Supplemental Material: Memorizing without overfitting: Bias, variance, and interpolation in over-parameterized models

Jason W. Rocks[1] and Pankaj Mehta[1,2]

[1] *Department of Physics, Boston University, Boston, Massachusetts 02215, USA*
[2] *Faculty of Computing and Data Sciences, Boston University, Boston, Massachusetts 02215, USA*

CONTENTS

S1. CAVITY DERIVATIONS

In this section, we provide detailed derivations of all closed-form solutions for both models. We begin by setting up the theoretical framework and providing some useful approximations before deriving solutions for our two models. These calculations follow the general procedure laid out in Ref. 49.

A. Notation Conventions

- We define M as the number of points in the training data set, N_f as the number of input features, and N_p as the number of fit parameters/hidden features. We define the ratios $\alpha_f = N_f/M$ and $\alpha_p = N_p/M$.
- Unless otherwise specified, the type of symbol used for an index label (e.g., Δy_a) or as a summation index (e.g., $\sum_a$) implies its range. The symbols a, b, or c imply ranges over the training data points from 1 to M, the symbols j, k, or l imply ranges over the input features from 1 to N_f, and the symbols J, K, or L imply ranges over the fit parameters/hidden features from 1 to N_p.
- The notation $\mathrm{E}_x[\cdot]$, $\mathrm{Var}_x[\cdot]$ and $\mathrm{Cov}_x[\cdot,\cdot]$ represent the mean, variance, and covariance, respectively, with respect to one or more random variables x. A lack of subscript implies averages taken with respect to the total ensemble distribution, i.e., taken over all possible sources of randomness. A subscript 0 implies averages taken with respect to random variables containing one or more 0-valued indices (e.g., X_{a0}, X_{0j}, W_{0J}, or W_{j0}).
- We use the notation $\mathcal{O}(\cdot)$ to represent standard "Big-O" notation, indicating an upper bound on the limiting scaling behavior of a quantity with respect to the argument.

B. Theoretical Setup

For completeness, we begin by reproducing some of our theoretical setup from the main text. We consider data points $(y, \vec{\mathbf{x}})$, each consisting of a label y and a vector $\vec{\mathbf{x}}$ of N_f input features. The labels are related to the input features via the teacher model

$$y(\vec{\mathbf{x}}) = y^*(\vec{\mathbf{x}}; \vec{\boldsymbol{\beta}}) + \varepsilon, \tag{S1}$$

where ε is the label noise and $y^*(\vec{\mathbf{x}}; \vec{\boldsymbol{\beta}})$ are the true labels which depend on a vector $\vec{\boldsymbol{\beta}}$ of "ground truth" parameters. We consider features and label noise that are independently and identically distributed, drawn from a normal distributions with zero mean and variances σ_X^2/N_f and σ_ε^2, respectively, so that

$$\mathrm{E}[x_{a,j}] = 0, \qquad \mathrm{Cov}[x_{a,j}, x_{b,k}] = \frac{\sigma_X^2}{N_f}\delta_{ab}\delta_{jk} \tag{S2}$$

$$\mathrm{E}[\varepsilon_a] = 0, \qquad \mathrm{Cov}[\varepsilon_a, \varepsilon_b] = \sigma_\varepsilon^2\delta_{ab} \tag{S3}$$

for two data points $\vec{\mathbf{x}}_a$ and $\vec{\mathbf{x}}_b$ with label noise ε_a and ε_b.

We also restrict ourselves to a tacher model of the form

$$y^*(\vec{\mathbf{x}}; \vec{\boldsymbol{\beta}}) = \frac{\sigma_\beta \sigma_X}{\langle f' \rangle} f\left(\frac{\vec{\mathbf{x}} \cdot \vec{\boldsymbol{\beta}}}{\sigma_X \sigma_\beta}\right) \tag{S4}$$

where the function f is an arbitrary nonlinear function and $\langle f' \rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}h e^{-\frac{h^2}{2}} f'(h)$ with prime notation used to indicate a derivative. We assume the ground truth parameters are independent of all other random parameters and are also normally distributed with zero mean and variance σ_β^2,

$$\mathrm{E}[\beta_k] = 0, \qquad \mathrm{Cov}[\beta_j, \beta_k] = \sigma_\beta^2 \delta_{jk}. \tag{S5}$$

We consider a training set of M data points, $\mathcal{D} = \{(y_b, \vec{\mathbf{x}}_b)\}_{b=1}^M$. We organize each input feature vector into the rows of an observation matrix X of size $M \times N_f$.

We consider a linear student model,

$$\hat{y}(\vec{\mathbf{x}}) = \vec{\mathbf{z}}(\vec{\mathbf{x}}) \cdot \hat{\mathbf{w}}, \tag{S6}$$

where $\hat{\mathbf{w}}$ is a vector of N_p fit parameters. The values of the fit parameters are determined by minimizing the loss function

$$L(\hat{\mathbf{w}}) = \frac{1}{2}\sum_b \Delta y_b^2 + \frac{\lambda}{2}\sum_K \hat{w}_K^2, \tag{S7}$$

where we have defined the residual label error as $\Delta y_a = y_a - \hat{y}_a$. Taking the gradient of the loss with respect to the fit parameters and setting it to zero results in a system of N_p equations for the N_p fit parameter,

$$0 = \frac{\partial L(\hat{\mathbf{w}})}{\partial \hat{w}_J} = -\sum_b \Delta y_b Z_{bJ} + \lambda \hat{w}_J. \tag{S8}$$

Note that the regularization term ensures that this system of equations always has a unique solution.

C. Central Limit Approximation

Frequently in these derivations, we encounter large sums of statistically independent random variables. In the thermodynamic limit, we utilize the central limit theorem to approximate these sums as a single random variable defined by only a mean and a variance. Here, we derive expressions for the approximate forms of three different types of sums that will be needed. In the following derivations N and N' are considered to be thermodynamically large variables.

First, we define a length-N vector $\vec{\mathbf{a}}$ of random variables a_j that are normally distributed with zero mean and variance σ_a^2/N,

$$\mathrm{E}[a_j] = 0, \qquad \mathrm{Cov}[a_j, a_k] = \frac{\sigma_a^2}{N}\delta_{jk}. \tag{S9}$$

The first sum we approximate is the dot product $\vec{\mathbf{c}}\cdot\vec{\mathbf{a}}$ where $\vec{\mathbf{c}}$ is a length-N vector with elements c_j that are independent of $\vec{\mathbf{a}}$. In the thermodynamic limit, this sum approximates to

$$\sum_k c_k a_k \approx \sigma z, \qquad \sigma^2 = \frac{\sigma_a^2}{N}\sum_k c_k^2, \tag{S10}$$

where z is a normally distributed variable with zero mean and unit variance. To derive this we simply evaluate the mean and variance of this sum with respect to $\vec{\mathbf{a}}$,

$$\mathrm{E}_{\vec{\mathbf{a}}}\left[\sum_k c_k a_k\right] = \sum_k x_k \mathrm{E}_{\vec{\mathbf{a}}}[a_k] = 0 \tag{S11}$$

$$\mathrm{Var}_{\vec{\mathbf{a}}}\left[\sum_k c_k a_k\right] = \sum_k c_k^2 \mathrm{Var}_{\vec{\mathbf{a}}}[a_k] = \frac{\sigma_a^2}{N}\sum_k c_k^2. \tag{S12}$$

The second sum we consider is the product $\vec{\mathbf{a}}^T A\vec{\mathbf{a}}$, where A is an $N \times N$ matrix whose elements are independent of $\vec{\mathbf{a}}$,

$$\sum_{jk} A_{jk} a_j a_k \approx \frac{\sigma_a^2}{N}\sum_k A_{kk}. \tag{S13}$$

To derive this, we evaluate the mean of this sum with respect to $\vec{\mathbf{a}}$ to be

$$\mathrm{E}_{\vec{\mathbf{a}}}\left[\sum_{jk} A_{jk} a_j a_k\right] = \sum_{jk} A_{jk}\mathrm{E}_{\vec{\mathbf{a}}}[a_j a_k] = \frac{\sigma_a^2}{N}\sum_k A_{kk}. \tag{S14}$$

To calculate the variance, we use Wick's theorem to derive the fourth moment of the elements of $\vec{\mathbf{a}}$,

$$\mathrm{E}[a_j a_k a_l a_m] = \sigma_a^4(\delta_{jk}\delta_{lm} + \delta_{jl}\delta_{km} + \delta_{jm}\delta_{kl}). \tag{S15}$$

Applying this identity, we find the variance to be

$$\begin{aligned}\mathrm{Var}_{\vec{\mathbf{a}}}\left[\sum_{jk} A_{jk}a_j a_k\right] &= \sum_{jklm} A_{jk}A_{lm}\mathrm{Cov}_{\vec{\mathbf{a}}}[a_j a_k, a_l a_m] \\ &= \sum_{jklm} A_{jk}A_{lm}(\mathrm{E}_{\vec{\mathbf{a}}}[a_j a_k a_l a_m] - \mathrm{E}_{\vec{\mathbf{a}}}[a_j a_k]\mathrm{E}_{\vec{\mathbf{a}}}[a_l a_m]) \\ &= 2\frac{\sigma_a^4}{N^2}\sum_{jk} A_{jk}^2 \\ &= 2\frac{\sigma_a^4}{N^2}\sum_{i} \sigma_i^2 \\ &\approx \mathcal{O}\Big(\frac{1}{N}\Big).\end{aligned} \tag{S16}$$

In the second-to-last line, we have rewritten the trace over $A^T A$ in terms of the eigenvalues σ_i of A. If each eigenvalue is $\mathcal{O}(1)$, then the total variance is $\mathcal{O}(1/N)$ and can be neglected.

For the third sum, we define an additional vector $\vec{\mathbf{b}}$ of length N' whose elements b_J are independent of $\vec{\mathbf{a}}$ with zero mean and variance σ_b^2/N',

$$\mathrm{E}[b_J] = 0, \qquad \mathrm{Cov}[b_J, b_K] = \frac{\sigma_b^2}{N'}\delta_{JK}. \tag{S17}$$

The third sum we approximate is the product $\vec{\mathbf{a}}^T B\vec{\mathbf{b}}$, where B is a $N \times N'$ rectangular matrix whose elements are independent of both $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$,

$$\sum_{jK} B_{jK}a_j b_K \approx 0. \tag{S18}$$

To derive this, we take the mean with respect to both $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$,

$$\mathrm{E}_{\vec{\mathbf{a}},\vec{\mathbf{b}}}\left[\sum_{jK} B_{jK}a_j b_K\right] = \sum_{jK} B_{jK}\mathrm{E}_{\vec{\mathbf{a}}}[a_j]\mathrm{E}_{\vec{\mathbf{b}}}[b_K] = 0, \tag{S19}$$

and also evaluate the variance to be

$$\begin{aligned}\mathrm{Var}_{\vec{\mathbf{a}},\vec{\mathbf{b}}}\left[\sum_{jK} B_{jK}a_j b_K\right] &= \sum_{jKlM} B_{jK}B_{lM}\mathrm{Cov}_{\vec{\mathbf{a}},\vec{\mathbf{b}}}[a_j b_K, a_l b_M] \\ &= \sum_{jKlM} B_{jK}B_{lM}\big(\mathrm{E}_{\vec{\mathbf{a}}}[a_j a_l]\mathrm{E}_{\vec{\mathbf{b}}}[b_K b_M] - \mathrm{E}_{\vec{\mathbf{a}}}[a_j]\mathrm{E}_{\vec{\mathbf{b}}}[b_K]\mathrm{E}_{\vec{\mathbf{a}}}[a_l]\mathrm{E}_{\vec{\mathbf{b}}}[b_M]\big) \\ &= \frac{\sigma_a^2\sigma_b^2}{NN'}\sum_{jK} B_{jK}^2 \\ &= \frac{\sigma_a^2\sigma_b^2}{NN'}\sum_{i} \sigma_i^2 \\ &\approx \mathcal{O}\Big(\frac{1}{N}\Big).\end{aligned} \tag{S20}$$

Analogous to the variance of the previous sum, in the second-to-last line we have decomposed B in terms of its singular values σ_i. If each singular value of $\mathcal{O}(1)$, then the total variance is $\mathcal{O}(1/N)$ and can be neglected. Since the mean is also zero, we find that all sums of this form can be ignored in the thermodynamic limit.

D. Nonlinear Function Statistics

Here, we show how the labels and hidden features can each be decomposed into linear and nonlinear components that are statistically independent of one another. We also derive the statistical properties of the resulting nonlinear components.

1. Integral Identities

First, we derive some useful integral identities for the expectation values of the nonlinear functions encountered in this work. In this section, we consider input features that are correlated, but collectively follow a multivariate normal distribution with mean zero and covariance matrix $\Sigma_{\vec{\mathbf{x}}}$,

$$\mathrm{E}[\vec{\mathbf{x}}] = 0, \qquad \mathrm{Cov}[\vec{\mathbf{x}}, \vec{\mathbf{x}}^T] = \Sigma_{\vec{\mathbf{x}}}, \tag{S21}$$

where the covariance is normalized so that $\mathrm{Tr}\,\Sigma_{\vec{\mathbf{x}}} = \sigma_X^2$. Throughout the rest this work, we usually consider independent input features where $\Sigma_{\vec{\mathbf{x}}} = \frac{\sigma_X^2}{N_f} I_{N_f}$. We also define a pair of random vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$, each of length N_f, whose elements are independent of $\vec{\mathbf{x}}$ with mean and variances

$$\begin{aligned} \mathrm{E}[a_j] = 0, \qquad & \mathrm{Cov}[a_j, a_k] = \sigma_a^2 \delta_{jk} \\ \mathrm{E}[b_j] = 0, \qquad & \mathrm{Cov}[b_j, b_k] = \sigma_b^2 \delta_{jk} \\ & \mathrm{Cov}[a_j, b_k] = 0. \end{aligned} \tag{S22}$$

Defining $g(h)$ as an arbitrary function and taking the thermodynamic limit, we will utilize the following three approximate identities:

$$\mathrm{E}_{\vec{\mathbf{x}}}\left[g\left(\frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{\sigma_X\sigma_a}\right)\right] \approx \langle g\rangle \tag{S23}$$

$$\mathrm{E}_{\vec{\mathbf{x}}}\left[\vec{\mathbf{x}} g\left(\frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{\sigma_X\sigma_a}\right)\right] \approx \frac{\Sigma_{\vec{\mathbf{x}}}\vec{\mathbf{a}}}{\sigma_X\sigma_a}\langle g'\rangle \tag{S24}$$

$$\mathrm{E}_{\vec{\mathbf{x}}}\left[g\left(\frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{\sigma_X\sigma_a}\right) g\left(\frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{b}}}{\sigma_X\sigma_b}\right)\right] \approx \langle g\rangle^2 \approx \mathrm{E}_{\vec{\mathbf{x}}}\left[g\left(\frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{\sigma_X\sigma_a}\right)\right]\mathrm{E}_{\vec{\mathbf{x}}}\left[g\left(\frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{b}}}{\sigma_X\sigma_b}\right)\right], \tag{S25}$$

where we have defined the integrals

$$\langle g\rangle = \frac{1}{\sqrt{2\pi}}\int \mathrm{d}h e^{-\frac{h^2}{2}} g(h), \qquad \langle g'\rangle = \frac{1}{\sqrt{2\pi}}\int \mathrm{d}h e^{-\frac{h^2}{2}} g'(h). \tag{S26}$$

Each of the above averages is evaluated with respect to the distribution of input features, where we define the differential over all elements of a vector of input features $\vec{\mathbf{x}}$ as,

$$\mathcal{D}\vec{\mathbf{x}} = \frac{\mathrm{d}\vec{\mathbf{x}}}{\sqrt{(2\pi)^{N_f}\det\Sigma_{\vec{\mathbf{x}}}}} e^{-\frac{1}{2}\vec{\mathbf{x}}^T\Sigma_{\vec{\mathbf{x}}}^{-1}\vec{\mathbf{x}}}. \tag{S27}$$

Next, we derive the identity in Eq. (S23),

$$\begin{aligned} \mathrm{E}_{\vec{\mathbf{x}}}\left[g\left(\frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{\sigma_X\sigma_a}\right)\right] &= \int \mathcal{D}\vec{\mathbf{x}} g\left(\frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{\sigma_X\sigma_a}\right) \\ &= \int \mathcal{D}\vec{\mathbf{x}}\mathrm{d}h g(h)\delta\left(h - \frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{\sigma_X\sigma_a}\right) \\ &= \int \mathcal{D}\vec{\mathbf{x}}\mathrm{d}h \frac{\mathrm{d}\tilde{h}}{2\pi} g(h) e^{i\tilde{h}\left(h - \frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{\sigma_X\sigma_a}\right)} \\ &= \int \mathrm{d}h \frac{\mathrm{d}\tilde{h}}{2\pi} g(h) e^{i\tilde{h}h} \int \frac{\mathrm{d}\vec{\mathbf{x}}}{\sqrt{(2\pi)^{N_f}\det\Sigma_{\vec{\mathbf{x}}}}} e^{-\frac{1}{2}\vec{\mathbf{x}}^T\Sigma_{\vec{\mathbf{x}}}^{-1}\vec{\mathbf{x}} - i\tilde{h}\frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{\sigma_X\sigma_a}} \\ &= \int \mathrm{d}h \frac{\mathrm{d}\tilde{h}}{2\pi} g(h) e^{i\tilde{h}h - \frac{\tilde{h}^2}{2}\frac{\vec{\mathbf{a}}^T\Sigma_{\vec{\mathbf{x}}}\vec{\mathbf{a}}}{\sigma_X^2\sigma_a^2}}. \end{aligned} \tag{S28}$$

At this point, we approximate the sum in the exponential using the central limit theorem. With proper rescaling, we apply Eq. (S13) to find

$$\frac{\vec{\mathbf{a}}^T\Sigma_{\vec{\mathbf{x}}}\vec{\mathbf{a}}}{\sigma_X^2\sigma_a^2} = \frac{1}{\sigma_X^2\sigma_a^2}\sum_{jk}(N_f\Sigma_{\vec{\mathbf{x}},jk})\left(\frac{a_j}{\sqrt{N_f}}\right)\left(\frac{a_k}{\sqrt{N_f}}\right) \approx 1, \tag{S29}$$

where we have identified $N_f \Sigma_{\vec{\mathbf{x}}}$ and $a_j/\sqrt{N_f}$ with A_{jk} and a_j, respectively, in Eq. (S10). Using this approximation, we proceed to find

$$\begin{aligned} \mathrm{E}_{\vec{\mathbf{x}}}\left[g\left(\frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{\sigma_X\sigma_a}\right)\right] &\approx \int \mathrm{d}h \frac{\mathrm{d}\tilde{h}}{2\pi} g(h) e^{i\tilde{h}h - \frac{\tilde{h}^2}{2}} \\ &= \frac{1}{\sqrt{2\pi}} \int \mathrm{d}h e^{-\frac{h^2}{2}} g(h) \\ &= \langle g \rangle . \end{aligned} \tag{S30}$$

To derive the remaining two identities in Eqs. (S24) and (S25), we follow analogous derivations. In particular, we note that Eq. (S25) implies that the two functions $g\left(\frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{\sigma_X\sigma_a}\right)$ and $g\left(\frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{b}}}{\sigma_X\sigma_b}\right)$ are statistically independent from one another in the thermodynamic limit.

