A Random Matrix Perspective on Mixtures of Nonlinearities in High Dimensions

Ben Adlam[*] [†]
Google Research

Jake Levinson[*] [†]
Google Research

Jeffrey Pennington[*]
Google Research

Abstract

One of the distinguishing characteristics of modern deep learning systems is their use of neural network architectures with enormous numbers of parameters, often in the millions and sometimes even in the billions. While this paradigm has inspired significant research on the properties of large networks, relatively little work has been devoted to the fact that these networks are often used to model large complex datasets, which may themselves contain millions or even billions of constraints. In this work, we focus on this high-dimensional regime in which both the dataset size and the number of features tend to infinity. We analyze the performance of random feature regression with features $F = f(WX + B)$ for a random weight matrix W and bias vector B, obtaining exact formulae for the asymptotic training and test errors for data generated by a linear teacher model. The role of the bias can be understood as parameterizing a distribution over activation functions, and our analysis directly generalizes to such distributions, even those not expressible with a traditional additive bias. Intriguingly, we find that a mixture of nonlinearities can improve both the training and test errors over the best single nonlinearity, suggesting that mixtures of nonlinearities might be useful for approximate kernel methods or neural network architecture design.

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

∗ Equal contribution. † Work done as a member of the Google AI Residency program (g.co/brainresidency).

1 INTRODUCTION

Our theoretical understanding of deep learning algorithms continues to lag behind their impressive practical successes (Krizhevsky et al., 2012; Hinton et al., 2012; Wu et al., 2016). One main challenge in building a fuller understanding stems from the fact that deep neural networks are complex nonlinear functions that employ millions or even billions of parameters (Shazeer et al., 2017). Traditional wisdom would suggest that to this parameter complexity corresponds to an optimization difficulty. Recent work, however, suggests that as the width of a network's hidden layers becomes large, the loss function simplifies and a theoretical analysis becomes tractable (Jacot et al., 2018; Chizat and Bach, 2018a; Mei et al., 2019; Rotskoff et al., 2019; Rotskoff and Vanden-Eijnden, 2018). In some scenarios, the simplification is such that throughout training the parameters of the model stay within an infinitesimal radius of their initial values, implying that much can be understood by studying the distribution over functions induced by the random initialization (Jacot et al., 2018; Chizat and Bach, 2018b; Lee et al., 2019).

Another challenge in understanding deep learning systems that they are often trained on very large, complex datasets. Even very large models, may not be large in comparison to the number of constraints they are designed to satisfy. Indeed, many important phenomena may become apparent only by examining the high-dimensional regime where the dataset size and width are both large and of the same order.

In this work, we focus on the high-dimensional regime and analyze the performance of a regression model trained on the random features $F = f(WX + B)$ for a random weight matrix W and bias vector B. We obtain an exact formula for the training error on a noisy autoencoding task and for the test error from fitting data labeled by a linear model in the limit that the width and dataset size both go to infinity. These results are determined by the resolvent of the kernel matrix $F^\top F$, whose properties we analyze via the *resolvent method* from random matrix theory. Our analysis also

provides an exact formula for the eigenvalue density of the kernel matrix, which may be of independent interest since it provides a characterization for how spectral properties of the data covariance matrix propagate through neural network layers at initialization.

1.1 Our contributions

The main contribution of our work is an exact characterization of the training error of a ridge-regularized random feature regression model on a noisy autoencoder task and the test error when the data are labeled by a linear-teacher model in the high-dimensional regime. This is one of the first non-trivial models to be solved exactly in the joint limit of large data and large width and provides an interesting testing ground in which to analyze this regime. Some additional contributions include:

- An exact characterization of the spectral density of the random feature matrix $F = f(WX + B)$, extending prior results of Pennington and Worah (2017) to non-Gaussian data distributions and to non-zero bias distributions.
- One interpretation of the random additive bias is that it induces a distribution of activation functions parameterized by B, i.e. $f(Z; B) := f(Z + B)$. Our analysis trivially extends to *any* distribution of activation functions $f(\cdot\,; B)$ parameterized by B.
- We show that there exists a non-trivial distribution over activation functions that outperforms the best possible single activation function in terms of memorization capacity and test error when fitting data labeled by a linear-teacher model.
- Our method of proof introduces a surrogate "linearization" of F, F^{lin}, that possesses the same spectral information as F. This linearization is likely to be of further interest and utility in analyzing neural networks in high dimensions.

1.2 Related Work

Neural networks have been studied from the perspective of high-dimensional statistics in a number of recent works. However, prior work has focused on the bias-free case. Pennington and Worah (2017) studied the spectrum of the activation matrix $f(WX)$ for i.i.d. Gaussian data and derived an analytic expression for the training error of a ridge-regularized random feature model trained on pure noise. Hastie et al. (2019) studied ridgeless interpolation in high-dimensions for linear features as well as nonlinear random features of i.i.d. Gaussian data. In Louart et al. (2018), a deterministic equivalent for the resolvent of the kernel matrix $F^\top F$ is derived, which allowed for a characterization of the asymptotic training and test performance of linear ridge regression of random feature models.

Other work has investigated learning dynamics and generalization in the high-dimensional regime (Liao and Couillet, 2018a; Lampinen and Ganguli, 2018; Advani and Ganguli, 2016; Advani et al., 2020) as well as the spectra of more complicated objects such as the Hessian (Pennington and Bahri, 2017) and Fisher information matrix (Pennington and Worah, 2018). From the mathematical perspective, random matrix theory provides natural tools for analyzing the behavior of neural networks in the high-dimensional regime (Silverstein and Bai, 1995). Liao and Couillet (2018b) examined spectra for data drawn from Gaussian mixture models; see also El Karoui (2010) on the spectra of random kernel matrices.

Since an initial preprint of this paper was made available online, several papers have studied the generalization properties of random feature regression (all without a bias vector) in the high-dimensional limit. For isotropic covariates without bias, a precise characterization of the test error and double descent was described by many authors, including (Mei and Montanari, 2019; d'Ascoli et al., 2020; Adlam and Pennington, 2020a; Ba et al., 2019; Dhifallah and Lu, 2020). The origin of these peaks was explained via a detailed bias-variance decomposition by (Adlam and Pennington, 2020b; Lin and Dobriban, 2021; d'Ascoli et al., 2020). More general data distributions were studied by Liao et al. (2020) for random Fourier features, and in a series of works based on the Gaussian equivalence conjecture (Goldt et al., 2020, 2021), including d'Ascoli et al. (2021); Gerace et al. (2020); Loureiro et al. (2021).

The equivalence of the spectra of the nonlinear random feature matrix F and a "linearized" version of F (identified in this work, and concurrently by Péché et al. (2019)) has become a crucial step in analyzing the high-dimensional asymptotics of random feature methods. Some reliance on surrogate linearized Gaussian equivalent models underlies all of the above works, but, to the best of our knowledge, this conjectured equivalence remains unproven, except in the cases of spherical data or weights (Mei and Montanari, 2019), isotropic Gaussian covariates (Hu and Lu, 2020), and Fourier features (i.e. a specific nonlinearity) (Liao et al., 2020).

Significance of mixtures. A primary motivation for studying random features is their connection to neural networks, which typically use affine (as opposed to linear) transformations. We consider the following basic question: *What is the benefit of the additive bias*

terms in neural networks? Biases are commonly motivated as learnable thresholds for activating individual neurons, and there are certain contrived problems that cannot be adequately solved without them.

However, such scenarios bear little resemblance to practical models or real-world data, and the learnability of thresholds would seem to be irrelevant for highly-overparameterized models, where parameters stay close to their initial values (Lee et al., 2019; Chizat and Bach, 2018b; Jacot et al., 2018). It is therefore natural to wonder whether the bias might affect performance via a different mechanism that would persist in the high-dimensional regime. We describe biases as a method of parameterizing a distribution over activation functions, and examine their effect on training and test performance in simple high-dimensional regression tasks.

Note that naively appending a ones vector to X does not shortcut the derivation. Unfortunately, this leads to biases that are the same order as the weights, and so the effect disappears in the large dataset limit. Moreover, this modification on the data violates the assumptions of previous work; one way or another a nontrivial extension of the bias-free case must be done. We emphasize that our results are essentially a proof-of-principle, and we do not claim to have a method for determining the optimal mixtures for practice. The latter question has been examined empirically in many works (e.g. Manessi and Rozza (2018); Hagg et al. (2017); Agostinelli et al. (2014)), which consistently find that mixtures can outperform single nonlinearities.

2 PRELIMINARIES

Consider a dataset $X \in \mathbb{R}^{n_0 \times m}$ and the random feature matrix,

$$F = f(WX; B),$$

generated by a single hidden-layer network with i.i.d. Gaussian weights $W \in \mathbb{R}^{n_1 \times n_0}$ ($W_{ak} \sim \mathcal{N}(0, \sigma_W^2/n_0)$), activation function f, and biases $B = b\mathbf{1}_m^\top \in \mathbb{R}^{n_1 \times m}$ (for $b \in \mathbb{R}^{n_1}$). We regard the second argument of f as parametrizing (continuously or discretely) an ensemble of activation functions. We refer to B (or b) as the *bias*, in reference to the important special case $f(WX + B)$ (additive bias).

We assume that $f(\cdot; b)$ is differentiable almost everywhere and $\mathbb{E}\,|f(N; b)|^k$ for $N \sim \mathcal{N}(0, \sigma)$ is finite for all $1 \le k \le 4$, $\sigma > 0$, and $b \in \text{support}(\mu_B)$. When μ_B is a single Dirac mass at location $b_0 \in \mathbb{R}$, the activation function can be written as $f(WX; B) = f(WX; b_0) = g(WX)$ for some single-argument function g (for an additive bias, $g(WX) = f(WX + b_0)$). When this is the case, we say the model has a *single activation function*, as opposed to a mixture or distribution of activation functions.

The quantities of interest for our investigation are the kernel matrix, $\frac{1}{n_1}F^\top F$, and its *resolvent*,

$$G(z) := \left(\tfrac{1}{n_1}F^\top F - zI\right)^{-1}. \tag{1}$$

As we review in Sec. 4, the optimal regression coefficients of a linear model on the random features F are a simple function of this resolvent.

The high-dimensional regime that we study is the one in which the dataset size m, feature dimensionality n_0, and hidden layer width n_1 all go to infinity at the same rate. In particular, as is standard in the random matrix literature, we assume that we can parameterize the limit in terms of the dataset size m in such a way that there exist two positive constants,

$$\phi := \lim_{m\to\infty} \frac{n_0(m)}{m} \quad \text{and} \quad \psi := \lim_{m\to\infty} \frac{n_0(m)}{n_1(m)}. \tag{2}$$

Note that the resolvent is a random matrix, but as m grows large, its normalized trace becomes a deterministic quantity. In the limit that $m \to \infty$, this quantity is known as the *Stieltjes transform*,

$$s(z) := \lim_{m\to\infty} \tfrac{1}{m}\operatorname{tr} G(z). \tag{3}$$

Together with an auxiliary transform $\tilde{s}(z)$, defined below, these deterministic quantities completely characterize the asymptotic training error of kernel ridge regression on a noisy autoencoder task in this high-dimensional regime.

The Stieltjes transform frequently arises in random matrix theory as a way to encode the spectra of matrices. In particular, if λ_i are the eigenvalues of $\frac{1}{n_1}F^\top F$ and the empirical distribution of eigenvalues converges in distribution to some deterministic limiting density as $m \to \infty$,

$$\frac{1}{m}\sum_{i=1}^{m} \delta_{\lambda_i} \to \mu(\lambda)d\lambda, \tag{4}$$

then (with appropriate technical assumptions), the limiting spectral density itself can be recovered from the Stieltjes transform $s(z)$ via the inversion formula,

$$\mu(\lambda) = \lim_{\epsilon\to 0+} \frac{s(\lambda - i\epsilon) - s(\lambda + i\epsilon)}{2\pi i\epsilon}. \tag{5}$$

The Stieltjes transform then substitutes convergence in distribution for pointwise convergence for all z such that $\Im z > 0$.

2.1 Methods for computing the Stieltjes transform

We briefly review two standard methods for computing the Stieltjes transform $s(z)$, the *resolvent method* and the *moments method*.

The resolvent method is an approach for computing the Stieltjes transform based on the application of the Schur complement formula to the resolvent itself (or to a closely-related block matrix). Intuitively, as the matrix size becomes large, the minors of the matrix are similar in distribution to the larger matrix, and, moreover, the Cauchy interlacing theorem guarantees that their Stieltjes transforms are close as well. This allows for the derivation of a self-consistent equation (SCE) in which the Stieltjes transform appears on the left-hand side as the trace of the resolvent, and on the right-hand side as the trace of one of its minors.

The moments method is more combinatorial in nature and involves expanding the resolvent for large z and computing the traces of each term,

$$s(z) = \lim_{m\to\infty} \tfrac{1}{m}\operatorname{tr} G(z) = -\lim_{m\to\infty} \sum_{k=0}^{\infty} \frac{1}{n_1^k} \frac{\operatorname{tr}(F^\top F)^k}{z^{k+1}} . \tag{6}$$

The traces themselves are expanded out as

$$\operatorname{tr}(F^\top F)^k = \sum F_{a_1\alpha_1} F_{a_1\alpha_2} \cdots F_{a_k\alpha_1} , \tag{7}$$

where the sum runs over matrix indices $a_1, \ldots, a_k, \alpha_1, \ldots \alpha_k$. The essence of the moment method involves analyzing the asymptotic contribution of each term in the sum based on its combinatorial type and the details of F, and resumming the results to obtain $s(z)$.

We refer the reader to Erdos and Yau (2017); Tao (2012) for more details about these methods and additional background on random matrix theory.

3 RESULT FOR STIELTJES TRANSFORM

3.1 Main theorem

We make the following assumptions on the data matrix X and bias vector b:

1. $\left|\frac{1}{n_0}\sum_a X_{a\alpha}X_{a\beta} - \delta_{\alpha\beta}\sigma_X^2\right| \le C n_0^{\epsilon-1/2}$ for some positive constants $C > 0$ and $\epsilon < 1/100$ uniformly in α and β for all $n_0 > N$ for some constant N;
2. the empirical eigenvalue distribution of $\frac{1}{n_0}X^\top X$ converges in distribution to μ_X;
3. the empirical bias distribution converges also, i.e. $\frac{1}{n_1}\sum_{a=1}^{n_1}\delta_{b_a} \to \mu_B$ in distribution.

