\top - \Omega)G_{-i}\|_F\Big) \prec \frac{1}{p^{1/2}\lambda^2} \tag{55}$$

and $|\breve{G}_{ii} - \mathbf{E}\,\breve{G}_{ii}| \prec \frac{1}{p^{1/2}\lambda^{3/2}}$ from Eq. (46) we thus obtain

$$\Big\|\mathbf{E}\,G(\lambda) - M(\lambda, \langle\mathbf{E}\,\breve{G}\rangle)\Big\|_F \prec \frac{n}{p^{3/2}\lambda^3}, \quad M(\lambda, m) := \Big(\lambda\frac{n}{p}m\Omega + \lambda\Big)^{-1}. \tag{56}$$

Note that while $M(\lambda, \langle\mathbf{E}\,\breve{G}\rangle)$ is a deterministic matrix, it still depends on the expected trace of $\breve{G}$ explicitly. However, we claim that

$$|m - \langle\mathbf{E}\,\breve{G}\rangle| \lesssim \frac{m}{p\lambda^3}, \tag{57}$$

proving

$$\|\mathbf{E}\,G - M\|_F \lesssim \frac{1}{p^{1/2}\lambda^{3/2}} + \frac{n}{p^{3/2}\lambda^3} + \|M(\lambda, \langle\mathbf{E}\,\breve{G}\rangle) - M(\lambda, m)\|_F \lesssim \frac{1}{p^{1/2}\lambda^3}\Big(1 + \frac{n}{p} + \frac{p}{n}\Big). \tag{58}$$

Now Eq. (51) follows directly together with the concentration estimate Eq. (44). Finally, eq. (57) follows from

$$\begin{aligned} |m - \langle\mathbf{E}\,\breve{G}\rangle| &\le \lambda m|\langle\mathbf{E}\,\breve{G}\rangle|\left|\frac{1}{\lambda m} - \frac{1}{\lambda\langle\mathbf{E}\,\breve{G}\rangle}\right| = \lambda m|\langle\mathbf{E}\,\breve{G}\rangle|\Big|\langle\Omega M(\lambda, m(\lambda))\rangle - \langle\Omega M(\lambda, \langle\mathbf{E}\,\breve{G}\rangle)\rangle\Big| + O\Big(\frac{m}{p\lambda^2}\Big) \\ &\le |m - \langle\mathbf{E}\,\breve{G}\rangle|\lambda^2 m|\langle\mathbf{E}\,\breve{G}\rangle|\frac{n}{p}\langle\Omega M(\lambda, m(\lambda))\Omega M(\lambda, \langle\mathbf{E}\,\breve{G}\rangle)\rangle + O\Big(\frac{m}{p\lambda^2}\Big) \\ &\le |m - \langle\mathbf{E}\,\breve{G}\rangle|\frac{\langle\Omega\rangle}{\lambda + \langle\Omega\rangle} + O\Big(\frac{m}{p\lambda^2}\Big) \end{aligned} \tag{59}$$

due to

$$\begin{aligned} \langle\Omega M(\lambda, \langle\mathbf{E}\,\breve{G}\rangle)\rangle &= \frac{p}{n\lambda\langle\mathbf{E}\,\breve{G}\rangle}\Big(1 - \lambda\langle M(\lambda, \langle\mathbf{E}\,\breve{G}\rangle)\rangle\Big) \\ &= \frac{p}{n\lambda\langle\mathbf{E}\,\breve{G}\rangle}\Big(1 - \lambda\langle\mathbf{E}\,G\rangle\Big) + O\bigg(\frac{1}{p\lambda^3\langle\mathbf{E}\,\breve{G}\rangle}\bigg) = \frac{1}{\lambda\langle\mathbf{E}\,\breve{G}\rangle} - 1 + O\bigg(\frac{1}{p\lambda^3\langle\mathbf{E}\,\breve{G}\rangle}\bigg). \end{aligned} \tag{60}$$

and

$$\lambda^2 m\langle\mathbf{E}\,\breve{G}\rangle\frac{n}{p}\langle\Omega M(\lambda, m)\Omega M(\lambda, \langle\mathbf{E}\,\breve{G}\rangle)\rangle \le \lambda m\langle\Omega M\rangle = 1 - \lambda m \le \frac{\langle\Omega\rangle}{\lambda + \langle\Omega\rangle}. \tag{61}$$

□

Multi-Resolvent Deterministic Equivalents

The key for proving Theorem 3.1 is extending the anisotropic Marchenko-Pastur to mutli-resolvent expressions, which we summarize in the following proposition. For simplicity we carry the precise error term in the comparable regime only in the first statement, the other ones being similar.

Proposition A.4.

1. *For any $A \in \mathbf{R}^{k\times p}$ we have*

$$\frac{1}{\sqrt{kp}}\langle GXZ^\top A\rangle = \frac{\lambda mn}{\sqrt{kp}}\langle M\Phi A\rangle + O\bigg(\frac{n}{k^{1/2}p^{3/2}\lambda^3}\Big(1 + \frac{n}{p} + \frac{p}{n}\Big)\bigg) \tag{62}$$

2. *For any $A \in \mathbf{R}^{p\times p}$ we have more generally*

$$\langle AG\Omega G\rangle = \frac{\langle AM\Omega M\rangle}{1 - \frac{n}{p}(m\lambda)^2\langle\Omega M\Omega M\rangle} + O\bigg(\frac{\langle|A|^2\rangle^{1/2}}{p\lambda^7}\bigg) \tag{63a}$$

while for any $A, B \in \mathbf{R}^{p\times p}$ we have

$$\langle AGBG\rangle = \langle AMBM\rangle + \frac{n}{p}(m\lambda)^2\frac{\langle AM\Omega M\rangle\langle\Omega MBM\rangle}{1 - \frac{n}{p}(m\lambda)^2\langle\Omega M\Omega M\rangle} + O\bigg(\frac{\langle|A|^2\rangle^{1/2}\|B\|}{p\lambda^7}\bigg) \tag{63b}$$

3. *For any $A \in \mathbf{R}^{p\times p}$ we have*

$$\left\langle \frac{X^\top G\Omega G X A}{p} \right\rangle = \frac{\lambda^2 m^2 \langle \Omega M \Omega M\rangle}{1-\frac{n}{p}(m\lambda)^2\langle \Omega M\Omega M\rangle}\langle A\rangle + O\left(\frac{\langle |A|^2\rangle^{1/2}}{p\lambda^7}\right) \tag{64}$$

4. *Finally, for any $A \in \mathbf{R}^{p\times p}$ we have*

$$\left\langle \frac{ZX^\top G\Omega G X Z^\top A}{kp} \right\rangle = (m\lambda)^2 \frac{n}{k} \frac{\left\langle A\left(\left(\Psi - 2\frac{n}{p}\lambda m \Phi^\top M\Phi\right)\langle \Omega M\Omega M\rangle + \frac{n}{p}\Phi^\top M\Omega M\Phi\right)\right\rangle}{1-\frac{n}{p}(m\lambda)^2\langle\Omega M\Omega M\rangle} + O\left(\frac{\langle |A|^2\rangle^{1/2}}{p\lambda^7}\right) \tag{65}$$

Before turning to the proof of Proposition A.4, we demonstrate how Proposition A.4 implies Theorem 3.1.

Proof of Theorem 3.1. By applying Proposition A.4 to the terms of Eq. (2) we obtain

$$\begin{aligned}
\mathcal{E}_{\mathrm{gen}} &= \frac{\varphi_*^\top \Psi\varphi_*}{k} + \frac{\varphi_*^\top ZX^\top G\Omega G X Z^\top \varphi_*}{kp^2} + \frac{n}{p}\left\langle \frac{X^\top G\Omega G X\Sigma}{p}\right\rangle - 2\frac{\varphi_*^\top \Phi^\top G X Z^\top\varphi_*}{kp} \\
&= \frac{1}{k}\varphi_*\left(\Psi + (m\lambda)^2\frac{n}{p}\frac{\left(\Psi - 2\frac{n}{p}\lambda m\Phi^\top M\Phi\right)\langle\Omega M\Omega M\rangle + \frac{n}{p}\Phi^\top M\Omega M\Phi}{1-\frac{n}{p}(m\lambda)^2\langle \Omega M\Omega M\rangle} - 2\lambda m\frac{n}{p}\Phi^\top M\Phi\right)\varphi_* \\
&\quad + \langle\Sigma\rangle \frac{(\lambda m)^2\frac{n}{p}\langle M\Omega M\Omega\rangle}{1-\frac{n}{p}(\lambda m)^2\langle \Omega M\Omega M\rangle} + O\left(\frac{\|\varphi_*\|^2}{p^{1/2}\lambda^7}\right).
\end{aligned} \tag{66}$$

It remains to show that the matrix in the brackets can be simplified to the expression in Theorem 3.1. For the last term in the numerator of the fraction we use

$$m\lambda\frac{n}{p}M\Omega M = M - \lambda M^2, \tag{67}$$

so that the bracket, after simplifying, becomes

$$\frac{\Psi - m\lambda\frac{n}{p}\Phi^\top(M+\lambda M^2)\Phi}{1-\frac{n}{p}(m\lambda)^2\langle\Omega M\Omega M\rangle}, \tag{68}$$

just as claimed. □

Proof of Proposition A.4. We begin with the proof of Item 1. First note that $\langle GXZ^\top A\rangle$ is a Lipschitz function of the Gaussian randomness d used to construct X and Z. Indeed, denoting G, X, Z evaluated at another realization of the Gaussian randomness by G', X', Z' we have

$$\begin{aligned}
\langle GXZ^\top A\rangle - \langle G'X'(Z')^\top A\rangle &= \langle (G-G')XZ^\top A\rangle + \langle G'(X-X')Z^\top A\rangle + \langle G'X'(Z-Z')^\top A\rangle \\
&= O\left(\frac{\|X-X'\|_F\|X\|\|Z\|\langle|A|^2\rangle^{1/2}}{\lambda^{3/2}p} + \frac{(\|X-X'\|_F\|Z\| + \|X\|\|Z-Z'\|_F)\langle |A|^2\rangle^{1/2}}{p\lambda}\right),
\end{aligned} \tag{69}$$

so that on the high probability event (recall Eq. (48)) that $\|X\| \prec \sqrt{p}$, $\|Z\| \prec \sqrt{k}$ it follows that $\langle GXZ^\top A\rangle$ is Lipschitz with constant $\langle |A|^2\rangle^{1/2}/p\lambda^{3/2}$. By estimating the complement of this high probability event trivially we can conclude

$$\left|\frac{1}{\sqrt{kp}}\langle GXZ^\top A\rangle - \frac{1}{\sqrt{kp}}\langle \mathbf{E}\, GXZ^\top A\rangle\right| \prec \frac{\langle |A|^2\rangle^{1/2}}{p\lambda^{3/2}}. \tag{70}$$

