Asymptotics of Learning with Deep Structured (Random) Features

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

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? Despite 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 (Belkin et al., 2019) and benign overfitting (Bartlett et al., 2020) phenomena. As such, proxys for networks at initialisation, such as the Random Features (RF) model (Rahimi & Recht, 2007) have thus been the object of considerable theoretical attention, with their learning being asymptotically characterized in the two-layer case (Goldt et al., 2021; 2020; Gerace et al., 2021; Hu & Lu, 2022; Dhifallah & Lu, 2020; Mei & Montanari, 2022; Mei et al., 2022) and the deep case (Zavatone-Veth et al., 2022; Schröder et al., 2023; Bosch et al., 2023a; Zavatone-Veth & Pehlevan, 2023). With the exception of (Gerace et al., 2021; Mel & Pennington, 2022) (limited to two-layer networks) and (Zavatone-Veth & Pehlevan, 2023) (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 (Thamm et al., 2022; Martin & Mahoney, 2021). Of particular relevance to our work is a recent observation by (Guth et al., 2023) 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:

[*]Equal contribution [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. Correspondence to: Dominik Schröder <dschroeder@ethz.ch>, Hugo Cui <hugo.cui@epfl.ch>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

• 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 (Schröder et al., 2023) as a simple linear recursion depending on the weight matrices of each layer. These formulae extend similar results of (Cui et al., 2023; Schröder et al., 2023) 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.

Code The code for the numerical experiments described in Appendix C is openly available in this repository.

Related works — **Random features** (Rfs) were introduced in (Rahimi & Recht, 2007) 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 (Liao & Couillet, 2018; Pennington & Worah, 2019; Benigni & Péché, 2021). The test error was on the other hand characterized in (Mei & Montanari, 2022; Mei et al., 2022) for ridge regression, and extended to generic convex losses by (Gerace et al., 2021; Goldt et al., 2021; Dhifallah & Lu, 2020), and in (Liang & Sur, 2022; Loureiro et al., 2022; Bosch et al., 2023b) for other penalties. RFs have been studied as a model for networks in the lazy regime, see e.g. (Ghorbani et al., 2019; 2021; Yehudai & Shamir, 2019; Refinetti et al., 2021). The role of structure in the RF weights was discussed in (Gerace et al., 2024a) for rotationally invariant weights and (Mel & Pennington, 2022) for anisotropic Gaussian weights;
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 (Zavatone-Veth et al., 2022) for the case of unstructured weights and (Zavatone-Veth & Pehlevan, 2023) in the case of structured weights. For non-linear RFs, (Schröder et al., 2023) provides deterministic equivalents for the sample covariance matrices, and (Schröder et al., 2023; Bosch et al., 2023a) provide a tight characterization of the test error. The recent work of (Guth et al., 2023) 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 $\varphi_*\colon \mathbf{R}^d \to \mathbf{R}^k$ 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[1] in the sense $\mathbf{E}\,\varphi(\mathsf{x}_i) = 0$, $\mathbf{E}\,\varphi_*(\mathsf{x}_i) = 0$,
(ii) the feature covariances

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

have uniformly bounded spectral norm.
(iii) scalar Lipschitz functions of the feature matrices

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

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.

[1] This is a commonly used assumption which simplifies the analysis. Our techniques also apply to the case of non-zero mean, however doing so would add a rank-one component to the sample covariance matrix, considerably complicating the final expressions for the deterministic equivalents.

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 fixed $L \in \mathbb{N}$ and dimensions $d, p_1, \dots, p_L = p$, let $\varphi_1, \dots, \varphi_L \colon \mathbf{R} \to \mathbf{R}$ be Lipschitz-continuous *activation functions* $|\varphi_l(a) - \varphi_l(b)| \lesssim |a-b|$ applied entrywise, and let $W_1 \in \mathbf{R}^{p_1 \times d}, W_2 \in \mathbf{R}^{p_2 \times p_1}, \dots$ 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[2] $\varphi \colon \mathbf{R}^d \to \mathbf{R}^p, \varphi_* \colon \mathbf{R}^d \to \mathbf{R}^k$ 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[3] 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 (Guth et al., 2023), 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 $Z_1, \dots, Z_{l-1}$.

Note that Gaussian rainbow networks above can be seen as a generalization of the deep random features model studied in (Schröder et al., 2023; Bosch et al., 2023a; Fan & Wang, 2020), 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. For a centered random vector $x \in \mathbf{R}^d$ we denote its *sub-Gaussian norm* as $\|x\|_{\psi_2} := \inf_{\sigma \geq 0}\{\mathbf{E} \exp^{\langle v, x \rangle} \leq \exp^{\frac{\|v\|^2 \sigma^2}{2}} \ \forall\, v \in \mathbf{R}^d\}$.

3. Test error of Lipschitz feature models

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

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

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

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

we obtain the bounds

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

We also remark that $m(\lambda)$ depends on Ω only through its eigenvalues $\omega_1, \dots, \omega_p$, 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*

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

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

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

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[4]

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

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 (Bach, 2024; Dobriban & Wager, 2018; Bartlett et al., 2021; Cheng & Montanari, 2022) and the references therein.*

(a) We confirm Conjecture 1 of (Loureiro et al., 2022) under assumption 2.2. The expression for the error term in Theorem 3.1 matches the expression obtained in (Loureiro et al., 2022) for a Gaussian covariates teacher-student model.

(b) Independently and concurrently to the current work (Latourelle-Vigeant & Paquette, 2023) (partially confirming a conjecture made in (Louart et al., 2018b)) obtained similar results under different assumptions. Most importantly (Latourelle-Vigeant & Paquette, 2023) 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 (Mei et al., 2022; Adlam & Pennington, 2020) 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 $\Sigma \neq I$.

[4]This allows to identify the leading order of the generalization error as long as the ratio of the largest and smallest parameter is much smaller than the square-root of the smallest one.

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* (Adlam & Pennington, 2020; Latourelle-Vigeant & Paquette, 2023). 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[5] In comparison we believe that our approach is more direct and flexible.

[5]For instance (Adlam & Pennington, 2020) used block matrices with up to 16×16 blocks in order to evaluate the asymptotic test error.

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 (Adamczak, 2015a). The concentration step is very similar to analagous considerations in previous works (Chouard, 2022; Louart et al., 2018a) 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:

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[6].

Remark 3.4. *Tracial Marchenko-Pastur laws (case $A = I$ above) have a long history, going back to (Marčenko & Pastur, 1967) in the isotropic case $\Omega = I$, (Silverstein, 1995) in the general case with separable covariance $x = \sqrt{\Omega} z$ and (Bai & Zhou, 2008) under quadratic form concentration assumption. Anisotropic Marchenko-Pastur laws under various conditions and with varying precision have been proven e.g. in (Rubio & Mestre, 2011; Chouard, 2022; Louart et al., 2018b; Knowles & Yin, 2017).*

For the proof of Theorem 3.3 the resolvent $\breve{G} := (X^\top X/p + \lambda)^{-1} \in \mathbf{R}^{n\times n}$ 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.

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

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

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

in Frobenius norm, which, together with the relation $1 - \lambda\langle \breve{G}\rangle = \frac{p}{n}\big(1 - \lambda\langle G\rangle\big)$ between the traces of G and $\breve{G}$, yields a self-consistent equation for $\langle \breve{G}\rangle$. 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

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

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 (Knowles & Yin, 2017) we measure the error of the approximation Eq. (16) with respect to the Frobenius norm of the observable A. As in the case of unified local laws for Wigner matrices (Cipolloni et al., 2022) 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

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

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 $\frac{1}{n} X Z^\top = \frac{1}{n}\sum x_i z_i^\top \approx \Phi$, 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 $G\frac{XZ^\top}{n}$,

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

but the derivation of the multi-resolvent equivalents requires 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 careful 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

$$\begin{aligned}\varphi(\mathsf{x}) &= \varphi_L(W_L\varphi_{L-1}(\dots\varphi_1(W_1\mathsf{x})))\\ \varphi_*(\mathsf{x}) &= \widetilde{\varphi}_{\widetilde{L}}(V_{\widetilde{L}}\widetilde{\varphi}_{\widetilde{L}-1}(\dots\widetilde{\varphi}_1(V_1\mathsf{x})))\end{aligned} \tag{22}$$

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 we allow the k-th row of W_ℓ to be correlated only to the k-th row of $W_{\ell'}, V_{\ell'}$ as this allows for particularly simple expressions for the linearized covariances[7]. 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 $L \le \widetilde{L}$. Furthermore, for all $\ell \le L \le \widetilde{L}$, we assume

(a) All the internal widths p_ℓ of W_ℓ, V_ℓ agree,

(b) The rows w_ℓ, v_ℓ of W_ℓ, V_ℓ are i.i.d. with mean zero and

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

with $\|C_\ell\| + \|\widetilde{C}_\ell\| + \|\check{C}_\ell\| \lesssim 1$,

(c) Asymptotic orthogonality of the rows of W_ℓ, V_ℓ. Let w, w' be two independent copies of a row of W_ℓ. Then, $\langle w, w'\rangle \prec d^{-1/2}$, same for V_ℓ,

(d) The rows of W_ℓ, V_ℓ are sub-Gaussian random vectors:

$$\|w_\ell\|_{\psi_2} + \|v_\ell\|_{\psi_2} = O(d^{-1/2}), \tag{23}$$

(e) Centered activation functions $\varphi_\ell, \widetilde{\varphi}_\ell$ (see Assumption 2.2).

[7] The identity matrices in Eq. (24) 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.

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 $\Omega_\ell^{\rm lin}, \Phi_\ell^{\rm lin}, \Psi_\ell^{\rm lin}$ by the recursions

$$\begin{aligned}\Omega_\ell^{\rm lin} &= (\kappa_\ell^1)^2 W_\ell\Omega_{\ell-1}^{\rm lin}W_\ell^\top + (\kappa_\ell^*)^2\mathbb{I}_{p_\ell}\\ \Psi_\ell^{\rm lin} &= (\tilde\kappa_\ell^1)^2 V_\ell\Psi_{\ell-1}^{\rm lin}V_\ell^\top + (\tilde\kappa_\ell^*)^2\mathbb{I}_{p_\ell}\\ \Phi_\ell^{\rm lin} &= \kappa_\ell^1\tilde\kappa_\ell^1 W_\ell\Phi_{\ell-1}^{\rm lin}V_\ell^\top + (\check\kappa_\ell^*)^2\mathbb{I}_{p_\ell},\end{aligned} \tag{24}$$

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

$$\kappa_\ell^1 := \mathbf{E}\,\varphi_\ell'(N_\ell), \quad \widetilde{\kappa}_\ell^1 := \mathbf{E}\,\widetilde{\varphi}_\ell'(\widetilde{N}_\ell) \tag{25}$$

and

$$\begin{aligned}\kappa_\ell^* &= \sqrt{\mathbf{E}[\varphi_\ell(N_\ell)^2] - r_\ell(\kappa_\ell^1)^2}\\ \tilde\kappa_\ell^* &= \sqrt{\mathbf{E}[\widetilde{\varphi}_\ell(\widetilde{N}_\ell)^2] - \tilde r_\ell(\tilde\kappa_\ell^1)^2}\\ \check\kappa_\ell^* &= \sqrt{\mathbf{E}[\varphi_\ell(N_\ell)\widetilde{\varphi}_\ell(\widetilde{N}_\ell)] - \check r_\ell\kappa_\ell^1\tilde\kappa_\ell^1},\end{aligned} \tag{26}$$

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

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

Finally, for $\tilde L \ge \ell \ge L+1$, define

$$\Phi_\ell^{\rm lin} = \tilde\kappa_\ell^1\Phi_{\ell-1}^{\rm lin}V_\ell^\top, \tag{27}$$

with still $\tilde\kappa_\ell^1, \tilde\kappa_\ell^*$ just as before, and $\Psi_\ell^{\rm lin}$ with the same recursion (24).

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.*

$$\|\Omega - \Omega_L^{\rm lin}\|_F + \|\Psi - \Psi_{\tilde L}^{\rm lin}\|_F + \|\Phi - \Phi_{\tilde L}^{\rm lin}\|_F \lesssim 1 \tag{28}$$

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 $L = \widetilde{L} = 2$, we have*

$$\begin{aligned}&\|\Omega_1 - \Omega_1^{\rm lin}\|_F + \|\Psi_1 - \Psi_1^{\rm lin}\|_F + \|\Phi_1 - \Phi_1^{\rm lin}\|_F \prec 1,\\ &\|\Omega_2 - \Omega_2^{\rm lin}\|_F + \|\Psi_2 - \Psi_2^{\rm lin}\|_F + \|\Phi_2 - \Phi_2^{\rm lin}\|_F \prec 1.\end{aligned}$$

Remark 4.5 (Comparison). *The approach we take here is somewhat different from previous works (Schröder et al., 2023; Fan & Wang, 2020; Chouard, 2023) 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.13 and Theorem B.14 for the formal proof. In the proof, we crucially rely on the theory of Wiener chaos expansion and Stein's method (see (Nourdin & Peccati, 2012)). Gaussian Wiener chaos is a generalization of Hermite polynomial expansions, which previously have been used for approximate linearization (Fan & Wang, 2020; Schröder et al., 2023) 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

$$F = \mathbf{E}\, F + \sum_{p\geq 1} I_p\Big(\frac{\mathbf{E}\, D^p F}{p!}\Big), \tag{29}$$

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 $\varphi_1(w^\top \mathsf{x}), \varphi_1(u^\top \mathsf{x})$ we obtain, for instance

$$\begin{aligned}&\mathbf{E}\, \varphi_1(w^\top \mathsf{x})\varphi_1(v^\top \mathsf{x}) \\ &\quad = \sum_{p\geq 1} \frac{1}{p!}\, \mathbf{E}\, \varphi_1^{(p)}(w^\top \mathsf{x})\, \mathbf{E}\, \varphi_1^{(p)}(u^\top \mathsf{x}) \langle w, v\rangle^p,\end{aligned} \tag{30}$$

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

$$d_W(F, N) \lesssim \mathbf{E}|\mathbf{E}\, F^2 - \langle DF, -DL^{-1}F\rangle| \tag{31}$$

for the Wasserstein distance d_W, where

$$F := w^\top \phi_1(W\mathsf{x}), \quad N \sim \mathcal{N}(0, \mathbf{E}\, F^2), \tag{32}$$

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

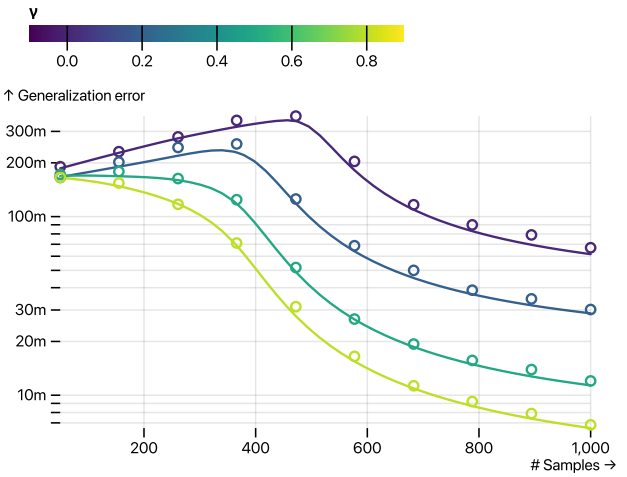

Figure 1. Test error for a target $\theta_*^\top \tanh(W_* x)$, when learning with a four-layer Gaussian rainbow network with feature map $\varphi(x) = \tanh(W_3 \tanh(W_2 \tanh(W_1 x)))$. All width were taken equal to the input dimension d, and the regularization employed is $\lambda = 10^{-4}$. The student weights are correlated across layers, with $W_1 = W_2$, and the covariance C_3 of W_3 depending on W_1 as $C_3 = (W_1 W_1^\top + {}^1\!/_2 \mathbb{I}_d)^{-1}$. Target/student correlations are also present, with $\check{C}_1 = {}^1\!/_2 \mathbb{I}_d$. 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. Circles : numerical simulations in $d = 1000$.