Theorem 1. *Define* $\sigma_Z := \sigma_W\sigma_X$, *the resolvent* $G(z) := \left(\frac{1}{n_1}F^\top F - zI\right)^{-1}$, *and the Gaussian expectations*

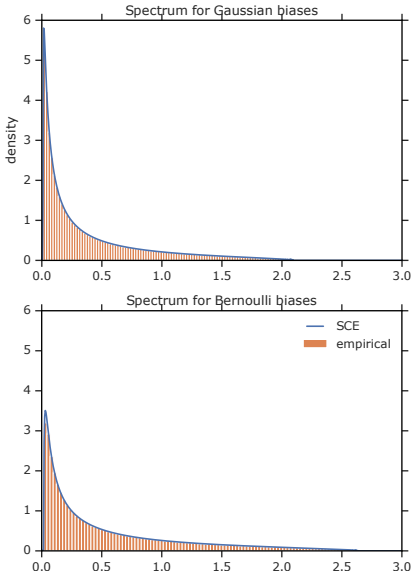

Figure 1: We get excellent agreement of theory and simulation for spectral densities for any bias distribution. We set $\phi = 1.5$, $\psi = 0.8$, $\sigma_X = \sigma_W = 1$, and $f = \mathrm{ReLU}$. Simulations are performed on matrices of size $m = 2^{14}$. **Top** Gaussian distribution over the biases, $\mathcal{N}(0,1)$. **Bottom** Bernoulli distribution over biases, Bern(0.5).

tions

$$\eta(b) := \mathbb{E}_{N\sim\mathcal{N}(0,\sigma_Z^2)}\left[f(N;b)^2\right] \tag{8}$$
$$\zeta(b) := \left(\mathbb{E}_{N\sim\mathcal{N}(0,\sigma_Z^2)}\left[Nf(N;b)/\sigma_Z\right]\right)^2 .$$

Then under the above assumptions and for all z such that $\Im z > 0$, the transforms

$$\tfrac{1}{m}\operatorname{tr} G(z) \quad \textit{and} \quad \tfrac{1}{m}\operatorname{tr}\left(\tfrac{1}{n_0}X^\top X G(z)\right) , \tag{9}$$

converge in probability to the unique solution, $s(z)$ and $\tilde{s}(z)$, of the Eq. (10) that map $\mathbb{C}^+$ to $\mathbb{C}^+$:

$$s(z) = \mathbb{E}_{S\sim\mu_X}\left[\frac{1}{C_0(z) + SC_1(z)}\right] \tag{10}$$
$$\tilde{s}(z) = \mathbb{E}_{S\sim\mu_X}\left[\frac{S}{C_0(z) + SC_1(z)}\right] ,$$

where

$$C_0(z) := -z + \mathbb{E}_{B\sim\mu_B}\left[\frac{\eta(B) - \zeta(B)}{D(B)}\right] , \tag{11}$$
$$C_1(z) := \mathbb{E}_{B\sim\mu_B}\left[\frac{\zeta(B)}{D(B)}\right] ,$$
$$D(b) := 1 + \frac{\psi}{\phi}\left(\zeta(b)\tilde{s}(z) + (\eta(b) - \zeta(b))\, s(z)\right) .$$

The proof is quite involved and is presented in the supplementary material. The basic idea is to derive

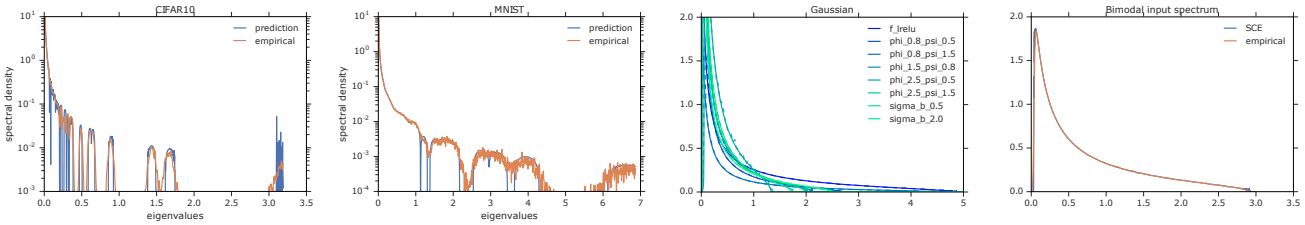

Figure 2: Empirical spectral densities agree with our predictions for varied data distributions and shape parameters. **Left**: One class from CIFAR (airplane), mean subtracted. **Center left**: Classes $\{0, 8\}$ from MNIST, mean subtracted. **Center right**: Gaussian input data, varying the NN parameter settings and activation function. **Right**: Input data with a bimodal spectrum. All plots used $f = \mathrm{ReLU}, \phi = 1.5, \psi = 0.8$ and $\sigma_W = \sigma_B = 1$, except for the indicated modified parameter. Empirical densities were smoothed using a Gaussian KDE.

a multivariate Gaussian random matrix model with the same correlation structure as F, then derive a self-consistent equation (SCE) using the resolvent method for this *linearized* version of F, denoted F^{lin}. As F^{lin} is potentially of independent interest, we write it here:

$$F^{\mathrm{lin}} := \mathcal{C}\Theta^1\Sigma + \left(\mathcal{V}^2 - \mathcal{C}^2\right)^{1/2}\Theta^2, \tag{12}$$

where (1) Θ^1 and Θ^2 are $n_1 \times m$ matrices with mean 0, variance $1/n_1$, i.i.d. Gaussian entries; (2) $\mathcal{V}$ and $\mathcal{C}$ are $n_1 \times n_1$ diagonal matrices with entries $\mathcal{V}_a^2 = \eta(b_a)$ and $\mathcal{C}_a^2 = \zeta(b_a)$; and (3) Σ is a matrix square root of $\frac{1}{n_0}X^\top X$. Note that while F and F^{lin} have the same asymptotic spectrum, the norm of their difference is large.

Remark 1. *The self-consistent equations consist of two* coupled *equations involving the Stieltjes transform $s(z)$ and an auxiliary object $\tilde{s}(z)$, (cf. Paul and Silverstein (2009, Eq. (2))), which we will see in Cor. 1 essentially measures the autoencoding capacity of the network.*

Remark 2. *Note that the self-consistent equations contain an expectation over the limiting spectral density of the input data. While the assumptions on the data matrix X in Thm. 1 are quite general, they may not be optimal. See Sec. 3.3, where we show strong agreement with empirical data from MNIST and CIFAR-10 and for a range of synthetic distributions. This suggests that the theorem may hold for even more general data distributions.*

3.2 Alternate representation and limiting results

When the data distribution is i.i.d. Gaussian, the expectations in Eq. (10) can be expressed in closed form, though one must be careful to choose the correct branch of the resulting function. For simplicity and future reference, we focus on the setting where $0 < \phi \leq \psi \leq 1$, in which case we have the coupled algebraic equations,

$$s(z) = \Big(C_1 - (C_0 + C_1)\phi + \sqrt{C_1^2 + 2(C_0 - C_1)C_1\phi + (C_0 + C_1)^2\phi^2}\Big) / \left(2C_0C_1\right), \tag{13}$$

$$\tilde{s}(z) = \frac{1 - C_0 s(z)}{C_1}. \tag{14}$$

When μ_B in Eq. (11) is trivial, i.e. a single Dirac mass, the result should reduce to the single activation function case with $F = f(WX)$, which was studied in Pennington and Worah (2017). Indeed, writing $\eta = \mathbb{E}_B[\eta(B)]$ and $\zeta = \mathbb{E}_B[\zeta(b)]$ for such a distribution, and eliminating $\tilde{s}(z)$ from Eqs. (13) and (14), we find that $s(z)$ satisfies the following quartic polynomial:

$$\begin{aligned} 0 = {} & (z^2\zeta^2\psi^2)s(z)^4 + (2z\zeta^2\psi(\psi - \phi)s(z)^3 \\ & + (\zeta^2(\psi - \phi)^2 + z\zeta\phi\psi + z\eta\phi^2\psi)s(z)^2 \\ & + (\zeta\phi(\psi - \phi) + \phi^2(z\phi + \eta(\psi - \phi))s(z) + \phi^3, \end{aligned} \tag{15}$$

which agrees with the result in Pennington and Worah (2017) upon identifying $s(z) = -(1 - \phi/\psi)/z - \phi/\psi G(z)$.

3.3 Spectral density estimates

The self-consistent equations in Thm. 1 can be solved numerically by iterating Eq. (10) until convergence, using numerical integration. By using the Stieltjes inversion formula, Eq. (5), we can extract predictions for the limiting eigenvalue density of $\frac{1}{n_1}F^\top F$. The results show close agreement with empirical spectral simulations from several interesting practical datasets and synthetic distributions, see Figs. 1 and 2.

4 STATISTICAL IMPLICATIONS

4.1 Memorization capacity for noisy autoencoders

First, we consider the problem of kernel ridge regression with random features given by $F = f(WX; B)$ and

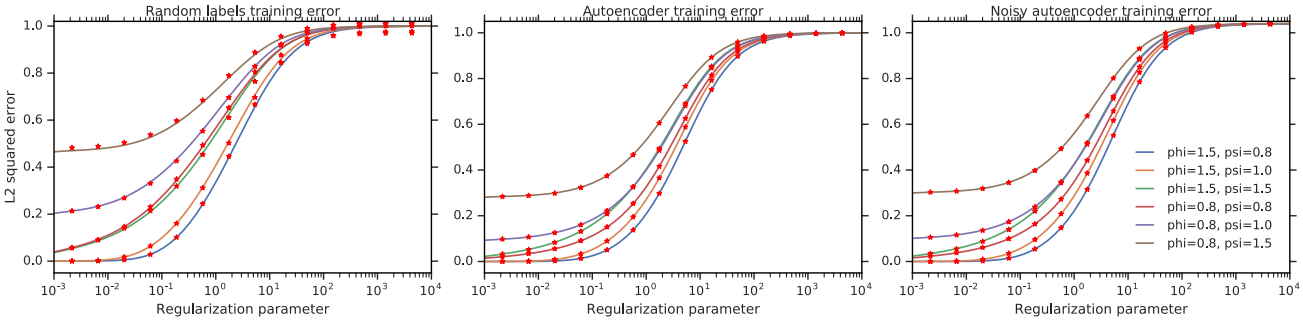

Figure 3: Comparisons of simulated ridge regression error and our theoretical prediction. We use ReLU with $\sigma_X = \sigma_W = \sigma_B = 1$ for all plots and vary the shape parameters ϕ and ψ. For simulations we use $m = 2^{13}$ throughout. Note we also normalize the activation function so that $\mathbb{E}_b[\eta(b)] = 1$. **Left**: Our predictions for ridge regression with random labels are solid lines. Simulated losses are the red stars. **Center**: Autoencoder error. **Right**: Noisy autoencoder error with $\sigma_\varepsilon = 0.2$.

noisy regression targets given by $Y = AX + \epsilon$ for some random $A \in \mathbb{R}^{n_2 \times n_0}$ and Gaussian noise $\epsilon \in \mathbb{R}^{n_2 \times m}$ such that $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ i.i.d. and independent of X, W, and B. As is common in the literature of high-dimensional statistics, we assume an isotropic prior on A such that $\mathbb{E}A^\top A = \frac{n_2}{n_0}\sigma_A^2 I$. Note that when $\sigma_\epsilon = 0$, we have the pure memorization setting studied by Pennington and Worah (2017). This problem provides a statistical interpretation for the role of the companion Stieltjes transform of Thm. 1. The training loss for this problem is a measure of the memorization capacity of the model, and we study the average training loss over an ensemble of problems defined by A and ϵ.

Corollary 1. *Let W_2^* be the minimizer for regularized training loss*

$$\mathcal{L}(W_2) = \tfrac{1}{n_2 m}||Y - W_2 F||^2 + \tfrac{n_1}{n_2 m}\gamma||W_2||^2\,, \qquad (16)$$

with random features $F = f(WX; B)$. Then, with the same assumptions as Thm. 1, the average training error, $E_{train} := \mathbb{E}_{A,\epsilon}\frac{1}{n_2 m}||Y - W_2^* F||^2$, converges to

$$-\gamma^2 \frac{d}{d\gamma}\left(\sigma_A^2 \tilde{s}(-\gamma) + \sigma_\epsilon^2 s(-\gamma)\right). \qquad (17)$$

In particular, we see that the derivative of the Stieltjes transform $s'(-\gamma)$ measures the capacity to learn noisy labels, whereas $\tilde{s}'(-\gamma)$ measures pure autoencoding capacity. See Fig. 3 for a comparison between these theoretical predictions and simulation. Under stronger assumptions on A, the training loss consider as a random variable (without taking expectation over A and ϵ) can be shown to converge in probability.