For the expectation we write out $XZ^\top$ and use eq. (49) we obtain

$$\frac{1}{\sqrt{kp}}GXZ^\top = \frac{1}{\sqrt{kp}}\sum_i Gx_iz_i^\top = \frac{1}{\sqrt{kp}}\sum_i \lambda\check{G}_{ii}G_ix_iz_i^\top. \tag{71}$$

With

$$\begin{aligned}
\frac{\lambda}{\sqrt{kp}}\mathbf{E}\sum_i(\check{G})_{ii}\langle G_i x_i z_i^\top A\rangle &= \frac{\lambda}{\sqrt{kp}}\sum_i\Big((\mathbf{E}\,\check{G}_{ii})\langle\mathbf{E}\,G_i x_i z_i^\top A\rangle + O\Big(\sqrt{\operatorname{Var}\check{G}_{ii}}\sqrt{\operatorname{Var}\langle G_i x_i z_i^\top A\rangle}\Big)\Big)\\
&= \frac{\lambda}{\sqrt{kp}}\sum_i(\mathbf{E}\,\check{G}_{ii})\langle\mathbf{E}\,G_i\Phi A\rangle + O\Big(\frac{n}{k^{1/2}p^{3/2}\lambda^2}\Big)\\
&= \frac{\lambda m n}{\sqrt{kp}}\langle M\Phi A\rangle + O\Big(\frac{n}{k^{1/2}p^{3/2}\lambda^3}\Big(1+\frac{n}{p}+\frac{p}{n}\Big)\Big)
\end{aligned} \tag{72}$$

due to Eq. (58), $\operatorname{Var}\check{G}_{ii}\lesssim\frac{1}{p\lambda^3}$ and

$$\operatorname{Var}\langle G_{-i}x_i z_i^\top A\rangle \lesssim \frac{1}{p^2}\mathbf{E}_{-i}\|AG_{-i}\|_F^2 + \operatorname{Var}_{-i}\langle G_{-i}\Phi A\rangle \lesssim \frac{\langle|A|^2\rangle}{p\lambda^3} \tag{73}$$

by eq. (44), this concludes the proof of Item 1.

We now turn to the proof of Item 2. First note that by Lipschitz concentration we have

$$|\langle AGBG - \mathbf{E}\,AGBG\rangle| \lesssim \frac{\|A\|\langle|B|^2\rangle^{1/2}}{p\lambda^{5/2}} \tag{74}$$

due to

$$|\langle AGBG\rangle - \langle AG'BG'\rangle| \le |\langle A(G-G')BG\rangle| + |\langle AG'B(G-G')\rangle| \le 2\frac{\|A\|\|B\|_F}{p\lambda}\|G-G'\|_F \tag{75}$$

and eq. (43).

It is useful to expand G around M as in

$$G = M + \lambda M\Omega G\frac{n}{p}\langle m-\check{G}\rangle - M\frac{XX^\top}{p}G + \lambda M\Omega G\frac{n}{p}\langle\check{G}\rangle = M - M\frac{XX^\top}{p}G + \lambda\langle\check{G}\rangle M\Omega G\frac{n}{p} + O\Big(\frac{1}{p\lambda^3}\Big)M\Omega G \tag{76}$$

using eq. (57) in the second step. Consequently we obtain

$$\begin{aligned}
\langle GAGB\rangle &= \langle MAGB\rangle - \langle M\frac{XX^\top}{p}GAGB\rangle + \frac{n\lambda}{p}\langle\check{G}\rangle\langle M\Omega GAGB\rangle + O\Big(\frac{\langle|A|^2\rangle^{1/2}\langle|B|^2\rangle^{1/2}}{p\lambda^6}\Big)\\
&= \langle MAMB\rangle - \frac{1}{p}\sum_i(\langle Mx_i x_i^\top GAGB\rangle - \lambda\check{G}_{ii}\langle M\Omega GAGB\rangle) + O\Big(\frac{\|B\|\langle|A|^2\rangle^{1/2}}{p\lambda^6}\Big)\\
&= \langle MAMB\rangle - \frac{\lambda}{p}\sum_i\check{G}_{ii}(\langle Mx_i x_i^\top G_{-i}AG_{-i}B\rangle - \langle M\Omega GAGB\rangle) + O\Big(\frac{\|B\|\langle|A|^2\rangle^{1/2}}{p\lambda^6}\Big)\\
&\quad + \frac{\lambda^2}{p}\sum_i\check{G}_{ii}^2\frac{x_i^\top G_{-i}AG_{-i}x_i}{p}\frac{x_i^\top G_{-i}BMx_i}{p},
\end{aligned} \tag{77}$$

using eq. (49) in the third step. The second term of eq. (77) can be estimated in expectation using

$$\begin{aligned}
\frac{\lambda}{p}\mathbf{E}\sum_i\check{G}_{ii}\langle Mx_i x_i^\top G_{-i}AG_{-i}B\rangle &= \frac{\lambda}{p}\sum_i\Big((\mathbf{E}\,\check{G}_{ii})\langle\mathbf{E}\,M\Omega G_{-i}AG_{-i}B\rangle + O\Big(\sqrt{\operatorname{Var}\check{G}_{ii}}\sqrt{\operatorname{Var}\langle Mx_i x_i^\top G_{-i}AG_{-i}B\rangle}\Big)\Big)\\
&= \frac{\lambda}{p}\sum_i(\mathbf{E}\,\check{G}_{ii})\langle\mathbf{E}\,M\Omega GAGB\rangle + O\Big(\frac{n\|B\|\langle|A|^2\rangle^{1/2}}{p^2\lambda^4}\Big)\\
&= \frac{\lambda}{p}\mathbf{E}\sum_i\check{G}_{ii}\langle M\Omega GAGB\rangle + O\Big(\frac{n\|B\|\langle|A|^2\rangle^{1/2}}{p^2\lambda^4}\Big)
\end{aligned} \tag{78}$$

since $\operatorname{Var}\check{G}_{ii}\lesssim\frac{1}{p\lambda^3}$,

$$\operatorname{Var}\langle Mx_i x_i^\top G_{-i}AG_{-i}B\rangle \lesssim \frac{1}{p^2}\mathbf{E}_{-i}\|G_{-i}AG_{-i}BM\|_F^2 + \operatorname{Var}_{-i}\langle M\Omega G_{-i}AG_{-i}B\rangle \lesssim \frac{\|B\|^2\langle|A|^2\rangle}{p\lambda^6}\Big(1+\frac{1}{p\lambda}\Big). \tag{79}$$

and

$$\|G - G_i\| \lesssim \frac{1}{p\lambda^2}, \quad \langle M\Omega G_{-i} A G_{-i} B\rangle = \langle M\Omega GAGB\rangle + O\Big(\frac{\|B\|\langle |A|^2\rangle^{1/2}}{p\lambda^4}\Big). \tag{80}$$

For the last term of Eq. (77) we have

$$\begin{aligned} \frac{x_i^\top G_{-i} A G_{-i} x_i}{p} &= \langle \Omega G_{-i} A G_{-i}\rangle + O\Big(\frac{1}{p}\|G_{-i} A G_{-i}\|_F\Big) = \langle \Omega GAG\rangle + O(\frac{\langle |A|^2\rangle^{1/2}}{p^{1/2}\lambda^2}) \\ \frac{x_i^\top G_{-i} BM x_i}{p} &= \langle \Omega G_{-i} BM\rangle + O\Big(\frac{1}{p}\|G_{-i} BM\|_F\Big) = \langle \Omega MBM\rangle + O(\frac{\langle |B|^2\rangle^{1/2}}{p^{1/2}\lambda^2}), \end{aligned} \tag{81}$$

so that with

$$\begin{aligned} \mathbf{E}\, \breve{G}_{ii}^2 \frac{x_i^\top G_{-i} A G_{-i} x_i}{p} \frac{x_i^\top G_{-i} BM x_i}{p} &= (\mathbf{E}\, \breve{G}_{ii}^2)\Big(\mathbf{E}\, \frac{x_i^\top G_{-i} A G_{-i} x_i}{p}\Big)\Big(\mathbf{E}\, \frac{x_i^\top G_{-i} BM x_i}{p}\Big) + O\Big(\frac{\langle |A|^2\rangle^{1/2}\langle |B|^2\rangle^{1/2}}{p\lambda^7}\Big) \\ &= (\mathbf{E}\, \breve{G}_{ii}^2)\langle \mathbf{E}\, \Omega GAG\rangle\langle \mathbf{E}\, \Omega GBM\rangle + O\Big(\frac{\langle |A|^2\rangle^{1/2}\langle |B|^2\rangle^{1/2}}{p\lambda^7}\Big) \end{aligned} \tag{82}$$

and

$$\frac{1}{p}\sum_i \breve{G}_{ii}^2 = \frac{n}{p}m^2 + 2\frac{n}{p}m\langle \breve{G} - m\rangle + \frac{1}{p}\sum_i (\breve{G}_{ii} - m)^2 = \frac{n}{p}m^2 + O\Big(\frac{1}{p\lambda^5} + \frac{n}{p^2\lambda^3}\Big) \tag{83}$$

we arrive at

$$\frac{\lambda^2}{p}\, \mathbf{E} \sum_i \breve{G}_{ii}^2 \frac{x_i^\top G_{-i} A G_{-i} x_i}{p} \frac{x_i^\top G_{-i} BM x_i}{p} = \frac{n}{p}(m\lambda)^2 \langle \Omega MBM\rangle\langle \mathbf{E}\, \Omega GAG\rangle + O\Big(\frac{\langle |A|^2\rangle^{1/2}\langle |B|^2\rangle^{1/2}}{p\lambda^6}\Big). \tag{84}$$

Choosing $B = \Omega$ it follows that

$$\langle GAG\Omega\rangle(1 - \frac{n}{p}\lambda^2 m^2\langle \Omega M\Omega M\rangle) = \langle MAM\Omega\rangle + O\Big(\frac{\langle |A|^2\rangle^{1/2}}{p\lambda^6}\Big), \tag{85}$$

so that the final claim Item 2 follows upon division.