2. *Label Decomposition*

By defining the ground truth parameters as shown below, we are able to decompose the labels into linear and nonlinear components,

$$y(\vec{\mathbf{x}}) = \vec{\mathbf{x}}\cdot\vec{\boldsymbol{\beta}} + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}) + \varepsilon, \qquad \vec{\boldsymbol{\beta}} \equiv \Sigma^{-1}_{\vec{\mathbf{x}}} \mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}, y^*(\vec{\mathbf{x}})], \tag{S31}$$

where $\delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}) \equiv y^*(\vec{\mathbf{x}}) - \vec{\mathbf{x}}\cdot\vec{\boldsymbol{\beta}}$ and the covariance matrix of the input features $\Sigma_{\vec{\mathbf{x}}} = \mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}, \vec{\mathbf{x}}^T]$ is assumed to be invertible.

We prove that the linear and nonlinear terms are statistically independent with respect the input features $\vec{\mathbf{x}}$ as follows:

$$\begin{aligned} \mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}\cdot\vec{\boldsymbol{\beta}}, \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}})] &= \mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}\cdot\vec{\boldsymbol{\beta}}, y(\vec{\mathbf{x}}) - \vec{\mathbf{x}}\cdot\vec{\boldsymbol{\beta}}] \\ &= \vec{\boldsymbol{\beta}}\cdot\mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}, y(\vec{\mathbf{x}})] - \vec{\boldsymbol{\beta}}\cdot\mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}, \vec{\mathbf{x}}^T]\vec{\boldsymbol{\beta}} \\ &= \vec{\boldsymbol{\beta}}\cdot\Sigma_{\vec{\mathbf{x}}}\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}\cdot\Sigma_{\vec{\mathbf{x}}}\vec{\boldsymbol{\beta}} \\ &= 0. \end{aligned} \tag{S32}$$

Furthermore, we can show that the ground truth parameters as defined in Eq. (S31) coincide with those of the teacher model,

$$y^*(\vec{\mathbf{x}}) = \frac{\sigma_\beta\sigma_X}{\langle f' \rangle} f\left(\frac{\vec{\mathbf{x}}\cdot\vec{\boldsymbol{\beta}}}{\sigma_X\sigma_\beta}\right). \tag{S33}$$

To do this, we evaluate the covariance in Eq. (S31) and use the identity in Eq. (S24) to find

$$\begin{aligned} \vec{\boldsymbol{\beta}} &= \Sigma^{-1}_{\vec{\mathbf{x}}}\mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}, y^*(\vec{\mathbf{x}})] \\ &= \Sigma^{-1}_{\vec{\mathbf{x}}}\frac{\sigma_\beta\sigma_X}{\langle f' \rangle}\mathrm{E}_{\vec{\mathbf{x}}}\left[\vec{\mathbf{x}} f\left(\frac{\vec{\mathbf{x}}\cdot\vec{\boldsymbol{\beta}}}{\sigma_X\sigma_\beta}\right)\right] \\ &\approx \Sigma^{-1}_{\vec{\mathbf{x}}}\frac{\sigma_\beta\sigma_X}{\langle f' \rangle}\frac{\Sigma_{\vec{\mathbf{x}}}\vec{\boldsymbol{\beta}}}{\sigma_X\sigma_\beta}\langle f' \rangle \\ &= \vec{\boldsymbol{\beta}}. \end{aligned} \tag{S34}$$

So we see that the definitions are consistent with one another.

Next, we calculate the variance of the nonlinear components of the labels $\delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}})$. To do this, we first calculate the mean of the squared true labels using the identity in Eq. (S23),

$$\mathrm{E}_{\vec{\mathbf{x}}}\left[(y^*(\vec{\mathbf{x}}))^2\right] = \frac{\sigma^2_\beta\sigma^2_X}{\langle f' \rangle^2}\mathrm{E}_{\vec{\mathbf{x}}}\left[f\left(\frac{\vec{\mathbf{x}}\cdot\vec{\boldsymbol{\beta}}}{\sigma_X\sigma_\beta}\right)^2\right] \approx \sigma^2_\beta\sigma^2_X\frac{\langle f^2 \rangle}{\langle f' \rangle^2}. \tag{S35}$$

Using this result and the fact that the linear and nonlinear components of the labels are independent, we find the variance of the nonlinear components in the thermodynamic limit to be

$$\mathrm{Var}_{\vec{\mathbf{x}}}[\delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}})] = \sigma_\beta^2 \sigma_X^2 \frac{\langle f^2 \rangle - \langle f' \rangle^2}{\langle f' \rangle^2}. \tag{S36}$$

Furthermore, it is clear that the nonlinear components for two independent data points $\vec{\mathbf{x}}_a$ and $\vec{\mathbf{x}}_b$ will also be independent.

Since there are no other random variables present in the above variance, we summarize the statistical properties of the nonlinear components of the labels for independent data points $\vec{\mathbf{x}}_a$ and $\vec{\mathbf{x}}_b$ with full ensemble averages, giving us

$$\mathrm{E}[\delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_a)] = 0, \qquad \mathrm{Cov}[\delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_a), \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_b)] = \sigma^2_{\delta y^*} \delta_{ab}, \tag{S37}$$

where we have defined the variance $\sigma^2_{\delta y^*}$ of the nonlinear components as

$$\begin{aligned} \sigma^2_{\delta y^*} &= \sigma_\beta^2 \sigma_X^2 \Delta f, \qquad \Delta f = \frac{\langle f^2 \rangle - \langle f' \rangle^2}{\langle f' \rangle^2} \\ \langle f^2 \rangle &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dh e^{-\frac{h^2}{2}} f^2(h), \qquad \langle f' \rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dh e^{-\frac{h^2}{2}} f'(h). \end{aligned} \tag{S38}$$

3. *Hidden Feature Decomposition*

Similar to the decomposition of the labels, we decompose the hidden features into linear and nonlinear components that are statistically independent with respect to the input features by defining W as follows:

$$\vec{\mathbf{z}}(\vec{\mathbf{x}}) = \frac{\mu_Z}{\sqrt{N_p}} \vec{\mathbf{1}} + W^T \vec{\mathbf{x}} + \delta \vec{\mathbf{z}}_{\mathrm{NL}}(\vec{\mathbf{x}}), \qquad W \equiv \Sigma^{-1}_{\vec{\mathbf{x}}} \mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}, \vec{\mathbf{z}}^T(\vec{\mathbf{x}})], \tag{S39}$$

where we have defined the nonlinear component as $\delta \vec{\mathbf{z}}_{\mathrm{NL}}(\vec{\mathbf{x}}) \equiv \vec{\mathbf{z}}(\vec{\mathbf{x}}) - \frac{\mu_Z}{\sqrt{N_p}} \vec{\mathbf{1}} - W^T \vec{\mathbf{x}}$. We have also defined the mean as $\mu_z / \sqrt{N_p} \vec{\mathbf{1}}$ where $\vec{\mathbf{1}}$ is a length-N_p vector of ones.

We prove that the linear and nonlinear terms are statistically independent with respect to the input features $\vec{\mathbf{x}}$ as follows:

$$\begin{aligned} \mathrm{Cov}_{\vec{\mathbf{x}}}[W^T \vec{\mathbf{x}}, \delta \vec{\mathbf{z}}_{\mathrm{NL}}(\vec{\mathbf{x}})^T] &= \mathrm{Cov}_{\vec{\mathbf{x}}}[W^T \vec{\mathbf{x}}, \vec{\mathbf{z}}^T(\vec{\mathbf{x}}) - \frac{\mu_Z}{\sqrt{N_f}} \vec{\mathbf{1}}^T - \vec{\mathbf{x}}^T W] \\ &= W^T \mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}, \vec{\mathbf{z}}^T(\vec{\mathbf{x}})] - \frac{\mu_Z}{\sqrt{N_p}} W^T \mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}, \vec{\mathbf{1}}^T] - W^T \mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}, \vec{\mathbf{x}}^T] W \\ &= W^T \Sigma_{\vec{\mathbf{x}}} W - W^T \Sigma_{\vec{\mathbf{x}}} W \\ &= 0. \end{aligned} \tag{S40}$$

We also show that W as defined in Eq. (S39) coincides with that in the definition of the hidden features,

$$\vec{\mathbf{z}}(\vec{\mathbf{x}}) = \frac{1}{\langle \varphi' \rangle} \frac{\sigma_W \sigma_X}{\sqrt{N_p}} \varphi\left(\frac{\sqrt{N_p}}{\sigma_W \sigma_X} W^T \vec{\mathbf{x}} \right). \tag{S41}$$

As in the previous section, we evaluate the covariance in Eq. (S39) and use the identity in Eq. (S24) to find

$$\begin{aligned} W &= \Sigma^{-1}_{\vec{\mathbf{x}}} \mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}, \vec{\mathbf{z}}^T(\vec{\mathbf{x}})] \\ &= \Sigma^{-1}_{\vec{\mathbf{x}}} \frac{1}{\langle \varphi' \rangle} \frac{\sigma_W \sigma_X}{\sqrt{N_p}} \mathrm{E}_{\vec{\mathbf{x}}}\left[\vec{\mathbf{x}} \varphi\left(\frac{\sqrt{N_p}}{\sigma_W \sigma_X} \vec{\mathbf{x}}^T W \right) \right] \\ &\approx \Sigma^{-1}_{\vec{\mathbf{x}}} \frac{1}{\langle \varphi' \rangle} \frac{\sigma_W \sigma_X}{\sqrt{N_p}} \frac{\sqrt{N_p}}{\sigma_W \sigma_X} \Sigma_{\vec{\mathbf{x}}} W \langle \varphi' \rangle \\ &= W. \end{aligned} \tag{S42}$$

So we see that the definitions are consistent with one another.

Next, we calculate the covariance of the nonlinear components of the hidden features $\delta\vec{\mathbf{z}}_{\mathrm{NL}}(\vec{\mathbf{x}})$ with respect to the full ensemble distribution. To do this, we first calculate the mean of each hidden feature using the identity in Eq. (S23),

$$\mathrm{E}_{\vec{\mathbf{x}}}[z_J(\vec{\mathbf{x}})] = \frac{1}{\langle\varphi'\rangle}\frac{\sigma_W\sigma_X}{\sqrt{N_p}}\mathrm{E}_{\vec{\mathbf{x}}}\left[\varphi\left(\frac{\sqrt{N_p}}{\sigma_W\sigma_X}\sum_k W_{kJ}x_k\right)\right] \approx \frac{\sigma_W\sigma_X}{\sqrt{N_p}}\frac{\langle\varphi\rangle}{\langle\varphi'\rangle}, \tag{S43}$$

from which we see that

$$\mu_z = \sigma_W\sigma_X\frac{\langle\varphi\rangle}{\langle\varphi'\rangle}. \tag{S44}$$

Similarly, we calculate the mean of the square of each hidden feature,

$$\mathrm{E}_{\vec{\mathbf{x}}}\left[z_J^2(\vec{\mathbf{x}})\right] = \frac{1}{\langle\varphi'\rangle^2}\frac{\sigma_W^2\sigma_X^2}{N_p}\mathrm{E}_{\vec{\mathbf{x}}}\left[\varphi^2\left(\frac{\sqrt{N_p}}{\sigma_W\sigma_X}\sum_k W_{kJ}x_k\right)\right] \approx \frac{\sigma_W^2\sigma_X^2}{N_p}\frac{\langle\varphi^2\rangle}{\langle\varphi'\rangle^2}. \tag{S45}$$

Using these two results and the independence of the linear and nonlinear components of the hidden features, we calculate the variance of the nonlinear component of each hidden feature to be

$$\mathrm{Var}_{\vec{\mathbf{x}}}[\delta z_{\mathrm{NL},J}(\vec{\mathbf{x}})] = \frac{\sigma_W^2\sigma_X^2}{N_p}\frac{\langle\varphi^2\rangle - \langle\varphi\rangle^2 - \langle\varphi'\rangle^2}{\langle\varphi'\rangle^2}. \tag{S46}$$

Finally, we calculate the mean of the product of two different hidden features $J \neq K$ for the same input features using the identity in Eq. (S25),

$$\mathrm{E}_{\vec{\mathbf{x}}}[z_J(\vec{\mathbf{x}})z_K(\vec{\mathbf{x}})] \approx \mathrm{E}_{\vec{\mathbf{x}}}[z_J(\vec{\mathbf{x}})]\mathrm{E}_{\vec{\mathbf{x}}}[z_K(\vec{\mathbf{x}})]. \tag{S47}$$

We observe that different hidden features are independent in the thermodynamic limit.

Since there are no other random variables present in any of the formulas, we summarize the statistical properties of the nonlinear components of the hidden features for independent data points $\vec{\mathbf{x}}_a$ and $\vec{\mathbf{x}}_b$ with full ensemble averages, giving us

$$\mathrm{E}[\delta z_{\mathrm{NL},J}(\vec{\mathbf{x}}_a)] = 0, \qquad \mathrm{Cov}[\delta z_{\mathrm{NL},J}(\vec{\mathbf{x}}_a), \delta z_{\mathrm{NL},K}(\vec{\mathbf{x}}_b)] = \frac{\sigma_{\delta z}^2}{N_P}\delta_{ab}\delta_{JK}, \tag{S48}$$

where we have defined the variance $\sigma_{\delta z}^2$ of the nonlinear components as

$$\begin{gathered}\sigma_{\delta z}^2 = \sigma_W^2\sigma_X^2\Delta\varphi, \qquad \Delta\varphi = \frac{\langle\varphi^2\rangle - \langle\varphi\rangle^2 - \langle\varphi'\rangle^2}{\langle\varphi'\rangle^2} \\ \langle\varphi\rangle = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dh e^{-\frac{h^2}{2}}\varphi(h), \qquad \langle\varphi^2\rangle = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dh e^{-\frac{h^2}{2}}\varphi^2(h), \qquad \langle\varphi'\rangle = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dh e^{-\frac{h^2}{2}}\varphi'(h).\end{gathered} \tag{S49}$$

E. General Solutions

Next, we derive the forms of the general solutions reported in Eq. (18)-(21) in the results section of the main text. First, we consider the training error. Recall this error takes the form

$$\mathcal{E}_{\mathrm{train}} = \frac{1}{M}\sum_b (\Delta y_b)^2. \tag{S50}$$

Taking the ensemble average, we express the training error as

$$\langle\mathcal{E}_{\mathrm{train}}\rangle = \langle\Delta y^2\rangle, \tag{S51}$$

where we have defined the mean of the squared label errors as

$$\langle \Delta y^2 \rangle = \mathrm{E}\left[\frac{1}{M} \sum_b \Delta y_b^2 \right]. \tag{S52}$$

Next, we consider the test error, bias and variance. To evaluate these quantities, we first decompose the predicted label for an arbitrary test data point $(y, \vec{\mathbf{x}})$ using the hidden feature decomposition in Eq. (S39),

$$\hat{y}(\vec{\mathbf{x}}) = \vec{\mathbf{x}} \cdot \hat{\boldsymbol{\beta}} + \frac{\mu_Z}{\sqrt{N_p}} \sum_J \hat{w}_J + \delta \vec{\mathbf{z}}_{\mathrm{NL}}(\vec{\mathbf{x}}) \cdot \hat{\mathbf{w}}, \tag{S53}$$

where we have defined the estimated ground truth parameters $\hat{\boldsymbol{\beta}} \equiv W \hat{\mathbf{w}}$. It is interesting to note that we can identify a definition for these parameters analagous to those of the ground truth parameters $\vec{\boldsymbol{\beta}}$ in Eq. (S31),

$$\hat{\boldsymbol{\beta}} \equiv \Sigma_{\vec{\mathbf{x}}}^{-1} \mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}, \hat{y}(\vec{\mathbf{x}})]. \tag{S54}$$

Because we consider data with labels that have zero mean, the label predictions should also have zero mean with respect to $\vec{\mathbf{x}}$. Therefore, if the first and last terms in Eq. (S53) each have zero mean with respect to $\vec{\mathbf{x}}$, the second term should evaluate to zero to ensure $\hat{y}(\vec{\mathbf{x}})$ overall has zero mean. For linear regression, this is clearly the case since $\mu_Z = 0$, but we will later prove that $\sum_J \hat{w}_J = 0$ for the random nonlinear features model. For now, we will neglect this term for the remainder of this section.

We evaluate the test error on a data set, $\mathcal{D}' = \{(y'_b, \vec{\mathbf{x}}'_b)\}_{b=1}^{M'}$, sampled independently from the same distribution as the training set. Recall that the test error is defined as

$$\mathcal{E}_{\mathrm{test}} = \frac{1}{M'} \sum_b (\Delta y'_b)^2, \tag{S55}$$

where the sum ranges from 1 to M'. We note that the residual label error of each test data point is described by the same distribution in the ensemble. Therefore, once we have taken the average over the test data, the test error can be expressed as an average over a single arbitrary test data point $(y, \vec{\mathbf{x}})$. Using this fact and applying the label decomposition in Eq. (S31) and the predicted label decomposition in Eq. (S53), we find

$$\begin{aligned} \mathrm{E}_{\mathcal{D}'}[\mathcal{E}_{\mathrm{test}}] &= \mathrm{E}_{(y,\vec{\mathbf{x}})}\left[\left(y(\vec{\mathbf{x}}) - \hat{y}(\vec{\mathbf{x}}) \right)^2 \right] \\ &= \mathrm{E}_{(y,\vec{\mathbf{x}})}\left[\left(\vec{\mathbf{x}} \cdot \Delta\vec{\boldsymbol{\beta}} - \delta \vec{\mathbf{z}}_{\mathrm{NL}}(\vec{\mathbf{x}}) \cdot \hat{\mathbf{w}} + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}) + \varepsilon \right)^2 \right] \\ &= \frac{\sigma_X^2}{N_f} \sum_k \Delta\beta_k^2 + \frac{\sigma_{\delta z}^2}{N_p} \sum_K \hat{w}_K^2 + \sigma_{\delta y^*}^2 + \sigma_\varepsilon^2. \end{aligned} \tag{S56}$$

Next, we apply the remainder of the ensemble average to find

$$\langle \mathcal{E}_{\mathrm{test}} \rangle = \sigma_X^2 \langle \Delta\beta^2 \rangle + \sigma_{\delta z}^2 \langle \hat{w}^2 \rangle + \sigma_{\delta y^*}^2 + \sigma_\varepsilon^2, \tag{S57}$$

where we have defined the average of the squared parameter errors and squared fit parameters, respectively, as

$$\langle \Delta\beta^2 \rangle = \mathrm{E}\left[\frac{1}{N_f} \sum_k \Delta\beta_k^2 \right], \qquad \langle \hat{w}^2 \rangle = \mathrm{E}\left[\frac{1}{N_p} \sum_K \hat{w}_K^2 \right]. \tag{S58}$$

Next, recall that the squared bias is defined as

$$\mathrm{Bias}^2[\hat{y}(\vec{\mathbf{x}})] = (\mathrm{E}_{\mathcal{D}}[\hat{y}(\vec{\mathbf{x}})] - y^*(\vec{\mathbf{x}}))^2. \tag{S59}$$