Proof. The optimal weights for the regularized loss are given by

$$W_2^* = \tfrac{1}{n_1} Y G(\gamma) F^\top \quad \text{for} \quad G(\gamma) = (\tfrac{1}{n_1} F^\top F + \gamma I)^{-1},$$

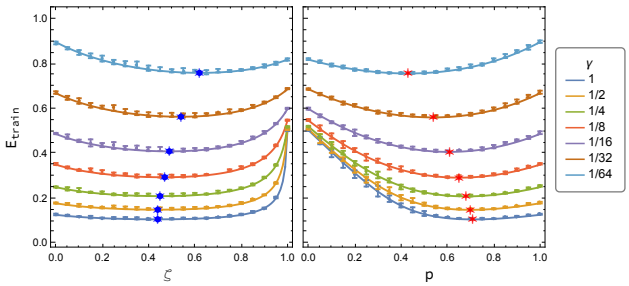

Figure 4: Performance on ridge-regularized noisy autoencoder with $\sigma_\epsilon = 1$, $\phi = 1/2$, and $\psi = 1/2$. Theoretical predictions for training error (solid lines) and 1σ error bars for empirical simulations of finite networks ($n_0 = 192$, $n_1 = 384$, $m = 384$) for various values of ridge regularization constant γ as the activation function varies. **Left:** a single activation function f is used. **Right:** the non-linearity is f_p, a Bernoulli(p)-mixture of a purely linear ($\zeta = 1$) and purely non-linear ($\zeta = 0$) function. Each simulation uses a randomly-chosen non-linearity having the specified values of ζ, demonstrating that E_{train} depends on the non-linearity solely through this constant. Red and blue stars denote minima.

resulting in training error

$$\begin{aligned} E_{\text{train}} &= \mathbb{E}_{A,\epsilon}\tfrac{1}{n_2 m}||Y - W_2^* F||^2 \qquad (18)\\ &= \gamma^2 \tfrac{1}{n_2 m}\mathbb{E}_{A,\epsilon}\text{tr}\left(Y^\top Y G(\gamma)^2\right)\\ &= \gamma^2\sigma_A^2 \tfrac{1}{n_0 m}\text{tr}\left(X^\top X G(\gamma)^2\right) + \gamma^2\sigma_\epsilon^2 \tfrac{1}{m}\text{tr}\left(G(\gamma)^2\right)\\ &\to \gamma^2\sigma_A^2 \tfrac{d}{d\gamma}\left(\tilde{s}(-\gamma)\right) + \gamma^2\sigma_\epsilon^2 \tfrac{d}{d\gamma}\left(s(-\gamma)\right)\\ &= -\gamma^2\left(\sigma_A^2 \tilde{s}'(-\gamma) + \sigma_\epsilon^2 s'(-\gamma)\right). \end{aligned}$$

□

Remark 3. *As in Pennington and Worah (2017), there is a scaling homogeneity in the E_{train}: an increase in the regularization constant γ can be compensated by a decrease in scale of W_2, which, in turn, can be compensated by increasing the scale of F, which is*

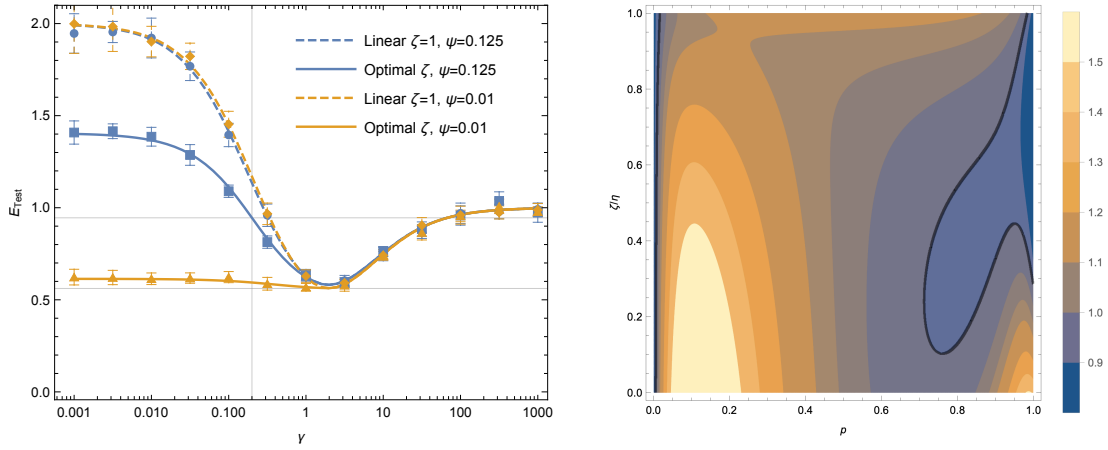

Figure 5: We plot E_{test} for random feature regression when the data are labeled by a linear model with $\phi = 1/2$ and $\sigma_\epsilon^2 = 2$. **Left**: Plots of E_{test} as a function of γ for both the linear activation function (dashed) and an optimal nonlinear activation function (solid). When $\psi = 1/100$ (orange) tuning ζ can almost overcome a too small γ and recover optimal E_{test}. When $\psi = 1/8$ (blue) tuning ζ reduces E_{test} over $\zeta = 1$, but the gap to optimal E_{test} is large. The dots and errors bars (± 1 s.d.) show simulations results for $n_0 = 1000$ over 10 trials. **Right**: Keeping $\psi = 1/8$, we show a contour plot of E_{test} for the binary mixture Eq. (21). We outline the region where the mixture improves over all single nonlinearities with a grey line. The best setting found in our grid search is on the far right of the plot at $p = 0.999$ and $\zeta(1) \approx 0.364$. Simulating the binary mixture for $n_0 = 1000$ over 10 trials estimated E_{test} as $0.785\,(\pm 0.035)$.

equivalent to increasing $\eta(b)$ and $\zeta(b)$. Owing to this homogeneity, we are free to choose a normalization of the activation function for which $\mathbb{E}_{b\sim\mu_B}[\eta(b)] = 1$.

4.2 Nonlinear mixtures can generalize better than single nonlinearities

The bias term in our random feature model can be viewed as one way of defining a distribution over activation functions. The choice of distribution, in general, affects the performance of the model on a given task—as quantified by the expectations in Eq. (11).

Using Cor.1, one can show that mixtures of activation function can improve the model's capacity over the best single nonlinearity (see Sec. E). Fascinatingly, this benefit of mixtures is not restricted to the training loss but extends to the test loss. Specifically, we define a data generating process where $\mathbf{x}_i$ are m i.i.d. samples from $\mathcal{N}(0, I_{n_0})$, and we label these points using a linear teacher such that $y_i = \beta^\top \mathbf{x}_i + \epsilon_i$, where β_i and ϵ_i are i.i.d. Gaussians of zero mean and variances $1/\sqrt{n_0}$ and σ_ϵ^2 respectively. We collect the data points $\mathbf{x}_i$ into the columns of a data matrix X and labels y_i into Y. Interestingly, in this high-dimensional limit the conditional distribution of a linear labeling function is asymptotically equivalent to a wide class of non-linear teacher neural networks (see Mei and Montanari (2019); Adlam and Pennington (2020a) for more details). Under our assumption on X, standard results from random matrix theory imply convergence of the spectrum of $X^\top X/n_0$ in distribution (Marčenko and Pastur, 1967; Silverstein and Bai, 1995). Moreover, Assumption 1. in Sec. 3.1 is easily verified to hold with for probability converging to 1.

Without loss of generality, we may assume that $\mathbb{E}\eta(b) = 1$, since any rescaling of the activation functions $f(\cdot; b)$ can be absorbed into W_2^*.

The average test loss for this model is defined as

$$E_{\text{test}} := \mathbb{E}_{\beta,\epsilon}\mathbb{E}_{\mathbf{x}}\left(\beta^\top \mathbf{x} - W_2^* f(W\mathbf{x}; B)\right)^2. \tag{19}$$

Using the results of Thm. 1 and additional calculations (that we defer to the appendix, see F), the test error can be characterized analytically.

Focusing first on the case of a single nonlinearity, i.e. when μ_B is a delta mass at 0, we can use the expression for the test error to find the optimal hyperparameters to minimize the test error. Unsurprisingly given the data generating process, optimal performance is achieved by effectively performing linear regression with regularization to match the SNR. This can be obtained from the random feature model in the limit that $n_1 \to \infty$ (so $\psi \to 0$), which has the effect of removing the randomness from W in the random feature kernel. More interestingly, the regularization in the model can be achieved with either the ridge parameter γ or the acti-

vation function. Any configuration satisfying

$$\frac{\gamma - \eta(0) + \zeta(0)}{\zeta(0)} = \sigma_\epsilon^2 \tag{20}$$

is optimal. When the activation is linear, $\zeta(0) = \eta(0) = 1$, and so $\gamma_{\text{opt}} = \sigma_\epsilon^2$, just as in a well-specified linear regression model.

However for a fixed nonlinear activation function, $1 = \eta(0) > \zeta(0)$, so Eq. (20) implies a smaller γ is optimal—suggesting an implicit regularizing effect of the nonlineaity. Strangely, the optimal γ can sometimes be negative in Eq. (20) to counteract over-regularization by the nonlinearity. Taking the opposite perspective and fixing γ, optimal performance is not always possible, as Eq. (20) is not necessarily satisfiable since we require $\zeta \in [0, 1]$. That said, there are situations where appropriately choosing the nonlinearity can completely compensate for suboptimal γ.

When $\psi > 0$, while the nonlinearity can still reduce the degradation in performance, a single activation function is no longer able to completely compensate for suboptimal γ. In such situations a mixture of activation functions can help compensate further. The goal here is not to identify "good" mixtures, since this will clearly be a dataset- and architecture-dependent question. Instead, we merely seek to demonstrate a proof-of-principle, namely that there exist non-trivial distributions over nonlinearities that can *provably* outperform the best possible single nonlinearity. For this analysis, we consider a simple but nontrivial distribution over activation functions: a Bernoulli mixture of two different functions. In more detail, $\mu_B = \text{Bernoulli(p)}$ and

$$f(z; b) := \begin{cases} \frac{x}{\sqrt{2-2p}} & \text{if } b = 0 \\ g(x) & \text{if } b = 1 \end{cases}, \tag{21}$$

where g is such that $\eta(1) = \mathbb{E}g(N)^2 = 1/2p$. Note that this implies $\mathbb{E}\eta(B) = 1$. We will optimize over $\zeta(1) = (\mathbb{E}\sigma_Z f'(N))^2$ and p.

To give a concrete example, we set $\phi = 1/2$, $\psi = 1/8$, $\sigma_\epsilon^2 = 2$, and $\gamma = 2/10$. In this setting, a single activation function is unable to achieve optimal performance (see Fig. 5). Specifically, the optimal test error of approximately 0.945 is achieved at $\zeta \approx 0.634$, which falls significantly short of the optimal test error of approximately 0.581 for $\gamma \approx 2.0$ and a linear activation function.

The mixture model from Eq. (21) can significantly reduce this gap. By performing a simple grid search over p and $\zeta(1)$, we find that $p = 0.999$ and $\zeta(1) \approx 0.364$ yields a test error of approximately 0.783. We illustrate these conclusions in Fig. 5, and provide empirical results from simulations confirming our findings.

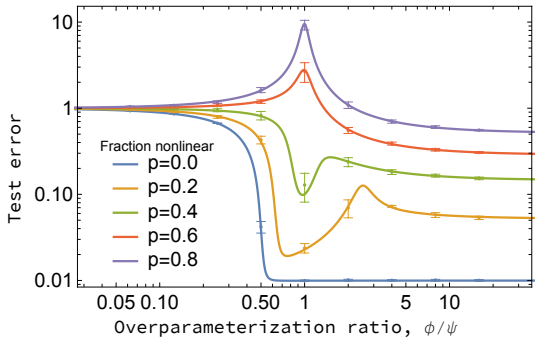

Figure 6: Theoretical predictions (solid lines) and empirical simulations with $n_0 = 128$ (markers) for the test error as a function of the overparameterization ratio $\phi/\psi = n_1/m$ for various mixtures of nonlinearities. Here $\phi = 1/2$, $\gamma = 10^{-3}$, $\sigma_\varepsilon^2 = 0.01$, and the nonlinearity is f_p, a Bernoulli(p)-mixture of a purely linear ($\zeta = 1$) and purely non-linear ($\zeta = 0$) function. As the fraction of nonlinearity p decreases, the double descent peak moves to the right and eventually disappears.

4.3 Nonlinear mixtures can shift the double descent peak

The setting of (bias-free) random feature ridge regression has proved a useful testing ground for studying the effect of overparameterization and the phenomenon of *double descent* (Belkin et al., 2019; Mei and Montanari, 2019; Adlam and Pennington, 2020a,b), whereby the test error first decreases, then increases to a peak, then decreases again as the number of parameters is increased. The origin of this peak has been investigated from various perspectives, and in simple random feature regression scenarios it seems to stem from the variance coming from the interaction between the random weights W and the data XAdlam and Pennington (2020b) and occurs when the number of random features equals the number of samples, i.e. $n_1 = m$.

In Fig. 6, we observe that double descent persists in the presence of nonlinear mixtures. Intriguingly, we find that the location of the peak depends on the mixture, shifting from $n_1 = m$ for a single nonlinearity to larger values of n_1 as the mixture becomes more linear, and eventually disappearing enitrely in the purely linear case.

5 CONCLUSIONS

In this work we studied the feature matrix $F = f(WX; B)$ where W is a random matrix with i.i.d. Gaussian entries. Under mild assumptions on X and B, we obtained an exact analytic formula, Eq. (10), that characterizes the Stieltjes transform of the spectral density of F. The result allowed us to describe the

exact training loss of a ridge-regularized noisy autoencoder and the test error from fitting data labeled by a linear model in the high-dimensional limit, providing one of the first closed-form solutions to a non-trivial model. We found excellent agreement between the asymptotic predictions of Eq. (10) and a variety of finite-dimensional empirical simulations.

We also advanced the interpretation of the bias B as one particular way of parameterizing a distribution of activation functions. Indeed, our derivations proceed completely unchanged whether this distribution is of the traditional additive form $f(\cdot + B)$ or the more general $f(\cdot; B)$. By examining the latter, we showed that there are configurations in which a non-trivial distribution over activation functions provably outperforms the best possible single activation function. This opens the door to future investigations regarding optimal methods for parameterizing distributions over activation functions for approximate kernel methods, and suggests the possibility that mixtures of nonlinearities could be a useful design consideration when constructing neural network architectures.

References

B. Adlam and J. Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR, 2020a.

B. Adlam and J. Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11022–11032. Curran Associates, Inc., 2020b. URL `https://proceedings.neurips.cc/paper/2020/file/7d420e2b2939762031eed0447a9be19f-Paper.pdf`.

M. Advani and S. Ganguli. Statistical mechanics of optimal convex inference in high dimensions. *Phys. Rev. X*, 6:031034, Aug 2016. doi: 10.1103/PhysRevX.6.031034. URL `https://link.aps.org/doi/10.1103/PhysRevX.6.031034`.

M. S. Advani, A. M. Saxe, and H. Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020. doi: 10.1016/j.neunet.2020.08.022. URL `https://doi.org/10.1016/j.neunet.2020.08.022`.

F. Agostinelli, M. Hoffman, P. Sadowski, and P. Baldi. Learning activation functions to improve deep neural networks. *arXiv preprint arXiv:1412.6830*, 2014.

J. Ba, M. Erdogdu, T. Suzuki, D. Wu, and T. Zhang. Generalization of two-layer neural networks: An asymptotic viewpoint. In *International conference on learning representations*, 2019.

Z. Bai and W. Zhou. Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, pages 425–442, 2008.

M. Banna and F. Merlevede. Limiting spectral distribution of large sample covariance matrices associated with a class of stationary processes. *Journal of Theoretical Probability*, 28(2):745–783, 2015.

M. Banna, F. Merlevède, and M. Peligrad. On the limiting spectral distribution for a large class of symmetric random matrices with correlated entries. *Stochastic Processes and their Applications*, 125(7):2700–2726, 2015.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Z. Burda, J. Jurkiewicz, and B. Wacław. Spectral moments of correlated wishart matrices. *Physical Review E*, 71(2):026111, 2005.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018a.

L. Chizat and F. Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018b.

S. d'Ascoli, L. Sagun, and G. Biroli. Triple descent and the two kinds of overfitting: Where & why do they appear? *Advances in Neural Information Processing Systems*, 33:3058–3069, 2020.