Turning to the proof of Item 3 we first note that by eq. (49) we have

$$\begin{aligned} \mathbf{E}\Big(\frac{X^\top G\Omega GX}{p}\Big)_{ii} &= \lambda^2\, \mathbf{E}\, \breve{G}_{ii}^2 \langle \mathbf{E}\, \Omega G_{-i}\Omega G_{-i}\rangle + O\Big(\frac{1}{p}\sqrt{\operatorname{Var} \breve{G}_{ii}^2}\sqrt{\operatorname{Var} x_i^\top G_{-i}\Omega G_{-i} x_i}\Big) \\ &= \lambda^2\, \mathbf{E}\, \breve{G}_{ii}^2 \frac{\langle \Omega M\Omega M\rangle}{1 - \frac{n}{p}(m\lambda)^2\langle \Omega M\Omega M\rangle} + O\Big(\frac{1}{p\lambda^7}\Big), \end{aligned} \tag{86}$$

so that by a Lipschitz concentration argument as in Eq. (74) we obtain for the diagonal part A_d of $A = A_d + A_o$ that

$$\Big\langle \frac{X^\top G\Omega GX}{p} A_d\Big\rangle = \frac{\lambda^2 m^2\langle \Omega M\Omega M\rangle}{1 - \frac{n}{p}(m\lambda)^2\langle \Omega M\Omega M\rangle}\langle A_d\rangle + O\Big(\frac{\langle |A_d|^2\rangle^{1/2}}{p\lambda^7}\Big). \tag{87}$$

For the off-diagonal part we use eq. (49) twice to obtain

$$\begin{aligned} \Big(\frac{X^\top G\Omega GX}{p}\Big)_{ij} &= \frac{\lambda^2 \breve{G}_{ii}\breve{G}_{jj}}{p} x_i^\top G_{-i}\Omega G_{-j} x_j \\ &= \frac{\lambda^2 \breve{G}_{ii}\breve{G}_{jj}}{p} x_i^\top G_{-ij}\Omega G_{-ij} x_j + \frac{\lambda^4 \breve{G}_{ii}^2\breve{G}_{jj}^2}{p^3} x_i^\top G_{-ij} x_j x_j^\top G_{-ij}\Omega G_{-ij} x_i x_i^\top G_{-ij} x_j \\ &\quad - \frac{\lambda^3 \breve{G}_{ii}^2\breve{G}_{jj}}{p^2} x_i^\top G_{-ij}\Omega G_{-ij} x_i x_i^\top G_{-ij} x_j - \frac{\lambda^3 \breve{G}_{ii}\breve{G}_{jj}^2}{p^2} x_i^\top G_{-ij} x_j x_j^\top G_{-ij}\Omega G_{-ij} x_j. \end{aligned} \tag{88}$$

The second term can be estimated trivially by $p^{-3/2}\lambda^{-4}$, while for the first, third and fourth terms the trivial estimates of $p^{-1/2}\lambda^{-2}$, $p^{-1}\lambda^{-3}$ and $p^{-1/2}\lambda^{-3}$ do not suffice. For those we use the expectation and decompose $\breve{G}_{ii} = m + (\breve{G}_{ii} - m)$, $\breve{G}_{jj} = m + (\breve{G}_{jj} - m)$ to obtain

$$\mathbf{E}\, \frac{\lambda^2 \breve{G}_{ii}\breve{G}_{jj}}{p} x_i^\top G_{-ij}\Omega G_{-ij} x_j = \mathbf{E}\, \frac{\lambda^2(\breve{G}_{ii} - m)(\breve{G}_{jj} - m)}{p} x_i^\top G_{-ij}\Omega G_{-ij} x_j = O\Big(\frac{1}{\lambda^2 p^{3/2}}\Big) \tag{89}$$

and

$$\mathbf{E}\,\frac{\lambda^3\check{G}_{ii}^2\check{G}_{jj}}{p^2}x_i^\top G_{-ij}\Omega G_{-ij}x_ix_i^\top G_{-ij}x_j=\mathbf{E}\,\frac{\lambda^3(\check{G}_{ii}^2\check{G}_{jj}-m^3)}{p^2}x_i^\top G_{-ij}\Omega G_{-ij}x_ix_i^\top G_{-ij}x_j=O\Big(\frac{1}{p^{3/2}\lambda^{7/2}}\Big) \tag{90}$$

using that, say, x_j is centered and independent of x_i, G_{-ij}. By combining these estimates we obtain

$$\mathbf{E}\Big|\Big(\frac{X^\top G\Omega GX}{p}\Big)_{ij}\Big|=O\Big(\frac{1}{p^{3/2}\lambda^4}\Big), \tag{91}$$

concluding the proof of Item 3.

We now turn to the proof of (4) which follows a similar strategy as the proof of Item 2. First we note that by a Lipschitz concentration argument as in Eq. (74) it is sufficient to approximate the expectation of $ZX^\top G\Omega GXZ^\top$. By writing out $ZX^\top$ and $XZ^\top$ and using eq. (49) twice we obtain

$$\begin{aligned}\frac{1}{kp}ZX^\top G\Omega GXZ^\top&=\frac{1}{kp}\sum_{ij}z_ix_i^\top G\Omega Gx_jz_j^\top\\&=\frac{1}{kp}\sum_i(\lambda\check{G}_{ii})^2z_ix_i^\top G_{-i}\Omega G_{-i}x_iz_i^\top+\frac{1}{kp}\sum_{i\neq j}(\lambda\check{G}_{ii})(\lambda\check{G}_{jj})z_ix_i^\top G_{-i}\Omega G_{-j}x_jz_j^\top.\end{aligned} \tag{92}$$

For the first term of Eq. (92) we have

$$\begin{aligned}\frac{n}{kp}\,\mathbf{E}\langle Az_ix_i^\top G_{-i}\Omega G_{-i}x_iz_i^\top\rangle&=\frac{n}{k^2p}(\mathbf{E}\,z_i^\top Az_i)(\mathbf{E}\,x_i^\top G_{-i}\Omega G_{-i}x_i)+O\Big(\frac{n}{k^2p}\sqrt{\operatorname{Var}z_i^\top Az_i}\sqrt{\operatorname{Var}x_i^\top G_{-i}\Omega G_{-i}x_i}\Big)\\&=\frac{n}{k}\langle A\Psi\rangle\,\mathbf{E}\langle\Omega G_{-i}\Omega G_{-i}\rangle+O\Big(\frac{n\langle|A|^2\rangle^{1/2}}{p^{1/2}k^{3/2}\lambda^2}\Big)\\&=\frac{n}{k}\langle A\Psi\rangle\frac{\langle\Omega M\Omega M\rangle}{1-\frac{n}{p}(m\lambda)^2\langle\Omega M\Omega M\rangle}+O\Big(\frac{n\langle|A|^2\rangle^{1/2}}{pk\lambda^3}(1+\sqrt{p/k})\Big)\end{aligned} \tag{93}$$

using Item 2 in the ultimate step. For the second term in the right hand side of Eq. (92) we expand both G_{-i} and G_{-j} around G_{-ij} using Eq. (49) to

$$\begin{aligned}\langle z_ix_i^\top G_{-i}\Omega G_{-j}x_jz_j^\top A\rangle&\approx\Big\langle z_ix_i^\top\Big(G_{-ij}-\lambda mG_{-ij}\frac{x_jx_j^\top}{p}G_{-ij}\Big)\Omega\Big(G_{-ij}-m\lambda G_{-ij}\frac{x_ix_i^\top}{p}G_{-ij}\Big)x_jz_j^\top A\Big\rangle\\&=\langle z_ix_i^\top G_{-ij}\Omega G_{-ij}x_jz_j^\top A\rangle+(\lambda m)^2\Big\langle z_ix_i^\top G_{-ij}\frac{x_jx_j^\top}{p}G_{-ij}\Omega G_{-ij}\frac{x_ix_i^\top}{p}G_{-ij}x_jz_j^\top A\Big\rangle\\&\quad-\lambda m\Big\langle z_ix_i^\top G_{-ij}\Omega G_{-ij}\frac{x_ix_i^\top}{p}G_{-ij}x_jz_j^\top A\Big\rangle-\lambda m\Big\langle z_ix_i^\top G_{-ij}\frac{x_jx_j^\top}{p}G_{-ij}\Omega G_{-ij}x_jz_j^\top A\Big\rangle.\end{aligned} \tag{94}$$

Here in the first line we replaced $(\check{G}_{-i})_{jj}$ and $(\check{G}_{-j})_{ii}$ by m which results in an error term negligible compared to the other error terms. The first term of Eq. (94) can, in expectation, be approximated by

$$\mathbf{E}\big\langle z_ix_i^\top G_{-ij}\Omega G_{-ij}x_jz_j^\top A\big\rangle=\mathbf{E}\langle\Phi^\top G_{-ij}\Omega G_{-ij}\Phi A\rangle=\frac{\langle\Phi^\top M\Omega M\Phi A\rangle}{1-\frac{n}{p}(m\lambda)^2\langle\Omega M\Omega M\rangle}+O\Big(\frac{\langle|A|^2\rangle^{1/2}}{p\lambda^7}\Big), \tag{95}$$

using Item 2 in the ultimate step. The third term of Eq. (94) can be approximated by

$$\begin{aligned}&\lambda m\,\mathbf{E}\Big\langle z_ix_i^\top G_{-ij}\Omega G_{-ij}\frac{x_ix_i^\top}{p}G_{-ij}x_jz_j^\top A\Big\rangle\\&=\frac{1}{kp}\lambda m(x_i^\top G_{-ij}\Phi Az_i)(x_i^\top G_{-ij}\Omega G_{-ij}x_i)\\&=\lambda m\,\mathbf{E}_{-ij}\Big(\langle\Phi^\top G_{-ij}\Phi A\rangle\langle\Omega G_{-ij}\Omega G_{-ij}\rangle+O\Big(\sqrt{\operatorname{Var}_i\frac{x_i^\top G_{-ij}\Phi Az_i}{k}}\sqrt{\operatorname{Var}_i\frac{x_i^\top G_{-ij}\Omega G_{-ij}x_i}{p}}\Big)\Big)\\&=\lambda m\frac{\langle\Phi^\top M\Phi A\rangle\langle\Omega M\Omega M\rangle}{1-\frac{n}{p}(m\lambda)^2\langle\Omega M\Omega M\rangle}+O\Big(\frac{\langle|A|^2\rangle}{\lambda^7p}\Big(1+\sqrt{\frac{p}{k}}\Big)\Big)\end{aligned} \tag{96}$$

and the fourth term is exactly the same by symmetry. Here in the ultimate step we used

$$\operatorname{Var}_i \frac{x_i^\top G_{-ij}\Omega G_{-ij}x_i}{p} \lesssim \frac{1}{p^2}\|G_{-ij}\Omega G_{-ij}\|_F^2 \lesssim \frac{1}{p\lambda^2}, \quad \operatorname{Var}_i \frac{x_i^\top G_{-ij}\Phi A z_i}{k} \lesssim \frac{\langle |A|^2\rangle}{\lambda k} \tag{97}$$

and Eq. (58) and Item 2. Finally, for the second term of Eq. (92) we use the simple bound

$$\begin{aligned}
&\left\langle z_i x_i^\top G_{-ij}\frac{x_j x_j^\top}{p}G_{-ij}\Omega G_{-ij}\frac{x_i x_i^\top}{p}G_{-ij}x_j z_j^\top A\right\rangle \\
&= \frac{1}{kp^2}(x_i^\top G_{-ij}x_j)(x_j^\top G_{-ij}\Omega G_{-ij}x_i)(x_i^\top G_{-ij}x_j)(z_j^\top A z_i) \\
&= O\left(\frac{1}{kp^2}\|G_{-ij}\|_F\|G_{-ij}\Omega G_{-ij}\|_F\|G_{-ij}\|_F\|A\|_F\right) = O\left(\frac{\langle |A|^2\rangle^{1/2}}{k^{1/2}p^{1/2}\lambda^4}\right).
\end{aligned} \tag{98}$$