Note that averaging over $\mathcal{D}$ implies averaging over only the features X and noise $\vec{\boldsymbol{\varepsilon}}$ of the training set. In order to compute this average correctly, we make use of the following trick: we reinterpret the squared average over $\mathcal{D}$ as two separate averages over uncorrelated training data sets. Now, instead of a single regression problem trained on a single data set $\mathcal{D}$, we consider two separate regression problems each trained independently on different training sets, $\mathcal{D}_1$ and $\mathcal{D}_2$, drawn from the same distribution with the same ground truth parameters $\vec{\boldsymbol{\beta}}$. These regression problems

will also share all other random variables including the test data point $(y, \vec{\mathbf{x}})$, W, etc. This allows us to express the squared bias as

$$\text{Bias}^2[\hat{y}(\vec{\mathbf{x}})] = \text{E}_{\mathcal{D}_1,\mathcal{D}_2}[(y(\vec{\mathbf{x}}) - \hat{y}_1(\vec{\mathbf{x}}))(y(\vec{\mathbf{x}}) - \hat{y}_2(\vec{\mathbf{x}}))], \tag{S60}$$

where we use subscripts 1 and 2 to denote quantities that result from training on data sets $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. From here, we average over the test data point to find

$$\begin{aligned}\text{E}_{\vec{\mathbf{x}}}\big[\text{Bias}^2[\hat{y}(\vec{\mathbf{x}})]\big] &= \text{E}_{\vec{\mathbf{x}},\mathcal{D}_1,\mathcal{D}_2}[(\hat{y}_1(\vec{\mathbf{x}}) - y^*(\vec{\mathbf{x}}))(\hat{y}_2(\vec{\mathbf{x}}) - y^*(\vec{\mathbf{x}}))] \\ &= \text{E}_{\mathcal{D}_1,\mathcal{D}_2}\left[\frac{\sigma_X^2}{N_f}\sum_k \Delta\beta_{1,k}\Delta\beta_{2,k}\right] + \text{E}_{\mathcal{D}_1,\mathcal{D}_2}\left[\frac{\sigma_{\delta z}^2}{N_p}\sum_K \hat{w}_{1,K}\hat{w}_{2,K}\right] + \sigma_{\delta y^*}^2.\end{aligned} \tag{S61}$$

Next, we average over the remainder of the ensemble variables, giving us

$$\langle\text{Bias}^2[\hat{y}(\vec{\mathbf{x}})]\rangle = \sigma_X^2\,\langle\Delta\beta_1\Delta\beta_2\rangle + \sigma_{\delta z}^2\,\langle\hat{w}_1\hat{w}_2\rangle + \sigma_{\delta y^*}^2, \tag{S62}$$

where we have defined the quantities,

$$\langle\Delta\beta_1\Delta\beta_2\rangle = \text{E}\left[\frac{1}{N_f}\sum_k \Delta\beta_{1,k}\Delta\beta_{2,k}\right], \qquad \langle\hat{w}_1\hat{w}_2\rangle = \text{E}\left[\frac{1}{N_p}\sum_K \hat{w}_{1,K}\hat{w}_{2,K}\right]. \tag{S63}$$

To derive the expression for the ensemble-averaged variance, we simply make use of the bias-variance decomposition in Eq. (17), giving us

$$\begin{aligned}\langle\text{Var}[\hat{y}(\vec{\mathbf{x}})]\rangle &= \langle\mathcal{E}_{\text{test}}\rangle - \langle\text{Bias}^2[\hat{y}(\vec{\mathbf{x}})]\rangle - \text{Noise} \\ &= \sigma_X^2\big(\langle\Delta\beta^2\rangle - \langle\Delta\beta_1\Delta\beta_2\rangle\big) + \sigma_{\delta z}^2\big(\langle\hat{w}^2\rangle - \langle\hat{w}_1\hat{w}_2\rangle\big).\end{aligned} \tag{S64}$$

Based on these expressions, we find that the training error, test error, bias, and variance depend on five key ensemble-averaged quantities: $\langle\Delta y^2\rangle$, $\langle\Delta\beta^2\rangle$, $\langle\hat{w}^2\rangle$, $\langle\Delta\beta_1\Delta\beta_2\rangle$, and $\langle\hat{w}_1\hat{w}_2\rangle$.

F. Linear Regression (No Basis Functions)

In linear regression without basis functions, the hidden features are the same as the input features,

$$\vec{\mathbf{z}}(\vec{\mathbf{x}}) = \vec{\mathbf{x}}. \tag{S65}$$

Using this definition for the hidden features, we decompose the equation for the gradient in Eq. (S8) into three sets of equations for the fit parameters, residual label errors, and residual parameters errors,

$$\begin{aligned}\lambda\hat{w}_j &= \sum_b \Delta y_b X_{bj} + \eta_j \\ \Delta y_a &= \sum_k \Delta\beta_k X_{ak} + \delta y_{\text{NL}}^*(\vec{\mathbf{x}}_a) + \varepsilon_a + \xi_a \\ \Delta\beta_j &= \beta_j - \hat{w}_j,\end{aligned} \tag{S66}$$

where we have also utilized Eq. (S31) to decompose the training labels into linear and nonlinear components. We have also added small auxiliary fields η_j and ξ_a to the two equations containing sums. We will use these extra fields to define perturbations about the solutions to these equations with the intent of setting the fields to zero by the end of the derivation.

1. *Cavity Expansion*

Next, we add an additional variable of each type, resulting in a total of $M+1$ data points and N_f+1 features. We specify each new variable using an index value of 0, giving us the new unknown quantities $\hat{w}_0$, Δy_0, and $\Delta\beta_0$. These new variables result in the addition of an extra term in each sum, giving us the equations

$$\begin{aligned}\lambda\hat{w}_j &= \sum_b \Delta y_b X_{bj} + \eta_j + \Delta y_0 X_{0j} \\ \Delta y_a &= \sum_k \Delta\beta_k X_{ak} + \delta y_{\text{NL}}^*(\vec{\mathbf{x}}_a) + \varepsilon_a + \xi_a + \Delta\beta_0 X_{a0}.\end{aligned} \tag{S67}$$

Each new variable is also described by a new equation,

$$\begin{aligned}\lambda\hat{w}_0 &= \sum_b \Delta y_b X_{b0} + \eta_0 + \Delta y_0 X_{00} \\ \Delta y_0 &= \sum_k \Delta\beta_k X_{0k} + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_0) + \varepsilon_0 + \xi_0 + \Delta\beta_0 X_{00} \\ \Delta\beta_0 &= \beta_0 - \hat{w}_0.\end{aligned} \tag{S68}$$

As a reminder, sums always start at an index value of 1. Therefore, we explicitly specify any terms with an index value of 0.

Now we take the thermodynamic limit in which M and N_f tend towards infinity, but their ratio α_f remains fixed. In this limit, we can interpret the extra terms in Eq. (S67) as small perturbations to the auxiliary fields since they each contain an element of X which has mean zero and an infinitesimal variance of $\mathcal{O}(1/N_f)$,

$$\delta\eta_j = \Delta y_0 X_{0j}, \qquad \delta\xi_a = \Delta\beta_0 X_{a0}. \tag{S69}$$

This allows us to expand each variable about its solution in the absence of the 0-indexed quantities, corresponding to the solution for M data points and N_f features,

$$\begin{aligned}\hat{w}_j &\approx \hat{w}_{j\setminus 0} + \sum_k \nu^{\hat{w}}_{jk}\delta\eta_k + \sum_b \chi^{\hat{w}}_{jb}\delta\xi_b \\ \Delta y_a &\approx \Delta y_{a\setminus 0} + \sum_k \nu^{\Delta y}_{ak}\delta\eta_k + \sum_b \chi^{\Delta y}_{ab}\delta\xi_b \\ \Delta\beta_j &\approx \Delta\beta_{j\setminus 0} + \sum_k \nu^{\Delta\beta}_{jk}\delta\eta_k + \sum_b \chi^{\Delta\beta}_{jb}\delta\xi_b.\end{aligned} \tag{S70}$$

We use subscripts with $\setminus 0$ to refer to the unperturbed solutions for each unknown quantity; that is, the solutions in the absence of the 0-indexed variables. We also define the susceptibility matrices as the following derivatives with respect to the auxiliary fields:

$$\begin{aligned}\nu^{\hat{w}}_{jk} &= \frac{\partial\hat{w}_j}{\partial\eta_k}, & \chi^{\hat{w}}_{jb} &= \frac{\partial\hat{w}_j}{\partial\xi_b}, \\ \nu^{\Delta y}_{ak} &= \frac{\partial\Delta y_a}{\partial\eta_k}, & \chi^{\Delta y}_{ab} &= \frac{\partial\Delta y_a}{\partial\xi_b}, \\ \nu^{\Delta\beta}_{jk} &= \frac{\partial\Delta\beta_j}{\partial\eta_k}, & \chi^{\Delta\beta}_{jb} &= \frac{\partial\Delta\beta_j}{\partial\xi_b}.\end{aligned} \tag{S71}$$

It is useful to note that the susceptibilities for the residual parameter errors are related to those for the fit parameters via a negative sign,

$$\nu^{\Delta\beta}_{jk} = -\nu^{\hat{w}}_{jk}, \qquad \chi^{\Delta\beta}_{jb} = -\chi^{\hat{w}}_{jb}. \tag{S72}$$

Therefore, we replace all susceptibilities for the residual parameter errors with their fit parameter counterparts. Substituting the expansions in Eq. (S70) into the equations for the 0-indexed variables, Eq. (S68), we arrive at the following equations:

$$\begin{aligned}\lambda\hat{w}_0 &= \sum_a \left(\Delta y_{a\setminus 0} + \sum_k \nu^{\Delta y}_{ak}\delta\eta_k + \sum_b \chi^{\Delta y}_{ab}\delta\xi_b\right) X_{a0} + \eta_0 + X_{00}\Delta y_0 \\ \Delta y_0 &= \sum_j \left(\Delta\beta_{j\setminus 0} - \sum_k \nu^{\hat{w}}_{jk}\delta\eta_k - \sum_b \chi^{\hat{w}}_{jb}\delta\xi_b\right) X_{0j} + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_0) + \varepsilon_0 + \xi_0 + \Delta\beta_0 X_{00}.\end{aligned} \tag{S73}$$

Our next step is to simplify these equations by approximating the sums over large numbers of random variables.

2. *Central Limit Approximations*

Each of the sums in Eq. (S73) contains a thermodynamically large number of statistically uncorrelated terms. This means that each sum satisfies the conditions necessary to apply the central limit theorem, allowing us to express

each in terms of a single normally-distributed random variable described by just its mean and its variance. First, we approximate the two sums that contain one of the unperturbed unknown quantities, $\Delta y_{a\backslash 0}$ or $\Delta\beta_{j\backslash 0}$. In both sums, the unperturbed quantities are statistically independent of any elements of X with a 0-index such as X_{a0} or X_{0j}. Using this independence, we apply the identity in Eq. (S10) to find

$$\begin{aligned}
\sum_b \Delta y_{b\backslash 0} X_{b0} &\approx \sigma_{\hat{w}} z_{\hat{w}}, & \sigma_{\hat{w}}^2 &= \sigma_X^2 \alpha_f^{-1} \langle \Delta y^2 \rangle, & \langle \Delta y^2 \rangle &= \frac{1}{M} \sum_b \Delta y_{b\backslash 0}^2 \\
\sum_k \Delta \beta_{k\backslash 0} X_{0k} &\approx \sigma_{\Delta y} z_{\Delta y}, & \sigma_{\Delta y}^2 &= \sigma_X^2 \langle \Delta \beta^2 \rangle, & \langle \Delta \beta^2 \rangle &= \frac{1}{N_f} \sum_k \Delta \beta_{k\backslash 0}^2,
\end{aligned} \tag{S74}$$

where $\sigma_{\hat{w}}^2$ and $\sigma_{\Delta y}^2$ are the total variances of the two sums and $z_{\hat{w}}$ are $z_{\Delta y}$ are random variables with zero mean and unit variance. It is straightforward to show that $z_{\hat{w}}$ and $z_{\Delta y}$ are statistically independent since the two sums are independent with respect to the zero-indexed elements of X.

Note that we have used the same notation, $\langle \Delta y^2 \rangle$ and $\langle \Delta \beta^2 \rangle$, for the two averages that defined previously in Sec. S1 E even though they each lack an ensemble average. In doing so, we have made the ansatz that these sums will converge to their ensemble averages in the thermodynamic limit. This assumption is typical of the cavity method.

Next, we approximate the sums that include either of the square susceptibility matrices, $\chi_{ab}^{\Delta y}$ or $\nu_{jk}^{\hat{w}}$. Similar to the unperturbed unknown quantities, we use the property that the susceptibilities are statistically independent of the elements of X with 0-valued indices. Applying the identity in Eq. (S13), we find

$$\begin{aligned}
\sum_{ab} \chi_{ab}^{\Delta y} X_{a0} X_{b0} &\approx \sigma_X^2 \alpha_f^{-1} \chi, & \chi &= \frac{1}{M} \sum_b \chi_{bb}^{\Delta y} \\
\sum_{jk} \nu_{jk}^{\hat{w}} X_{0j} X_{0k} &\approx \sigma_X^2 \nu, & \nu &= \frac{1}{N_f} \sum_k \nu_{kk}^{\hat{w}},
\end{aligned} \tag{S75}$$

where χ and ν can be interpreted as a pair of scalar susceptibilities.

The remainder of the sums contain rectangular susceptibility matrices which follow the form in Eq. (S18). Therefore, each of these sums is expected to be small in the thermodynamic limit and can be neglected.

3. *Self-consistency Equations*

Applying the approximations from the previous section to Eq. (S73), we obtain the following set of self-consistency equations for the 0-indexed variables, $\hat{w}_0$, Δy_0, and $\Delta\beta_0$:

$$\begin{aligned}
\lambda \hat{w}_0 &\approx \sigma_{\hat{w}} z_{\hat{w}} + \Delta\beta_0 \sigma_X^2 \alpha_f^{-1} \chi + \eta_0 \\
\Delta y_0 &\approx \sigma_{\Delta y} z_{\Delta y} - \Delta y_0 \sigma_X^2 \nu + \delta y_{\mathrm{NL}}^*(\vec{\mathbf{x}}_0) + \varepsilon_0 + \xi_0 \\
\Delta\beta_0 &\approx \beta_0 - \hat{w}_0.
\end{aligned} \tag{S76}$$

In these equations, we have also dropped terms proportional to X_{00} since this quantity has zero mean and a variance which goes to zero in the thermodynamic limit. Next, we solve these three equations for the 0-indexed variables, giving us

$$\begin{aligned}
\hat{w}_0 &= \frac{\beta_0 \sigma_X^2 \alpha_f^{-1} \chi + \sigma_{\hat{w}} z_{\hat{w}} + \eta_0}{\lambda + \sigma_X^2 \alpha_f^{-1} \chi} \\
\Delta y_0 &= \frac{\sigma_{\Delta y} z_{\Delta y} + \delta y_{\mathrm{NL}}^*(\vec{\mathbf{x}}_0) + \varepsilon_0 + \xi_0}{1 + \sigma_X^2 \nu} \\
\Delta\beta_0 &= \frac{\beta_0 \lambda - \sigma_{\hat{w}} z_{\hat{w}} - \eta_0}{\lambda + \sigma_X^2 \alpha_f^{-1} \chi}.
\end{aligned} \tag{S77}$$

Note that all random variables within each of the above equations are statistically independent from one another.

Next, we make the approximation that in the thermodynamic limit, each of the unknown quantities is "self-averaging." In other words, we assume that an average over a set of non-0-indexed variables is equivalent to taking an ensemble average of the single corresponding 0-indexed variable.

This allows us to use Eq. (S77) to find a set of self-consistent equations for the scalar susceptibilities by evaluating the appropriate derivatives with respect to the 0-indexed auxiliary fields and performing ensemble averages,

$$\begin{aligned}\chi &= \frac{1}{M}\sum_b \chi_{bb}^{\Delta y} \approx \mathrm{E}\left[\chi_{00}^{\Delta y}\right] = \mathrm{E}\left[\frac{\partial \Delta y_0}{\partial \xi_0}\right] = \frac{1}{1+\sigma_X^2\nu} \\ \nu &= \frac{1}{N_f}\sum_k \nu_{kk}^{\hat{w}} \approx \mathrm{E}\left[\nu_{00}^{\hat{w}}\right] = \mathrm{E}\left[\frac{\partial \hat{w}_0}{\partial \eta_0}\right] = \frac{1}{\lambda + \sigma_X^2\alpha_f^{-1}\chi}.\end{aligned} \tag{S78}$$

Furthermore, we find the following self-consistent equations for the quantities, $\langle \hat{w}^2 \rangle$, $\langle \Delta y^2 \rangle$, and $\langle \Delta\beta^2 \rangle$ by taking the appropriate expectation values of the 0-indexed quantities, plugging in the forms of the scalar susceptibilities, and setting the auxiliaries fields to zero:

$$\begin{aligned}\langle \hat{w}^2 \rangle &= \frac{1}{N_f}\sum_k \hat{w}_{k\backslash 0}^2 \approx \mathrm{E}\left[\hat{w}_0^2\right] = \nu^2\left(\sigma_\beta^2\sigma_X^4\alpha_f^{-2}\chi^2 + \sigma_X^2\alpha_f^{-1}\langle \Delta y^2 \rangle\right) \\ \langle \Delta y^2 \rangle &= \frac{1}{M}\sum_b \Delta y_{b\backslash 0}^2 \approx \mathrm{E}\left[\Delta y_0^2\right] = \chi^2\left(\sigma_X^2\langle \Delta\beta^2 \rangle + \sigma_{\delta y^*}^2 + \sigma_\varepsilon^2\right) \\ \langle \Delta\beta^2 \rangle &= \frac{1}{N_f}\sum_k \Delta\beta_{k\backslash 0}^2 \approx \mathrm{E}\left[\Delta\beta_0^2\right] = \nu^2\left(\sigma_\beta^2\lambda^2 + \sigma_X^2\alpha_f^{-1}\langle \Delta y^2 \rangle\right).\end{aligned} \tag{S79}$$

Note that we have also defined the mean squared fit parameter size $\langle \hat{w}^2 \rangle$. In addition, each of the three mean squared quantities can be interpreted as a full ensemble average. These self-consistent equations, along with those for the scalar susceptibilities, capture almost all behavior of our model of linear regression in the thermodynamic limit.

4. *Solutions with Finite Regularization ($\lambda \sim \mathcal{O}(1)$)*

Next, we derive the solutions when the regularization parameter λ is finite. By combing the two scalar susceptibilities in Eq. (S78), we find a quadratic equation for χ,

$$\chi^2 + \left[(\alpha_f - 1) + \bar{\lambda}\alpha_f\right]\chi - \bar{\lambda}\alpha_f = 0, \tag{S80}$$

where we have defined the dimensionless regularization parameter

$$\bar{\lambda} = \frac{\lambda}{\sigma_X^2}. \tag{S81}$$

Solving Eq. (S80), we find two solutions:

$$\chi = \frac{1}{2}\left[1 - \alpha_f\left(1+\bar{\lambda}\right) \pm \sqrt{\left[1-\alpha_f\left(1+\bar{\lambda}\right)\right]^2 + 4\alpha_f\bar{\lambda}}\right]. \tag{S82}$$

Using these solutions we can also find similar solutions for ν. Next, we solve Eq. (S79) to find closed-form solutions for $\langle \hat{w}^2 \rangle$, $\langle \Delta y^2 \rangle$, and $\langle \Delta\beta^2 \rangle$:

$$\begin{pmatrix} \langle \hat{w}^2 \rangle \\ \langle \Delta y^2 \rangle \\ \langle \Delta\beta^2 \rangle \end{pmatrix} = \begin{pmatrix} 1 & -\sigma_X^2\alpha_f^{-1}\nu^2 & 0 \\ 0 & 1 & -\sigma_X^2\chi^2 \\ 0 & -\sigma_X^2\alpha_f^{-1}\nu^2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_\beta^2\sigma_X^4\alpha_f^{-2}\chi^2\nu^2 \\ (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)\chi^2 \\ \bar{\lambda}^2\sigma_\beta^2\sigma_X^2\nu^2 \end{pmatrix}. \tag{S83}$$

In combination with the solutions for χ and ν, these solutions are exact in the thermodynamic limit.