S. d'Ascoli, M. Gabrié, L. Sagun, and G. Biroli. On the interplay between data structure and loss function in classification problems. *Advances in Neural Information Processing Systems*, 34, 2021.

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

S. d'Ascoli, M. Refinetti, G. Biroli, and F. Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR, 2020.

N. El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38, 01 2010. doi: 10.1214/08-AOS648.

N. El Karoui et al. Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *Annals of Applied Probability*, 19(6):2362–2405, 2009.

L. Erdos. The matrix dyson equation and its applications for random matrices. *arXiv preprint arXiv:1903.10060*, 2019.

L. Erdos and H.-T. Yau. A dynamical approach to random matrix theory. *Courant Lecture Notes in Mathematics*, 28, 2017.

F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová. Generalisation error in learning with random features and the hidden manifold model, 2020.

S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mézard, and L. Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. *Proceedings of Machine Learning Research vol*, 145:1–46, 2021.

G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

A. Hagg, M. Mensing, and A. Asteroth. Evolving parsimonious networks by mixing activation functions. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 425–432, 2017.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

J. W. Helton, S. A. McCullough, and V. Vinnikov. Noncommutative convexity arises from linear matrix inequalities. *Journal of Functional Analysis*, 240: 105–191, 2006.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

H. Hu and Y. M. Lu. Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669*, 2020.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 8580–8589, USA, 2018. Curran Associates Inc. URL `http://dl.acm.org/citation.cfm?id=3327757.3327948`.

A. Jacot, B. Simsek, F. Spadaro, C. Hongler, and F. Gabriel. Kernel alignment risk estimator: Risk prediction from training data. *Advances in Neural Information Processing Systems*, 33:15568–15578, 2020.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

A. K. Lampinen and S. Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. In *International Conference on Learning Representations*, 2018.

J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Z. Liao and R. Couillet. The dynamics of learning: A random matrix approach. In *35th International Conference on Machine Learning, PMLR*, 2018a.

Z. Liao and R. Couillet. On the Spectrum of Random Features Maps of High Dimensional Data. In *International Conference on Machine Learning (ICML 2018)*, Stockholm, Sweden, July 2018b. URL `https://hal.archives-ouvertes.fr/hal-01954933`. 13 pages (with Supplementary Material), 10 figure, ICML 2018.

Z. Liao, R. Couillet, and M. W. Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. *Advances in Neural Information Processing Systems*, 33:13939–13950, 2020.

L. Lin and E. Dobriban. What causes the test error? going beyond bias-variance via anova. *Journal of Machine Learning Research*, 22(155):1–82, 2021.

Z. Lixin. Spectral analysis of large dimentional random matrices. 2007.

C. Louart, Z. Liao, R. Couillet, et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

B. Loureiro, C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mézard, and L. Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model, 2021.

F. Manessi and A. Rozza. Learning combinations of activation functions. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 61–66. IEEE Computer Society, 2018.

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

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

S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

J. A. Mingo and R. Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

D. Paul and A. Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.

D. Paul and J. W. Silverstein. No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *Journal of Multivariate Analysis*, 100(1):37–57, 2009.

S. Péché et al. A note on the pennington-worah distribution. *Electronic Communications in Probability*, 24, 2019.

J. Pennington and Y. Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning*, pages 2798–2806, 2017.

J. Pennington and P. Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2017.

J. Pennington and P. Worah. The spectrum of the fisher information matrix of a single-hidden-layer neural network. In *Advances in Neural Information Processing Systems*, pages 5410–5419, 2018.

G. Rotskoff, S. Jelassi, J. Bruna, and E. Vanden-Eijnden. Global convergence of neuron birth-death dynamics. In *International Conference on Machine Learning*, 2019.

G. M. Rotskoff and E. Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv preprint arXiv:1805.00915*, 2018.

N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural language models using sparsely gated mixtures of experts. *ICLR*, 2017. URL `http://arxiv.org/abs/1701.06538`.

J. W. Silverstein and Z. Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192, 1995.

T. Tao. *Topics in random matrix theory*, volume 132. American Mathematical Society Providence, RI, 2012.

N. Tripuraneni, B. Adlam, and J. Pennington. Covariate shift in high-dimensional random feature regression. *arXiv preprint arXiv:2111.08234*, 2021a.

N. Tripuraneni, B. Adlam, and J. Pennington. Overparameterization improves robustness to covariate shift in high dimensions. *Advances in Neural Information Processing Systems*, 34, 2021b.

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

P. Yaskov. The universality principle for spectral distributions of sample covariance matrices. *arXiv preprint arXiv:1410.5190*, 2014.

Supplementary Materials:
A Random Matrix Perspective on Mixtures of Nonlinearities in High Dimensions

A OUTLINE FOR THE PROOF

The main goal of the supplementary material is to prove Thm. 1. This is achieved in three steps. The first step is to derive the leading order correlation structure (or expected kernel) for F (Lem. 2), and specify a multivariate Gaussian model with the same correlation structure (Subsec. B.1). We refer to this Gaussian model as a *linearized model*, denoted F_{lin}, since it removes the nonlinearity f from the random part of our random matrix ensemble. The matrix F is interesting as the data induce covariances among columns and the biases among rows, and both these must be captured in the linearized model. Random matrices with covariances of this type have been studied before when they can be written as $A^{1/2}XBXA^{1/2}$ for X with i.i.d. entries and A and B nonnegative definite Hermitian (Lixin, 2007; Burda et al., 2005; Paul and Silverstein, 2009; El Karoui et al., 2009), and it is known they lead to coupled, functional self-consistent equations (Paul and Aue, 2014) such as Eq. (10).

Since F is not of the form $A^{1/2}XBXA^{1/2}$, the second step is to show that the limiting Stieltjes transform for the kernel matrix $F^\top F$ is equivalent to a multivariate Gaussian random matrix model with an identical correlation structure, i.e. equivalent to F_{lin}. Results of this form are common in random matrix theory, where under correlation assumptions on the entries the matrix Dyson equation determines the limiting behavior of the resolvent (Erdos, 2019). Moreover, this deterministic equation only depends on the second order moments through a super operator. Thus, under some assumptions on the correlation structure, random matrices with the same second order moments have the same asymptotic spectra.

These assumptions on the correlation structure come in many forms: Banna and Merlevede (2015); Banna et al. (2015) have very general conditions when the matrix entries can be written as functions of an i.i.d. random field. Cor. 1.1 of Bai and Zhou (2008) gives conditions based on the concentration of quadratic forms that imply the limiting Stieltjes transform only depends on $t_{\alpha\beta} := \mathbb{E}F_{a\alpha}F_{a\beta}$.[1] While this result requires that the rows of the matrix are i.i.d., the argument has been generalized from the identically distributed setting in Yaskov (2014). Again the condition is based on the concentration of quadratic forms, see Remark 1 and Assumptions (A3) and (A4) of Yaskov (2014).

The more general result of Yaskov (2014) is required for the following reason: Either B can be consider as random or it can be conditioned on in all expectations. If B is taken to be random and b_a are drawn i.i.d. from μ_B, then the rows of F are also i.i.d., so the setting of Bai and Zhou (2008) applies. However, the concentration of the quadratic forms fails for $t_{\alpha\beta} := \mathbb{E}_B\mathbb{E}_W[F_{i\alpha}F_{i\beta}|B]$. Instead all expectations must be considered conditional on B, and since B is independent of W, it is sufficient to take B deterministic such that $\frac{1}{n_1}\sum_{a=1}^{n_1}\delta_{b_a} \to \mu_B$. Defining $t^{(a)}_{\alpha\beta} := \mathbb{E}[F_{a\alpha}F_{a\beta}|B]$, one can show concentration of the required quadratic forms, but the rows of F are no longer identically distributed and Bai and Zhou (2008) does not apply. To this end, we consider all expectations in the proof to be taken over W and be conditional on B and X, i.e. $\mathbb{E}[\cdot] \equiv \mathbb{E}[\cdot|B, X]$.

The third a final step of the proof is a derivation of a self-consistent equation specifying the limiting Stieltjes transform of the Gaussian model. We do this using the resolvent method (see Erdos and Yau (2017) for an introduction). While the application of this method to multivariate Gaussian covariance matrices is standard, we note the form of the SCE is new and that we include the derivation here for completeness.

B CORRELATION STRUCTURE OF F

To begin we derive an asymptotic form for correlation structure of F. For simplicity of presentation, we derive the results for $f(\cdot, B) \equiv f(\cdot + B)$, but the argument generalizes directly. We recall that W is an $n_1 \times n_0$ random matrix with mean 0, variance σ_W^2/n_0, i.i.d. Gaussian entries. Define $c_{\alpha\beta} := \sum_k X_{k\alpha}X_{k\beta}/n_0$. Recall that X is assumed to satisfy

$$c_{\alpha\beta} - \delta_{\alpha\beta}\sigma_X^2 = \mathcal{O}\left(n_0^{\varepsilon-1/2}\right) \tag{S1}$$

for all α and β.

[1] Note this cannot depend on a!

Observe that

$$Z_{a\alpha} := \sum_{k=1}^{n_0} W_{ak} X_{k\alpha} \text{ and } Z_{b\beta} := \sum_{k=1}^{n_0} W_{bk} X_{k\beta} \tag{S2}$$

are jointly Gaussian. Moreover,

$$\mathbb{E} Z_{a\alpha} = 0 \quad \text{and} \quad \mathbb{E} Z_{a\alpha} Z_{b\beta} = \sigma_W^2 c_{\alpha\beta} \mathbf{1}(a = b). \tag{S3}$$

Note in particular, that $Z_{a\alpha}$ and $Z_{b\beta}$ are independent if $a \neq b$. For convenience, we normalize $c_{\alpha\beta}$ and define $\tilde{c}_{\alpha\beta} := c_{\alpha\beta}/\sigma_X^2$ and $\varepsilon_\alpha := \tilde{c}_{\alpha\alpha} - 1$, and note $\tilde{c}_{\alpha\beta} = \mathcal{O}(n^{\varepsilon - 1/2})$ for $\alpha \neq \beta$ and $\varepsilon_\alpha = \mathcal{O}(n^{\varepsilon - 1/2})$ for all α.

In order to specify the correlations, we employ transforms of the activation function f. The transforms are Gaussian integrals of f at different locations and scales. Let $N \sim \mathcal{N}(0,1)$ and recall $\sigma_Z = \sigma_X \sigma_W$, then define

$$\xi_0(x) := \mathbb{E}\left[f\left(\sigma_Z N + x\right)\right], \quad \xi_1(x) := \mathbb{E}\left[N f\left(\sigma_Z N + x\right)\right], \quad \xi_2(x) := \mathbb{E}\left[\frac{(\sigma_Z^2 - 1)N^2}{2\sigma_Z^2} f\left(\sigma_Z N + x\right)\right], \tag{S4}$$

$$\eta(x) \equiv \eta_0(x) := \mathbb{E}\left[f(\sigma_Z N + x)^2\right], \quad \eta_2(x) := \mathbb{E}\left[\frac{(\sigma_Z^2 - 1)N^2}{2\sigma_Z^2} f\left(\sigma_Z N + x\right)^2\right], \quad \text{and} \quad \zeta(x) := \xi_1(x)^2 \tag{S5}$$

Lemma 1. *Suppose ε is a small constant (i.e. $|\varepsilon| < 1/10$) and that $N \sim \mathcal{N}(0,1)$, then*

$$\mathbb{E}_N f\left(\sigma_Z \sqrt{1+\varepsilon} N + b\right) = \xi_0(b) + \xi_2(b)\varepsilon + \mathcal{O}(\varepsilon^2) \tag{S6}$$

and

$$\mathbb{E}_N f\left(\sigma_Z \sqrt{1+\varepsilon} N + b\right)^2 = \eta(b) + \eta_2(b)\varepsilon + \mathcal{O}(\varepsilon^2) \tag{S7}$$

Proof. A natural approach to proving Lem. 1 is to Taylor expand f point-wise at $\sigma_Z N + b$ in the small quantity $\sigma_Z\left(\sqrt{1+\varepsilon} - 1\right) N + b$, and then take expectation over N. However, this argument requires additional regularity assumptions on f. Instead, we can write

$$\mathbb{E}_N f\left(\sigma_Z \sqrt{1+\varepsilon} N + b\right) = \int_{\mathbb{R}} f(z) \phi_\varepsilon(z) dz, \tag{S8}$$

where ϕ_ε is the p.d.f. of $\mathcal{N}(b, \sigma_Z^2(1+\varepsilon))$, and then Taylor expand ϕ_ε in ε about 0. Note when $\varepsilon = 0$, ϕ_0 is the p.d.f. of $\mathcal{N}(b, \sigma_Z^2)$ the distribution of $\sigma_Z N$ for $N \sim \mathcal{N}(0,1)$.

Note that the function $\varepsilon \mapsto \phi_\varepsilon(z)$ is C^∞ in an open interval I containing all values of ε allowed in the lemma statement and for all $z \in \mathbb{R}$. Moreover, this function and its derivatives are bounded uniformly in z over the same open interval I, since we assume $\sigma_X > 0$ and $\sigma_W > 0$.[2] Thus, for all z, we have the Taylor expansion of $\phi_\varepsilon(z)$:

$$\phi_\varepsilon(z) = \phi_0(z) + \varepsilon \frac{(z-b)^2 - \sigma_Z^2}{2\sigma_Z^2} \phi_0(z) + \varepsilon^2 R_\varepsilon(z), \tag{S9}$$

where $R_\varepsilon(z)$ is a remainder term. Using the Lagrange form of the remainder, we can write[3]

$$R_\varepsilon(z) = \frac{\phi_{\varepsilon'}^{(2)}(z)}{2} \varepsilon^2 = \varepsilon^2 p_{\varepsilon'}(z) \phi_{\varepsilon'}(z) \tag{S10}$$

for some $|\varepsilon'| \leq \varepsilon$, where $p_\varepsilon(z)$ is a degree-4 polynomial with ε-dependent coefficients that are finite for ε in the open interval $I \ni \varepsilon'$.