By combining all the above estimates we conclude the proof of Item 4. □

B Linearization of population covariance

B.1 Technical background

In this section we state several definitions and propositions from [59], that will be used further in our arguments. Let $x \in \mathbf{R}^d$ be a mean-zero Gaussian vector with covariance $\mathbf{E}\, xx^T = I$. Let $X = \{X(v) := v^\top x, \text{ for } v \in \mathbf{R}^d\}$ be a collection of jointly Gaussian centered random variables. Note that $\mathbf{E}\, X(g)X(h) = g^\top h$. The theory of *Wiener chaos*, which will be introduced shortly, can be used to study functions on the probability space $(\Omega, \mathcal{F}, P)$, where $\mathcal{F}$ is generated by X. For our needs, we only state the results for the explicit construction of X, however, note that the results from [59] are about general separable Hilbert spaces.

Following ([59], Definition 2.2.3), we write $\mathcal{H}_n$ to denote the closed linear subspace of $L^2(\Omega, \mathcal{F}, P)$ generated by the random variables of type $H_n(X(h)), h \in \mathbf{R}^d =: \mathfrak{H}, \|h\| = 1$. We call $\mathcal{H}_n$, the nth Wiener chaos.

Definition B.1. Let $L^2(\Omega, \mathfrak{H}^{\tilde{\otimes} p})$ be the space of functions $f : \mathbf{R}^{d\times p} \to \mathbf{R}$, such that f is square-integrable and

$$f(a_1, \ldots, a_p) = \frac{1}{p!}\sum_{\sigma\in S_p} f(a_{\sigma(1)}, \ldots, a_{\sigma(p)}). \tag{99}$$

Let $\mathscr{S}$ denote the set of all random variables of the form $f(X(h_1), \ldots, X(h_m))$, where $f : \mathbf{R}^m \to \mathbf{R}$ is a C^∞-function.

Definition B.2 ([59], Definition 2.3.2)**.** Let $F \in \mathscr{S}$ and $p \geq 1$ be an integer. The pth Malliavin derivative of F (with respect to X) is the element of $L^2(\Omega, \mathfrak{H}^{\tilde{\otimes} p})$, defined by

$$D^p F := \sum_{i_1,\ldots,i_p=1}^{m} \frac{\partial^p f}{\partial x_{i_1}\ldots\partial x_{i_p}}(X(h_1), \ldots, X(h_m))h_{i_1}\otimes\ldots\otimes h_{i_p}. \tag{100}$$

Proposition B.3 ([59], Proposition 2.3.7)**.** *Let $\phi : \mathbf{R}^m \to \mathbf{R}$ be a continuously differentiable function with bounded partial derivatives. Suppose that $F = (F_1, \ldots, F_m)$ is a random vector whose components are functions with derivatives in $L^q(\gamma)$, for some $q \geq 1$. Then, derivative of $\phi(F)$ also lies in $L^q(\gamma)$ and*

$$D\phi(F) = \sum_{i=1}^{m} \frac{\partial \phi}{\partial x_i}(F)DF_i. \tag{101}$$

Definition B.4 ([59], Definition 2.5.2)**.** We define $\delta^p u$ as the unique element of L^2 satisfying

$$\mathbf{E}[F\delta^p(u)] = E[\langle D^p F, u\rangle_{\mathfrak{H}^{\otimes p}}].$$

Definition B.5 ([59], Definition 2.7.1)**.** Let $p \geq 1$ and $f \in \mathfrak{H}^{\tilde{\otimes} p}$. The p-th multiple integral of f with respect to X is defined by $I_p(f) = \delta^p(f)$.

Proposition B.6 ([59], Proposition 2.7.5)**.** *Fix integers* $1 \le q \le p$ *and* $f \in \mathfrak{H}^{\tilde{\otimes} p}$ *and* $g \in \mathfrak{H}^{\tilde{\otimes} q}$*. We have*

$$\mathbf{E}\, I_p(f) I_q(g) = \delta_{pq} p! \langle f, g \rangle_{\mathfrak{H}^{\otimes p}} \tag{102}$$

Theorem B.7 ([59], Theorem 2.7.7)**.** *Let* $f \in \mathfrak{H}$ *be such that* $\|f\|_{\mathfrak{H}} = 1$*. Then, for any integer* $p \ge 1$*, we have*

$$H_p(X(f)) = I_p(f^{\otimes p}), \tag{103}$$

where H_p *is the* p*-th Hermite polynomial.*

Corollary B.8 ([59], Corollary 2.7.8)**.** *Every* $F \in L^2(\Omega)$ *can be expanded as*

$$F = \mathbf{E}\, F + \sum_{p=1}^{\infty} I_p(f_p), \tag{104}$$

for some unique collection of kernels $f_p \in \mathfrak{H}^{\tilde{\otimes} p}$*,* $p \ge 1$*. Moreover, if* $F \in C^\infty$*, then for all* $p \ge 1$*,*

$$f_p = \frac{1}{p!}\, \mathbf{E}\, D^p F. \tag{105}$$

Theorem B.9 ([59], Theorem 5.1.5)**.** *Let* $F \in C^\infty$ *be a square-integrable function. Let* $\mathbf{E}\, F = 0$ *and* $\mathbf{E}\, F^2 = \sigma^2 > 0$ *and* $N \sim \mathcal{N}(0, \sigma^2)$*. Let* $h : \mathbf{R} \to \mathbf{R}$ *be* C^2 *with* $\|h''\|_\infty < \infty$*. Then,*

$$|\mathbf{E}\, h(N) - \mathbf{E}\, h(F)| \le \frac{1}{2} \|h''\|_\infty\, \mathbf{E}\left[|\langle DF, -DL^{-1}F \rangle - \sigma^2| \right]. \tag{106}$$

For our application, we need the following expansion: for smooth odd functions f, and matrix $W \in \mathbf{R}^{k \times d}$, we can write

$$f(Wx)_i = f(w_i^\top x) = \sum_{p \ge 1} \frac{\mathbf{E}\, f^{(p)}((WW^\top)_{ii}^{1/2} N)}{p!} I_p(w_i^{\otimes p}), \tag{107}$$

where $w_i \in \mathbf{R}^d$ is the i-th row of W. Here without loss of generality we assume that x has i.i.d. entries, the general case of covariance Ω_0 then follows upon redefining $W_1 \mapsto W_1 \sqrt{\Omega_0}$.

Lemma B.10 (Weak correlation)**.** *Let* $b \ge 1$ *be an fixed integer. Let* $h_0, h_1, \ldots, h_b$ *be a collection of functions. Then, if* $\langle w_i, w_j \rangle \lesssim d^{-1/2}$ *for* $i \ne j$*, we have that*

$$\mathbf{E}\left[h_0(u^\top \varphi_1(W^1 x)) \prod_{i=1}^{b} h_i(w_i^\top x) \right] = \mathbf{E}\, h_0(u^\top \varphi_1(W^1 x)) \prod_{i=1}^{b} \mathbf{E}\, h_i(w_i^\top x) + O(d^{-1/2}). \tag{108}$$

Proof. The fact that $u_i \lesssim d^{-1/2}$ and $\varphi_1(w^\top x) \lesssim 1$ together with perturbation analysis imply that

$$\mathbf{E}\left[h_0(u^\top \varphi_1(W^1 x)) \prod_{i=1}^{b} h_i(w_i^\top x) \right] = \mathbf{E}\left[h_0\left(\sum_{k \ge b+1} u_k \varphi_1(w_k^\top x) \right) \prod_{i=1}^{b} h_i(w_i^\top x) \right] + O(d^{-1/2}). \tag{109}$$

Let $A := h_0\left(\sum_{k \ge b+1} u_k \varphi_1(w_k^\top x)\right)$ and $B := \prod_{i=1}^{b} h_i(w_i^\top x)$. Note that for any $p \ge 1$, $\langle \mathbf{E}\, D^p A, \mathbf{E}\, D^p B \rangle$ constitutes of terms $\langle w_i, w_j \rangle$ where $i \ne j$. Each of these inner products is of order $O(d^{-1/2})$ by our assumptions. Therefore, in total, $\langle \mathbf{E}\, D^p A, \mathbf{E}\, D^p B \rangle = O(d^{-p/2})$. This implies that

$$\mathbf{E}\left[h_0\left(\sum_{k \ge b+1} u_k \varphi_1(w_k^\top x) \right) \prod_{i=1}^{b} h_i(w_i^\top x) \right] = \mathbf{E}\left[h_0\left(\sum_{k \ge b+1} u_k \varphi_1(w_k^\top x) \right) \right] \mathbf{E}\left[\prod_{i=1}^{b} h_i(w_i^\top x) \right] + O(d^{-1/2}). \tag{110}$$

Similarly, it follows that $\mathbf{E} \prod_{i=1}^{b} h_i(w_i^\top x) = \prod_{i=1}^{b} \mathbf{E}\, h_i(w_i^\top x) + O(d^{-1/2})$ and finally, using perturbation analysis again, we conclude that

$$\mathbf{E}\left[h_0(u^\top \varphi_1(W^1 x)) \prod_{i=1}^{b} h_i(w_i^\top x) \right] = \mathbf{E}\, h_0(u^\top \varphi_1(W^1 x)) \prod_{i=1}^{b} \mathbf{E}\, h_i(w_i^\top x) + O(d^{-1/2}) \tag{111}$$

□

B.2 One layer linearization

Consider a mean-zero Gaussian random vector $x \in \mathbf{R}^d$ with covariance $\mathbf{E}\, xx^\top = I$, two weight matrices $W \in \mathbf{R}^{k\times d}, V \in \mathbf{R}^{k\times d}$ and two smooth odd functions φ, ψ applied entrywise to Wx, Vx. We assume that rows of W and V satisfy Assumption 4.1; in particular, they are i.i.d. mean-zero samples $(w_i, v_i) \sim (w, v)$. We define $C_w := C_1/k = \mathbf{E}\, ww^\top$, $C_v := \widetilde{C}_1/k = \mathbf{E}\, vv^\top$ and $C_{wv} := \breve{C}_1/k = \mathbf{E}\, wv^\top$. Note that we also have $\operatorname{Tr} C_w, \operatorname{Tr} C_v, \operatorname{Tr} C_{wv} \lesssim 1$.