5. *Solutions in Ridge-less Limit ($\lambda \to 0$)*

In order to make the solutions in the previous section easier to interpret, we take the ridge-less limit where $\lambda \to 0$. Based on the form of Eq. (S80), we make the ansatz that the lowest order contribution to χ is $\mathcal{O}(1)$ in small $\bar{\lambda}$. Accordingly, we expand χ in small $\bar{\lambda}$ up to $\mathcal{O}(\bar{\lambda})$ as

$$\chi \approx \chi_0 + \bar{\lambda}\chi_1. \tag{S84}$$

Substituting this approximation into the formula for χ in Eq. (S80), we find the following equation at $\mathcal{O}(1)$:

$$0 = \chi_0^2 + (\alpha_f - 1)\chi_0. \tag{S85}$$

Solving this equation, we find two solutions for χ_0,

$$\chi_0^{(1)} = 1 - \alpha_f, \qquad \chi_0^{(2)} = 0. \tag{S86}$$

We label each set of solutions for all quantities with a superscript (1) or (2). These two solutions correspond to the two solutions in the exact formula for χ in Eq. (S82). Next, we collect terms in Eq. (S80) at $\mathcal{O}(\bar{\lambda})$,

$$0 = 2\chi_0\chi_1 + (\alpha_f - 1)\chi_1 + \alpha_f(\chi_0 - 1). \tag{S87}$$

Solving, we obtain an equation for χ_1 in terms of χ_0,

$$\chi_1 = \frac{\alpha_f(1 - \chi_0)}{2\chi_0 + \alpha_f - 1}. \tag{S88}$$

Combining this equation with $\chi_0^{(2)} = 0$, we obtain the leading order term for solution (2),

$$\chi_1^{(2)} = \frac{\alpha_f}{\alpha_f - 1}. \tag{S89}$$

Next, we solve for the two solutions for ν. By inspecting the equation for ν in terms of χ in Eq. (S78), we make the ansatz that the lowest contribution of ν is $\mathcal{O}(1/\lambda)$,

$$\nu \approx \frac{1}{\bar{\lambda}}\nu_{-1} + \nu_0. \tag{S90}$$

Substituting the solutions for χ_0 and χ_1 into the equation for ν, we find that the solutions for ν_{-1} are

$$\nu_{-1}^{(1)} = 0, \qquad \nu_{-1}^{(2)} = \frac{1}{\sigma_X^2 + \frac{\sigma_X^2}{\alpha_f}\chi_1} = \frac{1}{\sigma_X^2}\frac{\alpha_f - 1}{\alpha_f}. \tag{S91}$$

Since $\nu_{-1}^{(1)}$ is zero, we also solve for the next order term for solution (1),

$$\nu_0^{(1)} = \frac{1}{\sigma_X^2}\frac{\alpha_f}{(1 - \alpha_f)}. \tag{S92}$$

For completion, we also find that we can continue with this procedure to derive

$$\nu_0^{(2)} = \frac{1}{\sigma_X^2}\frac{1}{(\alpha_f - 1)}. \tag{S93}$$

We also expand each of the ensemble-averaged quantities $\langle\hat{w}^2\rangle$, $\langle\Delta y^2\rangle$, and $\langle\Delta\beta^2\rangle$ in small $\bar{\lambda}$. We make the ansatz that each of these quantities is $\mathcal{O}(1)$ to lowest order with the next terms in the expansion at $\mathcal{O}(\bar{\lambda}^2)$:

$$\begin{aligned}
\langle\hat{w}^2\rangle &\approx \langle\hat{w}^2\rangle_0 + \bar{\lambda}^2\langle\hat{w}^2\rangle_2 \\
\langle\Delta y^2\rangle &\approx \langle\Delta y^2\rangle_0 + \bar{\lambda}^2\langle\Delta y^2\rangle_2 \\
\langle\Delta\beta^2\rangle &\approx \langle\Delta\beta^2\rangle_0 + \bar{\lambda}^2\langle\Delta\beta^2\rangle_2 .
\end{aligned} \tag{S94}$$

Solution (1): For the first set of solutions, the self-consistent equations, Eq. (S79), to lowest order, are

$$\begin{aligned}
\langle\hat{w}^2\rangle_0^{(1)} &= (\nu_0^{(1)})^2\left[\sigma_\beta^2\sigma_X^4\alpha_f^{-2}(\chi_0^{(1)})^2 + \sigma_X^2\alpha_f^{-1}\langle\Delta y^2\rangle_0^{(1)}\right] \\
\langle\Delta y^2\rangle_0^{(1)} &= (\chi_0^{(1)})^2\left[\sigma_X^2\langle\Delta\beta^2\rangle_0^{(1)} + \sigma_{\delta y^*}^2 + \sigma_\varepsilon^2\right] \\
\langle\Delta\beta^2\rangle_0^{(1)} &= (\nu_0^{(1)})^2\sigma_X^2\alpha_f^{-1}\langle\Delta y^2\rangle_0^{(1)} .
\end{aligned} \tag{S95}$$

Substituting the solutions for the susceptibilities into these equations and solving, we find

$$\begin{aligned}
\langle\hat{w}^2\rangle_0^{(1)} &= \sigma_\beta^2 + \frac{1}{\sigma_X^2}(\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)\frac{\alpha_f}{(1 - \alpha_f)} \\
\langle\Delta y^2\rangle_0^{(1)} &= (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)(1 - \alpha_f) \\
\langle\Delta\beta^2\rangle_0^{(1)} &= \frac{1}{\sigma_X^2}(\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)\frac{\alpha_f}{(1 - \alpha_f)}.
\end{aligned} \tag{S96}$$

Solution (2): For the second set of solutions, the self-consistent equations to lowest order, are

$$\begin{aligned}\langle \hat{w}^2 \rangle_0^{(2)} &= (\nu_{-1}^{(2)})^2 \Big[\sigma_\beta^2 \sigma_X^4 \alpha_f^{-2} (\chi_1^{(2)})^2 + \sigma_X^2 \alpha_f^{-1} \langle \Delta y^2 \rangle_2^{(2)} \Big] \\ \langle \Delta y^2 \rangle_0^{(2)} &= 0 \\ \langle \Delta \beta^2 \rangle_0^{(2)} &= (\nu_{-1}^{(2)})^2 \Big[\sigma_\beta^2 \sigma_X^4 + \sigma_X^2 \alpha_f^{-1} \langle \Delta y^2 \rangle_2^{(2)} \Big].\end{aligned} \tag{S97}$$

We see that we also need the solution for $\langle \Delta y^2 \rangle$ to next lowest order,

$$\langle \Delta y^2 \rangle_2^{(2)} = (\chi_1^{(2)})^2 \Big[\sigma_X^2 \langle \Delta \beta^2 \rangle_0^{(2)} + \sigma_{\delta y^*}^2 + \sigma_\varepsilon^2 \Big]. \tag{S98}$$

Substituting the solutions for the susceptibilities into these equations and solving, we find

$$\begin{aligned}\langle \hat{w}^2 \rangle_0^{(2)} &= \sigma_\beta^2 \frac{1}{\alpha_f} + \frac{1}{\sigma_X^2} (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{1}{(\alpha_f - 1)} \\ \langle \Delta y^2 \rangle_2^{(2)} &= \sigma_\beta^2 \sigma_X^2 \frac{\alpha_f}{(\alpha_f - 1)} + (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{\alpha_f^3}{(\alpha_f - 1)^3} \\ \langle \Delta \beta^2 \rangle_0^{(2)} &= \sigma_\beta^2 \frac{(\alpha_f - 1)}{\alpha_f} + \frac{1}{\sigma_X^2} (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{1}{(\alpha_f - 1)}.\end{aligned} \tag{S99}$$

Combined solutions: To determine when each of the two solutions applies, we use the fact that each of the ensemble-averaged quantities $\langle \hat{w}^2 \rangle$, $\langle \Delta y^2 \rangle$, and $\langle \Delta \beta^2 \rangle$ must always be positive by definition. Imposing this constraint, we find that solution (1) only applies when $\alpha_f < 1$, while solution (2) only applies when $\alpha_f > 1$. Combining these solutions, we arrive at the final forms for the three ensemble-averaged quantities in the $\lambda \to 0$ limit,

$$\begin{aligned}\langle \hat{w}^2 \rangle &= \begin{cases} \sigma_\beta^2 + \frac{(\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)}{\sigma_X^2} \frac{\alpha_f}{(1 - \alpha_f)} & \text{if} \quad N_f < M \\ \sigma_\beta^2 \frac{1}{\alpha_f} + \frac{(\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)}{\sigma_X^2} \frac{1}{(\alpha_f - 1)} & \text{if} \quad N_f > M \end{cases} \\ \langle \Delta y^2 \rangle &= \begin{cases} (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)(1 - \alpha_f) & \text{if} \quad N_f < M \\ \frac{\lambda^2}{\sigma_X^4} \Big[\sigma_\beta^2 \sigma_X^2 \frac{\alpha_f}{(\alpha_f - 1)} + (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{\alpha_f^3}{(\alpha_f - 1)^3} \Big] & \text{if} \quad N_f > M \end{cases} \\ \langle \Delta \beta^2 \rangle &= \begin{cases} \frac{(\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)}{\sigma_X^2} \frac{\alpha_f}{(1 - \alpha_f)} & \text{if} \quad N_f < M \\ \sigma_\beta^2 \frac{(\alpha_f - 1)}{\alpha_f} + \frac{(\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)}{\sigma_X^2} \frac{1}{(\alpha_f - 1)} & \text{if} \quad N_f > M. \end{cases}\end{aligned} \tag{S100}$$

For completeness, we also report solutions for the two scalar susceptibilities,

$$\chi = \begin{cases} 1 - \alpha_f & \text{if} \quad N_f < M \\ \frac{\lambda}{\sigma_X^2} \frac{\alpha_f}{(\alpha_f - 1)} & \text{if} \quad N_f > M \end{cases}, \qquad \nu = \begin{cases} \frac{1}{\sigma_X^2} \frac{\alpha_f}{(1 - \alpha_f)} & \text{if} \quad N_f < M \\ \frac{1}{\lambda} \frac{(\alpha_f - 1)}{\alpha_f} + \frac{1}{\sigma_X^2} \frac{1}{(\alpha_f - 1)} & \text{if} \quad N_f > M. \end{cases} \tag{S101}$$

Finally, we use these expressions to derive the training and test error according to the general solutions in Eqs. (S51) and (S57),

$$\langle \mathcal{E}_{\text{train}} \rangle = \begin{cases} (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)(1 - \alpha_f) & \text{if} \quad N_f < M \\ \frac{\lambda^2}{\sigma_X^4} \Big[\sigma_\beta^2 \sigma_X^2 \frac{\alpha_f}{(\alpha_f - 1)} + (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{\alpha_f^3}{(\alpha_f - 1)^3} \Big] & \text{if} \quad N_f > M \end{cases} \tag{S102}$$

$$\langle \mathcal{E}_{\text{test}} \rangle = \begin{cases} (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{1}{(1 - \alpha_f)} & \text{if} \quad N_f < M \\ \sigma_\beta^2 \sigma_X^2 \frac{(\alpha_f - 1)}{\alpha_f} + (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{\alpha_f}{(\alpha_f - 1)} & \text{if} \quad N_f > M. \end{cases} \tag{S103}$$

6. *Bias-Variance Decomposition*

Next, we derive the bias and variance. According to the general solutions in Eqs. (S62) and (S64), we only require the quantity $\langle \Delta\beta_1 \Delta\beta_2 \rangle$ since $\sigma_{\delta z}^2 = 0$. To find $\langle \Delta\beta_1 \Delta\beta_2 \rangle$, we use the formula for $\Delta\beta_0$ in Eq. (S77) to characterize its behavior when trained separately on two independent training sets, $\mathcal{D}_1$ and $\mathcal{D}_2$,

$$\begin{aligned}\Delta\beta_{1,0} &= \nu(\beta_0 \lambda - \sigma_{\hat{w}} z_{\hat{w}_1}) \\ \Delta\beta_{2,0} &= \nu(\beta_0 \lambda - \sigma_{\hat{w}} z_{\hat{w}_2}).\end{aligned} \tag{S104}$$

As a reminder, we use subscripts 1 and 2 to denote quantities that depend on one of the training sets. Note that while the random variables $z_{\hat{w}_1}$ and $z_{\hat{w}_2}$ are defined separately for the two regression problems, both equations share the same β_0. Multiplying these two equations together and using the self-averaging approximation, we find an expression for $\langle \Delta\beta_1 \Delta\beta_2 \rangle$,

$$\langle \Delta\beta_1 \Delta\beta_2 \rangle = \frac{1}{N_f} \sum_k \Delta\beta_{1,k} \Delta\beta_{2,} \approx \mathrm{E}[\Delta\beta_{1,0} \Delta\beta_{2,0}] = \nu^2 \big(\sigma_\beta^2 \lambda^2 + \mathrm{E}\big[\sigma_{\hat{w}}^2 z_{\hat{w}_1} z_{\hat{w}_2} \big] \big). \tag{S105}$$

Next, we calculate the expectation value of the product $z_{\hat{w}_1} z_{\hat{w}_2}$ and find that it evaluates to zero as a result of the statistical independence of the two design matrices, X_1 and X_2,

$$\begin{aligned} \mathrm{E}\big[\sigma_{\hat{w}}^2 z_{\hat{w}_1} z_{\hat{w}_2} \big] &\approx \mathrm{E}\left[\sum_{ab} \Delta y_{1,a\setminus 0} \Delta y_{2,b\setminus 0} X_{1,a0} X_{2,b0} \right] \\ &= \sum_{ab} \mathrm{E}\big[\Delta y_{1,a\setminus 0} \Delta y_{2,b\setminus 0} \big] \mathrm{E}[X_{1,a0} X_{2,b0}] \\ &= 0. \end{aligned} \tag{S106}$$

Substituting this solution into Eq. (S105), we find an expression for $\langle \Delta\beta_1 \Delta\beta_2 \rangle$,

$$\langle \Delta\beta_1 \Delta\beta_2 \rangle = \sigma_\beta^2 \lambda^2 \nu^2. \tag{S107}$$

Inserting the solutions for ν, we find in the $\lambda \to 0$ limit that

$$\langle \Delta\beta_1 \Delta\beta_2 \rangle = \begin{cases} \frac{\lambda^2}{\sigma_X^2} \sigma_\beta^2 \frac{\alpha_f^2}{(1-\alpha_f)^2} & \text{if} \quad N_f < M \\ \sigma_\beta^2 \frac{(\alpha_f - 1)^2}{\alpha_f^2} & \text{if} \quad N_f > M. \end{cases} \tag{S108}$$

From this, we arrive at the final expressions for the model bias and variance in the $\lambda \to 0$ limit,

$$\begin{aligned} \langle \mathrm{Bias}^2[\hat{y}(\vec{\mathbf{x}})] \rangle &= \begin{cases} \sigma_{\delta y^*}^2 & \text{if} \quad N_f < M \\ \sigma_\beta^2 \sigma_X^2 \frac{(\alpha_f - 1)^2}{\alpha_f^2} + \sigma_{\delta y^*}^2 & \text{if} \quad N_f > M \end{cases} \\ \langle \mathrm{Var}[\hat{y}(\vec{\mathbf{x}})] \rangle &= \begin{cases} (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{\alpha_f}{(1-\alpha_f)} & \text{if} \quad N_f < M \\ \sigma_\beta^2 \sigma_X^2 \frac{(\alpha_f - 1)}{\alpha_f^2} + (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{1}{(\alpha_f - 1)} & \text{if} \quad N_f > M. \end{cases} \end{aligned} \tag{S109}$$

G. Random Nonlinear Features Model (Two-layer Nonlinear Neural Network)

In the random nonlinear features model, the student model takes the form

$$\vec{\mathbf{z}}(\vec{\mathbf{x}}) = \frac{1}{\langle \varphi' \rangle} \frac{\sigma_W \sigma_X}{\sqrt{N_p}} \varphi\left(\frac{\sqrt{N_p}}{\sigma_W \sigma_X} W^T \vec{\mathbf{x}} \right), \tag{S110}$$

where the elements of the random transformation matrix W are identically and independently distributed, drawn from a normal distribution with zero mean and variance σ_W^2 / N_p,

$$\mathrm{E}[W_{jJ}] = 0, \qquad \mathrm{Cov}[W_{jJ}, W_{kK}] = \frac{\sigma_W^2}{N_p} \delta_{jk} \delta_{JK}. \tag{S111}$$

We also assume that the elements of W are statistically independent of the ground truth parameters $\vec{\beta}$, the label noise $\vec{\varepsilon}$, the features X, etc.

For this model, we again decompose the equation for the gradient in Eq. (S62) for the above set of hidden features. To perform the cavity method, our aim is to construct a set of equations that are linear in the random matrices W

and X. This results in four different sets of linear equations,

$$\begin{aligned}
\lambda \hat{w}_J &= \sqrt{M}\alpha_p^{-\frac{1}{2}}\mu_z \langle \Delta y \rangle + \sum_k \hat{u}_k W_{kJ} + \sum_b \Delta y_b \delta z_{\mathrm{NL},J}(\vec{\mathbf{x}}_b) + \eta_J \\
\hat{u}_j &= \sum_b \Delta y_b X_{bj} + \psi_j \\
\Delta y_a &= -\sqrt{N_p}\mu_z \langle \hat{w} \rangle + \sum_k \Delta\beta_k X_{ak} - \sum_K \hat{w}_K \delta z_{\mathrm{NL},K}(\vec{\mathbf{x}}_a) + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_a) + \varepsilon_a + \xi_a \\
\Delta\beta_j &= \beta_j - \sum_K \hat{w}_K W_{jK} + \zeta_j,
\end{aligned} \tag{S112}$$

where we have decomposed the training labels according to Eq. (S31) and the hidden features according to Eq. (S39). We have also added a different auxiliary field to each equation and have defined the means of the residual label errors and fit parameter, respectively, as

$$\langle \Delta y \rangle = \frac{1}{M}\sum_b \Delta y_b, \qquad \langle \hat{w} \rangle = \frac{1}{N_p}\sum_K \hat{w}_K. \tag{S113}$$

Furthermore, we have included an additional set of variables $\hat{\mathbf{u}} = X^T \Delta\vec{\mathbf{y}}$, which we will also have to solve for to obtain closed form solutions.