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 (Cui et al., 2023; Schröder et al., 2023; Bosch et al., 2023a; Fan & Wang, 2020; Chouard, 2023) for the case of weights which are both unstructured and independent, and iteratively build upon earlier results for the two-layer case developed in (Mei & Montanari, 2022; Gerace et al., 2021; Goldt et al., 2021; Hu & Lu, 2022). In fact, the expressions leveraged in these works can be recovered as a special case for $C_\ell = \tilde{C}_\ell = \mathbb{I}_{p_\ell}$ (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 $\theta_*^\top \tanh(V\mathsf{x})$. To illustrate our result, we con-

sider both target/student correlations $\check{C}_1 = {}^1\!/_2\mathbb{I}_d$, 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, $C_3 = (W_1 W_1^\top + {}^1\!/_2\mathbb{I}_d)^{-1}$. In order to have structured weights, the covariances $\widetilde{C}_1, C_1, C_2$ 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 (Schröder et al., 2023), 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 equivalent *linear stochastic* network $\varphi^{\text{lin}} = \psi_L \circ \cdots \circ \psi_1$, with

$$\psi_\ell(x) = \kappa_\ell^1 W_\ell x + \kappa_\ell^* \xi_\ell \tag{33}$$

where $\xi_\ell \sim \mathcal{N}(0, \mathbb{I}_{p_\ell})$ a stochastic noise. This equivalent viewpoint has proven fruitful in yielding insights on the implicit bias of RFs (Schröder et al., 2023; Jacot et al., 2020) and on the fundamental limitations of deep networks in the proportional regime (Cui et al., 2023). 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 (Jacot et al., 2020), 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 (Ba et al., 2022) 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 (Kingma & Ba, 2014) 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

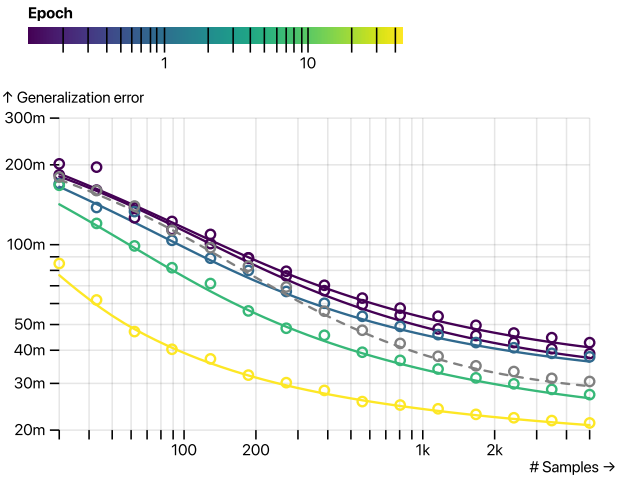

Figure 2. Test error when training the readout layer only of a relu-activated three-layer neural network during training, using the `Tensorflow` implementation of the Adam (Kingma & Ba, 2014) optimizer, over 120 epochs with batch size 128. (dashed): ridge regression. The data is sampled from a Gaussian distribution with mean and variance matching the distribution of MNIST images. In all training procedures, the regularization parameter has been numerically optimized. Solid lines represent the theoretical prediction of Theorem 3.1, dots represent numerical experiments. For more details we refer to Appendix C.3.

$$\varphi^g(x) = W_{\text{eff.}} x + C_{\text{eff.}}^{1/2} \xi \tag{34}$$

where $W_{\text{eff.}} \in \mathbf{R}^{p \times d}$ is proportional to the product of all the weight matrices

$$W_{\text{eff.}} = \left(\prod_{\ell=1}^{L} \kappa_\ell^1 \right) \hat{W}_L \hat{W}_{L-1} \dots \hat{W}_1 \tag{35}$$

and $\xi \sim \mathcal{N}(0, \mathbb{I}_p)$ is a stochastic noise colored by the covariance

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

Note that the effective linear network (34) simply corresponds to the composition of the equivalent stochastic linear

layers (33). A very similar expression for the covariance of the effective structured noise (36) appeared in (Schröder et al., 2023) for the random case with unstructured and untrained random weights. The effective linear model (34) 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 (34). Indeed, the interplay between these two matrices – both depending on the weights $\hat{W}_\ell$ – defines the inductive bias of the trained network in the high-dimensional regime, which ultimately determines the generalization properties of the network. To see this explicitly, consider the ridge regression problem on the effective linear features in eq. (34). Changing variables $\beta = C_{\text{eff.}}^{1/2}\theta/\sqrt{p}$ and assuming for simplicity that $C_{\text{eff.}}$ is invertible, this yields the following effective problem:

$$\min_{\beta\in\mathbf{R}^p} \sum_{i\in[n]} \left(\mathsf{y}_i - \beta^\top (C_{\text{eff.}}^{-1/2} W_{\text{eff.}} \mathsf{x}_i + \xi_i)\right)^2 + p\lambda\beta^\top C_{\text{eff.}}^{-1}\beta.$$

In this basis, the effective linear features has two components: an irreducible isotropic noise ξ_i and a term $C_{\text{eff.}}^{-1/2} W_{\text{eff.}} \mathsf{x}_i$ controlling the (linear) representation of the training data. A key difference with respect to the deep unstructured case of (Schröder et al., 2023) is that the effective ℓ_2-regularization is anisotropic.

For a (typically employed) random isotropic initialization, the initial network is equivalent to a unstructured dRF. In particular, the unstructured dRF inductive bias (Jacot et al., 2020) is not aligned to the target, treating all directions equally. In terms of generalization, since the effective linear features are noisy, this implies that the initial generalization error is lower-bounded by the best ridge estimator (Schröder et al., 2023). As the network is trained, the weights $\hat{W}_\ell$ adapt to the target, implying that even in the regime where the linearization in eq. (34) holds, the effective linear problem eq. (34) can regularize different directions adaptively, potentially outperforming the ridge regression baseline. The fact that the optimal regularization for ridge regression on linear feature maps might be anisotropic has been explored in detail in (Wu & Xu, 2020).

The learning of beneficial inductive biases over training is illustrated by Fig. 2 for synthetic data. Despite the fact that all represented feature maps are effectively just linear feature maps, they can still encode very different biases, yielding different phenomenology. In particular, remark that, when trained over a sufficient number of epochs, the trained feature map outperforms by ridge regression on the whole range of probed sample complexities – suggesting the trained weights $W_{\text{eff.}}, C_{\text{eff.}}$ learned some form of helpful inductive bias, and allow for a more performant linear model. A similar qualitative behaviour can also be observed in real data sets, as illustrated in Fig.3.

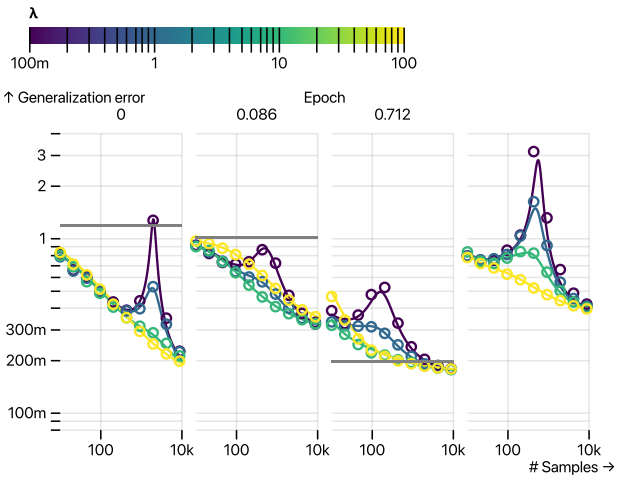

Figure 3. Test error when re-training the readout layer only of an Adam-optimized relu-activated three-layer neural network, trained on a regression task on MNIST. Labels are $+1$ (resp. -1) for even (resp. odd) digits. Solid lines represent the theoretical prediction of Theorem 3.1, dots represent numerical experiments on the real dataset. Different colors indicate different reguarization strengths λ. Different panels correspond to different training times. All details are provided in App. C.2.

6. Concluding remarks

Real data — We observe that the theoretical predictions of Theorem 3.1 also capture the learning curves of trained networks on some *real* datasets, when retraining the readout only using ridge regression, provided the features covariances Ω, Φ, Ψ are estimated from data. Fig. 3 contrasts the theoretical characterization of Theorem 3.1 with numerical experiments on MNIST (LeCun et al., 1998), for a three-neural network optimized with Adam (Kingma & Ba, 2014), revealing overall good agreement. All experimental details are reported in Appendix C.2. Note that closely related observations have also been made in (Loureiro et al., 2022).

Limitations — Our results provides an insight in the inductive bias of trained deep rainbow networks. However, as discussed in (Guth et al., 2023), this only captures a subset of neural networks. Understanding the boundaries of applicability of the Gaussian rainbow framework (and hence of our theory) is an interesting problem. A recent line of work investigating the properties of two-layer neural networks after a single step of training (Ba et al., 2022; Dandi et al., 2023; Moniri et al., 2023; Cui et al., 2024) provides a first clue. These works show that with an aggressive learning rate the hidden-layer weights can be approximated by a spiked random matrix model. Investigating under which conditions the asymptotic performance is equivalent to a structured Gaussian model is an interesting venue for future research. A similar problem was studied in the context of structured inputs in (Pesce et al., 2023; Gerace et al., 2024b).

Acknowledgement

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 Swiss National Science Foundation grant SMArtNet (grant number 212049).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Adamczak, R. A note on the hanson-wright inequality for random vectors with dependencies. *Electron. Commun. Prob.*, 20, 2015a. doi: 10.1214/ECP.v20-3829. URL https://doi.org/10.1214/ECP.v20-3829.

Adamczak, R. a. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electron. Commun. Probab.*, 20:no. 72, 13, 2015b. doi: 10.1214/ECP.v20-3829. URL https://doi.org/10.1214/ECP.v20-3829.

Adlam, B. and Pennington, J. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 74–84. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/adlam20a.html.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 37932–37946. Curran Associates, Inc., 2022.

Bach, F. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024. doi: 10.1137/23M1558781. URL https://doi.org/10.1137/23M1558781.

Bai, Z. and Zhou, W. Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, 18(2):425–442, 2008. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24308489.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. USA*, 117(48):30063–30070, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907378117. URL https://doi.org/10.1073/pnas.1907378117.

Bartlett, P. L., Montanari, A., and Rakhlin, A. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Belkin, M., Ma, S., Mandal, S., Belkin, M., and Hsu, D. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. U. S. A.*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116. URL https://doi.org/10.1073/pnas.1903070116.

Benigni, L. and Péché, S. Eigenvalue distribution of some nonlinear models of random matrices. *Electron. J. Probab.*, 26:Paper No. 150, 37, 2021. doi: 10.1214/21-ejp699.

Bosch, D., Panahi, A., and Hassibi, B. Precise asymptotic analysis of deep random feature models. In Neu, G. and Rosasco, L. (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4132–4179. PMLR, 12–15 Jul 2023a. URL https://proceedings.mlr.press/v195/bosch23a.html.

Bosch, D., Panahi, A., Ozcelikkale, A., and Dubhashi, D. Random features model with general convex regularization: A fine grained analysis with precise asymptotic learning curves. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11371–11414. PMLR, 25–27 Apr 2023b. URL https://proceedings.mlr.press/v206/bosch23a.html.

Cheng, C. and Montanari, A. Dimension free ridge regression. *arXiv preprint arXiv:2210.08571*, 2022.

Chouard, C. Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure. *arXiv preprint arXiv:2211.13044*, 2022.

Chouard, C. Deterministic equivalent of the conjugate kernel matrix associated to artificial neural networks. *arXiv preprint arXiv:2306.05850*, 2023.

Cipolloni, G., Erdős, L., and Schröder, D. Rank-uniform local law for Wigner matrices. *Forum Math., Sigma*, 10, 2022. doi: 10.1017/fms.2022.86. URL https://doi.org/10.1017/fms.2022.86.

Cui, H., Krzakala, F., and Zdeborova, L. Bayes-optimal learning of deep random networks of extensive-width. In *International Conference on Machine Learning*, pp. 6468–6521. PMLR, 2023.

Cui, H., Pesce, L., Dandi, Y., Krzakala, F., Lu, Y. M., Zdeborová, L., and Loureiro, B. Asymptotics of feature learning in two-layer networks after one gradient-step. *arXiv preprint arXiv:2402.04980*, 2024.

Dandi, Y., Krzakala, F., Loureiro, B., Pesce, L., and Stephan, L. Learning two-layer neural networks, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.

Dhifallah, O. and Lu, Y. M. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020.

Dobriban, E. and Wager, S. High-dimensional asymptotics of prediction: ridge regression and classification. *Ann. Statist.*, 46(1):247–279, 2018. ISSN 0090-5364. doi: 10.1214/17-AOS1549. URL https://doi.org/10.1214/17-AOS1549.

Fan, Z. and Wang, Z. 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. ISBN 9781713829546.

Gerace, F., Loureiro, B., Krzakala, F., Mézard, M., and Zdeborová, L. Generalisation error in learning with random features and the hidden manifold model*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124013, dec 2021. doi: 10.1088/1742-5468/ac3ae6. URL https://dx.doi.org/10.1088/1742-5468/ac3ae6.

Gerace, F., Krzakala, F., Loureiro, B., Stephan, L., and Zdeborová, L. Gaussian universality of perceptrons with random labels. *Phys. Rev. E*, 109: 034305, Mar 2024a. doi: 10.1103/PhysRevE.109.034305. URL https://link.aps.org/doi/10.1103/PhysRevE.109.034305.