Let N_w, N_v be jointly Gaussian mean-zero random variables, such that

$$\mathbf{E}\, N_w^2 = \operatorname{Tr} C_w, \quad \mathbf{E}\, N_v^2 = \operatorname{Tr} C_v, \quad \mathbf{E}\, N_w N_v = \operatorname{Tr} C_{wv}. \tag{112}$$

Define

$$\begin{aligned} \Phi_1 &= \mathbf{E}\, \varphi(Wx)\psi(Vx)^\top, \\ \Phi_1^{\text{lin}} &= (\mathbf{E}\, \varphi'(N_w))(\mathbf{E}\, \psi'(N_v))WV^\top + [\mathbf{E}\, \varphi(N_w)\psi(N_v) - (\mathbf{E}\, \varphi'(N_w))(\mathbf{E}\, \psi'(N_v))(\mathbf{E}\, N_w N_v)]I. \end{aligned} \tag{113}$$

Proposition B.11. *We have that, with high probability,* $\|\Phi_1 - \Phi_1^{\text{lin}}\|_F = O(1)$.

Proof. Using a Wiener chaos expansion (eq. (107)), we can write

$$\varphi(Wx)_i = \sum_{p\geq 1} \frac{\mathbf{E}\, \varphi^{(p)}((WW^\top)_{ii}^{1/2}N)}{p!} I_p(Wx)_i, \quad \psi(Vx)_j = \sum_{p\geq 1} \frac{\mathbf{E}\, \psi^{(p)}((VV^\top)_{jj}^{1/2}N)}{p!} I_p(Vx)_j \tag{114}$$

where $N \sim \mathcal{N}(0,1)$ and $I_p(Wx), I_q(Vx)$ are random vectors with covariance

$$\mathbf{E}\, I_p(Wx) I_q(Vx)^\top = p!\delta_{pq}(WV^\top)^{\odot p} \tag{115}$$

with $A^{\odot p}$ denoting the p-th entrywise (Hadamard) power. Thus we have the identity

$$\mathbf{E}\, \varphi(Wx)_i \psi(Vx)_j = \sum_{p\geq 1} \frac{1}{p!} (\mathbf{E}\, \varphi^{(p)}((WW^\top)_{ii}^{1/2}N))(WV^\top)_{ij}^p (\mathbf{E}\, \psi^{(p)}((VV^\top)_{jj}^{1/2}N)). \tag{116}$$

From concentration of quadratic forms assumption for w, v, it follows that

$$(WW^\top)_{ii} = \operatorname{Tr} C_w + O(d^{-1/2}), \quad (VV^\top)_{jj} = \operatorname{Tr} C_v + O(d^{-1/2}), \tag{117}$$

$$(WV^\top)_{ij} = \delta_{ij} \operatorname{Tr} C_{wv} + O(d^{-1/2}). \tag{118}$$

From perturbation analysis, we can write

$$\mathbf{E}\, \varphi^{(p)}((WW^\top)_{ii}^{1/2}N) = \mathbf{E}\, \varphi^{(p)}(\sqrt{\operatorname{Tr} C_w} N) + O(d^{-1/2}) = \mathbf{E}\, \varphi^{(p)}(N_w) + O(d^{-1/2}), \tag{119}$$

and similarly $\mathbf{E}\, \psi^{(p)}((VV^\top)_{jj}^{1/2}N) = \mathbf{E}\, \psi^{(p)}(N_v) + O(d^{-1/2})$.

off-diagonal entries Here, for $p \geq 2$, we have that $(WV^\top)_{ij}^p = O(d^{-p/2})$. Therefore,

$$\mathbf{E}\, \varphi(Wx)_i \psi(Vx)_j = \mathbf{E}\, \varphi'(N_w)\psi'(N_v)(WV^\top)_{ij} + O(d^{-1}) = (\Phi_1^{\text{lin}})_{ij} + O(d^{-1}). \tag{120}$$

diagonal entries We rewrite the infinite sum as

$$\begin{aligned} \mathbf{E}\, \varphi(Wx)_i \psi(Vx)_i &= \sum_{p\geq 1} \frac{1}{p!} (\mathbf{E}\, \varphi^{(p)}((WW^\top)_{ii}^{1/2}N))(WV^\top)_{ii}^p (\mathbf{E}\, \psi^{(p)}((VV^\top)_{ii}^{1/2}N)) \\ &= \sum_{p\geq 1} \frac{[\mathbf{E}\, \varphi^{(p)}(\sqrt{\operatorname{Tr} C_w} N)][\mathbf{E}\, \psi^{(p)}(\sqrt{\operatorname{Tr} C_v} N)]}{p!} (\operatorname{Tr} C_{wv})^p + O(d^{-1/2}) \\ &= \mathbf{E}\, \varphi(N_w)\psi(N_v) + O(d^{-1/2}) = (\Phi_1^{\text{lin}})_{ii} + O(d^{-1/2}). \end{aligned} \tag{121}$$

Summing up over all entries, we conclude that $\|\Phi_1 - \Phi_1^{\text{lin}}\|_F = O(1)$. □

Note that in case of independent N_v, N_w (i.e., independent v, w) the second term of Φ_1^{lin} vanishes and in case of $W = V$, $f \equiv g$ this reduces to

$$\Phi_1^{\text{lin}} = (\mathbf{E}\, \varphi'(N_w))^2 WW^\top + [\mathbf{E}\, \varphi(N_w)^2 - (\mathbf{E}\, \varphi'(N_w))^2 \operatorname{Tr} C_w]I. \tag{122}$$

B.3 Two layer case

We now consider the simplest 2-layer example

$$\varphi_2(W^2\varphi_1(W^1x)), \quad \psi(Vx) \tag{123}$$

with general matrices $W^1, W^2, V \in \mathbf{R}^{k\times d}$ and smooth odd functions $\varphi_1, \varphi_2, \psi$. We assume that rows of W^1, W^2 and V satisty Assumption 4.1; in particular, they are mean-zero i.i.d. samples $(w_i^1, w_i^2, v_i) \sim (w^1, w^2, v)$. Define $C_{1,w} := C_1/k = \mathbf{E}\, w^1(w^1)^\top, C_{2,w} := C_2/k = \mathbf{E}\, w^2(w^2)^\top$, $C_v := \widetilde{C}_1/k = \mathbf{E}\, vv^\top$ and $C_{wv} := \breve{C}_1/k = \mathbf{E}\, w^1v^\top$. Define

$$\begin{aligned}
&\Phi_1 := \mathbf{E}\,\varphi_1(W^1x)\psi(Vx)^\top,\\
&\Phi_2 := \mathbf{E}\,\varphi_2(W^2\varphi_1(W^1x))\psi(Vx)^\top,\\
&\Phi_1^{\text{lin}} := (\mathbf{E}\,\varphi_1'(N_{1,w}))(\mathbf{E}\,\psi'(N_{1,v}))WV^\top + [\mathbf{E}\,\varphi_1(N_{1,w})\psi(N_{1,v}) - (\mathbf{E}\,\varphi_1'(N_{1,w}))(\mathbf{E}\,\psi'(N_{1,v}))(\mathbf{E}\, N_{1,w}N_{1,v})]I,\\
&\Phi_2^{\text{lin}} := (\mathbf{E}\,\varphi_2'(N_{2,w}))W\Phi_1^{\text{lin}},\\
&\Omega_1 := \mathbf{E}\,\varphi_1(W^1x)\varphi_1(W^1x)^\top,\\
&\Omega_1^{\text{lin}} := (\mathbf{E}\,\varphi_1'(N_{1,w}))^2W^1(W^1)^\top + [\mathbf{E}\,\varphi_1(N_{1,w})^2 - (\mathbf{E}\,\varphi_1'(N_{1,w}))^2\operatorname{Tr} C_1]I,\\
&\Omega_2 := \mathbf{E}\,\varphi_2(W^2\varphi_1(W^1x))\varphi_2(W^2\varphi_1(W^1x))^\top,\\
&\Omega_2^{\text{lin}} := (\mathbf{E}\,\varphi_2'(N_{2,w}))^2W^2\Omega_1^{\text{lin}}(W^2)^\top + [\mathbf{E}\,\varphi_2(N_{2,w})^2 - (\mathbf{E}\,\varphi_2'(N_{2,w}))^2\operatorname{Tr} C_1]I,
\end{aligned} \tag{124}$$

where $N_{1,w}, N_{2,w}, N_v$ are zero-mean jointly Gaussian.

$$\mathbf{E}\, N_{1,w}^2 = \operatorname{Tr}(C_{1,w}) \quad \mathbf{E}\, N_v^2 = \operatorname{Tr}(C_v) \quad \mathbf{E}\, N_{1,w}N_v = \operatorname{Tr}(C_{wv}) \quad \mathbf{E}\, N_{2,w}^2 = \operatorname{Tr}(C_{2,w}\Omega_1^{\text{lin}}) \tag{125}$$

Theorem B.12. *We have that, with high probability, (i) $\|\Phi_2 - \Phi_2^{\text{lin}}\|_F = O(1)$ and (ii) $\|\Omega_2 - \Omega_2^{\text{lin}}\|_F = O(1)$.*

We split the proof into following lemmas:

Lemma B.13 (Diagonal entries of Ω_2)**.** *For row v of matrix W^2, with high probability,*

$$\mathbf{E}\,\varphi_2(v^\top\varphi_1(W^1x))^2 = \mathbf{E}\,\varphi_2(N_{2,w})^2 + O(d^{-1/2}). \tag{126}$$