1. *Cavity Expansion*

Next, we add an additional variable of each type, resulting in a total of $M+1$ data points, N_f+1 input features and N_p+1 fit parameters/hidden features. Each additional variable is represented using an index value of 0, written as $\hat{w}_0$, $\hat{u}_0$, Δy_0, and $\Delta\beta_0$. After including these new unknown quantities, the four equations become

$$\begin{aligned}
\lambda \hat{w}_J &= \sqrt{M}\alpha_p^{-\frac{1}{2}}\mu_z \langle \Delta y \rangle + \sum_k \hat{u}_k W_{kJ} + \sum_b \Delta y_b \delta z_{\mathrm{NL},J}(\vec{\mathbf{x}}_b) + \eta_J + \hat{u}_0 W_{0J} + \Delta y_0 \delta z_{\mathrm{NL},J}(\vec{\mathbf{x}}_0) \\
\hat{u}_j &= \sum_b \Delta y_b X_{bj} + \psi_j + \Delta y_0 X_{0j} \\
\Delta y_a &= -\sqrt{N_p}\mu_z \langle \hat{w} \rangle + \sum_k \Delta\beta_k X_{ak} - \sum_K \hat{w}_K \delta z_{\mathrm{NL},K}(\vec{\mathbf{x}}_a) + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_a) + \varepsilon_a + \xi_a + \Delta\beta_0 X_{a0} - \hat{w}_0 \delta z_{\mathrm{NL},0}(\vec{\mathbf{x}}_a) \\
\Delta\beta_j &= \beta_j - \sum_K \hat{w}_K W_{jK} + \zeta_j - \hat{w}_0 W_{j0},
\end{aligned} \tag{S114}$$

with each new variable described by a new equation,

$$\begin{aligned}
\lambda \hat{w}_0 &= \sqrt{M}\alpha_p^{-\frac{1}{2}}\mu_z \langle \Delta y \rangle + \sum_k \hat{u}_k W_{k0} + \sum_b \Delta y_b \delta z_{\mathrm{NL},0}(\vec{\mathbf{x}}_b) + \eta_J + \hat{u}_0 W_{00} + \Delta y_0 \delta z_{\mathrm{NL},0}(\vec{\mathbf{x}}_0) \\
\hat{u}_0 &= \sum_b \Delta y_b X_{b0} + \psi_0 + \Delta y_0 X_{00} \\
\Delta y_0 &= -\sqrt{N_p}\mu_z \langle \hat{w} \rangle + \sum_k \Delta\beta_k X_{0k} - \sum_K \hat{w}_K \delta z_{\mathrm{NL},K}(\vec{\mathbf{x}}_0) + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_0) + \varepsilon_0 + \xi_0 + \Delta\beta_0 X_{00} - \hat{w}_0 \delta z_{\mathrm{NL},0}(\vec{\mathbf{x}}_0) \\
\Delta\beta_0 &= \beta_0 - \sum_K \hat{w}_K W_{0K} + \zeta_0 - \hat{w}_0 W_{00}.
\end{aligned} \tag{S115}$$

Now we take the thermodynamic limit in which M, N_f, and N_p tend towards infinity, but their ratios, $\alpha_f = N_f/M$ and $\alpha_p = N_p/M$, remain fixed. We interpret the extra terms in Eq. (S114) as small perturbations to the auxiliary fields,

$$\begin{aligned}
\delta\eta_J &= \hat{u}_0 W_{0J} + \Delta y_0 \delta z_{\mathrm{NL},J}(\vec{\mathbf{x}}_0), \qquad & \delta\psi_j &= \Delta y_0 X_{0j}, \\
\delta\xi_a &= \Delta\beta_0 X_{a0} - \hat{w}_0 \delta z_{\mathrm{NL},0}(\vec{\mathbf{x}}_a), \qquad & \delta\zeta_j &= -\hat{w}_0 W_{j0},
\end{aligned} \tag{S116}$$

allowing us to expand each unknown quantity about its solution in the absence of the 0-indexed variables, which correspond to the solutions for M data points, N_f input features, and N_p fit parameters,

$$\begin{aligned}
\hat{w}_J &\approx \hat{w}_{J\backslash 0} + \sum_K \nu_{JK}^{\hat{w}} \delta\eta_K + \sum_k \phi_{Jk}^{\hat{w}} \delta\psi_k + \sum_b \chi_{Jb}^{\hat{w}} \delta\xi_b + \sum_k \omega_{Jk}^{\hat{w}} \delta\zeta_k \\
\hat{u}_j &\approx \hat{u}_{j\backslash 0} + \sum_K \nu_{jK}^{\hat{u}} \delta\eta_K + \sum_k \phi_{jk}^{\hat{u}} \delta\psi_k + \sum_b \chi_{jb}^{\hat{u}} \delta\xi_b + \sum_k \omega_{jk}^{\hat{u}} \delta\zeta_k \\
\Delta y_a &\approx \Delta y_{a\backslash 0} + \sum_K \nu_{aK}^{\Delta y} \delta\eta_K + \sum_k \phi_{ak}^{\Delta y} \delta\psi_k + \sum_b \chi_{ab}^{\Delta y} \delta\xi_b + \sum_k \omega_{ak}^{\Delta y} \delta\zeta_k \\
\Delta\beta_j &\approx \Delta\beta_{j\backslash 0} + \sum_K \nu_{jK}^{\Delta\beta} \delta\eta_K + \sum_k \phi_{jk}^{\Delta\beta} \delta\psi_k + \sum_b \chi_{jb}^{\Delta\beta} \delta\xi_b + \sum_k \omega_{jk}^{\Delta\beta} \delta\zeta_k.
\end{aligned} \tag{S117}$$

We define each of the susceptibility matrices as a derivative of a variable with respect to an auxiliary fields,

$$\begin{aligned}
\nu_{JK}^{\hat{w}} &= \frac{\partial \hat{w}_J}{\partial \eta_K}, & \phi_{Jk}^{\hat{w}} &= \frac{\partial \hat{w}_J}{\partial \psi_k}, & \chi_{Jb}^{\hat{w}} &= \frac{\partial \hat{w}_J}{\partial \xi_b}, & \omega_{Jk}^{\hat{w}} &= \frac{\partial \hat{w}_J}{\partial \zeta_k}, \\
\nu_{jK}^{\hat{u}} &= \frac{\partial \hat{u}_j}{\partial \eta_K}, & \phi_{jk}^{\hat{u}} &= \frac{\partial \hat{u}_j}{\partial \psi_k}, & \chi_{jb}^{\hat{u}} &= \frac{\partial \hat{u}_j}{\partial \xi_b}, & \omega_{jk}^{\hat{u}} &= \frac{\partial \hat{u}_j}{\partial \zeta_k}, \\
\nu_{aK}^{\Delta y} &= \frac{\partial \Delta y_a}{\partial \eta_K}, & \phi_{ak}^{\Delta y} &= \frac{\partial \Delta y_a}{\partial \psi_k}, & \chi_{ab}^{\Delta y} &= \frac{\partial \Delta y_a}{\partial \xi_b}, & \omega_{ak}^{\Delta y} &= \frac{\partial \Delta y_a}{\partial \zeta_k}, \\
\nu_{jK}^{\Delta\beta} &= \frac{\partial \Delta\beta_j}{\partial \eta_K}, & \phi_{jk}^{\Delta\beta} &= \frac{\partial \Delta\beta_j}{\partial \psi_k}, & \chi_{jb}^{\Delta\beta} &= \frac{\partial \Delta\beta_j}{\partial \xi_b}, & \omega_{jk}^{\Delta\beta} &= \frac{\partial \Delta\beta_j}{\partial \zeta_k}.
\end{aligned} \tag{S118}$$

Next, we substitute the expansions in Eq. (S117) into the 0-indexed equations in Eq. (S115). We then aim to approximate each of the resulting sums in these expanded equations.

2. *Central Limit Approximations*

We approximate each of the sums containing one of the unperturbed quantities, $\hat{w}_{J\backslash 0}$, $\hat{u}_{j\backslash 0}$, $\Delta y_{a\backslash 0}$, or $\Delta\beta_{j\backslash 0}$, using the central limit theorem. Because the unperturbed quantities in each of these sums are statistically independent of all elements of both X and W with a 0-valued index, we are able to apply the identity in Eq. (S10) to find

$$\begin{aligned}
\sum_k \hat{u}_{k\backslash 0} W_{k0} + \sum_b \Delta y_{b\backslash 0} \delta z_{\mathrm{NL},0}(\vec{\mathbf{x}}_b) &\approx \sigma_{\hat{w}} z_{\hat{w}}, & \sigma_{\hat{w}}^2 &= \sigma_W^2 \frac{\alpha_f}{\alpha_p} \langle \hat{u}^2 \rangle + \sigma_{\delta z}^2 \alpha_p^{-1} \langle \Delta y^2 \rangle \\
\sum_b \Delta y_{b\backslash 0} X_{b0} &\approx \sigma_{\hat{u}} z_{\hat{u}}, & \sigma_{\hat{u}}^2 &= \sigma_X^2 \alpha_f^{-1} \langle \Delta y^2 \rangle \\
\sum_k \Delta\beta_{k\backslash 0} X_{0k} - \sum_K \hat{w}_{K\backslash 0} \delta z_{\mathrm{NL},K}(\vec{\mathbf{x}}_0) &\approx \sigma_{\Delta y} z_{\Delta y}, & \sigma_{\Delta y}^2 &= \sigma_X^2 \langle \Delta\beta^2 \rangle + \sigma_{\delta z}^2 \langle \hat{w}^2 \rangle \\
\sum_K \hat{w}_{K\backslash 0} W_{0K} &\approx \sigma_{\Delta\beta} z_{\Delta\beta}, & \sigma_{\Delta\beta}^2 &= \sigma_W^2 \langle \hat{w}^2 \rangle,
\end{aligned} \tag{S119}$$

where $z_{\hat{w}}$, $z_{\hat{u}}$, $z_{\Delta y}$, and $z_{\Delta\beta}$ are all independent random variables with zero mean and unit variance. We also define the following averages:

$$\langle \hat{w}^2 \rangle = \frac{1}{N_p} \sum_K \hat{w}_{K\backslash 0}^2, \qquad \langle \hat{u}^2 \rangle = \frac{1}{N_f} \sum_k \hat{u}_{k\backslash 0}^2, \qquad \langle \Delta y^2 \rangle = \frac{1}{M} \sum_b \Delta y_{b\backslash 0}^2, \qquad \langle \Delta\beta^2 \rangle = \frac{1}{N_f} \sum_k \Delta\beta_{k\backslash 0}^2. \tag{S120}$$

Next, we approximate each of the sums containing a square susceptibility matrix. Using the fact that all of the susceptibility matrices are statistically independent of all elements of both X, W, and $\delta\vec{\mathbf{z}}_{\mathrm{NL}}(\vec{\mathbf{x}})$ with a 0-valued index,

we apply the identity in Eq. (S13) to find

$$\begin{aligned}
\sum_{jk} \omega_{jk}^{\hat{u}} W_{j0} W_{k0} &\approx \sigma_W^2 \frac{\alpha_f}{\alpha_p} \omega, & \omega &= \frac{1}{N_f} \sum_k \omega_{kk}^{\hat{u}} \\
\sum_{ab} \chi_{ab}^{\Delta y} X_{a0} X_{b0} &\approx \sigma_X^2 \alpha_f^{-1} \chi, & \chi &= \frac{1}{M} \sum_b \chi_{bb}^{\Delta y} \\
\sum_{jk} \phi_{jk}^{\Delta\beta} X_{0j} X_{0k} &\approx \sigma_X^2 \phi, & \phi &= \frac{1}{N_f} \sum_k \phi_{kk}^{\Delta\beta} \\
\sum_{JK} \nu_{JK}^{\hat{w}} W_{0J} W_{0K} &\approx \sigma_W^2 \nu, & \nu &= \frac{1}{N_p} \sum_K \nu_{KK}^{\hat{w}}.
\end{aligned} \tag{S121}$$

Similarly, we approximate the two additional sums containing the nonlinear components of hidden features,

$$\begin{aligned}
\sum_{JK} \nu_{JK}^{\hat{w}} \delta z_{\mathrm{NL},J}(\vec{\mathbf{x}}_0) \delta z_{\mathrm{NL},K}(\vec{\mathbf{x}}_0) &\approx \sigma_{\delta z}^2 \nu \\
\sum_{ab} \chi_{ab}^{\Delta y} \delta z_{\mathrm{NL},0}(\vec{\mathbf{x}}_a) \delta z_{\mathrm{NL},0}(\vec{\mathbf{x}}_b) &\approx \sigma_{\delta z}^2 \alpha_p^{-1} \chi,
\end{aligned} \tag{S122}$$

with $\sigma_{\delta z}$ defined in Eq. (S49).

Finally, each of the remaining sums contains a rectangular susceptibility matrix, so according to the identity in Eq. (S18), is approximately zero in the thermodynamic limit.

3. *Self-consistency Equations*

Next, we substitute the expansions in Eq. (S117) into Eq. (S115) and apply the approximations from the previous sections, resulting in a set of self-consistent equations for $\hat{w}_0$, $\hat{u}_0$, Δy_0, and $\Delta\beta_0$,

$$\begin{aligned}
\lambda \hat{w}_0 &\approx \sqrt{M} \alpha_p^{-\frac{1}{2}} \mu_z \langle y \rangle + \sigma_{\hat{w}} z_{\hat{w}} - \hat{w}_0 \left(\sigma_W^2 \frac{\alpha_f}{\alpha_p} \omega + \sigma_{\delta z}^2 \alpha_p^{-1} \chi \right) + \eta_0 \\
\hat{u}_0 &\approx \sigma_{\hat{u}} z_{\hat{u}} + \Delta\beta_0 \sigma_X^2 \alpha_f^{-1} \chi + \psi_0 \\
\Delta y_0 &\approx -\sqrt{N_p} \mu_z \langle \hat{w} \rangle + \sigma_{\Delta y} z_{\Delta y} + \Delta y_0 \left(\sigma_X^2 \phi - \sigma_{\delta z}^2 \nu \right) + \delta y_{\mathrm{NL}}^*(\vec{\mathbf{x}}_0) + \varepsilon_0 + \xi_0 \\
\Delta\beta_0 &\approx \beta_0 - \sigma_{\Delta\beta} z_{\Delta\beta} - \hat{u}_0 \sigma_W^2 \nu + \zeta_0.
\end{aligned} \tag{S123}$$

We have also made use of the fact that the terms including X_{00} or W_{00} are infinitesimally small in the thermodynamic limit with zero mean and variances of $\mathcal{O}(1/N_f)$ and $\mathcal{O}(1/N_p)$, respectively. Solving these equations for the 0-indexed variables, we find

$$\begin{aligned}
\hat{w}_0 &= \frac{\sqrt{M} \alpha_p^{-\frac{1}{2}} \mu_z \langle y \rangle + \sigma_{\hat{w}} z_{\hat{w}} + \eta_0}{\lambda + \sigma_W^2 \frac{\alpha_f}{\alpha_p} \omega + \sigma_{\delta z}^2 \alpha_p^{-1} \chi} \\
\hat{u}_0 &= \frac{\sigma_{\hat{u}} z_{\hat{u}} + \psi_0 + \sigma_X^2 \alpha_f^{-1} \chi (\beta_0 - \sigma_{\Delta\beta} z_{\Delta\beta} + \zeta_0)}{1 + \sigma_W^2 \sigma_X^2 \alpha_f^{-1} \chi \nu} \\
\Delta y_0 &= \frac{-\sqrt{N_p} \mu_z \langle \hat{w} \rangle + \sigma_{\Delta y} z_{\Delta y} + \delta y_{\mathrm{NL}}^*(\vec{\mathbf{x}}_0) + \varepsilon_0 + \xi_0}{1 - \sigma_X^2 \phi + \sigma_{\delta z}^2 \nu} \\
\Delta\beta_0 &= \frac{\beta_0 - \sigma_{\Delta\beta} z_{\Delta\beta} + \zeta_0 - \sigma_W^2 \nu^2 (\sigma_{\hat{u}} z_{\hat{u}} + \psi_0)}{1 + \sigma_W^2 \sigma_X^2 \alpha_f^{-1} \chi \nu}.
\end{aligned} \tag{S124}$$

We then derive a set of self-consistent equations for the scalar susceptibilities by taking appropriate derivatives of

these variables with respect to the auxiliary fields,

$$\begin{aligned}
\nu &= \frac{1}{N_p}\sum_K \nu_{KK}^{\hat{w}} \approx \mathrm{E}\left[\nu_{00}^{\hat{w}}\right] = \mathrm{E}\left[\frac{\partial \hat{w}_0}{\partial \eta_0}\right] = \frac{1}{\lambda + \sigma_W^2 \frac{\alpha_f}{\alpha_p}\omega + \sigma_{\delta z}^2 \alpha_p^{-1}\chi} \\
\omega &= \frac{1}{N_f}\sum_k \omega_{kk}^{\hat{u}} \approx \mathrm{E}\left[\omega_{00}^{\hat{u}}\right] = \mathrm{E}\left[\frac{\partial \hat{u}_0}{\partial \zeta_0}\right] = \frac{\sigma_X^2 \alpha_f^{-1}\chi}{1 + \sigma_W^2 \sigma_X^2 \alpha_f^{-1}\chi\nu} \\
\chi &= \frac{1}{M}\sum_b \chi_{bb}^{\Delta y} \approx \mathrm{E}\left[\chi_{00}^{\Delta y}\right] = \mathrm{E}\left[\frac{\partial \Delta y_0}{\partial \xi_0}\right] = \frac{1}{1 - \sigma_X^2\phi + \sigma_{\delta z}^2\nu} \\
\phi &= \frac{1}{N_f}\sum_k \phi_{kk}^{\Delta\beta} \approx \mathrm{E}\left[\phi_{00}^{\Delta\beta}\right] = \mathrm{E}\left[\frac{\partial \Delta\beta_0}{\partial \psi_0}\right] = -\frac{\sigma_W^2\nu}{1 + \sigma_W^2\sigma_X^2\alpha_f^{-1}\chi\nu}.
\end{aligned} \tag{S125}$$

For convenience, we also introduce a fifth scalar susceptibility,

$$\kappa = \frac{1}{N_f}\sum_k \phi_{kk}^{\hat{u}} = \frac{1}{N_f}\sum_k \omega_{kk}^{\Delta\beta} \approx \mathrm{E}\left[\frac{\partial \hat{u}_0}{\partial \psi_0}\right] = \mathrm{E}\left[\frac{\partial \Delta\beta_0}{\partial \zeta_0}\right] = \frac{1}{1 + \sigma_W^2\sigma_X^2\alpha_f^{-1}\chi\nu}. \tag{S126}$$

Using this formula for κ, we re-express the four other susceptibilities as

$$\begin{aligned}
\omega &= \sigma_X^2 \alpha_f^{-1}\chi\kappa \\
\phi &= -\sigma_W^2\nu\kappa \\
\nu &= \frac{1}{\lambda + \sigma_W^2\sigma_X^2\alpha_p^{-1}\chi(\kappa + \Delta\varphi)} \\
\chi &= \frac{1}{1 + \sigma_W^2\sigma_X^2\nu(\kappa + \Delta\varphi)}.
\end{aligned} \tag{S127}$$

Next, we find self-consistent equations for the averages of the fit parameter and residual label errors,

$$\begin{aligned}
\langle \hat{w} \rangle &= \frac{1}{N_p}\sum_K \hat{w}_K \approx \mathrm{E}[\hat{w}_0] = \nu\sqrt{M}\alpha_p^{-\frac{1}{2}}\mu_z \langle \Delta y \rangle \\
\langle \Delta y \rangle &= \frac{1}{M}\sum_b \Delta y_b \approx \mathrm{E}[\Delta y_0] = -\chi\sqrt{N_p}\mu_z \langle \hat{w} \rangle ,
\end{aligned} \tag{S128}$$

where we have set the auxiliary fields to zero. Solving these equations, it is clear that both averages are zero,

$$\langle \hat{w} \rangle = 0, \qquad \langle \Delta y \rangle = 0. \tag{S129}$$

Finally, we square and average each of Eq. (S124) to find self-consistent equations for the four ensemble-averaged squared quantities (again setting the auxiliary fields to zero),

$$\begin{aligned}
\langle \hat{w}^2 \rangle &= \frac{1}{N_p}\sum_K \hat{w}_{K\backslash 0}^2 \approx \mathrm{E}\left[\hat{w}_0^2\right] = \nu^2\left(\sigma_W^2 \frac{\alpha_f}{\alpha_p}\langle \hat{u}^2 \rangle + \sigma_{\delta z}^2 \alpha_p^{-1}\langle \Delta y^2 \rangle\right) \\
\langle \hat{u}^2 \rangle &= \frac{1}{N_f}\sum_k \hat{u}_{k\backslash 0}^2 \approx \mathrm{E}\left[\hat{u}_0^2\right] = \kappa^2\sigma_X^2\alpha_f^{-1}\langle \Delta y^2 \rangle + \omega^2\left(\sigma_\beta^2 + \sigma_W^2\langle \hat{w}^2 \rangle\right) \\
\langle \Delta y^2 \rangle &= \frac{1}{M}\sum_b \Delta y_{b\backslash 0}^2 \approx \mathrm{E}\left[\Delta y_0^2\right] = \chi^2\left(\sigma_X^2\langle \Delta\beta^2 \rangle + \sigma_{\delta z}^2\langle \hat{w}^2 \rangle + \sigma_{\delta y^*}^2 + \sigma_\varepsilon^2\right) \\
\langle \Delta\beta^2 \rangle &= \frac{1}{N_f}\sum_k \Delta\beta_{k\backslash 0}^2 \approx \mathrm{E}\left[\Delta\beta_0^2\right] = \kappa^2\left(\sigma_\beta^2 + \sigma_W^2\langle \hat{w}^2 \rangle\right) + \phi^2\sigma_X^2\alpha_f^{-1}\langle \Delta y^2 \rangle .
\end{aligned} \tag{S130}$$