Gerace, F., Krzakala, F., Loureiro, B., Stephan, L., and Zdeborová, L. Gaussian universality of perceptrons with random labels. *Phys. Rev. E*, 109: 034305, Mar 2024b. doi: 10.1103/PhysRevE.109.034305. URL https://link.aps.org/doi/10.1103/PhysRevE.109.034305.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Limitations of lazy training of two-layers neural network. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. When do neural networks outperform kernel methods? *J. Stat. Mech. Theory Exp.*, 2021(12):Paper No. 124009, 110, 2021. doi: 10.1088/1742-5468/ac3a81.

Goldt, S., Mézard, M., Krzakala, F., and Zdeborová, L. Modeling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model. *Phys. Rev. X*, 10(4), 2020. doi: 10.1103/PhysRevX.10.041044. URL https://doi.org/10.1103/PhysRevX.10.041044.

Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mezard, M., and Zdeborová, L. 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, pp. 426–471, 2021.

Guth, F., Ménard, B., Rochette, G., and Mallat, S. A rainbow in deep network black boxes. *arXiv preprint arXiv:2305.18512*, 2023.

Hu, H. and Lu, Y. M. Universality Laws for High-Dimensional Learning with Random Features. *IEEE Trans. Inf. Theory*, 2022. doi: 10.1109/TIT.2022.3217698.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Implicit regularization of random feature models. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4631–4640. PMLR, 13–18 Jul 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Knowles, A. and Yin, J. Anisotropic local laws for random matrices. *Probab. Theory Related Fields*, 169(1-2):257–352, 2017. ISSN 0178-8051. doi: 10.1007/s00440-016-0730-4.

Latourelle-Vigeant, H. and Paquette, E. Matrix dyson equation for correlated linearizations and test error of random features regression. *arXiv preprint arXiv:2312.09194*, 2023.

LeCun, Y., Bottou, L., Haffner, P., Bengio, Y., LeCun, Y., LeCun, Y., and Bottou, L. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86 (11):2278–2323, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Liang, T. and Sur, P. A precise high-dimensional asymptotic theory for boosting and minimum-ℓ1-norm interpolated classifiers. *Ann. Statist.*, 50(3):1669–1695, 2022. ISSN 0090-5364. doi: 10.1214/22-aos2170.

Liao, Z. and Couillet, R. On the spectrum of random features maps of high dimensional data. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3063–3071. PMLR, 10–15 Jul 2018.

Louart, C., Liao, Z., and Couillet, R. A random matrix approach to neural networks. *Ann. Appl. Probab.*, 28(2): 1190–1248, 2018a. doi: 10.1214/17-AAP1328. URL https://doi.org/10.1214/17-AAP1328.

Louart, C., Liao, Z., and Couillet, R. A random matrix approach to neural networks. *Ann. Appl. Probab.*, 28 (2):1190–1248, 2018b. ISSN 1050-5164. doi: 10.1214/17-AAP1328. URL https://doi.org/10.1214/17-AAP1328.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mézard, M., and Zdeborová, L. Learning curves of generic features maps for realistic datasets with a teacher-student model. *J. Stat. Mech. Theory Exp.*, 2022(11): Paper No. 114001, 78, 2022. doi: 10.1088/1742-5468/ac9825.

Martin, C. H. and Mahoney, M. W. 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. URL http://jmlr.org/papers/v22/20-410.html.

Marčenko, V. A. and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, apr 1967. doi: 10.1070/SM1967v001n04ABEH001994. URL https://dx.doi.org/10.1070/SM1967v001n04ABEH001994.

Mei, S. and Montanari, A. The generalization error of random features regression: precise asymptotics and the double descent curve. *Comm. Pure Appl. Math.*, 75(4):667–766, 2022. ISSN 0010-3640. doi: 10.1002/cpa.22008. URL https://doi.org/10.1002/cpa.22008.

Mei, S., Misiakiewicz, T., and Montanari, A. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Appl. Comput. Harmon. Anal.*, 59:3–84, 2022. ISSN 1063-5203. doi: 10.1016/j.acha.2021.12.003.

Mel, G. and Pennington, J. Anisotropic random feature regression in high dimensions. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JfaWawZ8BmX.

Moniri, B., Lee, D., Hassani, H., and Dobriban, E. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.

Nourdin, I. and Peccati, G. *Normal approximations with Malliavin calculus: from Stein's method to universality*, volume 192. Cambridge University Press, 2012.

Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. *J. Stat. Mech. Theory Exp.*, (12):124005, 14, 2019. doi: 10.1088/1742-5468/ab3bc3. URL https://doi.org/10.1088/1742-5468/ab3bc3.

Pesce, L., Krzakala, F., Loureiro, B., and Stephan, L. Are Gaussian data all you need? The extents and limits of universality in high-dimensional generalized linear estimation. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27680–27708. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/pesce23a.html.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Refinetti, M., Goldt, S., Krzakala, F., and Zdeborová, L. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8936–8947. PMLR, 18–24 Jul 2021.

Rubio, F. and Mestre, X. Spectral convergence for a general class of random matrices. *Statistics & Probability Letters*, 81(5):592–602, 2011. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2011.01.004. URL https://www.sciencedKirect.com/science/article/pii/S0167715211000113.

Schröder, D., Cui, H., Dmitriev, D., and Loureiro, B. Deterministic equivalent and error universality of deep random features learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30285–30320. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/schroder23a.html.

Silverstein, J. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995. ISSN 0047-259X. doi: https://doi.org/10.1006/jmva.1995.1083. URL https://www.sciencedirect.com/science/article/pii/S0047259X85710834.

Thamm, M., Staats, M., and Rosenow, B. Random matrix analysis of deep neural network weight matrices. *Phys. Rev. E*, 106(5):Paper No. 054124, 15, 2022. ISSN 2470-0045. doi: 10.1103/physreve.106.054124. URL https://doi.org/10.1103/physreve.106.054124.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wu, D. and Xu, J. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10112–10123. Curran Associates, Inc., 2020.

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

Zavatone-Veth, J. and Pehlevan, C. Learning curves for deep structured gaussian feature models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 42866–42897. Curran Associates, Inc., 2023.

Zavatone-Veth, J. A., Pehlevan, C., and Tong, W. L. Contrasting random and learned features in deep Bayesian linear regression. *Phys. Rev. E*, 105(6), 2022. doi: 10.1103/PhysRevE.105.064118.

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

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

A sufficent condition for Lipschitz concentration is that that the columns $x_i = \varphi(\mathsf{x}_i)$ are Lipschitz functions of Gaussian random vectors x_i of bounded covariance $\Omega_0 := \mathbf{E}\, \mathsf{x}_i \mathsf{x}_i^\top$, c.f. Remark 2.4. Indeed, let $\widetilde{\varphi}(\mathsf{g}) := \varphi(\sqrt{\Omega_0}\mathsf{g})$ and consider standard Gaussian vectors $\mathsf{g}_i, \dots, \mathsf{g}_n$. 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 $\mathsf{g}_1, \dots, \mathsf{g}_n$ into $\mathsf{g} \in \mathbf{R}^{np}$ and write $X = X(\mathsf{g}) = (\widetilde{\varphi}(\mathsf{g}_1), \dots, \widetilde{\varphi}(\mathsf{g}_n))$. 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 $\mathsf{g} \mapsto f(X(\mathsf{g}))$ is Lipschitz due to

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

Resolvent concentration

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

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

The two resolvents are related by the identity

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

Both resolvents $G, \breve{G}$ 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

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

implying

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

Therefore we obtain that

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

from Theorem A.2,

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

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

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

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 (Adamczak, 2015b).

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 (Vershynin, 2010).

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

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

where the denominators can be simplified using

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

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

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

so that

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

Using the bounds

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

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

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). w □

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*[8]

$$\frac{1}{\sqrt{kp}}\langle GXZ^\top A\rangle = \frac{\lambda mn}{\sqrt{kp}}\langle M\Phi A\rangle + O\Big(\frac{n}{k^{1/2}p^{3/2}\lambda^3}\Big(1 + \frac{n}{p} + \frac{p}{n}\Big)\Big) \tag{59}$$

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

$$\langle AG\Omega G\rangle = \frac{\langle AM\Omega M\rangle}{1 - \frac{n}{p}(m\lambda)^2\langle \Omega M\Omega M\rangle} + O\Big(\frac{\langle |A|^2\rangle^{1/2}}{p\lambda^7}\Big) \tag{60a}$$

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

$$\langle AGBG\rangle = \langle AMBM\rangle + \frac{n}{p}(m\lambda)^2\frac{\langle AM\Omega M\rangle\langle \Omega MBM\rangle}{1 - \frac{n}{p}(m\lambda)^2\langle \Omega M\Omega M\rangle} + O\Big(\frac{\langle |A|^2\rangle^{1/2}\|B\|}{p\lambda^7}\Big) \tag{60b}$$

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

$$\Big\langle \frac{X^\top G\Omega GXA}{p}\Big\rangle = \frac{\lambda^2 m^2\langle \Omega M\Omega M\rangle}{1 - \frac{n}{p}(m\lambda)^2\langle \Omega M\Omega M\rangle}\langle A\rangle + O\Big(\frac{\langle |A|^2\rangle^{1/2}}{p\lambda^7}\Big) \tag{61}$$

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

$$\Big\langle \frac{ZX^\top G\Omega GXZ^\top A}{kp}\Big\rangle = (m\lambda)^2\frac{n}{k}\frac{\Big\langle A\Big(\big(\Psi - 2\frac{n}{p}\lambda m\Phi^\top M\Phi\big)\langle \Omega M\Omega M\rangle + \frac{n}{p}\Phi^\top M\Omega M\Phi\Big)\Big\rangle}{1 - \frac{n}{p}(m\lambda)^2\langle \Omega M\Omega M\rangle} + O\Big(\frac{\langle |A|^2\rangle^{1/2}}{p\lambda^7}\Big) \tag{62}$$

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

[8] In a slight abuse of notation we use the $O(\cdots)$ notation in the sense of "$\prec$"

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

$$\begin{aligned}\mathcal{E}_{\rm gen} &= \frac{\varphi_*^\top \Psi \varphi_*}{k} + \frac{\varphi_*^\top Z X^\top G \Omega G X Z^\top \varphi_*}{kp^2} + \frac{n}{p}\left\langle \frac{X^\top G \Omega G X \Sigma}{p}\right\rangle - 2\frac{\varphi_*^\top \Phi^\top G X Z^\top \varphi_*}{kp} \\ &= \frac{1}{k}\varphi_*\left(\Psi + (m\lambda)^2 \frac{n}{p}\frac{\left(\Psi - 2\frac{n}{p}\lambda m \Phi^\top M \Phi\right)\langle \Omega M \Omega M\rangle + \frac{n}{p}\Phi^\top M \Omega M \Phi}{1 - \frac{n}{p}(m\lambda)^2\langle \Omega M \Omega M\rangle} - 2\lambda m \frac{n}{p}\Phi^\top M \Phi\right)\varphi_* \\ &\quad + \langle \Sigma\rangle \frac{(\lambda m)^2 \frac{n}{p}\langle M\Omega M\Omega\rangle}{1 - \frac{n}{p}(\lambda m)^2\langle \Omega M \Omega M\rangle} + O\left(\frac{\|\varphi_*\|^2}{p^{1/2}\lambda^7}\right).\end{aligned} \tag{63}$$

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

$$m\lambda \frac{n}{p} M\Omega M = M - \lambda M^2, \tag{64}$$

so that the bracket, after simplifying, becomes

$$\frac{\Psi - m\lambda \frac{n}{p}\Phi^\top (M + \lambda M^2)\Phi}{1 - \frac{n}{p}(m\lambda)^2\langle \Omega M \Omega M\rangle}, \tag{65}$$

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

$$\begin{aligned}\langle GXZ^\top A\rangle - \langle G'X'(Z')^\top A\rangle &= \langle (G-G')XZ^\top A\rangle + \langle G'(X-X')Z^\top A\rangle + \langle G'X'(Z-Z')^\top A\rangle \\ &= O\left(\frac{\|X-X'\|_F \|X\| \|Z\| \langle |A|^2\rangle^{1/2}}{\lambda^{3/2}p} + \frac{(\|X-X'\|_F\|Z\| + \|X\|\|Z-Z'\|_F)\langle |A|^2\rangle^{1/2}}{p\lambda}\right),\end{aligned} \tag{66}$$

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

$$\left|\frac{1}{\sqrt{kp}}\langle GXZ^\top A\rangle - \frac{1}{\sqrt{kp}}\langle \mathbf{E}\, GXZ^\top A\rangle\right| \prec \frac{\langle |A|^2\rangle^{1/2}}{p\lambda^{3/2}}. \tag{67}$$

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

$$\frac{1}{\sqrt{kp}} GXZ^\top = \frac{1}{\sqrt{kp}}\sum_i G x_i z_i^\top = \frac{1}{\sqrt{kp}}\sum_i \lambda \breve{G}_{ii} G_i x_i z_i^\top. \tag{68}$$

With

$$\begin{aligned}\frac{\lambda}{\sqrt{kp}}\mathbf{E}\sum_i (\breve{G})_{ii}\langle G_i x_i z_i^\top A\rangle &= \frac{\lambda}{\sqrt{kp}}\sum_i \left((\mathbf{E}\,\breve{G}_{ii})\langle \mathbf{E}\, G_i x_i z_i^\top A\rangle + O\left(\sqrt{\operatorname{Var}\breve{G}_{ii}}\sqrt{\operatorname{Var}\langle G_i x_i z_i^\top A\rangle}\right)\right) \\ &= \frac{\lambda}{\sqrt{kp}}\sum_i (\mathbf{E}\,\breve{G}_{ii})\langle \mathbf{E}\, G_i \Phi A\rangle + O\left(\frac{n}{k^{1/2}p^{3/2}\lambda^2}\right) \\ &= \frac{\lambda m n}{\sqrt{kp}}\langle M\Phi A\rangle + O\left(\frac{n}{k^{1/2}p^{3/2}\lambda^3}\left(1 + \frac{n}{p} + \frac{p}{n}\right)\right)\end{aligned} \tag{69}$$

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

$$\operatorname{Var}\langle G_{-i}x_i z_i^\top A\rangle \lesssim \frac{1}{p^2}\mathbf{E}_{-i}\|AG_{-i}\|_F^2 + \operatorname{Var}_{-i}\langle G_{-i}\Phi A\rangle \lesssim \frac{\langle |A|^2\rangle}{p\lambda^3} \tag{70}$$

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

$$|\langle AGBG - \mathbf{E}\, AGBG\rangle| \lesssim \frac{\|A\|\langle |B|^2\rangle^{1/2}}{p\lambda^{5/2}} \tag{71}$$

due to

$$|\langle AGBG\rangle - \langle AG'BG'\rangle| \le |\langle A(G-G')BG\rangle| + |\langle AG'B(G-G')\rangle| \le 2\frac{\|A\|\|B\|_F}{p\lambda}\|G-G'\|_F \tag{72}$$

and eq. (43).