Lemma B.14 (Off-diagonal entries Ω_2)**.** *If u and z are independent rows, such that $u_i \lesssim d^{-1/2}$ and $z_i \lesssim d^{-1/2}$, we have with high probability,*

$$\begin{aligned}
&\mathbf{E}\,\varphi_2(u^\top\varphi_1(Wx))\psi_2(z^\top\psi_1(Vx))\\
&\quad = \mathbf{E}\,\varphi_2'(u^\top\varphi_1(Wx))\,\mathbf{E}\,\psi_2'(z^\top\psi_1(Vx))u^\top\,\mathbf{E}\left[\varphi_1(Wx)\psi_1(Vx)^\top\right] z + O(d^{-1}).
\end{aligned} \tag{127}$$

Lemma B.15 (Entries of Φ_2)**.** *For rows u, v of matrices W^2, V respectively, with high probability*

$$\mathbf{E}\,\varphi_2(u^\top\varphi_1(W^1x))\psi(v^\top x) = \mathbf{E}\big[\varphi_2'(u^\top\varphi_1(W^1x))\big]u^\top\,\mathbf{E}\,\varphi_1(W^1x)\psi(v^\top x) + O(d^{-1}). \tag{128}$$

Proof of Theorem B.12. Lemma B.15 implies that

$$\|\Phi_2 - \mathbf{E}\big[\varphi_2'(u^\top\varphi_1(Wx))\big]W^2\Phi_1\|_F = O(1). \tag{129}$$

Note that, since from Proposition B.11 $\|\Phi_1 - \Phi_1^{\text{lin}}\|_F = O(1)$ and since $\|W^2\| = O(1)$, we have that

$$\|\mathbf{E}\big[\varphi_2'(u^\top\varphi_1(Wx))\big]W^2\Phi_1 - \mathbf{E}\big[\varphi_2'(u^\top\varphi_1(Wx))\big]W^2\Phi_1^{\text{lin}}\|_F = O(1), \tag{130}$$

therefore, by triangle inequality,

$$\|\Phi_2 - \mathbf{E}\big[\varphi_2'(u^\top\varphi_1(Wx))\big]W^2\Phi_1^{\text{lin}}\|_F = O(1). \tag{131}$$

Finally, note that by simple chaos expansion, $\mathbf{E}\big[\varphi_2'(u^\top\varphi_1(Wx))\big] = \mathbf{E}\,\varphi_2'(N_{2,w})$, therefore $\|\Phi_2 - \Phi_2^{\text{lin}}\|_F = O(1)$.

Furthermore, Lemma B.13 together with Lemma B.14 imply that

$$\|\Omega_2 - (\mathbf{E}\big[\varphi_2'(u^\top\varphi_1(Wx))\big])^2W^2\Omega_1(W^2)^\top - (\mathbf{E}\,\varphi_2(N_{2,w})^2 - (\mathbf{E}\big[\varphi_2'(u^\top\varphi_1(Wx))\big])^2)I\|_F = O(1). \tag{132}$$

Again, using the fact that $\|\Omega_1 - \Omega_1^{\text{lin}}\|_F = O(1)$, $\|W^2\| = O(1)$ and approximating $\mathbf{E}\,\varphi_2'(u^\top\varphi_1(Wx))$ with $\mathbf{E}\,\varphi_2'(N_{2,w})$, we obtain that $\|\Omega_2 - \Omega_2^{\text{lin}}\|_F = O(1)$. □

B.4 Proof of Lemma B.13

Let $F = v^\top \varphi_1(W^1 x)$. Our goal is to compute $\mathbf{E}\,\varphi_2(v^\top \varphi_1(W^1 x))^2 = \mathbf{E}\,\varphi_2(F)^2$. For simplicity we omit indices in φ_1 and W^1 and write just φ and W. We can decompose

$$F = v^\top \varphi(Wx) = \sum_p I_p\left(\frac{\mathbf{E}\,D^p F}{p!}\right) = \sum_p I_p\left(\sum_i \frac{v_i\,\mathbf{E}\,D^p \varphi(w_i^\top x)}{p!}\right) = \sum_p I_p\left(\sum_i \frac{v_i w_i^{\otimes p}\,\mathbf{E}\,\varphi^{(p)}(w_i^\top x)}{p!}\right). \tag{133}$$

Let $\varphi_i^p := \mathbf{E}\,\varphi^{(p)}(w_i^\top x)$. We obtain that

$$DF = \sum_{p\geq 1} p I_{p-1}\left(\sum_i \frac{v_i w_i^{\otimes p}\varphi_i^p}{p!}\right) \quad \text{and} \quad -DL^{-1}F = \sum_{q\geq 1} I_{q-1}\left(\sum_i \frac{v_i w_i^{\otimes q}\varphi_i^q}{q!}\right). \tag{134}$$

Lemma B.16.

$$\mathbf{E}|\langle DF, -DL^{-1}F\rangle - \mathbf{E}\,F^2| = O(d^{-1/2}). \tag{135}$$

Proof. Note that

$$I_{p-1}\left(\sum_i \frac{v_i w_i^{\otimes p}\varphi_i^p}{p!}\right) = \sum_i \frac{v_i \varphi_i^p I_{p-1}(w_i^{\otimes p-1}) w_i}{p!}, \tag{136}$$

which implies that, for some coefficients $c_{p,q}$,

$$\langle DF, -DL^{-1}F\rangle = \sum_{p,q\geq 1} c_{p,q} \sum_{i,j} \langle w_i, w_j\rangle v_i v_j \varphi_i^p \varphi_j^q I_{p-1}(w_i^{\otimes p-1}) I_{q-1}(w_j^{\otimes q-1}). \tag{137}$$

Now, using ([59], Theorem 2.7.10), we can rewrite

$$\begin{aligned} &I_{p-1}(w_i^{\otimes p-1}) I_{q-1}(w_j^{\otimes q-1}) \\ &= \sum_{r=0}^{p\wedge q-1} \langle w_i, w_j\rangle^r c_{r,p,q} I_{p+q-2(r+1)}(w_i^{\otimes p-1-r} \tilde{\otimes} w_j^{\otimes q-1-r}) \\ &= \sum_{\substack{s=|p-q| \\ 2 \text{ divides } (s-|p-q|)}}^{p+q-2} c_{r,p,q} \langle w_i, w_j\rangle^{(p+q-2-s)/2} I_s\left(w_i^{\otimes(p-q+s)} \tilde{\otimes} w_j^{\otimes(q-p+s)}\right), \end{aligned} \tag{138}$$

and plugging this expression back into eq. (137), we get

$$\langle DF, -DL^{-1}F\rangle = \sum_{s\geq 0} \sum_{\substack{|p-q|\leq s \\ 2 \text{ divides } (s-|p-q|) \\ p\wedge q \geq 1 + (s-|p-q|)/2}} \tilde{c}_{r,p,q} \sum_{i,j} \langle w_i, w_j\rangle^{(p+q-s)/2} v_i v_j \varphi_i^p \varphi_j^q I_s(w_i^{\otimes(s+p-q)/2} \tilde{\otimes} w_j^{\otimes(s+q-p)/2}). \tag{139}$$

In the sum above, term $s = 0$ corresponds to $\mathbf{E}\,F^2$. Let us collect all the terms corresponding to sth multiple integral. Note that given conditions on s, p, q, we always have that $(p + q - s)/2 \geq 1$, which is the power of the inner product $\langle w_i, w_j\rangle$ in the expression above. If we introduce $a := (p + q - s)/2$, then for fixed s, sth multiple integral I_s can be rewritten as follows:

$$I_s\left(\sum_{a\geq 1} \sum_{i,j} \langle w_i, w_j\rangle^a v_i v_j T_{ij}^s\right), \tag{140}$$

where T_{ij}^s is some s-dimensional tensor, which is a sum of tensor products of w_i and w_j, also containing combinatorial terms, and products of expectations of derivatives of f. Then, we can write

$$\mathbf{E}(\langle DF, -DL^{-1}F\rangle - \mathbf{E}\,F^2)^2 = \sum_{s\geq 1} \mathbf{E}\,I_s\left(\sum_{a\geq 1} \sum_{i,j} \langle w_i, w_j\rangle^a v_i v_j T_{ij}^s\right)^2. \tag{141}$$

Observe that

$$\mathbf{E}\,I_s\left(\sum_{a\geq 1} \sum_{i,j} \langle w_i, w_j\rangle^a v_i v_j T_{ij}^s\right)^2 = \sum_{a,a'\geq 1} \sum_{\substack{i,j \\ i',j'}} \langle w_i, w_j\rangle^a \langle w_{i'}, w_{j'}\rangle^{a'} v_i v_j v_{i'} v_{j'} \langle T_{ij}^s, T_{i'j'}^s\rangle, \tag{142}$$

and note that for some constant C (depending on combinatorial terms, and products of expectations of derivatives of f) $\langle T^s_{ij}, T^s_{i'j'}\rangle$ can be upper bounded by

$$\langle T^s_{ij}, T^s_{i'j'}\rangle \leq C(\langle w_i, w_{i'}\rangle + \langle w_i, w_{j'}\rangle + \langle w_j, w_{i'}\rangle + \langle w_j, w_{i'}\rangle)^s. \tag{143}$$

Now, we analyze each term of the summand in eq. (142) depending on s, a, a', i, i', j, j'. Define $N := |\{i, i', j, j'\}|$, the number of distinct indices among i, i', j and j'. Note that since entries of v are of order $v_i \lesssim d^{-1/2}$, we get that in total $v_i v_j v_{i'} v_{j'}$ contribute $O(d^{-2})$.

Case $N = 1$ Here, since there are in total d such terms, which immediately obtain $O(d^{-1})$ upper bound.

Case $N = 2$ There are in total $O(d^2)$ such terms. In this case, it must be that either (1) $i \neq j$ or (2) $i' \neq j'$ or (3) both $i = j$, $i' = j'$. Note that in the first two cases, we get that $\langle w_i, w_j\rangle^a \langle w_{i'}, w_{j'}\rangle^{a'} \lesssim d^{-1/2}$, which together with bound on v_i's gives $O(d^{2-2-1/2})$ contribution. If both $i = j$ and $i' = j'$, then necessarily $\langle T^s_{ij}, T^s_{i'j'}\rangle = O(d^{-1/2})$ and we arrive at the same conclusion.

Case $N = 3$**,** $\min(a, a', s) \geq 2$ There are in total $O(d^3)$ such terms. WLOG assume that $i = j$. Since $\min(a, a', s) \geq 2$, we get that $\langle w_{i'}, w_{j'}\rangle^{a'} = O(d^{-1})$ and $\langle T^s_{ij}, T^s_{i'j'}\rangle = O(d^{-1})$, which in total gives $O(d^{-1})$ contribution.