Now, we can use Eqs. (S8) and (S9) to see

$$\mathbb{E}_N f\left(\sigma_Z \sqrt{1+\varepsilon} N + b\right) = \mathbb{E}_N\left[f\left(\sigma_Z N + b\right)\right] + \varepsilon\, \mathbb{E}_N\left[f\left(\sigma_Z N + b\right) \frac{(\sigma_Z^2 - 1)N^2}{2\sigma_Z^2}\right] + \int_{\mathbb{R}} f(z) R_\varepsilon(z) dz. \tag{S11}$$

The first two term of the right-hand side of Eq. (S11) are as expected, and for the last term we see

$$\left| \int_{\mathbb{R}} f(z) R_\varepsilon(z) dz \right| \leq |\varepsilon|^2 \left(\int_{\mathbb{R}} |f(z)|^2 \phi_\varepsilon(z) dz \int_{\mathbb{R}} |p_\varepsilon(z)|^2 \phi_\varepsilon(z) dz \right)^{1/2} = \mathcal{O}(\varepsilon^2), \tag{S12}$$

by assumption on f. This proves Eq. (S6). An identical argument applied to $f(z)^2$ proves (S7). □

[2] In general, the derivatives look like $\phi_\varepsilon(z) p_\varepsilon(z)$ for some polynomial in p_ε with ε-dependent coefficients that are finite for ε in the open interval I.

[3] The derivative is with respect to ε in the display below.

Without loss of generality, $\mathbb{E}F_{a\alpha} = 0$. Eq. (S6) implies

$$\mathbb{E}F_{a\alpha} = \mathbb{E}_{N\sim\mathcal{N}(0,\sigma_W^2 c_{\alpha\alpha})}[f(N + b_a)] = \xi_0(b_a) + \xi_2(b_a)\varepsilon_\alpha + \mathcal{O}(\varepsilon_\alpha^2). \tag{S13}$$

Thus, the two leading order terms of $\mathbb{E}[F|B]$ are both low-rank and so cannot change the spectrum or Stieltjes transform at order 1 (for more detail see Bai and Zhou (2008)), and the remainder term has Frobenius norm bounded by $\mathcal{O}(n_0^{2\varepsilon})$ and so also cannot affect the spectrum or Stieltjes transform at order 1. In the notation above, we say $\xi_0(b) = 0$ for all b.

We now derive the correlation structure of F conditional on B (recall X is deterministic, but we may view this as conditioning on X).

Lemma 2. *To leading order the correlation structure of F is*

$$\mathbb{E}[F_{a\alpha}F_{b\beta}|B] = \begin{cases} 0 & \text{if } a \neq b \\ \eta(b_a) & \text{if } \alpha = \beta \text{ and } a = b \\ \zeta(b_a)c_{\alpha\beta} & \text{if } \alpha \neq \beta \text{ and } a = b \end{cases}. \tag{S14}$$

Proof. Using the observation in Eq. (S2), we repeatedly rewrite the expectations in $F^\top F$ as we can reduce the expectation over W to an expectation over at most two correlated Gaussian random variables. By independence (see Eq. (S3)), $\mathbb{E}F_{a\alpha}F_{b\beta} = \mathbb{E}F_{a\alpha} \cdot \mathbb{E}F_{b\beta} = 0$ for $a \neq b$. Using Lem. 1, we get

$$\mathbb{E}f(Z_{a\alpha})^2 = \eta(b_a) + \eta_2(b_a)\varepsilon_\alpha + \mathcal{O}(\varepsilon_\alpha^2). \tag{S15}$$

For the covariance calculation, $\mathbb{E}F_{a\alpha}F_{a\beta}$, we repeat the argument in the proof of Lem. 1 except we define $\varphi_{\varepsilon_1,\varepsilon_2,\rho}(z_1, z_2)$ as the bivariate p.d.f of

$$\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \sigma_Z^2\begin{pmatrix}1+\varepsilon_\alpha & \tilde{c}_{\alpha\beta}\\ \tilde{c}_{\alpha\beta} & 1+\varepsilon_\beta\end{pmatrix}\right). \tag{S16}$$

We then perform a multivate Taylor expansion about $(0, 0, 0)$ in $(\varepsilon_\alpha, \varepsilon_\beta, \tilde{c}_{\alpha\beta})$ to first-order. Note that under $\varphi_{0,0,0}$ the coordinates are independent and $\varphi_{0,0,0}(z_1, z_2) = \phi_0(z_1)\phi_0(z_2)$, so we can factor any integral over z_1 and z_2 using Fubini's theorem. We skip the details for the remainder term, as they are similar to before, but the leading order terms are

$$\begin{aligned}
\mathbb{E}F_{a\alpha}F_{a\beta} &= \int_{\mathbb{R}^2} f(z_1)f(z_2)\varphi_{\varepsilon_1,\varepsilon_2,\rho}(z_1, z_2)dz_1dz_2 && \text{(S17)}\\
&= \int_{\mathbb{R}^2} f(z_1)f(z_2)\varphi_{0,0,0}(z_1, z_2)dz_1dz_2 + \varepsilon_1\int_{\mathbb{R}^2}\frac{(z_1-b)^2-\sigma_Z^2}{2\sigma_Z^2}f(z_1)f(z_2)\varphi_{0,0,0}(z_1, z_2)dz_1dz_2 && \text{(S18)}\\
&\quad + \varepsilon_2\int_{\mathbb{R}^2}\frac{(z_2-b)^2-\sigma_Z^2}{2\sigma_Z^2}f(z_1)f(z_2)\varphi_{0,0,0}(z_1, z_2)dz_1dz_2 && \text{(S19)}\\
&\quad + \tilde{c}_{\alpha\beta}\int_{\mathbb{R}^2}\frac{(z_1-b)(z_2-b)}{\sigma_Z^2}f(z_1)f(z_2)\varphi_{0,0,0}(z_1, z_2)dz_1dz_2 + R, && \text{(S20)}
\end{aligned}$$

where $R = \mathcal{O}(n_0^{2\varepsilon-1})$. Using Fubini's theorem and the assumption that $\int_R f(z)\phi_0(z)dz = 0$, we find that

$$\mathbb{E}F_{a\alpha}F_{a\beta} = \tilde{c}_{\alpha\beta}\left(\int_{\mathbb{R}}\frac{z-b}{\sigma_Z}f(z)\phi_0(z)dz\right)^2 + \mathcal{O}(n_0^{2\varepsilon-1}). \tag{S21}$$

□

B.1 Linearized model

Define

$$F^{\text{lin}} := \mathcal{C}\Theta^1\Sigma + \left(\mathcal{V}^2 - \mathcal{C}^2\right)^{1/2}\Theta^2, \tag{S22}$$

where Θ^1 and Θ^2 are $n_1 \times m$ matrices that have i.i.d. Gaussian entries with mean 0 and variance $1/n_1$; 2) $\mathcal{V}$ and $\mathcal{C}$ are $n_1 \times n_1$ diagonal matrices with entries

$$\mathcal{V}_a^2 = \eta(b_a) \quad \text{and} \quad \mathcal{C}_a^2 = \zeta(b_a), \tag{S23}$$

for a bias vector b; 3) Σ is a matrix square root of $\frac{1}{n_0} X^\top X$.

A simple calculation shows that the entries of the matrix F^{lin} match the first and second mixed moments of F given in Lem. 2 up to residuals of size $\mathcal{O}(n_0^{2\varepsilon-1})$ in the off-diagonal entries and $\mathcal{O}(n_0^{\varepsilon-1/2})$ in the diagonal entries. Therefore, the matrix of residuals has Frobenius norm at most $\mathcal{O}(n_0^{2\varepsilon})$, which implies the limiting spectra of $[\mathbb{E}F_{1\alpha}F_{a\beta}]_{\alpha\beta}$ and $[\mathbb{E}F^{\text{lin}}{}_{1\alpha}F^{\text{lin}}{}_{a\beta}]_{\alpha\beta}$ agree to $\mathcal{O}(1)$. However, unlike F, F^{lin} is linear in the random matrices Θ^1 and Θ^2; in this sense, it is a linearization of F.

C VERIFYING ASSUMPTIONS (A3) AND (A4) OF Yaskov (2014)

We define

$$\Sigma_{\alpha\beta}^{(a)} := \mathbb{E}F_{a\alpha}F_{a\beta} = \mathbb{E}_{N_\alpha, N_\beta}\left[f(N_\alpha + b_a)f(N_\beta + b_a)\right] \tag{S24}$$

for

$$(N_\alpha, N_\beta) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \sigma_W^2 \begin{pmatrix} c_{\alpha\alpha} & c_{\alpha\beta} \\ c_{\beta\alpha} & c_{\beta\beta}\end{pmatrix}\right), \tag{S25}$$

whose leading order behavior we have understood in Lem. 2.

Assumption (A3). Define

$$\mathcal{Q}^{(a)} := \sum_{\alpha,\beta=1}^{m} F_{a\alpha}F_{a\beta}A_{\alpha\beta} - \operatorname{tr}\left(\Sigma^{(a)}A\right). \tag{S26}$$

By Chebyshev's inequality, suffices to show $\mathbb{E}\mathcal{Q}^{(a)} = 0$ and $\mathbb{E}|\mathcal{Q}^{(a)}|^2 = o(n_0^2)$ for all a uniformly in all real symmetric positive semi-definite A of bounded norm. The condition $\mathbb{E}\mathcal{Q}^{(a)} = 0$ is easily verified:

$$\mathbb{E}\mathcal{Q}^{(a)} = \sum_{\alpha,\beta=1}^{m} \mathbb{E}\left[F_{a\alpha}F_{a\beta}A_{\alpha\beta} - \Sigma_{\alpha\beta}^{(a)}A_{\beta\alpha}\right] = \sum_{\alpha,\beta=1}^{m} \mathbb{E}\left[F_{a\alpha}F_{a\beta} - \Sigma_{\alpha\beta}^{(a)}\right]A_{\beta\alpha} = 0. \tag{S27}$$

For the second condition, we can use the equivalent condition of Cor. 1.1 of Bai and Zhou (2008) for each choice of a.

Note that the definition of the multivariate Gaussian in Eq. (S25) easily extends to more indices, α, β, α', β', etc. In the following calculation it is useful to project out the correlated components of these Gaussian variables, i.e. to write

$$N_\alpha = N_\alpha^{(\alpha'\beta')} + C_\alpha^{\alpha'}N_{\alpha'} + C_\alpha^{\beta'}N_{\beta'}, \tag{S28}$$

where $N_\alpha^{(\alpha'\beta')}$, $N_{\alpha'}$, and $N_{\alpha'}$ are mean zero Gaussians such that $N_\alpha^{(\alpha'\beta')}$ is independent of $N_{\alpha'}$ and $N_{\beta'}$ and $C_\alpha^{\alpha'}$ and $C_\alpha^{\beta'}$ are constants. Moreover, our assumption on X implies that $C_\alpha^{\alpha'}$ and $C_\alpha^{\beta'}$ are $\mathcal{O}(n_0^{\varepsilon-1/2})$, or in words, the correlations are small. We also know that $\mathbb{V}[N_\alpha] - \mathbb{V}[N_\alpha^{(\alpha'\beta')}] = \mathcal{O}(n_0^{\varepsilon-1/2})$. This is exactly the same idea as in Sec. B but extended to functions that contain more entries of F. We note that Eq. (S28) is not necessarily unique.

For functions $\mathcal{F}$ and $\mathcal{G}$, the general principle is that $\mathbb{E}[\mathcal{F}(F_{a\alpha_1}, \ldots, F_{a\alpha_k})\mathcal{G}(F_{a\beta_1}, \ldots, F_{a\beta_l})]$ is close to zero when $\mathbb{E}\mathcal{F}(F_{a\alpha_1}, \ldots, F_{a\alpha_k}) \cdot \mathbb{E}\mathcal{G}(F_{a\beta_1}, \ldots, F_{a\beta_l}) = 0$ and $\{\alpha_1, \ldots, \alpha_k\} \cap \{\beta_1, \ldots, \beta_l\} = \emptyset$. More careful bookkeeping is required when higher-order cancellations are necessary to obtain the leading-order behavior.

The first condition of Cor. 1.1, Eq. (1.4), is straightforward to verify:

$$\mathbb{E}\left|F_{a\alpha}F_{a\beta} - \Sigma_{\alpha\beta}^{(a)}\right|^2 \le 4\mathbb{E}F_{a\alpha}^2F_{a\beta}^2 + 4(\Sigma_{\alpha\beta}^{(a)})^2, \tag{S29}$$

which is finite by assumption (see Sec. 2) and so is certainly $o(n_0)$ uniformly in α, β, and a.

The next condition, Eq. (1.5), is more involved. We have to show

$$\sum_{\Lambda} \left(\mathbb{E}\left(F_{a\alpha}F_{a\beta} - \Sigma^{(a)}_{\alpha\beta} \right) \left(F_{a\alpha'}F_{a\beta'} - \Sigma^{(a)}_{\alpha'\beta'} \right) \right)^2 = o(n_0^2) \tag{S30}$$

uniformly in a, where

$$\Lambda := \{(\alpha,\beta,\alpha',\beta') : 1 \le \alpha,\beta,\alpha',\beta' \le m\} \setminus \{(\alpha,\beta,\alpha',\beta') : \alpha = \alpha' \neq \beta = \beta' \text{ or } \alpha = \beta' \neq \alpha' = \beta\}.$$

We split the sum, Eq. (S30), into several pieces:

$$\sum_{\alpha} \left[\mathbb{E}\left(F_{a\alpha}^2 - \Sigma^{(a)}_{\alpha\alpha} \right)^2 \right]^2, \tag{S31}$$

$$4\sum_{\alpha\neq\beta} \left[\mathbb{E}\left(F_{a\alpha}F_{a\beta} - \Sigma^{(a)}_{\alpha\beta} \right) \left(F_{a\alpha}^2 - \Sigma^{(a)}_{\alpha\alpha} \right) \right]^2, \tag{S32}$$

$$\sum_{\alpha\neq\beta} \left[\mathbb{E}\left(F_{a\alpha}^2 - \Sigma^{(a)}_{\alpha\alpha} \right) \left(F_{a\beta}^2 - \Sigma^{(a)}_{\beta\beta} \right) \right]^2, \tag{S33}$$

$$2\sum_{\alpha,\beta,\alpha'}^{\text{distinct}} \left[\mathbb{E}\left(F_{a\alpha}^2 - \Sigma^{(a)}_{\alpha\alpha} \right) \left(F_{a\beta}F_{a\alpha'} - \Sigma^{(a)}_{\beta\alpha'} \right) \right]^2, \tag{S34}$$

$$4\sum_{\alpha,\beta,\alpha'}^{\text{distinct}} \left[\mathbb{E}\left(F_{a\alpha}F_{a\beta} - \Sigma^{(a)}_{\alpha\beta} \right) \left(F_{a\alpha}F_{a\alpha'} - \Sigma^{(a)}_{\alpha\alpha'} \right) \right]^2, \tag{S35}$$

$$\text{and} \quad \sum_{\alpha,\beta,\alpha'\beta'}^{\text{distinct}} \left[\mathbb{E}\left(F_{a\alpha}F_{a\beta} - \Sigma^{(a)}_{\alpha\beta} \right) \left(F_{a\alpha'}F_{a\beta'} - \Sigma^{(a)}_{\alpha'\beta'} \right) \right]^2. \tag{S36}$$

These are based on the six possible ways of partitioning the four indices. All indices in the above sums are distinct, and we will see that the addition correlations when some indices are equal are compensated by the lower combinatorial factor from the sum.