It is useful to expand G around M as in

$$G = M + \lambda M\Omega G\frac{n}{p}\langle m - \breve{G}\rangle - M\frac{XX^\top}{p}G + \lambda M\Omega G\frac{n}{p}\langle\breve{G}\rangle = M - M\frac{XX^\top}{p}G + \lambda\langle\breve{G}\rangle M\Omega G\frac{n}{p} + O\Big(\frac{1}{p\lambda^3}\Big)M\Omega G \tag{73}$$

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

$$\begin{aligned}
\langle GAGB\rangle &= \langle MAGB\rangle - \langle M\frac{XX^\top}{p}GAGB\rangle + \frac{n\lambda}{p}\langle\breve{G}\rangle\langle M\Omega GAGB\rangle + O\Big(\frac{\langle|A|^2\rangle^{1/2}\langle|B|^2\rangle^{1/2}}{p\lambda^6}\Big)\\
&= \langle MAMB\rangle - \frac{1}{p}\sum_i(\langle Mx_ix_i^\top GAGB\rangle - \lambda\breve{G}_{ii}\langle M\Omega GAGB\rangle) + O\Big(\frac{\|B\|\langle|A|^2\rangle^{1/2}}{p\lambda^6}\Big)\\
&= \langle MAMB\rangle - \frac{\lambda}{p}\sum_i\breve{G}_{ii}(\langle Mx_ix_i^\top G_{-i}AG_{-i}B\rangle - \langle M\Omega GAGB\rangle) + O\Big(\frac{\|B\|\langle|A|^2\rangle^{1/2}}{p\lambda^6}\Big)\\
&\quad + \frac{\lambda^2}{p}\sum_i\breve{G}_{ii}^2\frac{x_i^\top G_{-i}AG_{-i}x_i}{p}\frac{x_i^\top G_{-i}BMx_i}{p},
\end{aligned} \tag{74}$$

using eq. (49) in the third step. The second term of eq. (74) can be estimated in expectation using

$$\begin{aligned}
\frac{\lambda}{p}\mathbf{E}\sum_i\breve{G}_{ii}\langle Mx_ix_i^\top G_{-i}AG_{-i}B\rangle &= \frac{\lambda}{p}\sum_i\Big((\mathbf{E}\,\breve{G}_{ii})\langle\mathbf{E}\,M\Omega G_{-i}AG_{-i}B\rangle + O\Big(\sqrt{\operatorname{Var}\breve{G}_{ii}}\sqrt{\operatorname{Var}\langle Mx_ix_i^\top G_{-i}AG_{-i}B\rangle}\Big)\Big)\\
&= \frac{\lambda}{p}\sum_i(\mathbf{E}\,\breve{G}_{ii})\langle\mathbf{E}\,M\Omega GAGB\rangle + O\Big(\frac{n\|B\|\langle|A|^2\rangle^{1/2}}{p^2\lambda^4}\Big)\\
&= \frac{\lambda}{p}\mathbf{E}\sum_i\breve{G}_{ii}\langle M\Omega GAGB\rangle + O\Big(\frac{n\|B\|\langle|A|^2\rangle^{1/2}}{p^2\lambda^4}\Big)
\end{aligned} \tag{75}$$

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

$$\operatorname{Var}\langle Mx_ix_i^\top G_{-i}AG_{-i}B\rangle \lesssim \frac{1}{p^2}\,\mathbf{E}_{-i}\|G_{-i}AG_{-i}BM\|_F^2 + \operatorname{Var}_{-i}\langle M\Omega G_{-i}AG_{-i}B\rangle \lesssim \frac{\|B\|^2\langle|A|^2\rangle}{p\lambda^6}(1+\frac{1}{p\lambda}). \tag{76}$$

and

$$\|G - G_i\| \lesssim \frac{1}{p\lambda^2}, \quad \langle M\Omega G_{-i}AG_{-i}B\rangle = \langle M\Omega GAGB\rangle + O\Big(\frac{\|B\|\langle|A|^2\rangle^{1/2}}{p\lambda^4}\Big). \tag{77}$$

For the last term of Eq. (74) we have

$$\begin{aligned}
\frac{x_i^\top G_{-i}AG_{-i}x_i}{p} &= \langle\Omega G_{-i}AG_{-i}\rangle + O\Big(\frac{1}{p}\|G_{-i}AG_{-i}\|_F\Big) = \langle\Omega GAG\rangle + O(\frac{\langle|A|^2\rangle^{1/2}}{p^{1/2}\lambda^2})\\
\frac{x_i^\top G_{-i}BMx_i}{p} &= \langle\Omega G_{-i}BM\rangle + O\Big(\frac{1}{p}\|G_{-i}BM\|_F\Big) = \langle\Omega MBM\rangle + O(\frac{\langle|B|^2\rangle^{1/2}}{p^{1/2}\lambda^2}),
\end{aligned} \tag{78}$$

so that with

$$\mathbf{E}\,\breve{G}_{ii}^2\frac{x_i^\top G_{-i}AG_{-i}x_i}{p}\frac{x_i^\top G_{-i}BMx_i}{p} = (\mathbf{E}\,\breve{G}_{ii}^2)\Big(\mathbf{E}\,\frac{x_i^\top G_{-i}AG_{-i}x_i}{p}\Big)\Big(\mathbf{E}\,\frac{x_i^\top G_{-i}BMx_i}{p}\Big) + O\bigg(\frac{\langle|A|^2\rangle^{1/2}\langle|B|^2\rangle^{1/2}}{p\lambda^7}\bigg)$$
$$= (\mathbf{E}\,\breve{G}_{ii}^2)\langle\mathbf{E}\,\Omega GAG\rangle\langle\mathbf{E}\,\Omega GBM\rangle + O\bigg(\frac{\langle|A|^2\rangle^{1/2}\langle|B|^2\rangle^{1/2}}{p\lambda^7}\bigg) \tag{79}$$

and

$$\frac{1}{p}\sum_i \breve{G}_{ii}^2 = \frac{n}{p}m^2 + 2\frac{n}{p}m\langle\breve{G}-m\rangle + \frac{1}{p}\sum_i(\breve{G}_{ii}-m)^2 = \frac{n}{p}m^2 + O\bigg(\frac{1}{p\lambda^5}+\frac{n}{p^2\lambda^3}\bigg) \tag{80}$$

we arrive at

$$\frac{\lambda^2}{p}\,\mathbf{E}\sum_i \breve{G}_{ii}^2\frac{x_i^\top G_{-i}AG_{-i}x_i}{p}\frac{x_i^\top G_{-i}BMx_i}{p} = \frac{n}{p}(m\lambda)^2\langle\Omega MBM\rangle\langle\mathbf{E}\,\Omega GAG\rangle + O\bigg(\frac{\langle|A|^2\rangle^{1/2}\langle|B|^2\rangle^{1/2}}{p\lambda^6}\bigg). \tag{81}$$

Choosing $B=\Omega$ it follows that

$$\langle GAG\Omega\rangle(1-\frac{n}{p}\lambda^2m^2\langle\Omega M\Omega M\rangle) = \langle MAM\Omega\rangle + O\bigg(\frac{\langle|A|^2\rangle^{1/2}}{p\lambda^6}\bigg), \tag{82}$$

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

$$\mathbf{E}\Big(\frac{X^\top G\Omega GX}{p}\Big)_{ii} = \lambda^2\,\mathbf{E}\,\breve{G}_{ii}^2\langle\mathbf{E}\,\Omega G_{-i}\Omega G_{-i}\rangle + O\bigg(\frac{1}{p}\sqrt{\operatorname{Var}\breve{G}_{ii}^2}\sqrt{\operatorname{Var}x_i^\top G_{-i}\Omega G_{-i}x_i}\bigg)$$
$$= \lambda^2\,\mathbf{E}\,\breve{G}_{ii}^2\frac{\langle\Omega M\Omega M\rangle}{1-\frac{n}{p}(m\lambda)^2\langle\Omega M\Omega M\rangle} + O\bigg(\frac{1}{p\lambda^7}\bigg), \tag{83}$$

so that by a Lipschitz concentration argument as in Eq. (71) we obtain for the diagonal part A_d of $A=A_d+A_o$ that

$$\bigg\langle\frac{X^\top G\Omega GX}{p}A_d\bigg\rangle = \frac{\lambda^2m^2\langle\Omega M\Omega M\rangle}{1-\frac{n}{p}(m\lambda)^2\langle\Omega M\Omega M\rangle}\langle A_d\rangle + O\bigg(\frac{\langle|A_d|^2\rangle^{1/2}}{p\lambda^7}\bigg). \tag{84}$$

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

$$\Big(\frac{X^\top G\Omega GX}{p}\Big)_{ij} = \frac{\lambda^2\breve{G}_{ii}\breve{G}_{jj}}{p}x_i^\top G_{-i}\Omega G_{-j}x_j$$
$$= \frac{\lambda^2\breve{G}_{ii}\breve{G}_{jj}}{p}x_i^\top G_{-ij}\Omega G_{-ij}x_j + \frac{\lambda^4\breve{G}_{ii}^2\breve{G}_{jj}^2}{p^3}x_i^\top G_{-ij}x_jx_j^\top G_{-ij}\Omega G_{-ij}x_ix_i^\top G_{-ij}x_j$$
$$- \frac{\lambda^3\breve{G}_{ii}^2\breve{G}_{jj}}{p^2}x_i^\top G_{-ij}\Omega G_{-ij}x_ix_i^\top G_{-ij}x_j - \frac{\lambda^3\breve{G}_{ii}\breve{G}_{jj}^2}{p^2}x_i^\top G_{-ij}x_jx_j^\top G_{-ij}\Omega G_{-ij}x_j. \tag{85}$$

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

$$\mathbf{E}\,\frac{\lambda^2\breve{G}_{ii}\breve{G}_{jj}}{p}x_i^\top G_{-ij}\Omega G_{-ij}x_j = \mathbf{E}\,\frac{\lambda^2(\breve{G}_{ii}-m)(\breve{G}_{jj}-m)}{p}x_i^\top G_{-ij}\Omega G_{-ij}x_j = O\bigg(\frac{1}{\lambda^2p^{3/2}}\bigg) \tag{86}$$

and

$$\mathbf{E}\,\frac{\lambda^3\breve{G}_{ii}^2\breve{G}_{jj}}{p^2}x_i^\top G_{-ij}\Omega G_{-ij}x_ix_i^\top G_{-ij}x_j = \mathbf{E}\,\frac{\lambda^3(\breve{G}_{ii}^2\breve{G}_{jj}-m^3)}{p^2}x_i^\top G_{-ij}\Omega G_{-ij}x_ix_i^\top G_{-ij}x_j = O\bigg(\frac{1}{p^{3/2}\lambda^{7/2}}\bigg) \tag{87}$$

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

$$\mathbf{E}\left|\left(\frac{X^\top G\Omega G X}{p}\right)_{ij}\right| = O\left(\frac{1}{p^{3/2}\lambda^4}\right), \tag{88}$$

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. (71) it is sufficient to approximate the expectation of $ZX^\top G\Omega G X Z^\top$. By writing out $ZX^\top$ and $XZ^\top$ and using eq. (49) twice we obtain

$$\begin{aligned}\frac{1}{kp} ZX^\top G\Omega G X Z^\top &= \frac{1}{kp}\sum_{ij} z_i x_i^\top G\Omega G x_j z_j^\top \\ &= \frac{1}{kp}\sum_i (\lambda \breve{G}_{ii})^2 z_i x_i^\top G_{-i}\Omega G_{-i} x_i z_i^\top + \frac{1}{kp}\sum_{i\neq j}(\lambda\breve{G}_{ii})(\lambda\breve{G}_{jj}) z_i x_i^\top G_{-i}\Omega G_{-j} x_j z_j^\top.\end{aligned} \tag{89}$$

For the first term of Eq. (89) we have

$$\begin{aligned}\frac{n}{kp}\,\mathbf{E}\langle A z_i x_i^\top G_{-i}\Omega G_{-i} x_i z_i^\top\rangle &= \frac{n}{k^2 p}(\mathbf{E}\, z_i^\top A z_i)(\mathbf{E}\, x_i^\top G_{-i}\Omega G_{-i} x_i) + O\left(\frac{n}{k^2 p}\sqrt{\operatorname{Var} z_i^\top A z_i}\sqrt{\operatorname{Var} x_i^\top G_{-i}\Omega G_{-i} x_i}\right) \\ &= \frac{n}{k}\langle A\Psi\rangle\,\mathbf{E}\langle \Omega G_{-i}\Omega G_{-i}\rangle + O\left(\frac{n\langle |A|^2\rangle^{1/2}}{p^{1/2}k^{3/2}\lambda^2}\right) \\ &= \frac{n}{k}\langle A\Psi\rangle\frac{\langle \Omega M\Omega M\rangle}{1-\frac{n}{p}(m\lambda)^2\langle \Omega M\Omega M\rangle} + O\left(\frac{n\langle |A|^2\rangle^{1/2}}{pk\lambda^3}(1+\sqrt{p/k})\right)\end{aligned} \tag{90}$$

using Item 2 in the ultimate step. For the second term in the right hand side of Eq. (89) we expand both G_{-i} and G_{-j} around G_{-ij} using Eq. (49) to

$$\begin{aligned}\langle z_i x_i^\top G_{-i}\Omega G_{-j} x_j z_j^\top A\rangle &\approx \left\langle z_i x_i^\top\Big(G_{-ij} - \lambda m G_{-ij}\frac{x_j x_j^\top}{p} G_{-ij}\Big)\Omega\Big(G_{-ij} - m\lambda G_{-ij}\frac{x_i x_i^\top}{p} G_{-ij}\Big) x_j z_j^\top A\right\rangle \\ &= \big\langle z_i x_i^\top G_{-ij}\Omega G_{-ij} x_j z_j^\top A\big\rangle + (\lambda m)^2\left\langle z_i x_i^\top G_{-ij}\frac{x_j x_j^\top}{p} G_{-ij}\Omega G_{-ij}\frac{x_i x_i^\top}{p} G_{-ij} x_j z_j^\top A\right\rangle \\ &\quad - \lambda m\left\langle z_i x_i^\top G_{-ij}\Omega G_{-ij}\frac{x_i x_i^\top}{p} G_{-ij} x_j z_j^\top A\right\rangle - \lambda m\left\langle z_i x_i^\top G_{-ij}\frac{x_j x_j^\top}{p} G_{-ij}\Omega G_{-ij} x_j z_j^\top A\right\rangle.\end{aligned} \tag{91}$$