Case $N = 4$ Here, there are in total $O(d^4)$ such terms. If $\min(a, a') \geq 2$, then it total we obtain $O(d^{4-2-2-1/2}) = O(d^{-1/2})$ contribution. We start with the case $a = a' = 1$. Let

$$\begin{aligned}
X &= \sum_{\substack{i,j,i',j' \\ \text{all distinct}}} \langle w_i, w_j\rangle\langle w_{i'}, w_{j'}\rangle v_i v_j v_{i'} v_{j'} \langle T^1_{ij}, T^1_{i'j'}\rangle \qquad \text{with } \mathbf{E}_W X = 0, \\
X^2 &= \sum_{\substack{i_1,j_1,i'_1,j'_1 \\ \text{all distinct}}} \sum_{\substack{i_2,j_2,i'_2,j'_2 \\ \text{all distinct}}} \langle w_{i_1}, w_{j_1}\rangle\langle w_{i'_1}, w_{j'_1}\rangle\langle w_{i_2}, w_{j_2}\rangle\langle w_{i'_2}, w_{j'_2}\rangle v_{i_1} v_{j_1} v_{i'_1} v_{j'_1} v_{i_2} v_{j_2} v_{i'_2} v_{j'_2} \langle T^s_{i_1j_1}, T^s_{i'_1j'_1}\rangle\langle T^s_{i_2j_2}, T^s_{i'_2j'_2}\rangle.
\end{aligned} \tag{144}$$

Since w_a is independent from w_b for $a \neq b$, and all rows are mean-zero, we conclude that the only non-zero contribution comes from terms with pairings between indices. Therefore, we get that $\mathbf{E}_W X^2 = O(d^{-3})$, which implies that, with high probability, $X = O(d^{-1/2})$.

When $a = 1$ and $a' = 2$, by similar computation one obtains that $E_W X^2 = O(d^{-2})$, and therefore, with high probability, $X = O(d^{-1/2})$.

Case $N = 3$**,** $\min(a, a', s) = 1$ This case can be done similarly to the previous ones. □

Overall, we obtain that

$$\mathbf{E}(\langle DF, -DL^{-1}F\rangle - \mathbf{E}\, F^2)^2 = O(d^{-1/2}), \tag{145}$$

which, by Theorem B.9, using $h(x) = \varphi_2(x)^2$ that

$$\mathbf{E}\, \varphi_2(u^\top \varphi_1(W^1 x))^2 = \mathbf{E}\, \varphi_2(Z)^2 + O(d^{-1/2}), \tag{146}$$

where $\mathbf{E}\, Z^2 = u^\top\, \mathbf{E}\, \varphi_1(W^1 x)\, \mathbf{E}\, \varphi_1(W^1 x)^\top u$. By perturbation analysis, we obtain that

$$\mathbf{E}\, \varphi_2(u^\top \varphi_1(W^1 x))^2 = \mathbf{E}\, \varphi_2(N_{2,w})^2 + O(d^{-1/2}) \tag{147}$$

B.5 Proof of Lemma B.15

We restate the lemma:

Lemma B.17.

$$\mathbf{E}\, \varphi_2(u^\top \varphi_1(Wx))\psi(v^\top x) = \mathbf{E}\big[\varphi_2'(u^\top \varphi_1(Wx))\big] u^\top\, \mathbf{E}\, \varphi_1(Wx)\psi(v^\top x) + O(d^{-1}) \tag{148}$$

Our goal is to compute

$$\mathbf{E}\,\varphi_2(u^\top \varphi_1(Wx))\psi(v^\top x). \tag{149}$$

Let $F_i = \varphi_1(w_i^\top x)$ and $F = (F_1, \ldots, F_n)$, where w_i's are the rows of W. Note that we can view $\varphi_2(u^\top \varphi_1(Wx)) = \phi(F)$, for $\phi(F) = \varphi_2(\sum u_k F_k)$. Recall that $D^p F_k = \varphi^{(p)}(w_k^\top x) w_k^{\otimes p}$. Then, we have that

$$D\phi(F) = \sum_k \varphi_2'(u^\top \varphi_1(Wx)) u_k DF_k. \tag{150}$$

From Lemma B.10, it follows that

$$\mathbf{E}\,D\phi(F) = \sum_k \mathbf{E}\big[\varphi_2'(u^\top \varphi_1(Wx))\big] u_k\, \mathbf{E}\,DF_k + O(d^{-1/2}), \tag{151}$$

and we obtain that

$$\langle \mathbf{E}\,D\psi(v^\top x), \mathbf{E}\,D\varphi_2(u^\top \varphi_1(Wx))\rangle = \mathbf{E}\big[\varphi_2'(u^\top \varphi_1(Wx))\big] \sum_k u_k \langle \mathbf{E}\,D\psi(v^\top x), \mathbf{E}\,D\varphi_1(w_i^\top x)\rangle. \tag{152}$$

For the second derivative, we obtain that

$$D^2\phi(F) = \sum_{k,k'} \varphi_2''(u^\top \varphi_1(Wx)) u_k u_k' (DF_k) \otimes (DF_{k'}) + \sum_k \varphi_2'(u^\top \varphi_1(Wx)) u_k D^2 F_k. \tag{153}$$

Overall, for $\mathbf{E}\,D^p\phi(F)$, we obtain

$$\mathbf{E}\,D^p\phi(F) = \sum_k \mathbf{E}\left[\varphi_2'(u^\top \varphi_1(Wx)) u_k D^p F_k\right] + \mathbf{E}\,R \tag{154}$$

where R is the term containing higher derivatives of φ_2. This implies that

$$\langle \mathbf{E}\,D^p\phi(F), \mathbf{E}\,D^p\psi(v^\top x)\rangle = \sum_k u_k\, \mathbf{E}\left[f_2'(u^\top \varphi_1(Wx))\varphi^{(p)}(w_k^\top x)\right] \mathbf{E}\left[\psi^{(p)}(v^\top x)\right] \langle w_k, v\rangle^p + \langle \mathbf{E}\,R, \mathbf{E}\,D^p\psi(v^\top x)\rangle. \tag{155}$$

From Lemma B.10, we have that

$$\mathbf{E}\left[\varphi_2'(u^\top \varphi_1(Wx))\varphi^{(p)}(w_k^\top x)\right] = \mathbf{E}\left[\varphi_2'(u^\top \varphi_1(Wx))\right] \mathbf{E}\left[\varphi^{(p)}(w_k^\top x)\right] + Q, \tag{156}$$

where $Q = O(d^{-1/2})$. We now show that terms involving Q have $O(d^{-1})$ contribution to the final expression:

Lemma B.18.

$$\sum_k u_k \langle w_k, v\rangle^p = O(d^{-1/2}) \tag{157}$$

Proof. Since v can be correlated with at most one row of W by our assumptions, WLOG we assume that v is uncorrelated with all w_k for $k \geq 2$. Then we can write

$$\sum_k u_k \langle w_k, v\rangle^p = u_1 \langle w_1, v\rangle^p + \sum_{k\geq 2} u_k \langle w_k, v\rangle^p, \tag{158}$$

where the first term is $O(d^{-1/2})$ since $u_i = O(d^{-1/2})$. Next, if $p \geq 2$, we directly obtain that the second term is $O(d^{-1/2})$. For case $p = 1$, note since $\sum_{k\geq 2} u_k \langle w_k, v\rangle$ is the sum of independent mean-zero random variables (in the W space), we obtain that $\sum_k u_k \langle w_k, v\rangle = O(d^{-1/2})$. Altogether, we get that $\sum_k u_k \langle w_k, v\rangle^p = O(d^{-1/2})$ with high probability. □

Lemma B.18 implies that

$$\begin{aligned}
&\langle \mathbf{E}\,D^p\phi(F), \mathbf{E}\,D^p\psi(v^\top x)\rangle \\
&= \sum_k u_k\, \mathbf{E}\left[\varphi_2'(u^\top \varphi_1(Wx))\right] \mathbf{E}\left[\varphi^{(p)}(w_k^\top x)\right] \mathbf{E}\left[\psi^{(p)}(v^\top x)\right] \langle w_k, v\rangle^p + \langle \mathbf{E}\,R, \mathbf{E}\,D^p\psi(v^\top x)\rangle + O(d^{-1}) \\
&= \mathbf{E}\left[\varphi_2'(u^\top \varphi_1(Wx))\right] \sum_k u_k \langle \mathbf{E}\,D^p F_k, \mathbf{E}\,D^p\psi(v^\top x)\rangle + \langle \mathbf{E}\,R, \mathbf{E}\,D^p\psi(v^\top x)\rangle + O(d^{-1}).
\end{aligned} \tag{159}$$

Lemma B.19.

$$\langle \mathbf{E}\, R, \mathbf{E}\, D^p\psi(v^\top x)\rangle = O(d^{-1}). \tag{160}$$

Proof. Note that we can rewrite

$$\langle \mathbf{E}\, R, \mathbf{E}\, D^p\psi(v^\top x)\rangle = \sum_{\substack{\pi\vdash[p]\\ \pi\neq\{[p]\}}} \sum_{i_1,\ldots,i_{|\pi|}} \prod_{q=1}^{|\pi|} \left(u_{i_q}\langle v, w_{i_q}\rangle^{b(q)}\right), \tag{161}$$

where for π (partition of $[p]$) we denote $b(q)$ as the size of qth block. Let $\pi \neq \{[p]\}$ be some partition, $s := |\pi|$, and let $A_\pi := \sum_{i_1,\ldots,i_s}\prod_{q=1}^{s}\left(u_{i_q}\langle v, w_{i_q}\rangle^{b(q)}\right)$. If all blocks of π are of size at least 2, then naive estimate gives in total the contribution is $O(d^{-1})$. Now assume that there is a block of size 1. WLOG we assume that v is correlated with w_1 and $b(1) = 1$. If $i_1 = 1$, then let r be the number of indices among $s-1$ remaining indices, such that corresponding index is summed over $i_k \geq 2$. If $r \leq s-2$, then we arrive at the estimate $O(d^{r-s/2-r/2}) = O(d^{-1})$. Now, assume that $r = s-1$. If two indices coincide, then we arrive at the estimate $O(d^{(s-2)-s/2-(s-1)/2}) = O(d^{-1})$. Therefore, the remaining case is

$$B_\pi := \sum_{i_2\neq i_3\neq\ldots\neq i_s\neq 1} u_1\langle v, w_1\rangle \prod_{q=2}^{s} u_{i_q}\langle v, w_{i_q}\rangle^{b(q)}. \tag{162}$$