Eq. (S31) **is** $o(n_0^2)$**, since the summands are** $o(n_0)$**:** We have $\mathbb{E}\left(F_{a\alpha}^2 - \Sigma^{(a)}_{\alpha\alpha} \right)^2 \le 4\mathbb{E}F_{a\alpha}^4 + 4(\Sigma^{(a)}_{\alpha\alpha})^2$, which is $\mathcal{O}(1)$ by assumption and thus $o(n_0)$.

Eq. (S32) **is** $o(n_0^2)$**, since the summands are** $o(1)$**:** First consider the term

$$\mathbb{E}\left(F_{a\alpha}F_{a\beta} - \Sigma^{(a)}_{\alpha\beta} \right) \left(F_{a\alpha}^2 - \Sigma^{(a)}_{\alpha\alpha} \right) = \mathbb{E}[(f(N_\alpha + b_a)f(N_\beta + b_a) - \Sigma^{(a)}_{\alpha\beta})(f(N_\alpha + b_a)^2 - \Sigma^{(a)}_{\alpha\alpha})]$$

and note $N_\beta = N_\beta^{(\alpha)} + C_\beta^\alpha N_\alpha$ then Taylor expand in $C_\beta^\alpha N_\alpha$ to get

$$\begin{aligned} &\mathbb{E}\left(f(N_\alpha + b_a)f(N_\beta^{(\alpha)} + b_a) - \Sigma^{(a)}_{\alpha\beta} \right) \left(f(N_\alpha + b_a)^2 - \Sigma^{(a)}_{\alpha\alpha} \right) \\ &\quad + C_\beta^\alpha \mathbb{E}N_\alpha f'(N_\beta^{(\alpha)}) \left(f(N_\alpha + b_a)^2 - \Sigma^{(a)}_{\alpha\alpha} \right) + \mathcal{O}(n_0^{2\varepsilon-1}). \end{aligned}$$

The second term above is $\mathcal{O}(n_0^{\varepsilon-1/2})$ because of the C_β^α term. For the first term, note it is equal to

$$\mathbb{E}f(N_\beta^{(\alpha)} + b_a) \cdot \mathbb{E}\left(f(N_\alpha + b_a) \right) \left(f(N_\alpha + b_a)^2 - \Sigma^{(a)}_{\alpha\alpha} \right). \tag{S37}$$

Finally, $\mathbb{E}f(N_\beta^{(\alpha)} + b_a) = \mathcal{O}(n_0^{\varepsilon-1/2})$ since $C_\beta^\alpha = \mathcal{O}(n_0^{\varepsilon-1/2})$ and

$$0 = \mathbb{E}f(N_\beta + b_a) = \mathbb{E}f(N_\beta^{(\alpha)} + b_a) + C_\beta^\alpha \mathbb{E}N_\alpha f'(N_\beta^{(\alpha)}) + \mathcal{O}(n_0^{2\varepsilon-1}). \tag{S38}$$

So squaring and putting the above bounds together we find the summands are $\mathcal{O}(n_0^{2\varepsilon-1})$, which is $o(1)$.

Eq. (S33) **is** $o(n_0^2)$**, since the summands are** $o(1)$**:** Again we write $N_\beta = N_\beta^{(\alpha)} + C_\beta^\alpha N_\alpha$ and Taylor expand in $C_\beta^\alpha N_\alpha$:

$$\begin{aligned}&\mathbb{E}\left(f(N_\alpha+b_a)^2-\Sigma^{(a)}_{\alpha\alpha}\right)\cdot\mathbb{E}\left(f(N_\beta^{(\alpha)}+b_a)^2-\Sigma^{(a)}_{\beta\beta}\right)\\&\quad+2C_\beta^\alpha\mathbb{E}\left(N_\alpha f(N_\alpha+b_a)^2-\Sigma^{(a)}_{\alpha\alpha}\right)\cdot\mathbb{E}\left(f(N_\beta^{(\alpha)}+b_a)f'(N_\beta^{(\alpha)}+b_a)\right)+\mathcal{O}(n_0^{2\varepsilon-1}),\end{aligned}$$

where the first term is zero and the second is $\mathcal{O}(n_0^{\varepsilon-1/2})$ due to C_β^α. Therefore after squaring the summands are $\mathcal{O}(n_0^{2\varepsilon-1})$, which is $o(1)$.

Eq. (S34) **is** $o(n_0^2)$**, since the summands are** $o(n_0^{-1})$**:** We have

$$\mathbb{E}\left(F_{a\alpha}^2-\Sigma^{(a)}_{\alpha\alpha}\right)\left(F_{a\beta}F_{a\alpha'}-\Sigma^{(a)}_{\beta\alpha'}\right)=\mathbb{E}(f(N_\alpha+b_a)^2-\Sigma^{(a)}_{\alpha\alpha})(f(N_\beta+b_a)f(N_{\alpha'}+b_a)-\Sigma^{(a)}_{\beta\alpha'})\tag{S39}$$

and expand N_β and $N_{\alpha'}$ in N_α to get

$$\begin{aligned}&\mathbb{E}(f(N_\alpha+b_a)^2-\Sigma^{(a)}_{\alpha\alpha})\cdot\mathbb{E}(f(N_\beta^{(\alpha')}+b_a)f(N_{\alpha'}^{(\beta)}+b_a)-\Sigma^{(a)}_{\beta\alpha'})\\&\quad+\mathbb{E}N_\alpha(f(N_\alpha+b_a)^2-\Sigma^{(a)}_{\alpha\alpha})\left(C_\beta^\alpha f'(N_\beta^{(\alpha)}+b_a)f(N_{\alpha'}^{(\alpha)}+b_a)+C_{\alpha'}^\alpha f'(N_{\alpha'}^{(\alpha)}+b_a)f(N_\beta^{(\alpha)}+b_a)\right)\\&\quad+\mathcal{O}(n_0^{2\varepsilon-1}).\end{aligned}$$

The first term above is zero. The second term can be written

$$\begin{aligned}&C_\beta^\alpha\mathbb{E}N_\alpha(f(N_\alpha+b_a)^2-\Sigma^{(a)}_{\alpha\alpha})\cdot\mathbb{E}f'(N_\beta^{(\alpha)}+b_a)f(N_{\alpha'}^{(\alpha)}+b_a)\\&\quad+C_{\alpha'}^\alpha\mathbb{E}N_\alpha(f(N_\alpha+b_a)^2-\Sigma^{(a)}_{\alpha\alpha})\cdot\mathbb{E}f'(N_{\alpha'}^{(\alpha)}+b_a)f(N_\beta^{(\alpha)}+b_a).\end{aligned}\tag{S40}$$

Now we can expand $N_\beta^{(\alpha)}$ in $N_{\alpha'}$ as $N_\beta^{(\alpha)} = N_\beta^{(\alpha\alpha')} + C_\beta^{\alpha'}N_{\alpha'}$, where $N_\beta^{(\alpha\alpha')}$ and $N_{\alpha'}$ are independent. Then applying Lem. 2, we note

$$\mathbb{E}f'(N_\beta^{(\alpha)}+b_a)f(N_{\alpha'}^{(\alpha)}+b_a)=\mathbb{E}f'(N_\beta^{(\alpha\alpha')}+b_a)\cdot\mathbb{E}f(N_{\alpha'}^{(\alpha)}+b_a)+\mathcal{O}(n_0^{\varepsilon-1/2}).\tag{S41}$$

Applying Lem. 2 again shows $\mathbb{E}f(N_{\alpha'}^{(\alpha)}+b_a)=\mathcal{O}(n_0^{\varepsilon-1/2})$. Thus both terms in Eq. (S40) are $\mathcal{O}(n_0^{\varepsilon-1/2})$. Combining this with the fact that C_β^α and $C_{\alpha'}^\alpha$ are $\mathcal{O}(n_0^{\varepsilon-1/2})$, we conclude that Eq. (S39) is $\mathcal{O}(n_0^{2\varepsilon-1})$. Squaring shows the summands are $\mathcal{O}(n_0^{4\varepsilon-2})$, which is $o(n_0^{-1})$.

Eq. (S35) **is** $o(n_0^2)$**, since the summands are** $o(n_0^{-1})$**:** The argument is similar to that above for Eq. (S34). Except we expand as $N_\beta = N_\beta^{(\alpha)} + C_\beta^\alpha N_\alpha$ and $N_{\alpha'} = N_{\alpha'}^{(\alpha)} + C_{\alpha'}^\alpha N_\alpha$.

Eq. (S36) **is** $o(n_0^2)$**, since the summands are** $o(n_0^{-2})$**:** We expand $N_{\alpha'} = N_{\alpha'}^{(\alpha\beta)} + C_{\alpha'}^\alpha N_\alpha + C_{\alpha'}^\beta N_\beta$ and $N_{\beta'} = N_{\beta'}^{(\alpha\beta)} + C_{\beta'}^\alpha N_\alpha + C_{\beta'}^\beta N_\beta$, and then Taylor expand as before. We find

$$\begin{aligned}&\mathbb{E}\left(f(N_\alpha+b_a)f(N_\beta+b_a)-\Sigma^{(a)}_{\alpha\beta}\right)\left(f(N_{\alpha'}+b_a)f(N_{\beta'}+b_a)-\Sigma^{(a)}_{\alpha'\beta'}\right)\\&\quad=\mathbb{E}\left(f(N_\alpha+b_a)f(N_\beta+b_a)-\Sigma^{(a)}_{\alpha\beta}\right)\cdot\mathbb{E}\left(f(N_{\alpha'}^{(\alpha\beta)}+b_a)f(N_{\beta'}^{(\alpha\beta)}+b_a)-\Sigma^{(a)}_{\alpha'\beta'}\right)\\&\quad+\mathbb{E}\Bigg[\left(f(N_\alpha+b_a)f(N_\beta+b_a)-\Sigma^{(a)}_{\alpha\beta}\right)\Big((C_{\alpha'}^\alpha N_\alpha+C_{\alpha'}^\beta N_\beta)f(N_{\beta'}^{(\alpha\beta)}+b_a)f'(N_{\alpha'}^{(\alpha\beta)}+b_a)\\&\qquad+(C_{\alpha'}^\alpha N_\alpha+C_{\alpha'}^\beta N_\beta)f(N_{\alpha'}^{(\alpha\beta)}+b_a)f'(N_{\beta'}^{(\alpha\beta)}+b_a)\Big)\Bigg].\end{aligned}$$

The first term above is zero. For the second term, consider

$$C_{\alpha'}^\alpha\mathbb{E}N_\alpha\left(f(N_\alpha+b_a)f(N_\beta+b_a)-\Sigma^{(a)}_{\alpha\beta}\right)\cdot\mathbb{E}f(N_{\beta'}^{(\alpha\beta)}+b_a)f'(N_{\alpha'}^{(\alpha\beta)}+b_a)\tag{S42}$$

which is one of four similar terms that come from expanding the sums out of the second term above. We need to show Eq. (S42) is $\mathcal{O}(n^{3\varepsilon-3/2})$, and since the other 3 terms are of the same form as Eq. (S42) this will complete the argument. Taking $\mathbb{E}N_\alpha \left(f(N_\alpha+b_a)f(N_\beta+b_a)-\Sigma^{(a)}_{\alpha\beta}\right)$ first and expanding $N_\beta = N_\beta^{(\alpha)} + C_\beta^\alpha N_\alpha$, we see

$$\begin{aligned}\mathbb{E}N_\alpha \left(f(N_\alpha+b_a)f(N_\beta+b_a)-\Sigma^{(a)}_{\alpha\beta}\right) &= \mathbb{E}N_\alpha f(N_\alpha+b_a)f(N_\beta+b_a)\\ &= \mathbb{E}N_\alpha f(N_\alpha+b_a)\mathbb{E}f(N_\beta^{(\alpha)}+b_a) + C_\beta^\alpha \mathbb{E}N_\alpha^2 f(N_\alpha+b_a)f'(N_\beta^{(\alpha)}) + \mathcal{O}(n_0^{2\varepsilon-1})\\ &= \mathcal{O}(n_0^{\varepsilon-1/2}),\end{aligned}$$

since $\mathbb{E}f(N_\beta^{(\alpha)}+b_a) = \mathcal{O}(n_0^{\varepsilon-1/2})$ as above. Second, we consider $\mathbb{E}f(N_{\beta'}^{(\alpha\beta)}+b_a)f'(N_{\alpha'}^{(\alpha\beta)}+b_a)$ by expanding $N_{\beta'}^{(\alpha\beta)} = N_{\beta'}^{(\alpha\beta\alpha')} + C_{\beta'}^{\alpha'}N_{\alpha'}$

$$\mathbb{E}f(N_{\beta'}^{(\alpha\beta)}+b_a)f'(N_{\alpha'}^{(\alpha\beta)}+b_a) = \mathbb{E}f(N_{\beta'}^{(\alpha\beta\alpha')}+b_a)\mathbb{E}f'(N_{\alpha'}^{(\alpha\beta)}+b_a) + C_{\beta'}^{\alpha'}\mathbb{E}f'(N_{\alpha'}^{(\alpha\beta)}+b_a)N_{\alpha'}f' + \mathcal{O}(n_0^{2\varepsilon-1}),$$

which is $\mathcal{O}(n_0^{\varepsilon-1/2})$ since both $\mathbb{E}f(N_{\beta'}^{(\alpha\beta\alpha')}+b_a)$ and $C_{\beta'}^{\alpha'}$ are $\mathcal{O}(n_0^{\varepsilon-1/2})$ as before. Third, $C_{\alpha'}^{\alpha} = \mathcal{O}(n_0^{\varepsilon-1/2})$, so combining these three multiplicative factors we have Eq. (S42) is $\mathcal{O}(n_0^{3\varepsilon-3/2})$.

Finally, squaring we see the summands are $\mathcal{O}(n_0^{6\varepsilon-3})$, which is clearly $o(n_0^{-2})$.