Here in the first line we replaced $(\breve{G}_{-i})_{jj}$ and $(\breve{G}_{-j})_{ii}$ by m which results in an error term negligible compared to the other error terms. The first term of Eq. (91) can, in expectation, be approximated by

$$\mathbf{E}\big\langle z_i x_i^\top G_{-ij}\Omega G_{-ij} x_j z_j^\top A\big\rangle = \mathbf{E}\langle \Phi^\top G_{-ij}\Omega G_{-ij}\Phi A\rangle = \frac{\langle \Phi^\top M\Omega M\Phi A\rangle}{1-\frac{n}{p}(m\lambda)^2\langle \Omega M\Omega M\rangle} + O\left(\frac{\langle |A|^2\rangle^{1/2}}{p\lambda^7}\right), \tag{92}$$

using Item 2 in the ultimate step. The third term of Eq. (91) can be approximated by

$$\begin{aligned}&\lambda m\,\mathbf{E}\left\langle z_i x_i^\top G_{-ij}\Omega G_{-ij}\frac{x_i x_i^\top}{p} G_{-ij} x_j z_j^\top A\right\rangle \\ &= \frac{1}{kp}\lambda m (x_i^\top G_{-ij}\Phi A z_i)(x_i^\top G_{-ij}\Omega G_{-ij} x_i) \\ &= \lambda m\,\mathbf{E}_{-ij}\left(\langle \Phi^\top G_{-ij}\Phi A\rangle\langle \Omega G_{-ij}\Omega G_{-ij}\rangle + O\left(\sqrt{\operatorname{Var}_i\frac{x_i^\top G_{-ij}\Phi A z_i}{k}}\sqrt{\operatorname{Var}_i\frac{x_i^\top G_{-ij}\Omega G_{-ij} x_i}{p}}\right)\right) \\ &= \lambda m\frac{\langle \Phi^\top M\Phi A\rangle\langle \Omega M\Omega M\rangle}{1-\frac{n}{p}(m\lambda)^2\langle \Omega M\Omega M\rangle} + O\left(\frac{\langle |A|^2\rangle}{\lambda^7 p}\left(1+\sqrt{\frac{p}{k}}\right)\right)\end{aligned} \tag{93}$$

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

$$\mathrm{Var}_i \frac{x_i^\top G_{-ij}\Omega G_{-ij} x_i}{p} \lesssim \frac{1}{p^2}\|G_{-ij}\Omega G_{-ij}\|_F^2 \lesssim \frac{1}{p\lambda^2}, \quad \mathrm{Var}_i \frac{x_i^\top G_{-ij}\Phi A z_i}{k} \lesssim \frac{\langle |A|^2\rangle}{\lambda k} \tag{94}$$

and Eq. (58) and Item 2. Finally, for the second term of Eq. (89) we use the simple bound

$$\begin{aligned}
&\left\langle z_i x_i^\top G_{-ij}\frac{x_j x_j^\top}{p} G_{-ij}\Omega G_{-ij}\frac{x_i x_i^\top}{p} G_{-ij} x_j z_j^\top A\right\rangle \\
&= \frac{1}{kp^2}(x_i^\top G_{-ij}x_j)(x_j^\top G_{-ij}\Omega G_{-ij}x_i)(x_i^\top G_{-ij}x_j)(z_j^\top A z_i) \\
&= O\left(\frac{1}{kp^2}\|G_{-ij}\|_F\|G_{-ij}\Omega G_{-ij}\|_F\|G_{-ij}\|_F\|A\|_F\right) = O\left(\frac{\langle |A|^2\rangle^{1/2}}{k^{1/2}p^{1/2}\lambda^4}\right).
\end{aligned} \tag{95}$$

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 definition and propositions from (Nourdin & Peccati, 2012), 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 (Nourdin & Peccati, 2012) are about general separable Hilbert spaces.

Following ((Nourdin & Peccati, 2012), 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 $H_n(X(h)), h \in \mathbf{R}^d$, $\|h\| = 1$, where H_n is the n-th *Hermite polynomial*. We call $\mathcal{H}_n$, the n-th 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

$$f(a_1, \ldots, a_p) = \frac{1}{p!}\sum_{\sigma\in S_p} f(a_{\sigma(1)}, \ldots, a_{\sigma(p)}). \tag{96}$$

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 ((Nourdin & Peccati, 2012), 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

$$D^p F := \sum_{i_1,\ldots,i_p=1}^m \frac{\partial^p f}{\partial x_{i_1}\ldots\partial x_{i_p}}(X(h_1), \ldots, X(h_m)) h_{i_1}\otimes\ldots\otimes h_{i_p}. \tag{97}$$

Proposition B.3 ((Nourdin & Peccati, 2012), 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*

$$D\phi(F) = \sum_{i=1}^m \frac{\partial\phi}{\partial x_i}(F) DF_i. \tag{98}$$

Definition B.4 ((Nourdin & Peccati, 2012), 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 ((Nourdin & Peccati, 2012), Definition 2.7.1)**.** Let $p \geq 1$ and $f \in \mathfrak{H}^{\tilde{\otimes} p}$. The pth multiple integral of f with respect to X is defined by $I_p(f) = \delta^p(f)$.

Proposition B.6 ((Nourdin & Peccati, 2012), 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*

$$\mathbf{E}\, I_p(f) I_q(g) = \delta_{pq} p! \langle f, g \rangle_{\mathfrak{H}^{\otimes p}} \tag{99}$$

Theorem B.7 ((Nourdin & Peccati, 2012), 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*

$$H_p(X(f)) = I_p(f^{\otimes p}), \tag{100}$$

where H_p is the p-th Hermite polynomial.

Corollary B.8 ((Nourdin & Peccati, 2012), Corollary 2.7.8)**.** *Every $F \in L^2(\Omega)$ can be expanded as*

$$F = \mathbf{E}\, F + \sum_{p=1}^{\infty} I_p(f_p), \tag{101}$$

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 $p \ge 1$,

$$f_p = \frac{1}{p!}\, \mathbf{E}\, D^p F. \tag{102}$$

Theorem B.9 ((Nourdin & Peccati, 2012), 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,*

$$|\mathbf{E}\, h(N) - \mathbf{E}\, h(F)| \le \frac{1}{2} \|h''\|_\infty\, \mathbf{E}\left[|\langle DF, -DL^{-1}F \rangle - \sigma^2| \right]. \tag{103}$$

Finally, we use the following multivariate version of the previous theorem.

Theorem B.10 ((Nourdin & Peccati, 2012), Theorem 6.1.2)**.** *Fix $c \ge 2$, and let $F = (F_1, \dots, F_c)$ be a random vector such that $F_i \in \mathbb{D}^{1,4}$ with $\mathbf{E}\, F_i = 0$ for any i. Let $C \in \mathcal{M}_c(\mathbf{R})$ be a symmetric non-negative definite matrix, and let $N \sim \mathcal{N}(0, C)$. Then, for any $h : \mathbf{R}^c \to \mathbf{R}$ belonging to $\mathcal{C}^2$ such that $\|h''\|_\infty < \infty$,*

$$|\mathbf{E}\, h(F) - \mathbf{E}\, h(N)| \le \frac{c}{2} \|h''\|_\infty \sqrt{\sum_{i,j=1}^{c} \mathbf{E}\left[\left(C_{ij} - \langle DF_j, -DL^{-1}F_i \rangle_{\mathfrak{H}} \right)^2 \right]} \tag{104}$$

Remark B.11. *We believe there is a mistake in the original formulation of Theorem 6.1.2 in (Nourdin & Peccati, 2012). In particular, originally the expression on the right hand side did not contain c term.*

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

$$f(Wx)_i = f(w_i^\top x) = \sum_{p \ge 1} \frac{\mathbf{E}\, f^{(p)}((WW^\top)_{ii}^{1/2} N)}{p!} I_p(w_i^{\otimes p}), \tag{105}$$

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 $W_1 \mapsto W_1 \sqrt{\Omega_0}$. Let $(f_\ell) : \mathbf{R} \to \mathbf{R}$ be a sequence of smooth functions, (W^ℓ) be a sequence of matrices. We define a sequence of vectors x^i, such that $x^0 := x \sim \mathcal{N}(0, I)$, $x^{\ell+1} = f_{\ell+1}(W^{\ell+1} x^\ell)$.

Lemma B.12 (Weak correlation)**.** *Let $b \ge 1$ be a fixed integer. Let $h_0, h_1, \dots, h_b$ be a collection of functions. Then, we have that*

$$\mathbf{E}\left[h_0(u^\top f_1(W^1 x)) \prod_{i=1}^{b} h_i(w_i^\top x) \right] = \mathbf{E}\, h_0(u^\top f_1(W^1 x)) \prod_{i=1}^{b} \mathbf{E}\, h_i(w_i^\top x) + O(d^{-1/2}). \tag{106}$$

Proof. The fact that $u_i \lesssim d^{-1/2}$ and $f_1(w^\top x) \lesssim 1$ together with perturbation analysis imply that

$$\mathbf{E}\left[h_0(u^\top f_1(W^1 x)) \prod_{i=1}^{b} h_i(w_i^\top x) \right] = \mathbf{E}\left[h_0\left(\sum_{k \ge b+1} u_k f_1(w_k^\top x) \right) \prod_{i=1}^{b} h_i(w_i^\top x) \right] + O(d^{-1/2}). \tag{107}$$

Let $A \coloneqq h_0\left(\sum_{k\geq b+1} u_k f_1(w_k^\top x)\right)$ and $B \coloneqq \prod_{i=1}^b h_i(w_i^\top x)$. Note that for any $p \geq 1$, $\langle \mathbf{E}\, D^p A, \mathbf{E}\, D^p B\rangle$ constitutes of products of $\langle w_i, w_j\rangle$, where $i \neq j$. Each of these 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

$$\mathbf{E}\left[h_0\left(\sum_{k\geq b+1} u_k f_1(w_k^\top x)\right)\prod_{i=1}^b h_i(w_i^\top x)\right] = \mathbf{E}\left[h_0\left(\sum_{k\geq b+1} u_k f_1(w_k^\top x)\right)\right]\mathbf{E}\left[\prod_{i=1}^b h_i(w_i^\top x)\right] + O(d^{-1/2}). \quad (108)$$

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

$$\mathbf{E}\left[h_0(u^\top f_1(W^1 x))\prod_{i=1}^b h_i(w_i^\top x)\right] = \mathbf{E}\, h_0(u^\top f_1(W^1 x))\prod_{i=1}^b \mathbf{E}\, h_i(w_i^\top x) + O(d^{-1/2}) \quad (109)$$

□

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 $W \in \mathbf{R}^{k\times d}, V \in \mathbf{R}^{s\times d}$ and two smooth odd functions f, g applied entrywise to Wx, Vx. We assume that rows of W and V are mean-zero i.i.d. samples $(w_i, v_i) \sim (w, v)$, such that $C_w \coloneqq \mathbf{E}\, ww^\top$ and $C_v \coloneqq \mathbf{E}\, vv^\top$. Let $C_{wv} = \mathbf{E}\, wv^\top$ if $s = k$ and $C_{wv} = \mathbf{0}_{d\times d}$ (all-zero matrix) otherwise.

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

$$\mathbf{E}\, N_w^2 = \operatorname{Tr} C_w, \quad \mathbf{E}\, N_v^2 = \operatorname{Tr} C_v, \quad \mathbf{E}\, N_w N_v = \operatorname{Tr} C_{wv}. \quad (110)$$

Define

$$\begin{aligned}\Phi_1 &= \mathbf{E}\, f(Wx)g(Vx)^\top,\\ \Phi_1^{\text{lin}} &= (\mathbf{E}\, f'(N_w))(\mathbf{E}\, g'(N_v))WV^\top + [\mathbf{E}\, f(N_w)g(N_v) - (\mathbf{E}\, f'(N_w))(\mathbf{E}\, g'(N_v))(\mathbf{E}\, N_w N_v)]I.\end{aligned} \quad (111)$$

Proposition B.13. *We have that, with high probability,* $\|\Phi_1 - \Phi_1^{\text{lin}}\|_F = O(1)$.

Proof. Using a Wiener chaos expansion (eq. (105)), we can write

$$f(Wx)_i = \sum_{p\geq 1}\frac{\mathbf{E}\, f^{(p)}((WW^\top)_{ii}^{1/2}N)}{p!}I_p(Wx)_i, \quad g(Vx)_j = \sum_{p\geq 1}\frac{\mathbf{E}\, g^{(p)}((VV^\top)_{jj}^{1/2}N)}{p!}I_p(Vx)_j \quad (112)$$

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

$$\mathbf{E}\, I_p(Wx)I_q(Vx)^\top = p!\delta_{pq}(WV^\top)^{\odot p} \quad (113)$$

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

$$\mathbf{E}\, f(Wx)_i g(Vx)_j = \sum_{p\geq 1}\frac{1}{p!}(\mathbf{E}\, f^{(p)}((WW^\top)_{ii}^{1/2}N))(WV^\top)_{ij}^p(\mathbf{E}\, g^{(p)}((VV^\top)_{jj}^{1/2}N)). \quad (114)$$

From Theorem A.3 (note that $\|w\|_{\psi_2} \sim d^{-1/2}$ and same for v), and since $\operatorname{Tr} C_w \sim 1$, it follows that

$$(WW^\top)_{ii} = \operatorname{Tr} C_w + O(d^{-1/2}), \quad (VV^\top)_{jj} = \operatorname{Tr} C_v + O(d^{-1/2}), \quad (115)$$

$$(WV^\top)_{ij} = \delta_{ij}\operatorname{Tr} C_{wv} + O(d^{-1/2}). \quad (116)$$

From perturbation analysis, we can write

$$\mathbf{E}\, f^{(p)}((WW^\top)_{ii}^{1/2}N) = \mathbf{E}\, f^{(p)}(\sqrt{\operatorname{Tr} C_w}N) + O(d^{-1/2}) = \mathbf{E}\, f^{(p)}(N_w) + O(d^{-1/2}), \quad (117)$$

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

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

$$\mathbf{E}\, f(Wx)_i g(Vx)_j = \mathbf{E}\, f'(N_w) g'(N_v)(WV^\top)_{ij} + O(d^{-1}) = (\Phi_1^{\text{lin}})_{ij} + O(d^{-1}). \tag{118}$$

Diagonal entries If $s \neq k$, we have $(WV^\top)^p_{ii} = O(d^{-p/2})$ for $p \geq 2$, and thus obtain the same expression as in previous case. When $s = k$, we can rewrite the infinite sum as

$$\begin{aligned}
\mathbf{E}\, f(Wx)_i g(Vx)_i &= \sum_{p\geq 1} \frac{1}{p!} (\mathbf{E}\, f^{(p)}((WW^\top)^{1/2}_{ii} N))(WV^\top)^p_{ii} (\mathbf{E}\, g^{(p)}((VV^\top)^{1/2}_{ii} N)) \\
&= \sum_{p\geq 1} \frac{[\mathbf{E}\, f^{(p)}(\sqrt{\operatorname{Tr} C_w} N)][\mathbf{E}\, g^{(p)}(\sqrt{\operatorname{Tr} C_v} N)]}{p!} (\operatorname{Tr} C_{wv})^p + O(d^{-1/2}) \\
&= \mathbf{E}\, f(N_w) g(N_v) + O(d^{-1/2}) = (\Phi_1^{\text{lin}})_{ii} + O(d^{-1/2}).
\end{aligned} \tag{119}$$