4. *Solution with Finite Regularization ($\lambda \sim \mathcal{O}(1)$)*

We start by solving the equations for χ and ν in Eq. (S127) for κ and setting them equal,

$$\kappa = \frac{1 - \chi - \sigma_W^2\sigma_X^2\Delta\varphi\nu}{\sigma_W^2\sigma_X^2\chi\nu} = \frac{\alpha_p(1 - \lambda\nu) - \sigma_W^2\sigma_X^2\Delta\varphi\nu}{\sigma_W^2\sigma_X^2\chi\nu}, \tag{S131}$$

giving us a relation between ν and χ,

$$\nu = \frac{\chi + \alpha_p - 1}{\lambda \alpha_p}. \tag{S132}$$

Substituting κ from Eq. (S126) into χ from Eq. (S127), inserting the expression for ν we just found, and then multiplying out the denominators, we find a quartic equation for χ,

$$\begin{aligned} 0 = {} & \Delta\varphi\chi^4 + \left[2\Delta\varphi(\alpha_p - 1) + \alpha_p\bar{\lambda}\right]\chi^3 + \left[\Delta\varphi(\alpha_p - 1)^2 + ((1 + \Delta\varphi)\alpha_f + \alpha_p - 2)\alpha_p\bar{\lambda}\right]\chi^2 \\ & + \left[((1 + \Delta\varphi)\alpha_f - 1)(\alpha_p - 1) + \alpha_f\alpha_p\bar{\lambda}\right]\alpha_p\bar{\lambda}\chi - \alpha_f\alpha_p^2\bar{\lambda}^2, \end{aligned} \tag{S133}$$

where we have defined the dminesionless regularization parameter

$$\bar{\lambda} = \frac{\lambda}{\sigma_W^2\sigma_X^2}. \tag{S134}$$

Solving the quartic equation and solving for the remaining susceptibilities, we find exact solutions in the thermodynamic limit by solving Eq. (S130),

$$\begin{pmatrix} \langle\hat{w}^2\rangle \\ \langle\hat{u}^2\rangle \\ \langle\Delta y^2\rangle \\ \langle\Delta\beta^2\rangle \end{pmatrix} = \begin{pmatrix} 1 & -\sigma_W^2\frac{\alpha_f}{\alpha_p}\nu^2 & -\sigma_{\delta z}^2\alpha_p^{-1}\nu^2 & 0 \\ -\sigma_W^2\omega^2 & 1 & -\sigma_X^2\alpha_f^{-1}\kappa^2 & 0 \\ -\sigma_{\delta z}^2\chi^2 & 0 & 1 & -\sigma_X^2\chi^2 \\ -\sigma_W^2\kappa^2 & 0 & -\sigma_X^2\alpha_f^{-1}\phi^2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \sigma_\beta^2\omega^2 \\ (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)\chi^2 \\ \sigma_\beta^2\kappa^2 \end{pmatrix}. \tag{S135}$$

5. *Solutions in Ridge-less Limit ($\lambda \to 0$)*

In the ridge-less limit ($\lambda \to 0$), we make the ansatz that χ is $\mathcal{O}(1)$ in small $\bar{\lambda}$,

$$\chi \approx \chi_0 + \bar{\lambda}\chi_1. \tag{S136}$$

Using this approximation, Eq. (S133) gives us the following equation at $\mathcal{O}(1)$:

$$0 = \Delta\varphi\chi_0^4 + 2\Delta\varphi(\alpha_p - 1)\chi_0^3 + \Delta\varphi(\alpha_p - 1)^2\chi_0^2. \tag{S137}$$

This equation has two solutions for χ_0,

$$\chi_0^{(1)} = 1 - \alpha_p, \qquad \chi_0^{(2)} = 0, \tag{S138}$$

labeled by superscript (1) and (2).

At $\mathcal{O}(\bar{\lambda})$, we find the resulting equation to be uninformative after inserting either of the solutions for χ_0. However, the $\mathcal{O}(\bar{\lambda}^2)$ equation does provide unique solutions,

$$\begin{aligned} 0 = {} & \Delta\varphi\left(4\chi_0^3\chi_2 + 6\chi_0^2\chi_1^2\right) + 2\Delta\varphi(\alpha_p - 1)(3\chi_0^2\chi_2 + 3\chi_0\chi_1^2) + 3\alpha_p\chi_0^2\chi_1 + \Delta\varphi(\alpha_p - 1)^2(2\chi_0\chi_2 + \chi_1^2) \\ & + 2((1 + \Delta\varphi)\alpha_f + \alpha_p - 2)\alpha_p\chi_0\chi_1 + ((1 + \Delta\varphi)\alpha_f - 1)(\alpha_p - 1)\alpha_p\chi_1 + \alpha_f\alpha_p^2\chi_0 - \alpha_f\alpha_p^2. \end{aligned} \tag{S139}$$

Inserting $\chi_0^{(1)}$, this equation becomes

$$0 = \Delta\varphi(1 - \alpha_p)^2\chi_1^2 - \left[\alpha_p - (1 + \Delta\varphi)\alpha_f\right](1 - \alpha_p)\alpha_p\chi_1 - \alpha_f\alpha_p^3 \tag{S140}$$

with a pair of solutions,

$$\chi_1^{(1)} = \frac{1}{2\Delta\varphi\frac{(1-\alpha_p)}{\alpha_p}}\left[\alpha_p - (1 + \Delta\varphi)\alpha_f \pm \sqrt{\left[\alpha_p - (1 + \Delta\varphi)\alpha_f\right]^2 + 4\Delta\varphi\alpha_f\alpha_p}\right]. \tag{S141}$$

Similarly, inserting the second solution $\chi_0^{(2)}$, we find

$$0 = \Delta\varphi(\alpha_p - 1)^2\chi_1^2 - [1 - (1 + \Delta\varphi)\alpha_f](\alpha_p - 1)\alpha_p\chi_1 - \alpha_f\alpha_p^2 \tag{S142}$$

with a second pair of solutions,

$$\chi_1^{(2)} = \frac{1}{2\Delta\varphi\frac{(\alpha_p-1)}{\alpha_p}}\left[1-(1+\Delta\varphi)\alpha_f \pm \sqrt{\left[1-(1+\Delta\varphi)\alpha_f\right]^2+4\Delta\varphi\alpha_f}\right]. \tag{S143}$$

We note that the two solutions for χ_1 are qualitatively similar to the exact solution for χ in the model of linear regression, Eq. (S82). The implication is that the nonlinear nature of the activation function implicitly serves as a type of regularization via the quantity $\Delta\varphi$ which evaluates to zero in the linear limit.

Next, we solve for the solutions to ν. First, we make the ansatz

$$\nu \approx \frac{1}{\lambda}\nu_{-1} + \nu_0. \tag{S144}$$

Using Eq. (S132) and inserting the first solution for χ, we find

$$\nu_{-1}^{(1)} = 0 \tag{S145}$$

with the next order term

$$\nu_0^{(1)} = \frac{1}{\sigma_W^2\sigma_X^2}\frac{1}{2\Delta\varphi(1-\alpha_p)}\left[\alpha_p-(1+\Delta\varphi)\alpha_f \pm \sqrt{\left[\alpha_p-(1+\Delta\varphi)\alpha_f\right]^2+4\Delta\varphi\alpha_f\alpha_p}\right]. \tag{S146}$$

Similarly, the second solution for χ gives us

$$\nu_{-1}^{(2)} = \frac{1}{\sigma_W^2\sigma_X^2}\frac{(\alpha_p-1)}{\alpha_p}. \tag{S147}$$

For completion, we also find

$$\nu_0^{(2)} = \frac{\chi_1^{(2)}}{\alpha_p} = \frac{1}{2\Delta\varphi(\alpha_p-1)}\left[1-(1+\Delta\varphi)\alpha_f \pm \sqrt{\left[1-(1+\Delta\varphi)\alpha_f\right]^2+4\Delta\varphi\alpha_f}\right]. \tag{S148}$$

None of the remaining scalar susceptibilities have simple forms. Therefore, we use their representations in terms of ν and χ. Similarly, the solutions for $\langle\hat{w}^2\rangle$, $\langle\Delta y^2\rangle$, $\langle\hat{u}^2\rangle$, and $\langle\Delta\beta^2\rangle$ do not simplify significantly, but their limiting scaling behavior in terms of λ can still be determined. Using the fact that each of these quantities must be positive, it is straightforward to see that only two of the four solutions apply, depending on whether $\alpha_p > 1$ or $\alpha_p < 1$. The resulting solutions for χ and ν are then

$$\begin{aligned}
\chi &= \begin{cases} 1-\alpha_p & \text{if} \quad N_p < M \\ \frac{\lambda}{2\Delta\varphi\sigma_X^2\sigma_W^2}\frac{\alpha_p}{(\alpha_p-1)}\left[1-(1+\Delta\varphi)\alpha_f+\sqrt{\left[1-(1+\Delta\varphi)\alpha_f\right]^2+4\Delta\varphi\alpha_f}\right] & \text{if} \quad N_p > M \end{cases} \\
\nu &= \begin{cases} \frac{1}{2\Delta\varphi\sigma_X^2\sigma_W^2}\frac{1}{(1-\alpha_p)}\left[\alpha_p-(1+\Delta\varphi)\alpha_f+\sqrt{\left[\alpha_p-(1+\Delta\varphi)\alpha_f\right]^2+4\Delta\varphi\alpha_f\alpha_p}\right] & \text{if} \quad N_p < M \\ \frac{1}{\lambda}\frac{(\alpha_p-1)}{\alpha_p}+\frac{1}{2\Delta\varphi\sigma_X^2\sigma_W^2}\frac{1}{(\alpha_p-1)}\left[1-(1+\Delta\varphi)\alpha_f+\sqrt{\left[1-(1+\Delta\varphi)\alpha_f\right]^2+4\Delta\varphi\alpha_f}\right] & \text{if} \quad N_p > M \end{cases}
\end{aligned} \tag{S149}$$

In addition, $\langle\Delta y^2\rangle$ is $\mathcal{O}(1)$ in small λ when $\alpha_p < 1$ and $\mathcal{O}(\lambda^2)$ when $\alpha_p > 1$. The training and test error can be determined by substituting these susceptibilities into the equations for $\langle\hat{w}^2\rangle$, $\langle\Delta y^2\rangle$, $\langle\hat{u}^2\rangle$, and $\langle\Delta\beta^2\rangle$ in Eq. (S135) and then using the general solutions in Eqs. (S51) and (S57).

6. *Bias-Variance Decomposition*

To derive expressions for the bias and variance, according the general solutions in Eqs. (S62) and (S64), we need to calculate the covariance of the residual parameter errors $\langle\Delta\beta_1\Delta\beta_2\rangle$, as well as the covariance of the fit parameters $\langle\hat{w}_1\hat{w}_2\rangle$. As a reminder, the subscripts 1 and 2 refer to parameters resulting from fitting training sets $\mathcal{D}_1$ and $\mathcal{D}_2$ drawn independently from the same data distribution. We apply the self-consistent equations for the 0-indexed quantities, Eq. (S124), to the two data sets, giving us

$$\begin{aligned}
\hat{w}_{1,0} &= \nu\sigma_{\hat{w}}z_{\hat{w}_1} \\
\hat{u}_{1,0} &= \kappa\sigma_{\hat{u}}z_{\hat{u}_1} + \omega(\beta_0 - \sigma_{\Delta\beta}z_{\Delta\beta_1}) \\
\Delta y_{1,0} &= \chi(\sigma_{\Delta y}z_{\Delta y_1} + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_{1,0}) + \varepsilon_{1,0}) \\
\Delta\beta_{1,0} &= \kappa(\beta_0 - \sigma_{\Delta\beta}z_{\Delta\beta_1}) + \phi\sigma_{\hat{u}}z_{\hat{u}_1}
\end{aligned} \tag{S150}$$

and

$$\begin{aligned}
\hat{w}_{2,0} &= \nu\sigma_{\hat{w}} z_{\hat{w}_2} \\
\hat{u}_{2,0} &= \kappa\sigma_{\hat{u}} z_{\hat{u}_2} + \omega(\beta_0 - \sigma_{\Delta\beta} z_{\Delta\beta_2}) \\
\Delta y_{2,0} &= \chi(\sigma_{\Delta y} z_{\Delta y_2} + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_{2,0}) + \varepsilon_{2,0}) \\
\Delta\beta_{2,0} &= \kappa(\beta_0 - \sigma_{\Delta\beta} z_{\Delta\beta_2}) + \phi\sigma_{\hat{u}} z_{\hat{u}_2}.
\end{aligned} \tag{S151}$$

Multiplying these equations and using the self-averaging approximation, we find

$$\begin{aligned}
\langle \hat{w}_1 \hat{w}_2 \rangle &= \frac{1}{N_p} \sum_K \hat{w}_{1,K} \hat{w}_{2,K} \approx \mathrm{E}[\hat{w}_{1,0}\hat{w}_{2,0}] = \nu^2 \mathrm{E}\left[\sigma^2_{\hat{w}} z_{\hat{w}_1} z_{\hat{w}_2}\right] \\
\langle \hat{u}_1 \hat{u}_2 \rangle &= \frac{1}{N_f} \sum_k \hat{u}_{1,k} \hat{u}_{2,k} \approx \mathrm{E}[\hat{u}_{1,0}\hat{u}_{2,0}] = \kappa^2 \mathrm{E}\left[\sigma^2_{\hat{u}} z_{\hat{u}_1} z_{\hat{u}_2}\right] + \omega^2\left(\sigma^2_\beta + \mathrm{E}\left[\sigma^2_{\Delta\beta} z_{\Delta\beta_1} z_{\Delta\beta_2}\right]\right) \\
\langle \Delta y_1 \Delta y_2 \rangle &= \frac{1}{M} \sum_b \Delta y_{1,b} \Delta y_{2,b} \approx \mathrm{E}[\Delta y_{1,0}\Delta y_{2,0}] = \chi^2 \mathrm{E}\left[\sigma^2_{\Delta y} z_{\Delta y_1} z_{\Delta y_2}\right] \\
\langle \Delta\beta_1 \Delta\beta_2 \rangle &= \frac{1}{N_f} \sum_k \Delta\beta_{1,k} \Delta\beta_{2,k} \approx \mathrm{E}[\Delta\beta_{1,0}\Delta\beta_{2,0}] = \kappa^2\left(\sigma^2_\beta + \mathrm{E}\left[\sigma^2_{\Delta\beta} z_{\Delta\beta_1} z_{\Delta\beta_2}\right]\right) + \phi^2 \mathrm{E}\left[\sigma^2_{\hat{u}} z_{\hat{u}_1} z_{\hat{u}_2}\right].
\end{aligned} \tag{S152}$$

Next, we calculate each of the four resulting expectation values of products of random variables. The average of the product $z_{\hat{w}_1} z_{\hat{w}_2}$ is

$$\begin{aligned}
\mathrm{E}\left[\sigma^2_{\hat{w}} z_{\hat{w}_1} z_{\hat{w}_2}\right] &= \mathrm{E}\left[\left(\sum_j \hat{u}_{1,j\setminus 0} W_{j0} + \sum_a \Delta y_{1,a\setminus 0} \delta z_{\mathrm{NL},0}(\vec{\mathbf{x}}_{1,a})\right)\left(\sum_k \hat{u}_{2,k\setminus 0} W_{k0} + \sum_b \Delta y_{2,b\setminus 0} \delta z_{\mathrm{NL},0}(\vec{\mathbf{x}}_{2,b})\right)\right] \\
&= \sum_{jk} \mathrm{E}\left[\hat{u}_{1,j\setminus 0}\hat{u}_{2,k\setminus 0}\right] \mathrm{E}[W_{j0} W_{k0}] + \sum_{ab} \mathrm{E}\left[\Delta y_{1,a\setminus 0}\Delta y_{2,b\setminus 0}\right] \mathrm{E}[\delta z_{\mathrm{NL},0}(\vec{\mathbf{x}}_{1,a})] \mathrm{E}[\delta z_{\mathrm{NL},0}(\vec{\mathbf{x}}_{2,b})] \\
&= \frac{\sigma^2_W}{N_p} \sum_k \mathrm{E}\left[\hat{u}_{1,j\setminus 0}\hat{u}_{2,j\setminus 0}\right] \\
&\approx \sigma^2_W \frac{\alpha_f}{\alpha_p} \langle \hat{u}_1 \hat{u}_2 \rangle ,
\end{aligned} \tag{S153}$$

while the average of the product $z_{\Delta\beta_1} z_{\Delta\beta_2}$ results in

$$\begin{aligned}
\mathrm{E}\left[\sigma^2_{\Delta\beta} z_{\Delta\beta_1} z_{\Delta\beta_2}\right] &= \mathrm{E}\left[\sum_{JK} \hat{w}_{1,J\setminus 0} \hat{w}_{2,K\setminus 0} W_{0J} W_{0K}\right] \\
&= \sum_{JK} \mathrm{E}\left[\hat{w}_{1,J\setminus 0}\hat{w}_{2,K\setminus 0}\right] \mathrm{E}[W_{0J} W_{0K}] \\
&= \frac{\sigma^2_W}{N_p} \sum_K \mathrm{E}\left[\hat{w}_{1,K\setminus 0}\hat{w}_{2,K\setminus 0}\right] \\
&\approx \sigma^2_W \langle \hat{w}_1 \hat{w}_2 \rangle .
\end{aligned} \tag{S154}$$

We find that the other two products average to zero due to the independence of X_1 and X_2, giving us

$$\begin{aligned}
\mathrm{E}\left[\sigma^2_{\hat{u}} z_{\hat{u}_1} z_{\hat{u}_2}\right] &= \mathrm{E}\left[\sum_{ab} \Delta y_{1,a\setminus 0} \Delta y_{2,b\setminus 0} X_{1,a0} X_{2,b0}\right] \\
&= \sum_{ab} \mathrm{E}\left[\Delta y_{1,a\setminus 0}\Delta y_{2,b\setminus 0}\right] \mathrm{E}[X_{1,a0}] \mathrm{E}[X_{2,b0}] \\
&= 0
\end{aligned} \tag{S155}$$

and

$$\begin{aligned}\mathrm{E}\left[\sigma^2_{\Delta y} z_{\Delta y_1} z_{\Delta y_2}\right] &= \mathrm{E}\left[\left(\sum_j \Delta\beta_{1,j\setminus 0} X_{1,0j} - \sum_J \hat{w}_{1,J\setminus 0}\delta z_{\mathrm{NL},J}(\vec{\mathbf{x}}_{1,0})\right)\left(\sum_k \Delta\beta_{2,k\setminus 0} X_{2,0k} - \sum_K \hat{w}_{2,K\setminus 0}\delta z_{\mathrm{NL},K}(\vec{\mathbf{x}}_{2,0})\right)\right] \\ &= \sum_{jk} \mathrm{E}\left[\Delta\beta_{1,j\setminus 0}\Delta\beta_{2,k\setminus 0}\right]\mathrm{E}[X_{1,0j}]\mathrm{E}[X_{2,0k}] + \sum_{JK}\mathrm{E}\left[\hat{w}_{1,J\setminus 0}\hat{w}_{2,K\setminus 0}\right]\mathrm{E}[\delta z_{\mathrm{NL},J}(\vec{\mathbf{x}}_{1,0})]\mathrm{E}[\delta z_{\mathrm{NL},K}(\vec{\mathbf{x}}_{2,0})] \\ &= 0.\end{aligned} \tag{S156}$$