Assumption (A4). This is much easier to verify: using Lem. 2, we see

$$\begin{aligned}\mathrm{tr}(\Sigma^{(a)})^2 &= \sum_{\alpha,\beta}(\Sigma^{(a)}_{\alpha\beta})^2\\ &= \sum_{\alpha}(\Sigma^{(a)}_{\alpha\alpha})^2 + \sum_{\alpha\neq\beta}(\Sigma^{(a)}_{\alpha\beta})^2\\ &= n_0\eta(b_a)^2 + \zeta(b_a)^2\sum_{\alpha\neq\beta}c_{\alpha\beta}^2 + \mathcal{O}(n_0^{3\varepsilon+1/2})\\ &= \mathcal{O}(n_0^{2\varepsilon+1})\\ &= o(n_0^2)\end{aligned}$$

without any dependence on a.

D DERIVATION OF SELF-CONSISTENT EQUATION FOR F^{lin}

Theorem 2. *Let $s_m(z)$ be the Stieltjes transform of $\frac{1}{n_1}F^{lin\top}F^{lin}$, i.e. $\frac{1}{m}\mathrm{tr}G(z)$. Then with probability 1, as $m\to\infty$, for all z such that $\Im z>0$, $s_m(z)\to s(z)$, where $s(z)$ is the solution of the coupled equations*

$$s(z) = \mathbb{E}_{S\sim MP(\phi)}\left[\frac{1}{C_0(z)+SC_1(z)}\right] \quad \textit{and} \quad \tilde{s}(z) = \mathbb{E}_{S\sim MP(\phi)}\left[\frac{S}{C_0(z)+SC_1(z)}\right] \tag{S43}$$

with

$$C_0(z) := -z + \mathbb{E}_{B\sim\mu_B}\left[\frac{\eta(B)-\zeta(B)}{D(B)}\right], \quad C_1(z) := \mathbb{E}_{B\sim\mu_B}\left[\frac{\zeta(B)}{D(B)}\right], \tag{S44}$$

$$D(B) := 1 + \frac{\psi}{\phi}\left(\zeta(B)\tilde{s}(z) + (\eta(B)-\zeta(B))\,s(z)\right). \tag{S45}$$

We want to study the eigenvalues of $\frac{1}{n_1}F^{\mathrm{lin}\top}F^{\mathrm{lin}}$. Note without loss of generality it is sufficient to consider diagonal Σ, since we can diagonalize Σ using some orthogonal matrices O and O' as it is the square root of a positive definite matrix. Moreover, these orthogonal matrices, when applied to either Θ^1 or Θ^2 do not change their distributions. Thus, diagonalizing Σ, so that $\Sigma_{\alpha\alpha} = \sqrt{\lambda_\alpha^X}$, where λ_α^X are the eigenvalues of $X^\top X/n_0$, results in an equivalent matrix ensemble in distribution (see Silverstein and Bai (1995) for more detail). With this simplification, F^{lin} has independent entries given by

$$F^{\mathrm{lin}}{}_{a\alpha} = \mathcal{C}_a\Theta^1_{a\alpha}\sqrt{\lambda_\alpha^X} + \sqrt{\mathcal{V}_a^2-\mathcal{C}_a^2}\Theta^2_{a\alpha}. \tag{S46}$$

To make the derivation easier, we can partly linearize the problem by studying the matrix

$$H := \begin{bmatrix} -zI & {F^{\text{lin}}}^\top/\sqrt{n_1} \\ F^{\text{lin}}/\sqrt{n_1} & -I \end{bmatrix}. \tag{S47}$$

By the Schur complement formula, one easily finds

$$s_m(z) := \frac{1}{m}\text{tr}({F^{\text{lin}}}^\top F^{\text{lin}}/n_1 - zI) = \frac{1}{m}\sum_{\alpha=1}^{m} G_{\alpha\alpha}(z) \tag{S48}$$

$$\text{and} \quad z\tilde{s}_m(z) := \frac{1}{n_1}\text{tr}(F^{\text{lin}}{F^{\text{lin}}}^\top/n_1 - zI) = \frac{1}{n_1}\sum_{a=n_1+1}^{n_1+m} G_{aa}(z), \tag{S49}$$

where G is the *inverse* of H. Again by the Schur complement formula

$$\frac{1}{G_{\alpha\alpha}} = -z - \frac{1}{n_1}\sum_{a,b=1}^{n_1} {F^{\text{lin}}}_{a\alpha}{F^{\text{lin}}}_{b\alpha} G^{(\alpha)}_{m+a,m+b} \tag{S50}$$

$$\text{and} \quad \frac{1}{G_{aa}} = -1 - \frac{1}{n_1}\sum_{\alpha,\beta=1}^{m} {F^{\text{lin}}}_{a\alpha}{F^{\text{lin}}}_{a\beta} G^{(a)}_{\alpha\beta} \tag{S51}$$

for $\alpha \in \{1,\dots,m\}$ and $a \in \{m+1,\dots,m+n_1\}$ and $G^{(a)}$ is the inverse of the minor $H^{(a)}$. Since $G^{(\alpha)}$ is independent of $\Theta^1_{1\alpha},\dots,\Theta^1_{n1\alpha}$ and $\Theta^2_{1\alpha},\dots,\Theta^2_{n1\alpha}$, we see by taking the expectation over these variables that

$$\mathbb{E}\left[\frac{1}{n_1}\sum_{a,b=1}^{n_1} {F^{\text{lin}}}_{a\alpha}{F^{\text{lin}}}_{b\alpha} G^{(\alpha)}_{m+a,m+b}\right] = \frac{1}{n_1}\sum_{a=1}^{n_1}\left(\lambda^X_\alpha \zeta(b_a) + \eta(b_a) - \zeta(b_a)\right) G^{(\alpha)}_{m+a,m+a}. \tag{S52}$$

Moreover, standard concentration inequalities and the Ward identity (see Erdos and Yau (2017)) show

$$\left|\frac{1}{n_1}\sum_{a,b=1}^{n_1} {F^{\text{lin}}}_{a\alpha}{F^{\text{lin}}}_{b\alpha} G^{(\alpha)}_{m+a,m+b} - \mathbb{E}\left[\frac{1}{n_1}\sum_{a,b=1}^{n_1} {F^{\text{lin}}}_{a\alpha}{F^{\text{lin}}}_{b\alpha} G^{(\alpha)}_{m+a,m+b}\right]\right| \tag{S53}$$

$$\le Cm^\varepsilon\left(\frac{1}{m^2}\sum_{a,b}\left|G^{(\alpha)}_{m+a,m+b}\right|^2\right)^{1/2} \le Cm^\varepsilon\left(\frac{1}{m^2}\sum_{a}\frac{\Im G^{(\alpha)}_{m+a,m+a}}{\Im z}\right)^{1/2} \le \mathcal{O}(m^{\varepsilon-1/2}) \tag{S54}$$

with high probability. Similar bounds are easily obtained for $\sum_{\alpha,\beta=1}^m {F^{\text{lin}}}_{a\alpha}{F^{\text{lin}}}_{a\beta} G^{(a)}_{\alpha\beta}/n_1$.

We may also replace $G^{(\alpha)}$ with G at the expense of another small error that can be bounded using the Cauchy interlacing theorem: $\left|G^{(\alpha)}_{m+a,m+a} - G_{m+a,m+a}\right| \le \mathcal{O}(1/m)$. Using this control over these sums, we see

$$\frac{1}{G_{\alpha\alpha}} = -z - \sum_{a,b=1}^{n_1} {F^{\text{lin}}}_{a\alpha}{F^{\text{lin}}}_{b\alpha} G^{(\alpha)}_{ab} \tag{S55}$$

$$= -z - \frac{1}{n_1}\sum_{a=1}^{n_1}\left(\lambda^X_\alpha \zeta(b_a) + \eta(b_a) - \zeta(b_a)\right) G_{m+a,m+a} + \mathcal{O}(m^{\varepsilon-1/2}) \tag{S56}$$

and

$$\frac{1}{G_{aa}} = -1 - \sum_{\alpha,\beta=1}^{m} {F^{\text{lin}}}_{a\alpha}{F^{\text{lin}}}_{a\beta} G^{(a)}_{\alpha\beta} \tag{S57}$$

$$= -1 - \frac{1}{n_1}\sum_{\alpha=1}^{m}\left(\lambda^X_\alpha \zeta(b_a) + \eta(b_a) - \zeta(b_a)\right) G_{\alpha\alpha} + \mathcal{O}(m^{\varepsilon-1/2}). \tag{S58}$$

Finally, we invert Eq. (S57), multiply by $\lambda_\alpha^X \zeta(b_a) + \eta(b_a) - \zeta(b_a)$, and average over a to find

$$\frac{1}{n_1}\sum_{a=1}^{n_1}\left(\lambda_\alpha^X \zeta(b_a) + \eta(b_a) - \zeta(b_a)\right) G^{(\alpha)}_{m+a,m+a} \tag{S59}$$

$$= -\frac{1}{n_1}\sum_{a=1}^{n_1}\frac{\lambda_\alpha^X \zeta(b_a) + \eta(b_a) - \zeta(b_a)}{1+\frac{\psi}{\phi}\left(\zeta(b_a)\tilde{s}(z) + (\eta(b_a)-\zeta(b_a))\, s_m(z)\right) + \mathcal{O}(m^{\varepsilon-1/2})} \tag{S60}$$

$$= -\frac{1}{n_1}\sum_{a=1}^{n_1}\frac{\lambda_\alpha^X \zeta(b_a) + \eta(b_a) - \zeta(b_a)}{1+\frac{\psi}{\phi}\left(\zeta(b_a)\tilde{s}(z) + (\eta(b_a)-\zeta(b_a) s_m(z)\right)} + \mathcal{O}(m^{\varepsilon-1/2}) \tag{S61}$$

$$= -\mathbb{E}_{B\sim\mathcal{N}(0,\sigma_b^2)}\left[\frac{\lambda_\alpha^X \zeta(B) + \eta(B) - \zeta(B)}{1+\frac{\psi}{\phi}\left(\zeta(B)\tilde{s}(z) + (\eta(B)-\zeta(B)) s_m(z)\right)}\right] + o(1), \tag{S62}$$

where $\tilde{s}_m(z) = \frac{1}{m}\sum_\alpha \lambda_\alpha^X G_{\alpha\alpha}$, we Taylor expanded in the second step, and we used our assumption on B.

We can now invert Eq. (S55) and average over α to find

$$s_m(z) \to \mathbb{E}_{S\sim\mu_X}\left[\frac{1}{-z + \mathbb{E}_{B\sim\mathcal{N}(0,\sigma_b^2)}\left[\frac{S\zeta(B)+\eta(B)-\zeta(B)}{1+\frac{\psi}{\phi}(\zeta(B)\tilde{s}(z)+(\eta(B)-\zeta(B))s(z))}\right]}\right]. \tag{S63}$$

Similarly,

$$\tilde{s}_m(z) \to \mathbb{E}_{S\sim\mu_X}\left[\frac{S}{-z + \mathbb{E}_{B\sim\mathcal{N}(0,\sigma_b^2)}\left[\frac{S\zeta(B)+\eta(B)-\zeta(B)}{1+\frac{\psi}{\phi}(\zeta(B)\tilde{s}(z)+(\eta(B)-\zeta(B)s(z))}\right]}\right]. \tag{S64}$$

Note that the integral over S here is an integral over the limiting distribution of the data μ_X, for which we assume

$$\frac{1}{m}\sum_\alpha \delta_{\lambda_\alpha^X} \to \mu_X \tag{S65}$$

in distribution. In the case of i.i.d. Gaussian data, this is exactly given by the Marchenko-Pastur distribution.

E NONLINEAR MIXTURES CAN INCREASE MODEL CAPACITY OVER SINGLE NONLINEARITIES

In this section, we build on our results for the training loss on noisy autoencoder tasks to examine the benefits of nonlinear mixtures. For this analysis, we consider the simplest possible nontrivial distribution over activation functions: a Bernoulli mixture of two different functions. To each of these functions we associate two constants, η and ζ, which derive from Eq. (8) but have no B-dependence since each function is a *single nonlinearity*. Concretely, let

$$\eta = \mathbb{E}\left[f(N)^2\right] \quad \text{and} \quad \zeta = \mathbb{E}\left[\sigma_Z f'(N)\right]^2, \tag{S66}$$

where $N \sim \mathcal{N}(0, \sigma_Z)$. For the two functions themselves, we use (i) a "pure linear" activation function with $\eta = 1$, i.e. the identity function and (ii) a "pure nonlinear" (Hastie et al., 2019) activation function with $\eta = 1$ and $\zeta = 0$. (The particular purely nonlinear function in (ii) is irrelevant, as our theory predicts and our experiments confirm; see Fig. S1(a)). To be precise, we define for $p \sim$ Bernoulli(p),

$$f_p(x) := \begin{cases} x & \text{if } p = 0 \\ g_{\zeta=0}(x) & \text{if } p = 1 \end{cases}, \tag{S67}$$

where $g_{\zeta=0}$ is any function with $\eta = 1$ and $\zeta = 0$ (see, e.g., the functions in Fig. (3) of Pennington and Worah (2017)).

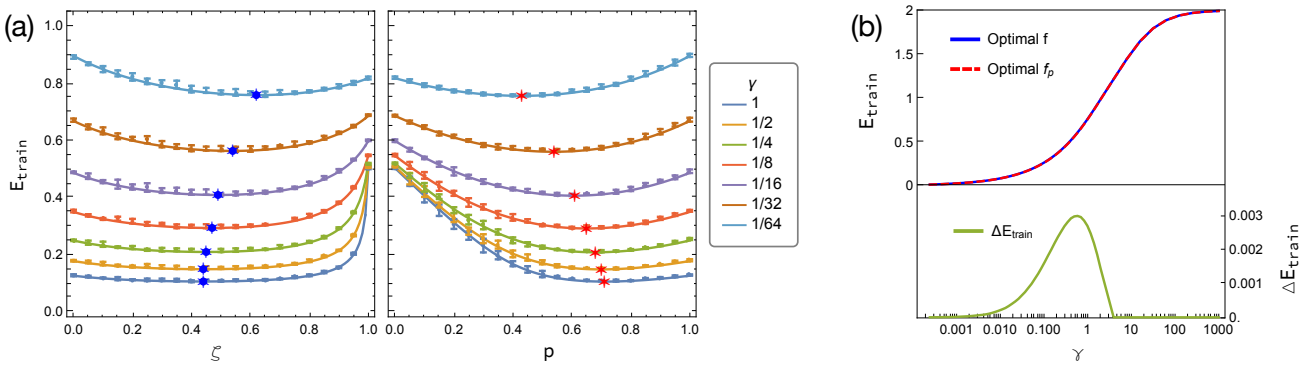

Figure S1: Performance on ridge-regularized noisy autoencoder with $\sigma_\epsilon = 1$, $\phi = 1/2$, and $\psi = 1/2$. (a) Theoretical predictions for training error (solid lines) and 1σ error bars for empirical simulations of finite networks ($n_0 = 192$, $n_1 = 384$, $m = 384$) for various values of ridge regularization constant γ as the activation function varies. In the left panel, a single activation function f is used. In the right panel, the non-linearity is f_p, a Bernoulli(p)-mixture of a purely linear ($\zeta = 1$) and purely non-linear ($\zeta = 0$) function. Each simulation uses a randomly-chosen non-linearity having the specified values of ζ, demonstrating that E_{train} depends on the non-linearity solely through this constant. Red and blue stars denote minima. (b) Training error as a function of γ for the optimal f and f_p, as determined in (a). The bottom panel shows the difference in training error, demonstrating that the optimal Bernoulli(p)-mixture of non-linearities has smaller training error than the best single non-linearity.