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

$$\Phi_1^{\text{lin}} = (\mathbf{E}\, f'(N_w))^2 WW^\top + [\mathbf{E}\, f(N_w)^2 - (\mathbf{E}\, f'(N_w))^2 \operatorname{Tr} C_w] I. \tag{120}$$

B.3. Two layer case

We now consider the 2-layer example

$$f_2(W^2 f_1(W^1 x)), \quad g_2(V^2 g_1(V^1 x)), \tag{121}$$

with smooth odd[9] functions f_1, f_2, g_1, g_2. We assume that the rows of W^1, W^2, V^1, V^2 are mean-zero i.i.d. samples $(w_i^1, w_i^2, v_i^1, v_i^2) \sim (w^1, w^2, v^1, v^2)$, such that $C_1 \coloneqq \mathbf{E}\, w^1 (w^1)^\top, C_2 \coloneqq \mathbf{E}\, w^2 (w^2)^\top, \tilde{C}_1 \coloneqq \mathbf{E}\, v^1 (v^1)^\top$, and $\tilde{C}_1 \coloneqq \mathbf{E}\, v^2 (v^2)^\top$. Let $\check{C}_1 = \mathbf{E}\, w^1 (v^1)^\top$ and $\check{C}_2 = \mathbf{E}\, w^2 (v^2)^\top$. Let $(N_1, \tilde{N}_1)$ be a zero-mean jointly Gaussian random variables:

$$(N_1, \tilde{N}_1) \sim \mathcal{N}\left(0, \begin{pmatrix} \operatorname{Tr}(C_1) & \operatorname{Tr}(\check{C}_1) \\ \operatorname{Tr}(\check{C}_1) & \operatorname{Tr}(\tilde{C}_1) \end{pmatrix}\right), \tag{122}$$

and define

$$\begin{aligned}
&\Phi_1 \coloneqq \mathbf{E}\, f_1(W^1 x) g_1(V^1 x)^\top, \\
&\Phi_1^{\text{lin}} = (\mathbf{E}\, f_1'(N_1))(\mathbf{E}\, g_1'(\widetilde{N}_1)) W^1 (V^1)^\top + [\mathbf{E}\, f_1(N_1) g_1(\widetilde{N}_1) - (\mathbf{E}\, f_1'(N_1))(\mathbf{E}\, g_1'(\widetilde{N}_1))(\mathbf{E}\, N_1 \widetilde{N}_1)] I.
\end{aligned} \tag{123}$$

Similarly, we define $\Omega_1, \Omega_1^{\text{lin}}$ with $V^1, g_1, \widetilde{N}_1$ replaced by W^1, f_1, N_1 and we define $\Psi_1, \Psi_1^{\text{lin}}$ with W^1, f_1, N_1 replaced by $V^1, g_1, \widetilde{N}_1$ (see Definition 4.2). Next, let $(N_2, \widetilde{N}_2)$ be a zero-mean jointly Gaussian random variables:

$$(N_2, \tilde{N}_2) \sim \mathcal{N}\left(0, \begin{pmatrix} \operatorname{Tr}(C_2 \Omega_1^{\text{lin}}) & \operatorname{Tr}(\check{C}_2 \Phi_1^{\text{lin}}) \\ \operatorname{Tr}(\check{C}_2 \Phi_1^{\text{lin}}) & \operatorname{Tr}(\tilde{C}_2 \Psi_1^{\text{lin}}) \end{pmatrix}\right), \tag{124}$$

and define

$$\begin{aligned}
&\Phi_2 \coloneqq \mathbf{E}\, f_2(W^2 f_1(W^1 x)) g_2(V^2 g_1(V^1 x))^\top, \\
&\Phi_2^{\text{lin}} = (\mathbf{E}\, f_2'(N_2))(\mathbf{E}\, g_2'(\widetilde{N}_2)) W^2 \Phi_1^{\text{lin}} (V^2)^\top + [\mathbf{E}\, f_2(N_2) g_2(\widetilde{N}_2) - (\mathbf{E}\, f_2'(N_2))(\mathbf{E}\, g_2'(\widetilde{N}_2))(\mathbf{E}\, N_2 \widetilde{N}_2)] I,
\end{aligned} \tag{125}$$

and, again, similarly $\Omega_2, \Omega_2^{\text{lin}}, \Psi_2$, and Ψ_2^{lin}.

Theorem B.14. *We have that* $\|\Phi_2 - \Phi_2^{\text{lin}}\|_F \prec 1$.

We split the proof into the following lemmas:

[9] For brevity we present the full proof in the case of odd activation functions. The argument for the general case (i.e., when only assuming that activation functions are centered w.r.t. Gaussian distribution) is similar, but requires more tedious estimates.

Lemma B.15 (Diagonal entries of Φ_2)**.** *For possibly correlated vectors u, z, we have the bound*

$$\left|\mathbf{E}\, f_2(u^\top f_1(W^1 x)) g_2(z^\top g_1(V^1 x)) - \mathbf{E}\, f_2(N_2) g_2(\tilde{N}_2)\right| \prec d^{-1/2}. \tag{126}$$

Lemma B.16 (Off-diagonal entries of Φ_2)**.** *If u and z are independent, we have*

$$\left|\mathbf{E}\, f_2(u^\top f_1(W^1 x)) g_2(z^\top g_1(V^1 x)) - (\mathbf{E}\, f_2'(N_1))(\mathbf{E}\, g_2'(N_2)) u^\top \Phi_1^{\mathrm{lin}} z\right| \prec d^{-1}. \tag{127}$$

Proof of Theorem B.14. The proof follows from Lemmas B.15 and B.16 upon summation over all entries. □

Proof of Theorem 4.4. The proof follows from Proposition B.13 and Theorem B.14. Note that results for Ω_i and Ψ_i can be obtained using aforementioned results only for f_i, W^i and only for g_i, V^i respectively. □

B.4. Proof of Lemma B.16

For simplicity of notation, we omit indices in W^1, V^1 and write W, V instead. We begin with showing that

$$\mathbf{E}\, f_2(u^\top f_1(Wx)) g_2(z^\top g_1(Vx)) = \mathbf{E}\left[f_2'(u^\top f_1(Wx))\right] \mathbf{E}\left[g_2'(z^\top g_1(Vx))\right] u^\top \Phi_1 z + O(d^{-1}), \tag{128}$$

for *independent* random vectors u and z. Recall that

$$f_2(u^\top f_1(Wx)) = \sum_{p\geq 1} \frac{1}{p!} I_p \left(\underbrace{\mathbf{E} \sum_{\pi \vdash [p]} f_2^{(|\pi|)}(u^\top f_1(Wx)) \widetilde{\bigotimes}_{B\in\pi} \left(\sum u_k f_1^{(|B|)}(w_i^\top x) w_i^{\otimes |B|} \right)}_{\mathbf{E}\, D^p f_2(u^\top f_1(Wx))} \right), \tag{129}$$

and that $\mathbf{E}\, f_2(u^\top f_1(Wx)) g_2(z^\top g_1(Vx)) = \sum_{p\geq 1} \frac{1}{p!} \langle \mathbf{E}\, D^p f_2(u^\top f_1(Wx)), \mathbf{E}\, D^p g_2(z^\top g_1(Vx)) \rangle$.

Let $f_2^p := \mathbf{E}\, f_2^{(p)}(u^\top f_1(Wx))$ and $g_2^p := \mathbf{E}\, g_2^{(p)}(z^\top g_1(Vx))$. Lemma B.12 implies that

$$\begin{aligned}
&\Big\langle \mathbf{E}\left[f_2^{(|\pi|)}(u^\top f_1(W^1 x)) \widetilde{\bigotimes}_{B\in\pi} \left(\sum_k u_k f_1^{(|B|)}(w_i^\top x) w_i^{\otimes |B|} \right) \right], \\
&\quad \mathbf{E}\left[g_2^{(|\pi'|)}(z^\top g_1(V^1 x)) \widetilde{\bigotimes}_{B\in\pi'} \left(\sum_k u_k f_1^{(|B|)}(w_i^\top x) w_i^{\otimes |B|} \right) \right] \Big\rangle \\
&= f_2^{|\pi|} g_2^{|\pi'|} \left\langle \widetilde{\bigotimes}_{B\in\pi} \left(\sum_k u_k f_1^{(|B|)}(w_i^\top x) w_i^{\otimes |B|} \right), \widetilde{\bigotimes}_{B'\in\pi'} \left(\sum_k z_k g_1^{(|B'|)}(v_i^\top x) v_i^{\otimes |B'|} \right) \right\rangle + O(d^{-1}).
\end{aligned} \tag{130}$$

Using it, we can write (denoting $f_{1i}^p := \mathbf{E}\, f_1^{(p)}(w_i^\top x)$ and $g_{1i}^p := \mathbf{E}\, g_1^{(p)}(v_i^\top x)$)

$$\begin{aligned}
&\langle \mathbf{E}\, D^p f_2(u^\top f_1(Wx)), \mathbf{E}\, D^p g_2(z^\top g_1(Vx)) \rangle \\
&= \sum_{\pi,\pi' \vdash [p]} \sum_{\substack{i_1,\dots,i_{|\pi|} \\ j_1,\dots,j_{|\pi'|}}} f_2^{|\pi|} g_2^{|\pi'|} \prod_{k=1}^{|\pi|} \left(f_{1i}^{b(k)} u_{i_k} \right) \prod_{k=1}^{|\pi'|} \left(g_{1i}^{b'(k)} z_{j_k} \right) \frac{1}{p!} \sum_{\sigma\in S_p} \prod_{q=1}^{p} \langle w_{i_{\pi(q)}}, v_{j_{\pi'(\sigma(q))}} \rangle + O(d^{-1}),
\end{aligned} \tag{131}$$

where $b(k)$ and $b'(k)$ denote the size of kth block in π and π' respectively. The term $\pi = \pi' = \{[p]\}$ corresponds to

$$\sum_{i,j} \mathbf{E}\, f_2'(u^\top f_1(Wx)) \mathbf{E}\, g_2'(z^\top g_1(Vx)) f_1^{(p)}(w_i^\top x) g_1^{(p)}(v_j^\top x) u_i z_j \langle w_i, v_j \rangle^p, \tag{132}$$

which, after summing over $p \geq 1$ is equal to

$$\mathbf{E}\left[f_2'(u^\top f_1(Wx))\right] \mathbf{E}\left[g_2'(z^\top g_1(Vx))\right] u^\top \mathbf{E}\left[f_1(W^\top x) g_1(V^\top x)\right] z. \tag{133}$$

Therefore, it remains to show that all the other terms contribute in total $O(d^{-1})$. Note that $f_2^{|\pi|} = O(1)$ and same for other derivatives.

Lemma B.17. *Fix $\pi, \pi' \vdash [p]$. Then,*

$$\sum_{\substack{i_1,\dots,i_{|\pi|}\\ j_1,\dots,j_{|\pi'|}}} \prod_{k=1}^{|\pi|} u_{i_k} \prod_{k=1}^{|\pi'|} z_{j_k} \prod_{q=1}^{p} \langle w_{i_{\pi(q)}}, v_{j_{\pi'(q)}} \rangle \prec d^{\frac{1}{2}(\min(|\pi|,|\pi'|)-p)}. \tag{134}$$

Proof. Without loss of generality, it is enough to show upper bound $\prec d^{\frac{1}{2}(|\pi'|-p)}$. By separating terms that depend on i_k for $k \in [|\pi|]$, we can rewrite

$$\sum_{\substack{i_1,\dots,i_{|\pi|}\\ j_1,\dots,j_{|\pi'|}}} \prod_{k=1}^{|\pi|} u_{i_k} \prod_{k=1}^{|\pi'|} z_{j_k} \prod_{q=1}^{p} \langle w_{i_{\pi(q)}}, v_{j_{\pi'(q)}} \rangle = \sum_{j_1,\dots,j_{|\pi'|}} \prod_{k=1}^{|\pi'|} z_{j_k} \prod_{k=1}^{|\pi|} \left(\sum_{i_k} u_{i_k} \prod_{q:\pi(q)=k} \langle w_{i_k}, v_{j_{\pi'(q)}} \rangle \right). \tag{135}$$

Note that $\sum_{i_k} u_{i_k} \prod_{q:\pi(q)=k} \langle w_{i_k}, v_{j_{\pi'(q)}} \rangle \prec d^{-b(k)/2}$, where $b(k)$ denotes the size of k-th block of π. Since $\sum_k b(k) = p$, we bound the total expression by $d^{|\pi'|-|\pi'|/2-p/2} = d^{\frac{1}{2}(|\pi'|-p)}$. □

Since we assume that f_1, g_1 are odd, terms with $|\pi| = p-1$ are equal to zero, and same with $|\pi'| = p-1$. When $\min(|\pi|, |\pi'|) \le p-2$, the previous lemma implies $O(d^{-1})$ total contribution. Therefore, the only case left is with $\pi = \pi' = \{\{1\}, \{2\}, \dots, \{p\}\}$. However, in this case it is easy to see that the final contribution is $O(d^{-p/2})$.

Next, note that perturbation analysis implies that $\mathbf{E}\, f_2'(u^\top f_1(W^1 x)) = \mathbf{E}\, f_2'(N_2) + O(d^{-1/2})$, same for g_2. Finally, using that $\|\Phi_1 - \Phi_1^{\text{lin}}\|_F \prec 1$, we obtain that

$$\left| \mathbf{E}\, f_2(u^\top f_1(W^1 x)) g_2(z^\top g_1(V^1 x)) - (\mathbf{E}\, f_2'(N_2))(\mathbf{E}\, g_2'(\tilde{N}_2)) u^\top \Phi_1^{\text{lin}} z \right| \prec d^{-1}, \tag{136}$$

which finishes the proof.

B.5. Proof of Lemma B.15

For convenience we restate several concentration results that follow from Assumption 4.1.

Lemma B.18. *Let $w_1, \dots, w_d$ be a collection of independent random vectors, such that for all $\|w_i\|_{\psi_2} = O(d^{-1/2})$ and $\|w_i\| = O(1)$ for all $i \in [d]$. Then*

(i) $\sum_i \langle w_1, w_i \rangle \prec 1$,

(ii) $\sum_{ij} \langle w_1, w_i \rangle \langle w_1, w_j \rangle \prec 1$.

Proof. Let $X = \sum_{i \ge 2} \langle w_1, w_i \rangle$. Note that $\|X\|_{\psi_2} = O(1)$, which, together with $\|w_1\| = O(1)$ implies (i).