Substituting these results into Eq. (S152), we find the self-consistent equations

$$\begin{aligned}\langle \hat{w}_1\hat{w}_2\rangle &= \nu^2\sigma_W^2\frac{\alpha_f}{\alpha_p}\langle \hat{u}_1\hat{u}_2\rangle \\ \langle \hat{u}_1\hat{u}_2\rangle &= \omega^2\left(\sigma_\beta^2 + \sigma_W^2\langle \hat{w}_1\hat{w}_2\rangle\right) \\ \langle \Delta y_1\Delta y_2\rangle &= 0 \\ \langle \Delta\beta_1\Delta\beta_2\rangle &= \kappa^2\left(\sigma_\beta^2 + \sigma_W^2\langle \hat{w}_1\hat{w}_2\rangle\right).\end{aligned} \tag{S157}$$

Solving these equations exactly in the thermodynamic limit, we find the expressions

$$\begin{aligned}\langle \hat{w}_1\hat{w}_2\rangle &= \frac{\sigma_\beta^2}{\sigma_W^2}\frac{\sigma_W^4\frac{\alpha_f}{\alpha_p}\omega^2\nu^2}{\left(1-\sigma_W^4\frac{\alpha_f}{\alpha_p}\omega^2\nu^2\right)} \\ \langle \hat{u}_1\hat{u}_2\rangle &= \sigma_\beta^2\frac{\omega^2}{\left(1-\sigma_W^4\frac{\alpha_f}{\alpha_p}\omega^2\nu^2\right)} \\ \langle \Delta\beta_1\Delta\beta_2\rangle &= \sigma_\beta^2\frac{\kappa^2}{\left(1-\sigma_W^4\frac{\alpha_f}{\alpha_p}\omega^2\nu^2\right)}.\end{aligned} \tag{S158}$$

S2. SPECTRAL DENSITIES OF KERNEL MATRICES

Here, we derive the spectral densities for the kernel matrix Z^TZ for each model. To do this, we use the technique laid out in Ref. 55. For any symmetric matrix A of size $N\times N$, the spectral density can be written in the form

$$\rho(x) = \frac{1}{\pi}\lim_{\varepsilon\to 0^+}\mathrm{Im}\,\frac{1}{N}\,\mathrm{Tr}\,G(x-i\varepsilon), \tag{S159}$$

where

$$G(z) = [zI_N - A]^{-1} \tag{S160}$$

is the Green's function.

In our case, we are interested in the case $A = Z^TZ$. From the cavity calculations, we observe that the susceptibility matrix

$$\nu^{\hat{w}}(\lambda) = \left[\lambda I_{N_p} + Z^TZ\right]^{-1} \tag{S161}$$

is related to the Green's function via the relation $G(z) = -\nu^{\hat{w}}(-z)$. This allows us to the express the spectral density in terms of ν,

$$\rho(x) = -\frac{1}{\pi}\lim_{\varepsilon\to 0^+}\mathrm{Im}\,\nu(-x+i\varepsilon). \tag{S162}$$

Therefore, all we will need to do is evaluate Eq. (S162) using the appropriate function $\nu(\lambda)$ for each model.

Sometimes, there will be some fraction of eigenvalues at zero. While the weight of this contribution can be directly calculated via Eq. (S162), sometimes it is easier to instead examine the susceptibility matrix $\chi^{\Delta y}$ in the limit $\lambda\to 0$, which becomes

$$\chi^{\Delta y} = I_M - Z\left[\lambda I_{N_p} + Z^TZ\right]^{-1}Z^T \approx I_M - ZZ^+. \tag{S163}$$

The matrix ZZ^+ is a projector, so its trace is the rank of Z^TZ. The trace of $\chi^{\Delta y}$ is then

$$\chi = \frac{1}{M}\,\mathrm{Tr}\,\chi^{\Delta y} = 1 - \frac{1}{M}\,\mathrm{rank}(Z^TZ). \tag{S164}$$

The fraction of eigenvalues at zero is then

$$f_{\mathrm{zero}} = 1 - \frac{1}{N_p}\,\mathrm{rank}(Z^TZ) = \frac{\chi + \alpha_p - 1}{\alpha_p}. \tag{S165}$$

A. Linear Regression

For linear regression, the kernel is a Wishart matrix of the form $A = X^TX$, where the elements of the matrix X are independent and identically distributed according to a normal distribution with zero mean, the expected eigenvalue spectrum is the Marchenko-Pastur distribution [51]. To show this, we start the self-consistency equations for the susceptibilities for this model,

$$\chi = \frac{1}{1+\bar{\nu}}, \qquad \bar{\nu} = \frac{1}{\bar{\lambda} + \alpha_f^{-1}\chi}, \tag{S166}$$

where we have non-dimensionalized ν and λ by defining

$$\bar{\nu} = \sigma_X^2\nu, \qquad \bar{\lambda} = \frac{\lambda}{\sigma_X^2}. \tag{S167}$$

Plugging χ into $\bar{\nu}$ and rearranging, we find a quadratic equation for $\bar{\nu}$,

$$\bar{\lambda}\bar{\nu}^2 + \left[\left(\alpha_f^{-1} - 1\right) + \bar{\lambda}\right]\bar{\nu} - 1 = 0. \tag{S168}$$

Solving this equation, we find

$$\nu(\lambda) = \frac{\sigma_X^2\left(1 - \alpha_f^{-1}\right) - \lambda \pm \sqrt{D(\lambda)}}{2\sigma_X^2\lambda}, \tag{S169}$$

where we have defined the discriminant

$$D(\lambda) = \left[\sigma_X^2\left(\alpha_f^{-1} - 1\right) + \lambda\right]^2 + 4\sigma_X^2\lambda. \tag{S170}$$

Next, we substitute the above solution for ν into Eq. (S162) and simplify to find to find

$$\rho(x) = \frac{1}{2\sigma_X^2}\left[\sigma_X^2\left(1 - \alpha_f^{-1}\right) \pm \mathrm{Re}\,\sqrt{D(-x)}\right]\frac{1}{\pi}\lim_{\epsilon\to 0^+}\frac{\epsilon}{(x^2+\epsilon^2)} \pm \frac{\mathrm{Im}\,\sqrt{D(-x)}}{2\pi\sigma_X^2 x}. \tag{S171}$$

We see that the first term contains the definition of a delta function evaluated at zero,

$$\delta(x) = \frac{1}{\pi}\lim_{\epsilon\to 0^+}\frac{\epsilon}{(x^2+\epsilon^2)}. \tag{S172}$$

This allows us to evaluate the the coefficient of this delta function at zero so that the first term in the spectrum becomes

$$\frac{1}{2\sigma_X^2}\left[\sigma_X^2\left(1 - \alpha_f^{-1}\right) \pm \mathrm{Re}\,\sqrt{D(0)}\right]\delta(x) = \max\left(0, 1 - \alpha_f^{-1}\right)\delta(x). \tag{S173}$$

We have chosen the signs of the solutions ($\pm$) so that the spectral density at zero is always non-negative. To simplify the second term in Eq. (S162), we need to find the interval over which $D(-x) < 0$. Solving $D(-x) = 0$,

$$D(-x) = x^2 - 2x\sigma_X^2\left(\alpha_f^{-1} + 1\right)x + \sigma_X^4\left(\alpha_f^{-1} - 1\right)^2 = 0, \tag{S174}$$

we find the limits of the interval to be

$$x_\pm = \sigma_X^2\left(1 \pm \sqrt{\alpha_f^{-1}}\right)^2. \tag{S175}$$

The second term in the spectrum then becomes

$$\pm \frac{\mathrm{Im}\,\sqrt{-(x_+ - x)(x - x_-)}}{2\pi\sigma_X^2 x} = \frac{\sqrt{(x_+ - x)(x - x_-)}}{2\pi\sigma_X^2 x}, \tag{S176}$$

where we have again chosen the plus sign so that the spectrum is always non-negative.

The complete spectrum is then written as

$$\rho(x) = \max\left(0, 1 - \alpha_f^{-1}\right)\delta(x) + \begin{cases} \frac{1}{2\pi\sigma_X^2 x}\sqrt{(x_{\max} - x)(x - x_{\min})} & \text{if} \quad x \in [x_{\min}, x_{\max}] \\ 0 & \text{otherwise} \end{cases} \tag{S177}$$

with

$$x_{\min} = \sigma_X^2\left(1 - \sqrt{\alpha_f^{-1}}\right)^2, \qquad x_{\max} = \sigma_X^2\left(1 + \sqrt{\alpha_f^{-1}}\right)^2. \tag{S178}$$

As expected, this is the Marchenko-Pastur distribution.

B. Random Nonlinear Features Model

Next, we derive the eigenvalue distribution for the kernel of the random nonlinear features model. Previously this result was derived in Ref. 52. To reproduce this analytic result, we start with three of the susceptibilities from the cavity derivation,

$$\bar{\nu} = \frac{1}{\bar{\lambda} + \alpha_p^{-1}\chi(\kappa + \Delta\varphi)}, \qquad \chi = \frac{1}{1 + \bar{\nu}(\kappa + \Delta\varphi)}, \qquad \kappa = \frac{1}{1 + \alpha_f^{-1}\chi\bar{\nu}}. \tag{S179}$$

where we have non-dimensionalized ν and λ by defining

$$\bar{\nu} = \sigma_W^2\sigma_X^2\nu, \qquad \bar{\lambda} = \frac{\lambda}{\sigma_W^2\sigma_X^2}. \tag{S180}$$

Solving each of the equations for χ and $\bar{\nu}$ for κ and then setting them equal we find

$$\kappa = \frac{1 - \chi - \Delta\varphi\bar{\nu}}{\chi\bar{\nu}} = \frac{\alpha_p(1 - \bar{\lambda}\bar{\nu}) - \Delta\varphi\bar{\nu}}{\chi\bar{\nu}}. \tag{S181}$$

From here, we find the following relation between χ and $\bar{\nu}$:

$$\chi = \alpha_p\bar{\lambda}\bar{\nu} - \alpha_p + 1. \tag{S182}$$

Next, we substitute κ into original equation for $\bar{\nu}$, solve for χ, and then substitute this result into Eq. (S182). If we then eliminate any denominators, we find a quartic equation for $\bar{\nu}$,

$$\begin{aligned} 0 = {} & \Delta\varphi(\alpha_p\bar{\lambda}\bar{\nu})^4 + \left[2\Delta\varphi(1 - \alpha_p) + \alpha_p\bar{\lambda}\right](\alpha_p\bar{\lambda}\bar{\nu})^3 \\ & + \left[\Delta\varphi(1 - \alpha_p)^2 + \alpha_p(\alpha_f(1 + \Delta\varphi) - \alpha_p + 1 - \alpha_p)\bar{\lambda}\right](\alpha_p\bar{\lambda}\bar{\nu})^2 \\ & + \alpha_p\left[(\alpha_f(1 + \Delta\varphi) - \alpha_p)(1 - \alpha_p) + \alpha_f\alpha_p\bar{\lambda}\right]\bar{\lambda}(\alpha_p\bar{\lambda}\bar{\nu}) - \alpha_f\alpha_p^3\bar{\lambda}^2. \end{aligned} \tag{S183}$$

Solving this quartic equation analytically is very involved, so instead we will solve this equation numerically for negative imaginary roots of $\nu(\lambda)$ with $\lambda = -x$, according to Eq. (S162). However, to find the interval over which the spectrum is positive, we rewrite the equation in general form for $\alpha_p\bar{\lambda}\bar{\nu}$,

$$0 = a_4(\alpha_p\bar{\lambda}\bar{\nu})^4 + a_3(\alpha_p\bar{\lambda}\bar{\nu})^3 + a_2(\alpha_p\bar{\lambda}\bar{\nu})^2 + a_1(\alpha_p\bar{\lambda}\bar{\nu}) + a_0, \tag{S184}$$

where the coefficients are

$$\begin{aligned} a_0 &= -\alpha_f \alpha_p^3 \bar{\lambda}^2 \\ a_1 &= \alpha_p \big[(\alpha_f(1+\Delta\varphi) - \alpha_p)(1-\alpha_p) + \alpha_f \alpha_p \bar{\lambda}\big] \bar{\lambda} \\ a_2 &= \Delta\varphi(1-\alpha_p)^2 + \alpha_p(\alpha_f(1+\Delta\varphi) - \alpha_p + 1 - \alpha_p)\bar{\lambda} \\ a_3 &= 2\Delta\varphi(1-\alpha_p) + \alpha_p \bar{\lambda} \\ a_4 &= \Delta\varphi. \end{aligned} \tag{S185}$$

The discriminant for a quartic equation is expressed in terms of these coefficients as

$$D(z) = R^2 - 4Q^3 \tag{S186}$$

with

$$\begin{aligned} R &= 2a_2^3 - 9a_1 a_2 a_3 + 27 a_0 a_3^2 + 27 a_1^2 a_4 - 72 a_0 a_2 a_4 \\ Q &= a_2^2 - 3a_1 a_3 + 12 a_0 a_4. \end{aligned} \tag{S187}$$

To find the limiting eigenvalues, we then solve the equation $D(\lambda) = 0$ (with $\lambda = -x$) numerically for the largest and smallest non-negative real roots.

To find the weight of the delta function component at zero, we use Eq. (S165) and the solutions for ν we found previously this model, giving us

$$f_{\text{zero}} = \max\left(0, 1 - \alpha_p^{-1}\right). \tag{S188}$$

S3. ACCURACY OF MINIMUM PRINCIPAL COMPONENT

In this section, we derive expressions for the predicted labels $\hat{y}$ as a function of projections of the data points along the minimum principal component $\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}(\vec{\mathbf{x}})$ used to assess model accuracy in Figs. 4 and 5. We seek two different predictions of the labels as a function of $\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}(\vec{\mathbf{x}})$: the labels $\hat{y}_{\text{train}}$ that result from a finite training set and the labels $\hat{y}_{\text{test}}$ that result from fitting to an average test set, or equivalently, the full data distribution (the limit of a training set of size $M \to \infty$ for fixed N_f and N_p).

Given a training set consisting of M data points $\mathcal{D} = \{(y_b, \vec{\mathbf{x}}_b)\}_{b=1}^M$ with corresponding hidden features $\vec{\mathbf{z}}_a = \vec{\mathbf{z}}(\vec{\mathbf{x}}_a)$, we start by decomposing the kernel matrix into n principal components $\hat{\mathbf{h}}_i$ with non-zero eigenvalues σ_i^2,

$$Z^T Z = \sum_{i=1}^{n} \sigma_i^2 \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^T. \tag{S189}$$

We define the principal components so that they form an orthonormal basis of (hidden) features,

$$\hat{\mathbf{h}}_i \cdot \hat{\mathbf{h}}_j = \delta_{ij}. \tag{S190}$$

We define the minimum component $\hat{\mathbf{h}}_{\min}$ as the principal component with the smallest non-zero eigenvalue $\sigma_{\min}^2$. Next, we define the empirical variance of $\hat{\mathbf{h}}_i \cdot \vec{\mathbf{z}}_a$ (holding $\hat{\mathbf{h}}_i$ fixed) within the training set and derive its relationship to the eigenvalue σ_i^2,

$$\begin{aligned} \text{Var}_{\vec{\mathbf{x}} \in \mathcal{D}}\left[\hat{\mathbf{h}}_i \cdot \vec{\mathbf{z}}(\vec{\mathbf{x}}) | \hat{\mathbf{h}}_i\right] &= \frac{1}{M} \sum_{a=1}^{M} (\hat{\mathbf{h}}_i \cdot \vec{\mathbf{z}}_a)^2 \\ &= \frac{1}{M} \hat{\mathbf{h}}_i^T Z^T Z \hat{\mathbf{h}}_i \\ &= \frac{\sigma_i^2}{M}. \end{aligned} \tag{S191}$$

Similarly, we define the empirical covariance of $\hat{\mathbf{h}}_i \cdot \vec{\mathbf{z}}_a$ and the labels y_a (again, holding $\hat{\mathbf{h}}_i$ fixed),

$$\begin{aligned} \text{Cov}_{\vec{\mathbf{x}} \in \mathcal{D}}\left[\hat{\mathbf{h}}_i \cdot \vec{\mathbf{z}}(\vec{\mathbf{x}}), y(\vec{\mathbf{x}}) | \hat{\mathbf{h}}_i\right] &= \frac{1}{M} \sum_{a=1}^{M} (\hat{\mathbf{h}}_i \cdot \vec{\mathbf{z}}_a) y_a \\ &= \frac{1}{M} \hat{\mathbf{h}}_i^T Z^T \vec{\mathbf{y}}. \end{aligned} \tag{S192}$$

Using the expression for the predicted labels in Eq. (3) and the exact solution for the fit parameters in the ridge-less limit in Eq. (10), we express the predicted label for an arbitrary test data point $\vec{\mathbf{z}}'$ in terms of the empirical variance and covariance as

$$\begin{aligned}\hat{y} &\approx \vec{\mathbf{z}}' \cdot (Z^T Z)^+ Z^T \vec{\mathbf{y}} \\ &= \vec{\mathbf{z}}' \cdot \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^T Z^T \vec{\mathbf{y}} \\ &= \sum_{i=1}^{n} \frac{\mathrm{Cov}_{\vec{\mathbf{x}}\in\mathcal{D}}\left[\hat{\mathbf{h}}_i \cdot \vec{\mathbf{z}}(\vec{\mathbf{x}}), y(\vec{\mathbf{x}}) | \hat{\mathbf{h}}_i\right]}{\mathrm{Var}_{\vec{\mathbf{x}}\in\mathcal{D}}\left[\hat{\mathbf{h}}_i \cdot \vec{\mathbf{z}}(\vec{\mathbf{x}}) | \hat{\mathbf{h}}_i\right]} (\hat{\mathbf{h}}_i \cdot \vec{\mathbf{z}}').\end{aligned} \tag{S193}$$

Dropping all terms except for the one containing the minimum component, we find the expression for the predicted labels as a function of $\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}'$ resulting from fitting the training data,

$$\hat{y}_{\text{train}}\left(\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}'\right) = \frac{\mathrm{Cov}_{\vec{\mathbf{x}}\in\mathcal{D}}\left[\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}, y(\vec{\mathbf{x}}) | \hat{\mathbf{h}}_{\min}\right]}{\mathrm{Var}_{\vec{\mathbf{x}}\in\mathcal{D}}\left[\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}} | \hat{\mathbf{h}}_{\min}\right]} \left(\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}'\right). \tag{S194}$$

To find this relationship for an average test set, we extend the empirical variance and covariance to consider an infinitely large data set, or equivalently, average over all possible data points $(y, \vec{\mathbf{x}})$ with hidden features $\vec{\mathbf{z}}(\vec{\mathbf{x}})$. However, we still hold $\hat{\mathbf{h}}_{\min}$ fixed since it is a result of the training set. The resulting relationship is

$$\hat{y}_{\text{test}}\left(\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}'\right) = \frac{\mathrm{Cov}_{\vec{\mathbf{x}}}\left[\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}(\vec{\mathbf{x}}), y(\vec{\mathbf{x}}) | \hat{\mathbf{h}}_{\min}\right]}{\mathrm{Var}_{\vec{\mathbf{x}}}\left[\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}(\vec{\mathbf{x}}) | \hat{\mathbf{h}}_{\min}\right]} \left(\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}'\right). \tag{S195}$$

According to Eq. (S191), we calculate the spread of the training data points along the minimum principal component,

$$\sigma_{\text{train}}^2 = \mathrm{Var}_{\vec{\mathbf{x}}\in\mathcal{D}}\left[\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}(\vec{\mathbf{x}}) | \hat{\mathbf{h}}_{\min}\right] = \frac{\sigma_{\min}^2}{M}. \tag{S196}$$

We find that it is related to the minimum eigenvalue of the kernel matrix. We also derive the true variance of data points along $\hat{\mathbf{h}}_{\min}$ for an average test set,

$$\sigma_{\text{test}}^2 = \mathrm{Var}_{\vec{\mathbf{x}}}\left[\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}(\vec{\mathbf{x}}) | \hat{\mathbf{h}}_{\min}\right]. \tag{S197}$$

In the next few sections, we derive this variance, along with the covariance with respect an average test set, or the full data distribution, for both models, along with expressions for $\hat{y}_{\text{test}}$.