The task of computing E_{train} for f_p is cumbersome but purely algebraic. To see how to proceed, notice that the expectations in Eq. (11) are simple for f_p:

$$C_0(z) = -z + \frac{p}{1 + \psi/\phi\, s(z)} \quad \text{and} \quad C_1(z) = \frac{1-p}{1 + \psi/\phi\, \tilde{s}(z)}\,. \tag{S68}$$

Plugging these equations into Eqs. (13) and (14), collecting terms and simplifying yields a set of coupled polynomial equations for $s(z)$ and $\tilde{s}(z)$. Taking the total derivative of these equations with respect to z yields two additional equations which can be solved to express $s'(z)$ and $\tilde{s}'(z)$ in terms of $s(z)$ and $\tilde{s}(z)$. Combining these results produces a polynomial system whose solution[4] encodes E_{train} through Eq. (17). Fig. S1(a) shows the result of this calculation in solid lines for various values of γ, while the 1σ error bars show empirical simulations with finite networks. The red stars in the figure show that for many values of γ, the optimal mixture percentage is intermediate, i.e. $0 < p < 1$.

The question is, does a non-trivial mixture actually outperform a single nonlinearity? First, we must understand the performance of the optimal single nonlinearity. We note that owing to the homogeneity of the the training loss in η, ζ, and γ, we can assume without loss of generality that $\eta = 1$. Therefore the entire effect of the nonlinearity should be encoded in the single constant ζ. In Fig. S1(a), we plot our theoretical prediction for E_{train} in solid lines and empirical simulations for finite networks as 1σ error bars. The activation function used for each simulation is chosen randomly, conditional on the value of ζ. So the good agreement in the left panel of Fig. S1(a) demonstrates not just the correctness of our theoretical result but also the fact that E_{train} depends on the activation function solely through the constant ζ. The blue stars in this figure indicate that the optimal single nonlinearity is neither purely linear ($\zeta = 1$), nor purely nonlinear ($\zeta = 0$), but rather something in between.

For this particular problem setup, the performance of the optimal single nonlinearity and the optimal Bernoulli mixture are rather close, as indicated by the top panel of Fig. S1(b). However, owing to our precise analytical formulation, we can evaluate the training loss to high precision and observe that there is indeed a difference in performance between the two models, as shown in the bottom panel of Fig. S1(b). This result establishes that there are some problems for which even the best single nonlinearity is outperformed by a mixture of nonlinearities.

[4]Special care must be taken in selecting the correct root of this equation, in accordance with the condition that $s(z) \sim -1/z$ for large $|z|$.

F DERIVATION OF THE TEST ERROR

The linearization from Sec. B.1 is a key ingredient in the calculation of the test error. In the bias free case, Adlam and Pennington (2020a) shows how it can be used with operator-valued free probability to provide an asymptotically exact prediction for the test error under the data generating assumption of Sec. 4.2. We briefly review this method here, highlighting the differences.

We use a slightly different linearization here than Sec. B.1 (which was more convenient for the derivation in Sec. D), but their correlation structure is easily verified to be identical. We set

$$F^{\text{lin}} := \mathcal{C}WX + \left(\mathcal{V}^2 - \mathcal{C}^2\right)^{1/2} \Theta. \tag{S69}$$

The test error is defined as

$$E_{\text{test}} = \mathbb{E}_{\beta,\epsilon}\mathbb{E}_{\mathbf{x}} \left(y(\mathbf{x}) - W_2^* f(W\mathbf{x}; B)\right)^2 . \tag{S70}$$

As in Sec. 4 of Adlam and Pennington (2020a), the test error can be expressed as the sum of terms E_1, E_2, and E_3, where

$$E_1 = \mathbb{E}_{\beta,\mathbf{x},\varepsilon}\text{tr}(y(\mathbf{x})y(\mathbf{x})^\top), \tag{S71}$$

$$E_2 = -2\mathbb{E}_{\beta,\mathbf{x},\varepsilon}\text{tr}(K_{\mathbf{x}}^\top K^{-1} Y^\top y(\mathbf{x})), \tag{S72}$$

$$E_3 = \mathbb{E}_{\beta,\mathbf{x},\varepsilon}\text{tr}(K_{\mathbf{x}}^\top K^{-1} Y^\top Y K^{-1} K_{\mathbf{x}}), \tag{S73}$$

$K = F^\top F + \gamma I$, $K_{\mathbf{x}} = F^\top f(W\mathbf{x}; B)$, $Y = \beta^\top X + \varepsilon$, and $y(\mathbf{x}) = \beta^\top \mathbf{x}$. Using these equalities the expectation over $\mathbf{x}$, β, and ϵ can be calculated to find

$$E_1 = 1 \tag{S74}$$

$$E_2 = E_{21} \tag{S75}$$

$$E_3 = E_{31} + E_{32}\,, \tag{S76}$$

where,

$$E_{21} = -2\frac{1}{n_0^{3/2} n_1}\mathbb{E}\text{tr}\left(X^\top W^\top \mathcal{C} F K^{-1}\right) \tag{S77}$$

$$E_{31} = \sigma_\varepsilon^2 \mathbb{E}\text{tr}\left(K^{-1}\Sigma_3 K^{-1}\right) \tag{S78}$$

$$E_{32} = \frac{1}{n_0}\mathbb{E}\text{tr}\left(K^{-1}\Sigma_3 K^{-1} X^\top X\right) \tag{S79}$$

and,

$$\Sigma_3 = \frac{1}{n_0 n_1^2} F^\top \mathcal{C} W W^\top \mathcal{C} F + \frac{1}{n_1^2} F^\top (\mathcal{V}^2 - \mathcal{C}^2) F\,. \tag{S80}$$

As was the case for the training error, we can apply the linearization $F \to F^{\text{lin}}$ without changing the limiting values of these asymptotic trace objects. A more detailed argument supporting these types of replacements can be found in (Tripuraneni et al., 2021a). To proceed further, we use iterated applications of the Schur complement formula to find that E_{21}, E_{31} and E_{32} can be expressed as the limiting traces of specific blocks of the inverses of certain matrices. For E_{21}, for example, the NCAlgebra Mathematica package (described in Helton et al., 2006) can be used to generate the following block matrix or *pencil*,

$$Q_{21} = \begin{pmatrix}
I_{n_0} & 0 & -\frac{\sqrt{n_0}W^\top}{\sqrt{n_1}} & 0 & 0 & 0 & 0 & 0 & 0 \\
-\frac{X^\top}{\sqrt{n_0}} & I_m & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & I_{n_1} & -\mathcal{C} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & I_{n_1} & -(\mathcal{V}^2 - \mathcal{C}^2)^{1/2} & 0 & -\mathcal{C} & 0 & 0 \\
0 & 0 & 0 & 0 & I_{n_1} & -\frac{\Theta}{\sqrt{n_1}} & 0 & 0 & 0 \\
\frac{X^\top}{\gamma\sqrt{n_0}} & 0 & 0 & 0 & 0 & I_m & 0 & 0 & \frac{\Theta^\top}{\gamma\sqrt{n_1}} \\
0 & 0 & 0 & 0 & 0 & 0 & I_{n_1} & -\frac{\sqrt{n_0}W}{\sqrt{n_1}} & 0 \\
0 & 0 & 0 & 0 & 0 & -\frac{X}{\sqrt{n_0}} & 0 & I_{n_0} & 0 \\
0 & 0 & 0 & -(\mathcal{V}^2 - \mathcal{C}^2)^{1/2} & 0 & 0 & 0 & 0 & I_{n_1}
\end{pmatrix}. \tag{S81}$$

Then, computing the inverse of $[Q_{21}]^\top$ via repeated applications of the Schur complement formula and taking block-wise traces, we readily find that

$$E_{21} = \frac{2}{\phi\gamma} G_{6,2}^{E_{21}} \,, \tag{S82}$$

where $G^{E_{21}} := \mathrm{id}_9 \otimes \frac{1}{m}\mathrm{tr}\,[(Q_{21})^\top]^{-1} \in M_9(\mathbb{C})$ is a scalar 9×9 matrix whose i,j entry $G_{i,j}^{E_{21}}$ is the normalized trace of the (i,j)-block of the inverse of $[(Q_{21})]^\top$.

Similar pencils can be produced for E_{31} and E_{32}, though the expressions are somewhat large and cumbersome so we omit them here. Given these pencils, matrix-valued self-consistent equations easily follow from the theory of operator-valued free probability (Mingo and Speicher, 2017). The calculations proceed in exactly the same fashion as in Sec. 4 of Adlam and Pennington (2020a) (see also (Tripuraneni et al., 2021b,a)), and the end result is in fact the same as Eq. (27) of Adlam and Pennington (2020a): asymptotically E_{test} is equal to the generalized cross-validation (GCV) metric of Golub et al. (1979). We remark that this correspondence to GCV strongly suggests a deeper underlying connection for random feature methods, and indeed similar results on the asymptotic correctness of the GCV error have been found in Hastie et al. (2019); Jacot et al. (2020). We leave the investigation of this general connection to future work.

G ADDITIONAL FIGURES

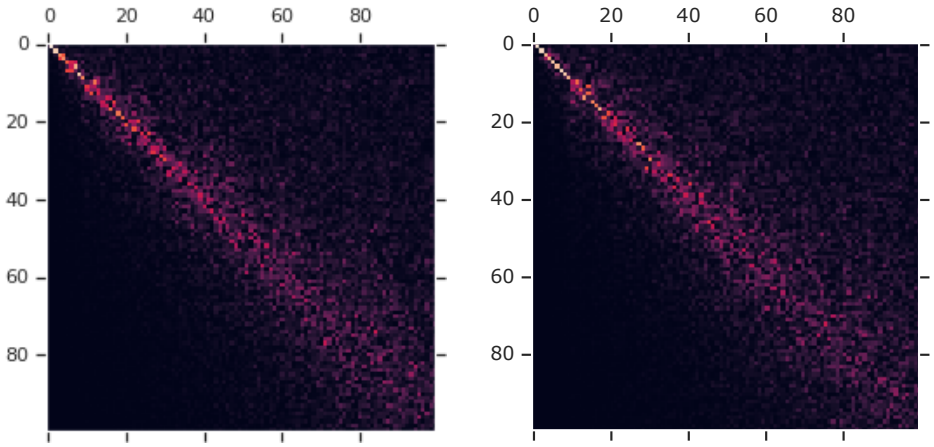

Figure S2: Comparison of right singular vectors of X and F for MNIST (left) and CIFAR10 (right). Entry ij shows $\mathbf{x}_i \cdot \mathbf{f}_j$. Although our theoretical results do not give predictions for how the singular vectors change, we found interesting behavior, with very little change to the largest singular vectors (which are nearly isolated in the spectrum), but more mixing of singular vectors in the dense part of the distribution.

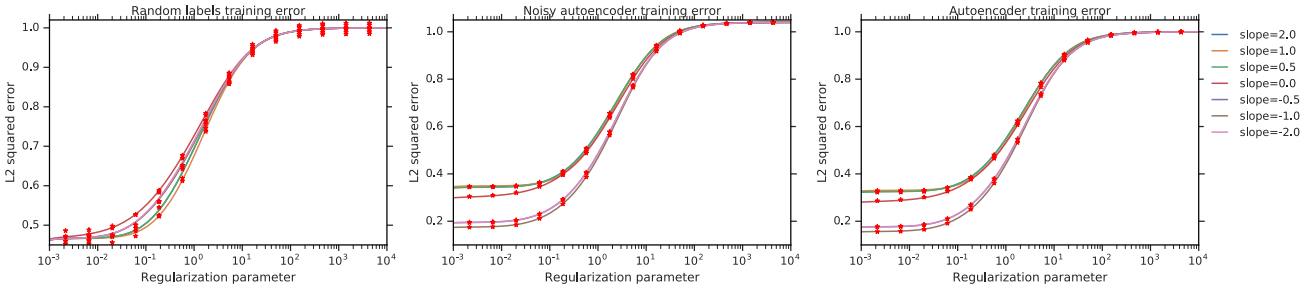

Figure S3: Comparisons of simulated ridge regression error and our theoretical prediction. Here we vary the activation function by changing the slope α in leaky ReLU. In particular $\alpha = -1$ is a linear function, $\alpha = 0$ is regular ReLU, and $\alpha = 1$ is the a scaled absolute value function. We normalize all functions so that $\mathbb{E}_b[\eta(b)] = 1$. We get excellent agreement with theory from only a single sample.

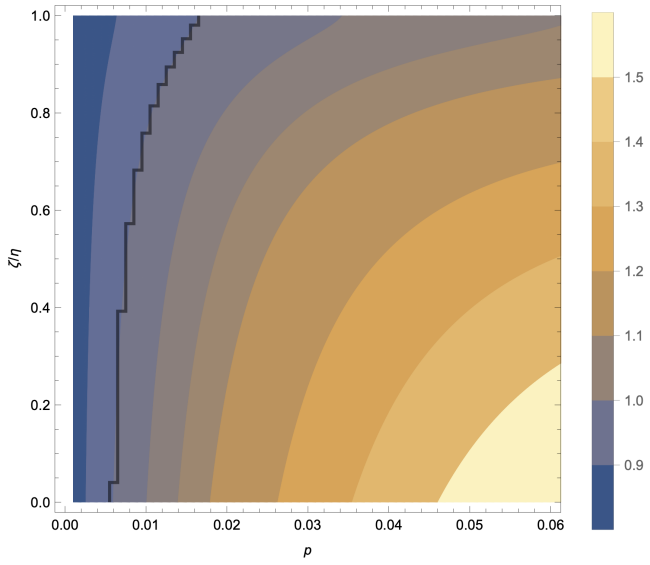

Figure S4: A close up of the left-hand side of Fig. 5 **Right** where p is small.