For (ii), note that $\sum_{ij} \langle w_1, w_i \rangle \langle w_1, w_j \rangle = \left(\sum_i \langle w_1, w_i \rangle\right)^2 \prec 1$ using (i). □

Let $F_1 = u^\top f_1(W^1 x)$ and $F_2 = z^\top g_1(V^1 x)$, where u, z may be correlated. For simplicity we omit indices in f_1, g_1, W^1, and V^1. Using Wiener chaos expansion, we obtain

$$F_1 = u^\top f(Wx) = \sum_{p \ge 1} I_p\left(\frac{\mathbf{E}\, D^p F_1}{p!}\right) = \sum_{p \text{ odd}} I_p\left(\sum_i \frac{u_i w_i^{\otimes p}\, \mathbf{E}\, f^{(p)}(w_i^\top x)}{p!}\right), \tag{137}$$

where the last equality uses the fact that $\mathbf{E}\, f^{(p)}(w_i^\top x) = 0$ for even p. Similarly, we can write

$$F_2 = z^\top g(Vx) = \sum_{p \text{ odd}} I_p\left(\sum_i \frac{z_i v_i^{\otimes p}\, \mathbf{E}\, g^{(p)}(v_i^\top x)}{p!}\right), \tag{138}$$

and denote $f_i^p := \mathbf{E}\, f^{(p)}(w_i^\top x), g_i^p := \mathbf{E}\, g^{(p)}(v_i^\top x)$. Next, we compute

$$DF_1 = \sum_{\text{odd } p} p I_{p-1}\left(\sum_i \frac{u_i w_i^{\otimes p} f_i^p}{p!}\right) \quad \text{and} \quad -DL^{-1}F_2 = \sum_{\text{odd } q} I_{q-1}\left(\sum_i \frac{z_i v_i^{\otimes q} g_i^q}{q!}\right) \tag{139}$$

Lemma B.19.

$$\mathbf{E}(\langle DF_1, -DL^{-1}F_2\rangle - \mathbf{E}\, F_1F_2)^2 = O(d^{-1}). \tag{140}$$

Proof. Since $I_{p-1}(w_i^{\otimes p}) = I_{p-1}(w_i^{\otimes p-1})w_i$, we can write

$$\langle DF_1, -DL^{-1}F_2\rangle = \sum_{\substack{\text{odd } p\\ \text{odd } q}} c_{pq} \sum_{i,j} \langle w_i, v_j\rangle u_i z_j f_i^p g_j^q I_{p-1}(w_i^{\otimes p-1}) I_{q-1}(w_j^{\otimes q-1}). \tag{141}$$

Using ((Nourdin & Peccati, 2012), Theorem 2.7.10), we obtain

$$\begin{aligned} &I_{p-1}(w_i^{\otimes p-1} I_{q-1}(w_j^{\otimes} q-1) \\ &= \sum_{r=0}^{p\wedge q-1} \langle w_i, v_j\rangle^r c_{rpq} I_{p+q-2(r+1)}(w_i^{\otimes p-1-r} \widetilde{\otimes} v_j^{\otimes q-1-r}) \\ &= \sum_{\substack{s=|p-q|\\ 2 \text{ divides } (s-|p-q|)}}^{p+q-2} c'_{spq} \langle w_i, v_j\rangle^{(p+q-2-s)/2} I_s(w_i^{\otimes p-q+s} \widetilde{\otimes} v_j^{q-p+s}). \end{aligned} \tag{142}$$

Therefore, we obtain

$$\langle DF_1, -DL^{-1}F_2\rangle = \sum_{s\geq 0} \sum_{\substack{|p-q|\leq s\\ 2 \text{ divides } (s-|p-q|)\\ p\wedge q \geq 1+(s-|p-q|)/2}} \tilde{c}_{r,p,q} \sum_{i,j} \langle w_i, v_j\rangle^{(p+q-s)/2} u_i z_j f_i^p f_j^q I_s(w_i^{\otimes (s+p-q)/2} \widetilde{\otimes} v_j^{\otimes (s+q-p)/2}). \tag{143}$$

The term $s = 0$ corresponds to $\mathbf{E}\, F_1F_2$. Since p and q must be odd in the non-zero terms of the sum, we obtain that s must be even. For $a \coloneqq (p+q-s)/2$, the s-th multiplie integral I_s can be rewritten as follows:

$$I_s\left(\sum_{a\geq 1}\sum_{i,j} \langle w_i, v_j\rangle u_i z_j T_{ij}^s\right), \tag{144}$$

where T_{ij}^s is a s-dimensional tensor, consisting of a sum of inner products of w_i and v_j, also containing combinatorial terms, and products of expectations of derivatives of f, g. We can write

$$\mathbf{E}\left(\langle DF_1, -DL^{-1}F_2\rangle - \mathbf{E}\, F_1F_2\right)^2 = \sum_{s\geq 2} \mathbf{E}\, I_s \left(\sum_{a\geq 1}\sum_{i,j} \langle w_i, v_j\rangle u_i z_j T_{ij}^s\right)^2. \tag{145}$$

Fix $s \geq 2$ and observe that

$$\mathbf{E}\, I_s \left(\sum_{a\geq 1}\sum_{i,j} \langle w_i, v_j\rangle u_i z_j T_{ij}^s\right)^2 = \sum_{a,a'\geq 1} \sum_{\substack{i,j\\ i',j'}} \langle w_i, v_j\rangle^a \langle w_{i'}, v_{j'}\rangle^{a'} u_i u_{i'} z_j z_{j'} \langle T_{ij}^s, T_{i'j'}^s\rangle, \tag{146}$$

and note that for some constant $C > 0$ (depending on combinatorial terms, and products of expectations of derivatives of f) $\langle T_{ij}^s, T_{i'j'}^s\rangle$ can be upper bounded by

$$\langle T_{ij}^s, T_{i'j'}^s\rangle \leq C(\langle w_i, w_{i'}\rangle + \langle w_i, v_{j'}\rangle + \langle v_j, w_{i'}\rangle + \langle v_j, v_{j'}\rangle)^s. \tag{147}$$

We analyze each term of the summand in Eq. (146) depending on a, a', i, i', j, j'. Let $N = |\{i, i', j, j'\}|$, the number of distinct indices among i, i', j, j'. Since entries of u and z are $O(d^{-1/2})$, we get that in total the term $u_i u_{i'} z_j z_{j'}$ contributes $O(d^{-2})$.

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

Case $N = 2$ There are $O(d^2)$ such terms. It must be that either (i) $i \neq j$ or (ii) $i' \neq j'$ or (iii) both $i = j$ and $i' = j'$. In the latter case, we obtain bound $O(d^{-1})$ since $s \geq 2$, and thus $\langle T^s_{ij}, T^s_{i'j'} \rangle = O(d^{-1})$. Otherwise, without loss of generality, assume that $i = i' = j'$ and $a = 1$. Here, Lemma B.18 (i) implies that the summand is $\prec 1/d$.

Case $N = 3$ If $i = j$, note that $\langle T^s_{ij}, T^s_{i'j'} \rangle \prec d^{-1}$ and we need to show that $\sum_{i \neq i' \neq j'} \langle w_{i'}, v_{j'} \rangle^a \prec d^2$. When $a \geq 2$ this follows from asymptotic orthogonality ($\langle w_{i'}, v_{j'} \rangle \prec d^{-1/2}$ for $i' \neq j'$, see Assumption 4.1), and otherwise from Lemma B.18 (i). If $i = i'$, we need to show that

$$\sum_{ijj'} \langle w_i, v_j \rangle^a \langle w_{i'}, v_{j'} \rangle^{a'} d^{-2} \prec \frac{1}{d}. \tag{148}$$

When $a \geq 2$ and $a' \geq 2$, this follows from asymptotic orthogonality. When $a = 1$ and $a' \geq 2$ (or vice versa), this follows from Lemma B.18 (i). Finally, when $a = 1$ and $a' = 1$, this follows from Lemma B.18 (ii). The remaining cases are identical to the covered ones.

Case $N = 4$ When $a \geq 2$ and $a' \geq 2$, the result follows trivially. When $a = 1$ and $a' \geq 2$, the result follows from Lemma B.18 (i). When $a = a' = 1$, the result follows again from Lemma B.18 (i) and noticing that

$$\frac{1}{d^2} \sum_{\substack{i \neq j \\ i' \neq j'}} \langle w_i, v_j \rangle \langle w_{i'}, v_{j'} \rangle = \left(\frac{1}{d} \sum_{i \neq j} \langle w_i, v_j \rangle \right)^2 \prec 1. \tag{149}$$

□

Next, using Theorem B.10 for $h(F_1, F_2) = f_2(F_1) g_2(F_2)$, we obtain that

$$\left| \mathbf{E}\, f_2(u^\top f_1(W^1 x)) g_2(z^\top g_1(V^1 x)) - \mathbf{E}\, f_2(G_1) g_2(G_2) \right| \prec d^{-1/2}, \tag{150}$$

where (G_1, G_2) is a jointly Gaussian random vector:

$$(G_1, G_2) \sim \mathcal{N}\left(0, \begin{pmatrix} \operatorname{Tr}\left[uu^\top \Omega_1\right] & \operatorname{Tr}\left[uz^\top \Phi_1\right] \\ \operatorname{Tr}\left[uz^\top \Phi_1\right] & \operatorname{Tr}\left[zz^\top \Psi_1\right] \end{pmatrix}\right). \tag{151}$$

Finally, using that $\|\Omega_1 - \Omega_1^{\text{lin}}\|_F \prec 1$ (same for Φ_1, Ψ_1) and perturbation analysis, we obtain that

$$\left| \mathbf{E}\, f_2(u^\top f_1(W^1 x)) g_2(z^\top g_1(V^1 x)) - \mathbf{E}\, f_2(N_1) g_2(N_2) \right| \prec d^{-1/2}. \tag{152}$$

B.6. Extending to $L \geq 3$

A natural question is to ask whether the same technique can be applied for a deeper networks. One possible direction is to apply Theorem B.10 for d-dimensional vector $(F_1, \dots F_d) \coloneqq (u_1^\top f_1(W^1 x), \dots, u_d^\top f_1(W^1 x))$, to approximate it by a Gaussian random vector $(N_1, \dots, N_d)$. Then, for example, the diagonal entries of Ω_3 can be written as $h(F_1, \dots, F_d) = f_3(\sum_k u_k f_2(F_k))^2$. If it is possible to derive that $f_3(\sum_k u_k f_2(F_k))^2 = f_3(\sum_k u_k f_2(N_k))$, then the problem is reduced to the 2 layered case, which can be treated as before.

However, it seems hard to apply Theorem B.10 to the d-dimensional vector, since this requires a much more careful error analysis. Recall that we only applied Theorem B.10 to 2-dimensional vectors. We leave the extension to $L, \widetilde{L} \geq 3$ as an interesting open question.

C. Details on numerics

C.1. Details of Fig. 1

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

$$\varphi_*(x) = \tanh\left(W_* x\right) \tag{153}$$

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

$$\tilde{C}_1 = \operatorname{diag}(\{k^{-0.3}\}_{1 \leq k \leq d}). \tag{154}$$

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

$$\varphi(x) = \tanh W_3(\tanh\left(W_2 \tanh(W_1 x)\right)) \tag{155}$$

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

$$W_1 = {}^1\!/_2 Z_1 \mathrm{diag}(\{k^{-\gamma/2}\}_{1\le k\le d}) + {}^1\!/_2 W_*, \tag{156}$$

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

$$W_3 = Z_3 C_3^{1/2} \tag{157}$$

where the covariance C_3 depends on the previous weights as

$$C_3 = (W_1 W_1^\top + {}^1\!/_2 \mathbb{I}_d)^{-1}. \tag{158}$$

C.2. MNIST Experiments

Data set We use the MNIST data set which we normalize by pixel-wise centering and global scaling to ensure unit variance. For each normalized image $x_i \in \mathbf{R}^{784}$ we define a label

$$y_i := \begin{cases} 1, & \text{if } x_i \text{ is an even digit,} \\ -1, & \text{if } x_i \text{ is an odd digit.} \end{cases}$$

We split the data set into four parts:

10% Test data I_{test}

25% Training data for the Adam optimizer I_{Adam},

25% Training data for regression I_{reg},

40% Data for approximating the (empirical) population covariance I_{emp}.

Neural network We then train a simple neural network of the form

$$x \mapsto \theta^\top \varphi(x), \quad \varphi(x) := \theta^\top \mathrm{relu}(W_2\,\mathrm{relu}(W_1 x)), \quad W_1 \in \mathbf{R}^{2352\times 784}, \quad W_2 \in \mathbf{R}^{2352\times 2352}, \quad \theta \in \mathbf{R}^{2352} \tag{159}$$

using the Adam optimizer over 120 epochs with a batch size of 128 using only the I_{Adam} split. During training we save the *feature maps* φ_t at various time steps t in order to study the training dynamics.

Feature ridge regression We then perform a ridge regression task using the features $\varphi_t(x_i)$ by minimizing

$$\theta(t, \lambda, I) := \arg\min_\theta \Big(\frac{1}{|I|} \sum_{i\in I} (y_i - \theta^\top \varphi_t(x_i))^2 + \lambda \|\theta\|^2 \Big) \tag{160}$$

for various random subsets $I \subset I_{\mathrm{reg}}$, and I_{test}, empirically estimate the *generalization error*

$$\mathcal{E}_{\mathrm{gen}}(t, \lambda, I)^2 := \frac{1}{|I_{\mathrm{test}}|} \sum_{i\in I_{\mathrm{test}}} (y_i - \theta(t, \lambda, I)^\top \varphi_t(x_i))^2. \tag{161}$$

Deterministic equivalent In order to compare $\mathcal{E}_{\text{gen}}(t)$ with the theoretical prediction from Theorem 3.1, we need to determine the covariance of the features ϕ_t as well as the label-feature covariance and the label variance. To do so, we use the I_{emp} part of the data to empirically estimate

$$\Omega_t := \frac{1}{|I_{\text{emp}}|} \sum_{i \in I_{\text{emp}}} \varphi_t(x_i)\varphi_t(x_i)^\top \in \mathbf{R}^{2352 \times 2352}, \quad \psi_t := \frac{1}{|I_{\text{emp}}|} \sum_{i \in I_{\text{emp}}} \varphi_t(x_i) y_i \in \mathbf{R}^{2352}, \quad \sigma^2 := \frac{1}{|I_{\text{emp}}|} \sum_{i \in I_{\text{emp}}} y_i^2 \in \mathbf{R} \tag{162}$$

and note that we expect this to be a reasonable approximation since $|I_{\text{emp}}| = 27805 \gg 2352$. Using these we have the formula

$$\mathcal{E}_{\text{gen}}^{\text{rmt}}(t, \lambda, n) := \frac{\sigma^2 - n\lambda m_t \psi_t^\top (M_t + \lambda M_t^2)\psi_t}{1 - n(m_t\lambda)^2 \operatorname{Tr} \Omega_t M_t \Omega_t M_t} \tag{163}$$

analogous to Eq. (12), where $m_t, M_t = m_t(\lambda, n), M_t(\lambda, n)$ are the solution to

$$\frac{1}{m_t(\lambda, n)} = \lambda + \operatorname{Tr} \Omega_t (1 + n m_t(\lambda, n)\Omega_t)^{-1}, \quad M_t(\lambda, n) := (\lambda + \lambda n m(\lambda, n)\Omega_t)^{-1}. \tag{164}$$

We observe in Figure 4 that $\mathcal{E}_{\text{gen}}^{\text{rmt}}$ is indeed an excellent approximation for $\mathcal{E}_{\text{gen}}$ throughout the training and for various choices of regularization. In Figure 5 we depict the interesting dynamics of the learning curves throughout the training process with a significant shift of the interpolation threshold to the left.