If exists $q > 1$, such that $b(q) \geq 1$, then from naive estimate we obtain bound $O(d^{s-1-s/2-(s-2)/2-1}) = O(d^{-1})$. Therefore, we can assume that all blocks of π are of size 1. By independence of rows, the W-expectation of B_π is zero and

$$\mathbf{E}_W\, B_\pi^2 = \sum_{i_2\neq i_3\neq\ldots\neq i_{|\pi|}\neq 1}\ \sum_{i'_2\neq i'_3\neq\ldots\neq i'_{|\pi|}\neq 1} u_1^2\langle v, w_1\rangle^2 \prod_{q=2}^{p} u_{i_q}u_{i'_q}\langle v, w_{i_q}\rangle\langle v, w_{i'_q}\rangle. \tag{163}$$

The only non-zero contributions come from pairings between i, i', which contribute $O(d^{(p-1)-p-(p-1)}) = O(d^{-1})$. The case $i_1 \neq 1$ follows by similar calculations. Overall, we proved that $\langle \mathbf{E}\, R, \mathbf{E}\, D^p\psi(v^\top x)\rangle = \sum_{\pi\neq\{[p]\}} O(d^{-1}) = O(d^{-1})$. □

Using Lemma B.19, we have that $\langle \mathbf{E}\, D^p\phi(F), \mathbf{E}\, D^p\psi(v^\top x)\rangle = \mathbf{E}\left[\varphi_2'(u^\top\varphi_1(Wx))\right]\sum_k u_k\langle \mathbf{E}\, D^pF_k, \mathbf{E}\, D^p\psi(v^\top x)\rangle + O(d^{-1})$, and this implies that

$$\begin{aligned}
&\mathbf{E}\,\varphi_2(u^\top\varphi_1(Wx))\psi(v^\top x)\\
&= \sum_{p\geq 1}\frac{1}{p!}\langle \mathbf{E}\, D^p\varphi_2(u^\top\varphi_1(Wx)), \mathbf{E}\, D^p\psi(v^\top x)\rangle\\
&= \sum_{p\geq 1}\frac{1}{p!}\sum_k u_k\,\mathbf{E}\big[\varphi_2'(u^\top\varphi_1(Wx))\big]\langle \mathbf{E}\, D^p\varphi_1(w_k^\top x), \mathbf{E}\, D^p\psi(v^\top x)\rangle + O(d^{-1})\\
&= \mathbf{E}\big[\varphi_2'(u^\top\varphi_1(Wx))\big]\sum_k u_k\,\mathbf{E}\,\varphi_1(w_k^\top x)\psi(v^\top x) + O(d^{-1})\\
&= \mathbf{E}\big[\varphi_2'(u^\top\varphi_1(Wx))\big]u^\top\,\mathbf{E}\,\varphi_1(Wx)\psi(v^\top x) + O(d^{-1}).
\end{aligned} \tag{164}$$

Therefore, we obtain that $\mathbf{E}\,\varphi_2(u^\top\varphi_1(Wx))\psi(v^\top x) = \mathbf{E}\big[\varphi_2'(u^\top\varphi_1(Wx))\big]u^\top\,\mathbf{E}\,\varphi_1(Wx)\psi(v^\top x) + O(d^{-1})$

B.6 Proof of Lemma B.14

For convenience, we restate the lemma.

Lemma B.20. *If u and z are independent, we have*

$$\mathbf{E}\,\varphi_2(u^\top\varphi_1(Wx))\psi_2(z^\top\psi_1(Vx)) = \mathbf{E}\,\varphi_2'(u^\top\varphi_1(Wx))\,\mathbf{E}\,\psi_2'(z^\top\psi_1(Vx))u^\top\,\mathbf{E}\left[\varphi_1(Wx)\psi_1(Vx)^\top\right]z + O(d^{-1}) \tag{165}$$

Here, our goal is to compute $\mathbf{E}\,\varphi_2(u^\top\varphi_1(Wx))\psi_2(z^\top\psi_1(Vx))$ with assumption that u and z are independent. Since $\mathbf{E}\, D^p\varphi_2(u^\top\varphi_1(Wx)) = \sum_k u_k w_k^{\otimes p}\,\mathbf{E}\left[\varphi_2'(u^\top\varphi_1(Wx))\varphi_1(w_k^\top x)\right]$, we can write

$$\begin{aligned}
&\langle \mathbf{E}\, D^p\varphi_2(u^\top\varphi_1(Wx)), \mathbf{E}\, D^p\psi_2(z^\top\psi_1(Vx))\rangle\\
&= \sum_{k,j}\langle w_k, v_j\rangle^p u_k z_j\,\mathbf{E}\left[\varphi_2'(u^\top\varphi_1(Wx))\varphi_1(w_k^\top x)\right]\mathbf{E}\left[\psi_2'(z^\top\psi_1(Vx))\psi_1(v_j^\top x)\right] + R,
\end{aligned} \tag{166}$$

where R is the term containing higher derivatives of φ_2, ψ_2. By computations similar to previous case, we can show that $R = O(d^{-1})$. Also, by weak correlation between the layers (also see previous case), we can show that

$$\begin{aligned}
&\langle \mathbf{E}\, D^p \varphi_2(u^\top \varphi_1(Wx)), \mathbf{E}\, D^p \psi_2(z^\top \psi_1(Vx)) \rangle \\
&\quad = \sum_{k,j} \langle w_k, v_j \rangle^p u_k z_j\, \mathbf{E}\left[\varphi_2'(u^\top \varphi_1(Wx)) \varphi_1^{(p)}(w_k^\top x) \right] \mathbf{E}\left[\psi_2'(z^\top \psi_1(Vx)) \psi_1^{(p)}(v_j^\top x) \right] + O(d^{-1}) \\
&\quad = \sum_{k,j} \langle w_k, v_j \rangle^p u_k z_j\, \mathbf{E}\left[\varphi_2'(u^\top \varphi_1(Wx)) \right] \mathbf{E}\left[\varphi_1^{(p)}(w_k^\top x) \right] \mathbf{E}\left[\psi_2'(z^\top \psi_1(Vx)) \right] \mathbf{E}\left[\psi_1^{(p)}(v_j^\top x) \right] + O(d^{-1})
\end{aligned} \tag{167}$$

After reverting chaos expansion, we obtain that

$$\mathbf{E}\, \varphi_2(u^\top \varphi_1(Wx)) \psi_2(z^\top \psi_1(Vx)) = \mathbf{E}\, \varphi_2'(u^\top \varphi_1(Wx))\, \mathbf{E}\, \psi_2'(z^\top \psi_1(Vx)) u^\top\, \mathbf{E}\left[\varphi_1(Wx) \psi_1(Vx)^\top \right] z + O(d^{-1}). \tag{168}$$

C Details on numerics

We provide in this Appendix more details on the experiments reported in Fig. 1 and Fig. 2.

C.1 Details of Fig. 1

Target We consider a two-layer structured RF teacher, with feature map

$$\varphi_*(x) = \tanh(W_* x) \tag{169}$$

where the weight $W_* = Z_* \tilde{C}_1^{\frac{1}{2}} \in \mathbb{R}^{d \times d}$ has covariance

$$\tilde{C}_1 = \mathrm{diag}(\{k^{-0.3}\}_{1 \le k \le d}). \tag{170}$$

Student We consider the task of learning this target with a four-layer RF student, with feature map

$$\varphi(x) = \tanh W_3(\tanh(W_2 \tanh(W_1 x))) \tag{171}$$

where, in order to introduce inter-layer and target/student weight correlations, we considered $W_2 = W_1$, with

$$W_1 = {}^1\!/_2 Z_1 \mathrm{diag}(\{k^{-\gamma/2}\}_{1 \le k \le d}) + {}^1\!/_2 W_*, \tag{172}$$

for $\gamma \in \{0.0, 0.2, 0.5, 0.8\}$. In other words, the covariance C_1 of W_1, W_2 is a sum of two power laws with decay γ and $0-3$. Finally, in order to introduce another form of correlation, we chose

$$W_3 = Z_3 C_3^{1/2} \tag{173}$$

where the covariance C_3 depends on the previous weights as

$$C_3 = (W_1 W_1^\top + {}^1\!/_2 \mathbb{I}_d)^{-1}. \tag{174}$$

C.2 Details on Fig. 2

In Fig. 2, we consider the task of learning a target corresponding to a structured three-layer RF

$$\varphi_*(x) = \theta_*^\top \tanh(W_2^* \mathrm{sign}(W_1^* x)), \tag{175}$$

where the weights $W_1^*, W_2^* \in \mathbb{R}^{d \times d}$ have covariance

$$C_1 = C_2 = \mathbb{I}_d + 8vv^\top, \tag{176}$$

Finally, the readout θ_* has i.i.d variance ${}^1\!/_d$ Gaussian entries. To learn this target, we consider training the three-layer feed-forward neural network

$$\theta^\top \tanh(W_2 \tanh(W_1 x)), \tag{177}$$

with trainable weights $W_1, W_2 \in \mathbb{R}^{d \times d}$. We consider the two-step training procedure

1. We first train the whole network end-to-end, using th `Pytorch` [65] implementation of full-batch Adam [61], for $T = 3000$ epochs, with learning rate $\eta = 10^{-3}$ and weight decay 10^{-5}. The training is done for $N = 1400$ training samples, with for $1 \leq \mu \leq n$ inputs $x^\mu \sim \mathcal{N}(0, \mathbb{I}_d)$, and the corresponding labels $\theta_*^\top \varphi_*(x^\mu)$. All these experiments were set in dimension $d = 1000$.

2. At the end of the training, one freezes the weight matrices to their trained values $\hat{W}_1, \hat{W}_2$, thereby obtaining the trained network feature map

$$\hat{\varphi}(x) = \tanh(\hat{W}_2 \tanh(\hat{W}_1 x). \tag{178}$$

   The readout weights θ are then re-trained alone, according to the ERM describe in the main text, with a fresh training set of $n = \alpha d$ samples. The ridge regularization strength λ is optimized over using cross-validation in this second step.

The generalization of the student network at the end of these two steps yields the blue curve in Fig. 2. We plot alongside the performance of the network with untrained first weights (green), i.e. when the first step is skipped. In this case, the feature map involves the weights at initialization, namely Gaussian random matrices, and corresponds to an unstructured dRF [12, 13]. Finally, the red curve corresponds to the performance of ridge regression directly on the inputs x, unprocessed by any feature map.