A. Linear Regression

In linear regression without basis functions, the input features and hidden features are identical such that $Z = X$. Using the decomposition of the labels in Eq. (S31), we find

$$\mathrm{Cov}_{\vec{\mathbf{x}}}\left[\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}(\vec{\mathbf{x}}), y(\vec{\mathbf{x}}) | \hat{\mathbf{h}}_{\min}\right] = \frac{\sigma_X^2}{N_f} \hat{\mathbf{h}}_{\min}^T \vec{\boldsymbol{\beta}} \tag{S198}$$

$$\mathrm{Var}_{\vec{\mathbf{x}}}\left[\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}(\vec{\mathbf{x}}) | \hat{\mathbf{h}}_{\min}\right] = \frac{\sigma_X^2}{N_f}. \tag{S199}$$

Using these results, the predicted labels for an average test set as a function of $\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}'$ are then

$$\hat{y}_{\text{test}}\left(\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}'\right) = \hat{\mathbf{h}}_{\min}^T \vec{\boldsymbol{\beta}} \left(\hat{\mathbf{h}}_{\min} \cdot \vec{\mathbf{z}}'\right), \tag{S200}$$

while the expected spread is

$$\sigma_{\text{test}}^2 = \frac{\sigma_X^2}{N_f}. \tag{S201}$$

B. Random Nonlinear Features Model

In the random nonlinear features model, we use the label decomposition in Eq. (S31) and the hidden feature decomposition in Eq. (S39), to find

$$\mathrm{Cov}_{\vec{\mathbf{x}}}\left[\hat{\mathbf{h}}_{\min}\cdot\vec{\mathbf{z}}(\vec{\mathbf{x}}), y(\vec{\mathbf{x}})|\hat{\mathbf{h}}_{\min}\right] = \frac{\sigma_X^2}{N_f}\hat{\mathbf{h}}_{\min}^T W^T\vec{\boldsymbol{\beta}} \tag{S202}$$

$$\mathrm{Var}_{\vec{\mathbf{x}}}\left[\hat{\mathbf{h}}_{\min}\cdot\vec{\mathbf{z}}(\vec{\mathbf{x}})|\hat{\mathbf{h}}_{\min}\right] = \frac{\sigma_X^2}{N_f}\hat{\mathbf{h}}_{\min}^T W^T W\hat{\mathbf{h}}_{\min} + \Delta\varphi\frac{\sigma_W^2\sigma_X^2}{N_p}. \tag{S203}$$

The predicted labels are then

$$\hat{y}_{\text{test}}\left(\hat{\mathbf{h}}_{\min}\cdot\vec{\mathbf{z}}'\right) = \frac{\hat{\mathbf{h}}_{\min}^T W^T\vec{\boldsymbol{\beta}}}{\left(\hat{\mathbf{h}}_{\min}^T W^T W\hat{\mathbf{h}}_{\min} + \Delta\varphi\sigma_W^2\frac{\alpha_f}{\alpha_p}\right)}\left(\hat{\mathbf{h}}_{\min}\cdot\vec{\mathbf{z}}'\right) \tag{S204}$$

and the expected spread is

$$\sigma_{\text{test}}^2 = \frac{\sigma_X^2}{N_f}\hat{\mathbf{h}}_{\min}^T W^T W\hat{\mathbf{h}}_{\min} + \Delta\varphi\frac{\sigma_W^2\sigma_X^2}{N_p}. \tag{S205}$$

S4. NUMERICAL SIMULATION DETAILS

In this section, we explain our procedures for generating numerical results.

A. General Details

In all plots of training error, test error, bias, and variance, each point (or pixel for $2d$ plots) is averaged over 1000 independent simulations, unless located exactly at a phase transition, in which case, each point is averaged over 150000 simulations. Small error bars are shown each plot, representing the error on the mean. We also scale the error in each plot by the variance of the labels $\sigma_y^2 = \sigma_\beta^2\sigma_X^2 + \sigma_{\delta y^*}^2 + \sigma_\varepsilon^2$. In all simulations, we use training and test sets of size $M = M' = 512$, a signal-to-noise ratio of $(\sigma_\beta^2\sigma_X^2 + \sigma_{\delta y^*}^2)/\sigma_\varepsilon^2 = 10$, and a regularization parameter of $\lambda = 10^{-6}$. We also use a linear teacher model $y^*(\vec{\mathbf{x}}) = \vec{\mathbf{x}}\cdot\vec{\boldsymbol{\beta}}$ ($\sigma_{\delta y^*}^2 = 0$) in most cases. In Fig. 5, we use a nonlinear teacher model with $f(h) = \tanh(h)$. In this case, we find that $\langle f'\rangle = 0.6057$ and $\langle f^2\rangle = 0.3943$, resulting in $\sigma_{\delta y^*}^2/\sigma_\beta^2\sigma_X^2 = \Delta f = 0.0747$.

To find the solution for a particular regression problem, we solve a different (but equivalent) system of equations depending on whether $N_p < M$ or $N_p > M$, allowing us to reduce the size of the linear system we need to solve. If $N_p < M$, we solve the system of N_p equations

$$\left[\lambda I_{N_p} + Z^T Z\right]\hat{\mathbf{w}} = Z^T\vec{\mathbf{y}} \tag{S206}$$

for the N_p unknown fit parameters $\hat{\mathbf{w}}$ where I_{N_p} is the $N_p \times N_p$ identity matrix. This equation is identical to that in Eq. (9) in the main text.

Alternatively, if $N_p > M$ we solve a system of M equations,

$$\left[\lambda I_M + ZZ^T\right]\hat{\mathbf{a}} = \vec{\mathbf{y}}, \tag{S207}$$

for the M unknowns $\hat{\mathbf{a}}$ where I_M is the $M \times M$ identity matrix. We then convert to fit parameters via the formula $\hat{\mathbf{w}} = Z^T\hat{\mathbf{a}}$.

B. Bias-Variance Decompositions

To efficiently calculate the ensemble-averaged bias and variance, we take inspiration from Eq. (S60). During each simulation, we independently generate two training data sets $\mathcal{D}_1$ and $\mathcal{D}_2$. Using the results from the first training set we calculate the training and test error. To calculate the bias, we also calculate the label predictions for both training sets for an identical test set, $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$, and record the residual label errors between these predictions and the true

labels of the test set $\vec{\mathbf{y}}^{*\prime}$ and record the product $(\hat{\mathbf{y}}_1 - \vec{\mathbf{y}}^{*\prime}) \cdot (\hat{\mathbf{y}}_2 - \vec{\mathbf{y}}^{*\prime})$. When averaged over many simulations, this quantity approximates the bias. We can then subtract this quantity from the average test error to find the variance. We follow an analogous procedure to find each contribution of the labels in Eq. (S31) to the bias and variance in Fig. 5. This is achieved by calculating the test error, bias and variance using only a single contribution from the labels at a time and setting the rest to zero.

C. Eigenvalue Decompositions of Kernel Matrices

For each of the numerical eigenvalue distributions for the kernel matrices presented in the main text, we choose $M = 4096$. We then average over the distributions for 10 independently sampled matrices when $\alpha_p = 1$ or $\alpha_p = 8$ and over 80 matrices when $\alpha_p = 1/8$. In this way, we ensure that the same number of non-zero eigenvalues is present in the part of the histograms corresponding to the bulk of the distributions (the distribution excluding the delta function at zero). For $M < N_p$ we calculate the eigenvalues of $Z^T Z$, while for $M > N_p$ we instead calculate the eigenvalues of ZZ^T since this matrix is smaller and contains the same non-zero eigenvalues. In the later case, we then manally append an additional $N_p - M$ zero-valued eigenvalues to the distribution.

D. Spread Along Mimimum Principal Components

For each the scatter plots in Figs. 4 and 5, we consider training and test sets of size $M = M' = 200$, with all other parameters specified in Sec. S4 A. We then calculate the principal component corresponding to the minimum eigenvalue numerically and use this to plot the relationship learned by the model for the training set, as detailed in Sec. S3. For the test set, we show the relationship for an average test set rather than the specific test shown, again using the formulas detailed in Sec. S3. We note that these formulas still require the minimum principal component calculated for the training set.

In Fig. 4, the spread of the the training set compared to a test set as a function of α_p is calculated using 100 simulations for each point. For each simulation, we record the ratio $\sigma_{\text{train}}/\sigma_{\text{test}}$ (see Sec. S3) and then average this quantity across simulations for each α_p.

S5. COMPLETE NUMERICAL RESULTS

In this section, we provide complete comparisons between the analytic and numerical results when lacking from the main text. For linear regression, comparisons to numerical results for the training error, test error, bias ,and variance are depicted in Fig. 2 in the main text.

For the random nonlinear features model, Fig. S1 provides comparisons to numerical results for the training error, test error, bias, and variance.

S6. NON-STANDARD BIAS-VARIANCE DECOMPOSITIONS

In Fig. S2, we show numerical results for the alternative definitions of bias described in Sec. VF for the random nonlinear features model. In the fixed design setting, the bias and variance are defined as

$$\begin{aligned} \text{Bias}_{\text{fd}}[\hat{y}(\vec{\mathbf{x}})] &= \text{E}_{\vec{\boldsymbol{\varepsilon}}}[\hat{y}(\vec{\mathbf{x}})] - y^*(\vec{\mathbf{x}}) \\ \text{Var}_{\text{fd}}[\hat{y}(\vec{\mathbf{x}})] &= \text{E}_{\vec{\boldsymbol{\varepsilon}}}[\hat{y}(\vec{\mathbf{x}})^2] - \text{E}_{\vec{\boldsymbol{\varepsilon}}}[\hat{y}(\vec{\mathbf{x}})]^2. \end{aligned} \tag{S208}$$

Alternatively, in the ensemble setting, the bias and variance are defined as

$$\begin{aligned} \text{Bias}_{\text{ens}}[\hat{y}(\vec{\mathbf{x}})] &= \text{E}_{X,\vec{\boldsymbol{\varepsilon}},W}[\hat{y}(\vec{\mathbf{x}})] - y^*(\vec{\mathbf{x}}) \\ \text{Var}_{\text{ens}}[\hat{y}(\vec{\mathbf{x}})] &= \text{E}_{X,\vec{\boldsymbol{\varepsilon}},W}[\hat{y}(\vec{\mathbf{x}})^2] - \text{E}_{X,\vec{\boldsymbol{\varepsilon}},W}[\hat{y}(\vec{\mathbf{x}})]^2. \end{aligned} \tag{S209}$$

We plot all four quantities with comparisons to the standard counterparts at fixed α_f in Figs. S2(a) and (b). We also show the full behavior as a function of both α_p and α_f for the four quantities in Figs. S2(c)-(f).

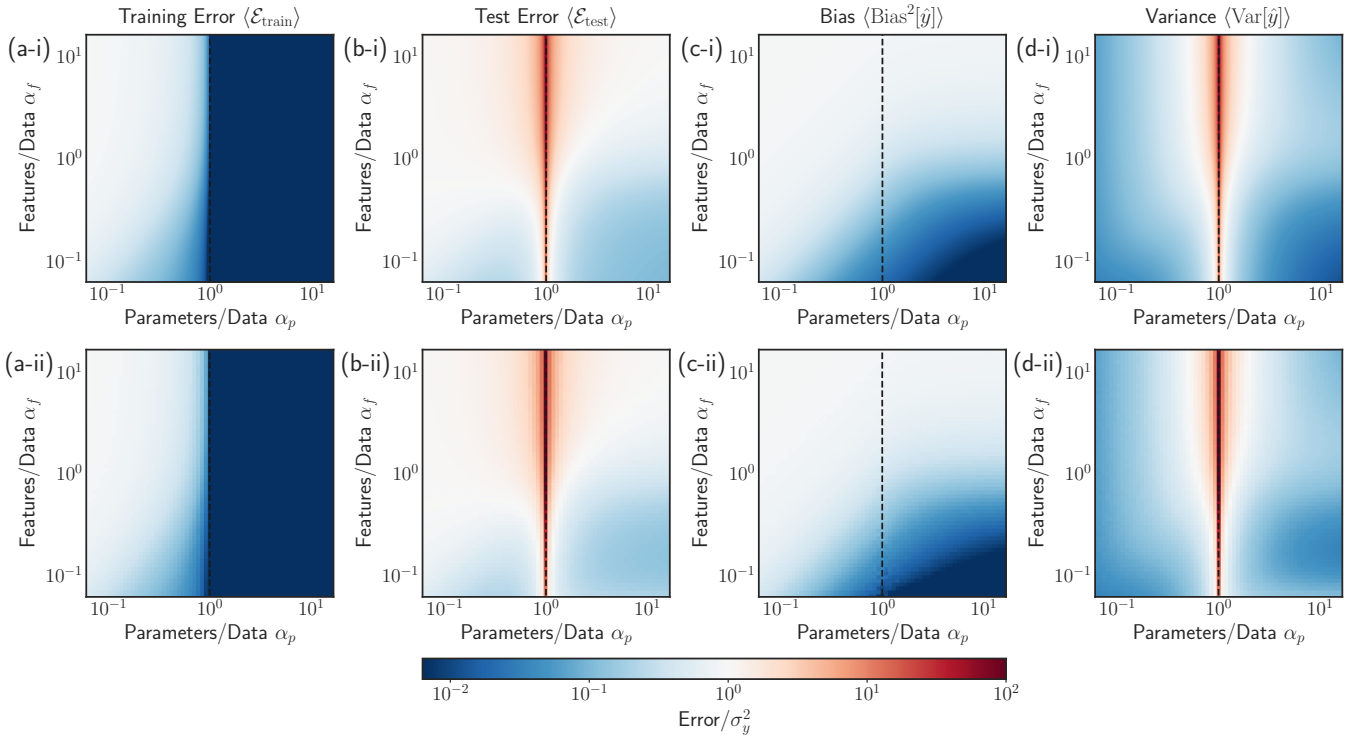

FIG. S1. **Comparison of analytic and numerical results for the random nonlinear features model: Training error and bias-variance decomposition.** (Top Row) Analytic solutions and (Bottom Row) numerical results are shown as a function of $\alpha_p = N_p/M$ and $\alpha_f = N_f/M$. Plotted are the ensemble-averaged (a) training error, (b) test error, (c) squared bias, and (d) variance. In each panel, a black dashed line marks the boundary between the under- and over-parameterized regimes at $\alpha_p = 1$.

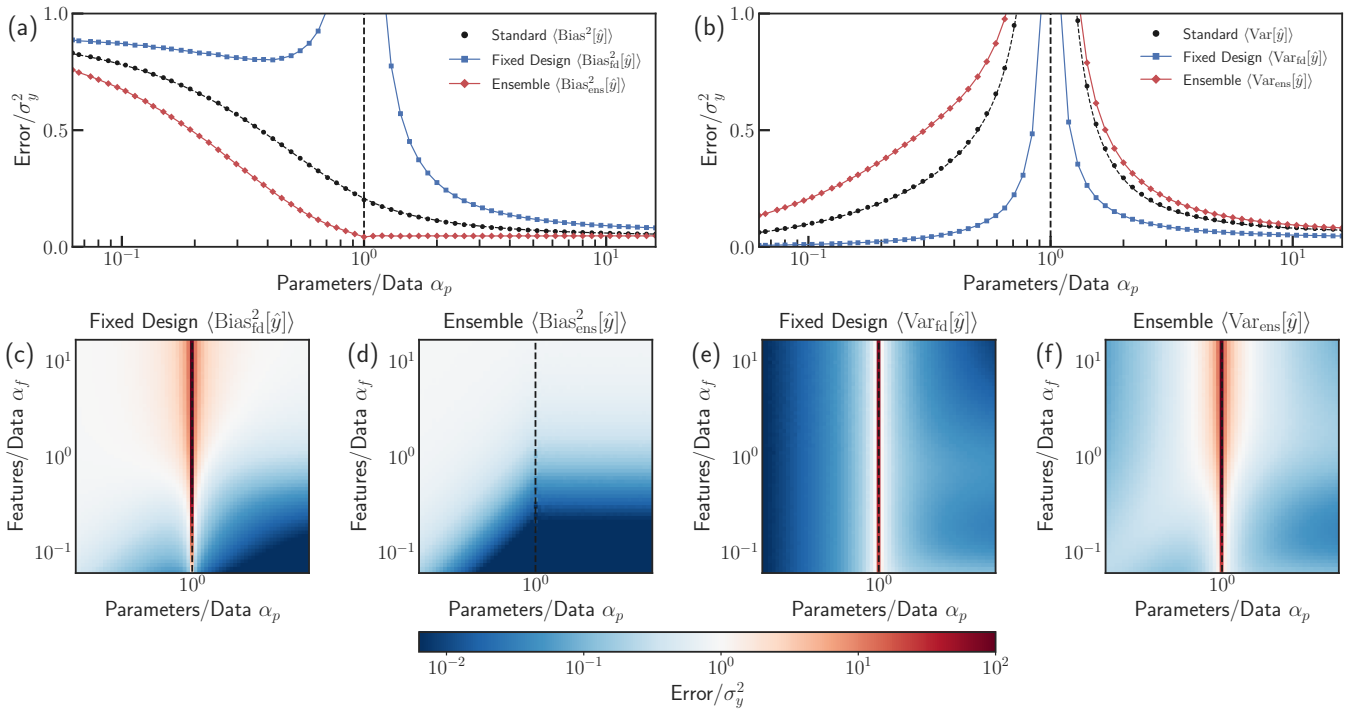

FIG. S2. **Numerical comparison of the bias-variance decompositions using three different definitions.** The **(a)** squared bias and **(b)** variance for the standard setting (black circles), fixed-design setting (blue squares) and ensemble setting (red diamonds) are shown for fixed $\alpha_f = 1/2$. Results are also shown as a function of $\alpha_p = N_p/M$ and $\alpha_f = N_f/M$ for the **(c)** squared fixed design bias, **(d)** squared ensemble bias, **(e)** fixed design variance, and **(f)** ensemble variance. In each panel, a black dashed line marks the boundary between the under- and over-parameterized regimes at $\alpha_p = 1$.