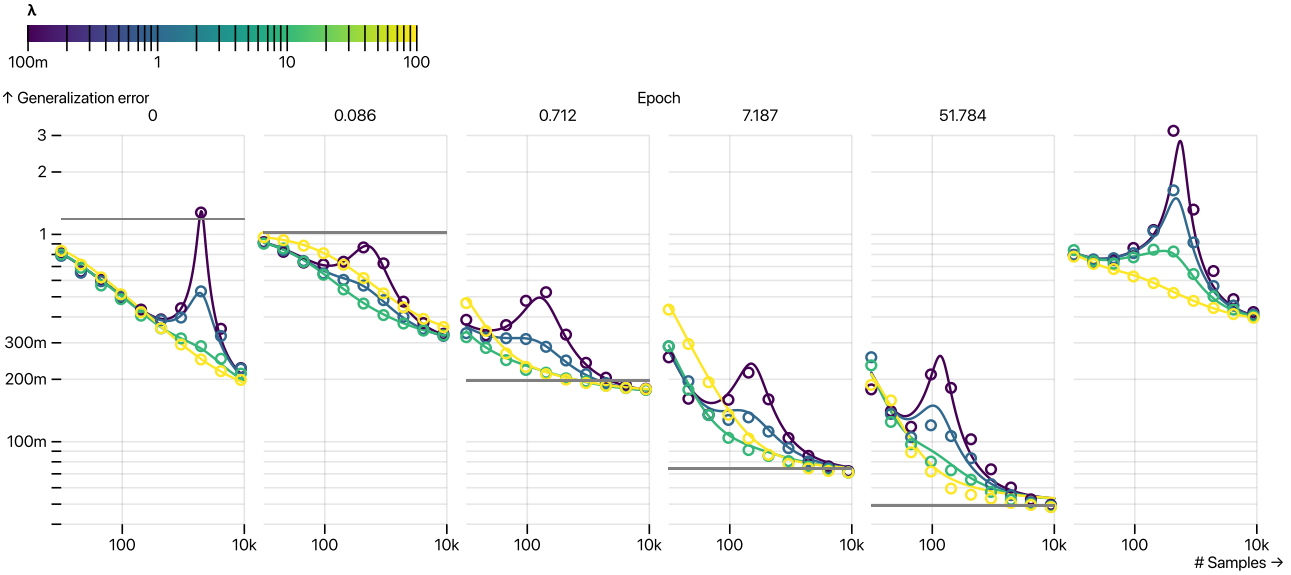

Figure 4. Plot of $\mathcal{E}_{\text{gen}}^{\text{rmt}}$, $\mathcal{E}_{\text{gen}}$ for various regularization parameters λ and time steps t (in "epoch.step" format). The horizontal lines represent the generalization error of the neural network, the curves $\mathcal{E}_{\text{gen}}^{\text{rmt}}$ and the dots $\mathcal{E}_{\text{gen}}$. The last pane contains a linear regression model for the sake of comparison. Interestingly, for this particular case already the random feature model outperforms linear regression.

Optimal regularization So far we have focused on fixed regularization parameters. Using the deterministic equivalent we can also find the optimal regularization parameter

$$\lambda_{\text{opt}}(t, n) := \arg\min_{\lambda} \mathcal{E}_{\text{gen}}^{\text{rmt}}(t, \lambda, n) \tag{165}$$

for each sample complexity n and time t by simply one-dimensional minimization. In Figure 6 we show the corresponding results. Interestingly ridge regression initially performs worse than the random feature regression also at optimal regularization. Then in the initial phase of training the performance of feature regression deteriorates before improving way beyond the initialization performance.

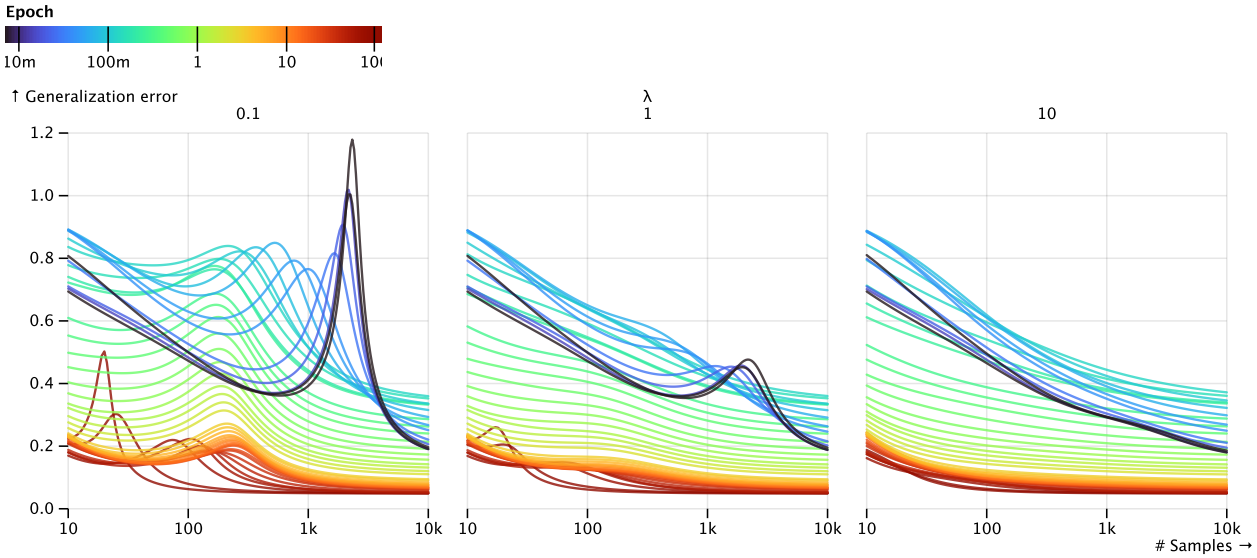

Figure 5. Dynamics of $\mathcal{E}_{\text{gen}}^{\text{rmt}}$ throughout the training

C.3. Synthetic MNIST experiments

We carried out similar experiments for synthetic data in order to empirically study the effect of population covariance linearization.

Data We generate Gaussian random vectors of zero mean and variance matching the variance of the normalized MNIST images described above. The synthetic labels are generated by a one hidden layer random feature network

$$\varphi_*(x) := \theta_* \tanh(W_* x), \quad W_* \in \mathbf{R}^{800\times 784}, \theta_* \in \mathbf{R}^{800}$$

for fixed but random W_*, θ_*.

Neural network We again train a simple neural network of the form

$$x \mapsto \theta^\top \varphi(x), \quad \varphi(x) := \theta^\top \operatorname{relu}(W_2 \operatorname{relu}(W_1 x)), \quad W_1 \in \mathbf{R}^{800\times 784}, \quad W_2 \in \mathbf{R}^{700\times 800}, \quad \theta \in \mathbf{R}^{700} \tag{166}$$

using the Adam optimizer over 50 epochs with a batch size of 128 using $40\,000$ samples. During training we save the *feature maps* φ_t at various time steps t in order to study the training dynamics.

Feature ridge regression We perform feature ridge regression on the trained features φ_t exactly as described above. The fact that the labels are now generated by a feature model now enables us to test the effect of population covariance linearization. In Figure 7 we observe that for random features the linearized deterministic equivalent is an excellent approximation for the empirically observed feature ridge regression error. However, during training the prediction deteriorates. We suspect that this effect is due to outlying eigevalues of the weight matrices which increasingly violate Assumption 4.1.

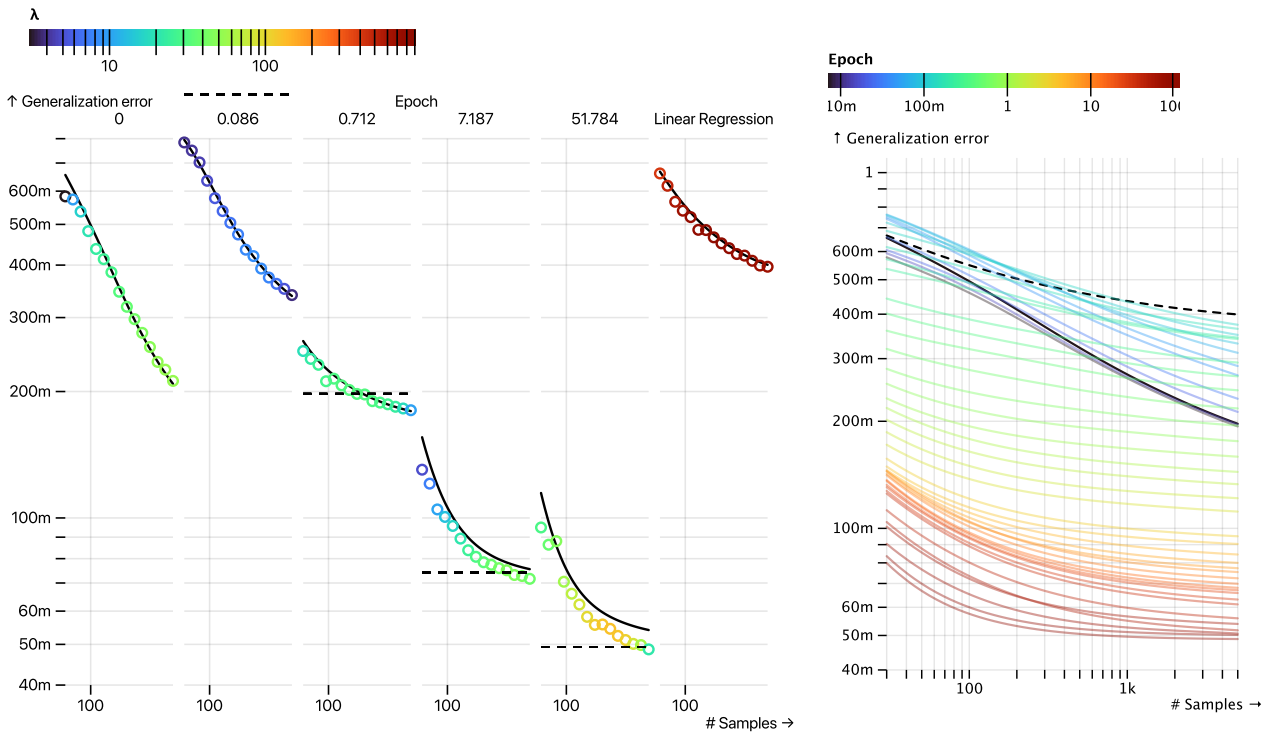

Figure 6. The left pane shows $\mathcal{E}_{\text{gen}}$ and $\mathcal{E}_{\text{gen}}^{\text{rmt}}$ throughout the training process at *optimal* regularization λ_{opt}. The colour of the dots encodes the value of λ_{opt}, while the dashed lines represent the generalization error of the neural network. The right pane shows the dynamics of $\mathcal{E}_{\text{gen}}^{\text{rmt}}$ throughout the training process, compared with linear regression.

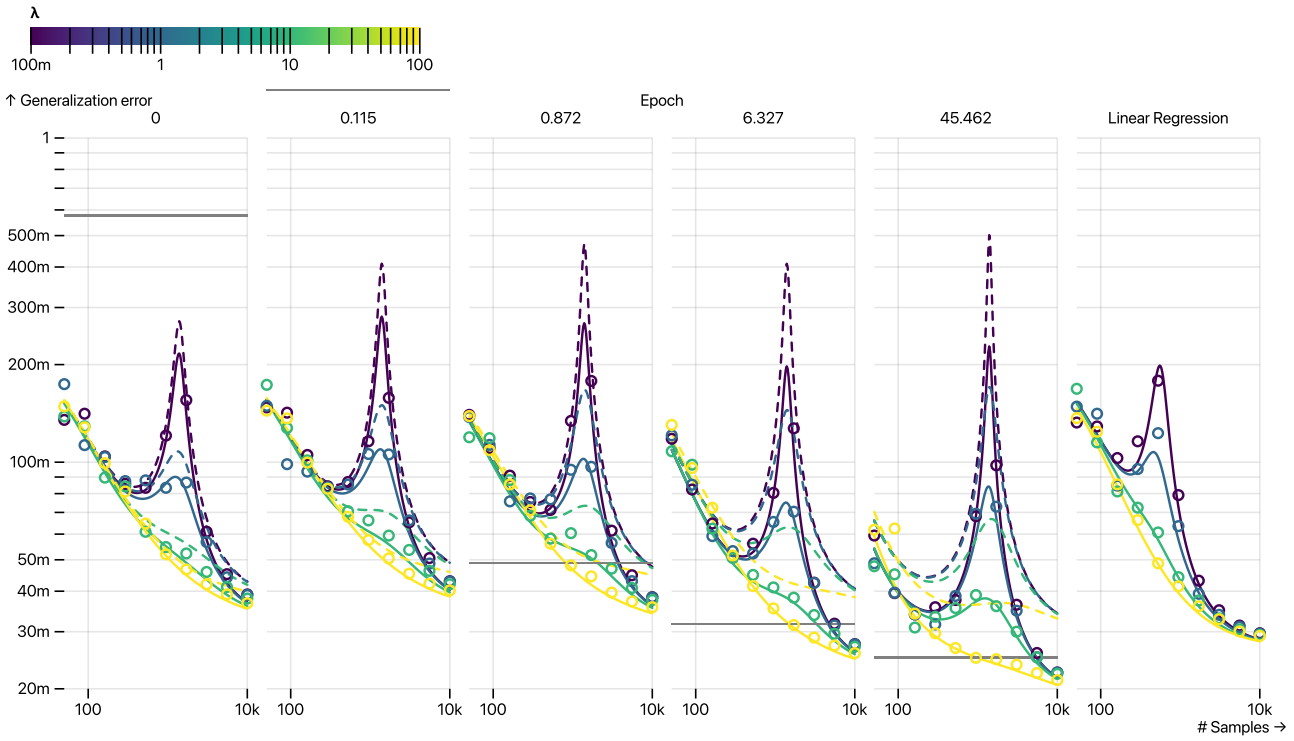

Figure 7. The solid and dashed lines represent $\mathcal{E}_{\text{gen}}^{\text{rmt}}$ using the empirical population covariances and the linearized population covariances, respectively. The dots represent the empirical $\mathcal{E}_{\text{gen}}$ while the horizontal line show the test error of the neural network during training. The right-most pane shows linear regression for comparison.