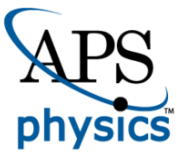

Bias-variance decomposition of overparameterized regression with random linear features

Jason W. Rocks and Pankaj Mehta

Phys. Rev. E **106**, 025304 — Published 4 August 2022

DOI: 10.1103/PhysRevE.106.025304

Bias-variance decomposition of overparameterized regression with random linear features

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*

In classical statistics, the bias-variance trade-off describes how varying a model's complexity (e.g., number of fit parameters) affects its ability to make accurate predictions. According to this trade-off, optimal performance is achieved when a model is expressive enough to capture trends in the data, yet not so complex that it overfits idiosyncratic features of the training data. Recently, it has become clear that this classic understanding of the bias-variance must be fundamentally revisited in light of the incredible predictive performance of "overparameterized models" – models that avoid overfitting even when the number of fit parameters is large enough to perfectly fit the training data. Here, we present results for one of the simplest examples of an overparameterized model: regression with random linear features (i.e. a two-layer neural network with a *linear* activation function). Using the zero-temperature cavity method, we derive analytic expressions for the training error, test error, bias, and variance. We show that the linear random features model exhibits three phase transitions: two different transitions to an interpolation regime where the training error is zero, along with an additional transition between regimes with large bias and minimal bias. Using random matrix theory, we show how each transition arises due to small nonzero eigenvalues in the Hessian matrix. Finally, we compare and contrast the phase diagram of the random *linear* features model to the random *nonlinear* features model and ordinary regression, highlighting the new phase transitions that result from the use of linear basis functions.

I. INTRODUCTION

One of the core concepts in modern statistics and supervised learning is the bias-variance decomposition. It states that the test error, the predictive performance of the model on new data, can be decomposed into three parts: bias, variance, and noise [1–3]. The bias captures errors due to underfitting, resulting from the inability of a statistical model to sufficiently express statistical relationships present in the data distribution. The variance, on the other hand, characterizes errors that result from "over-fitting" unrepresentative aspects of the training data set that do not generalize (e.g., label noise). Finally, the noise describes irreducible errors in a test data set due to randomness in the data generating process.

In classical statistics, the bias-variance trade-off suggests that optimal predictive performance is achieved by utilizing statistical models with intermediate model complexities, balancing errors due to bias and variance. While increasing a model's complexity (e.g., increasing the number of fit parameters) reduces bias, it comes at the price of increasing variance. One of the most interesting and surprising empirical results to emerge from deep learning over the last five years is the realization that this basic intuition is fundamentally incomplete; it does not apply to "overparameterized" models where the number of fit parameters is large enough to perfectly fit the training data (i.e. achieve zero error on the training data set) [4].

While the classic bias-variance trade-off still holds in the underparameterized regime (i.e., for models that have too few fit parameters to achieve zero training error), once a model's complexity is increased past the interpolation threshold – the point at which the training error goes to zero – the test error once again decreases. The resulting combination of a "U-shaped" test error in the underparameterized regime and the subsequent decrease in test error in the overparameterized regime is now commonly referred to as a "double-descent" curve [5, 6]. This double-descent behavior seems to be a generic property of all overparameterized supervised learning models and for this reason, has become a major area of research.

An important open question in the field is to understand the double-descent phenomena in terms of classical ideas of bias and variance. One fruitful approach has been to analyze analytically tractable models that exhibit the double-descent phenomena [7–35]. Among, the most popular of these models are linear regression (ridge regression without basis functions) and the random nonlinear features model (a two-layer neural network with an arbitrary nonlinear activation function where the top layer is trained and parameters for the intermediate layer are chosen to be random but fixed) [35]. Here, we build upon this previous work by examining a random features model for the special case of a linear activation function (i.e., the random linear features model). Using the zero temperature cavity method, we derive analytic expressions for the bias-variance decomposition and relate these results to the eigenvalue spectrum of the Hessian matrix.

The training and test errors of the random linear features have been computed analytically [8, 9, 15, 28, 34], with a subset of these studies attempting to carrying out bias-variance decompositions [9, 28, 34]. However, these studies use non-standard definitions of bias and variance that deviate from the traditional textbook definitions [1, 2]. This choice of definition can lead to qualitatively different and difficult to reconcile results. For ex-

ample, the authors of Ref. 9 find that the bias diverges at the interpolation threshold, while the authors of Ref. 28 find no such divergence (see Ref. 35 for an in-depth discussion). For this reason, here we utilize the standard definitions and carry out the bias-variance decomposition in a manner consistent with traditional definitions of these quantities in the underparameterized regime, allowing us to identify which properties stem from the model architecture versus random sampling of the data. In addition, we use the zero-temperature cavity method to provide an alternative derivation of the spectrum of the Hessian matrix of the random linear features model (i.e., the spectrum of a Wishart product matrix calculated previously in Ref. 36), allowing us to directly relate the eigenvalues of the Hessian to the double-descent phenomenon.

Summary of Major Results

We briefly summarize our major results:

- We derive analytic expressions for the test (generalization) error, training error, bias, and variance for the random linear features model with a nonlinear data distribution using the zero-temperature cavity method.
- We find that the behavior of this model is characterized by three distinct regimes: (i) an underparameterized regime with finite training error and large bias, (ii) a second underparameterized regime with minimal, constant bias, and (iii) an overparameterized, or interpolation, regime with zero training error.
- We find that the three regimes are separated by three phase transitions with two transitions to the interpolation regime, each characterized by a divergence in the test error, and one transition between the large bias and minimal bias underparameterized regimes. Importantly, we find that the variance, but not the bias, diverges at the phase transition to the interpolation regime.
- We explain how each phase transition arises as a result of small nonzero eigenvalues in the Hessian matrix and demonstrate how this phenomenon is captured by susceptibilities.
- We explain how the presence of linear features leads to an additional interpolation phase transition not present in an analogous model with nonlinear activation functions. We use random matrix theory to argue that the underlying reason for this difference is that nonlinear basis functions implicitly regularize small eigenvalues in the design matrix.

II. THEORETICAL SETUP

In this work, we focus on the supervised learning task of using relationships learned from a training data set, consisting of labels and associated input features, to accurately predict the labels of new data points from their input features. Here, we closely follow the theoretical formalism previously described in Ref. 35.

A. Data Distribution (Teacher Model)

We consider data points $(y, \vec{\mathbf{x}})$, each consisting of a continuous label y paired with a set of N_f continuous input features $\vec{\mathbf{x}}$. We assume that the relationship between the input features and labels (the data distribution or teacher model) can be expressed as

$$y(\vec{\mathbf{x}}) = y^*(\vec{\mathbf{x}}; \vec{\boldsymbol{\beta}}) + \varepsilon \tag{1}$$

where ε is the label noise. The unknown function $y^*(\vec{\mathbf{x}}; \vec{\boldsymbol{\beta}})$ represents the "true" labels and depends on a set of N_f "ground truth" parameters $\vec{\boldsymbol{\beta}}$, characterizing the correlations between the features and labels. Here, we restrict ourselves to a teacher model of the form

$$y^*(\vec{\mathbf{x}}; \vec{\boldsymbol{\beta}}) = \frac{\sigma_\beta \sigma_X}{\langle f' \rangle} f\left(\frac{\vec{\mathbf{x}} \cdot \vec{\boldsymbol{\beta}}}{\sigma_X \sigma_\beta}\right), \tag{2}$$

where the function f is an arbitrary nonlinear function and $\langle f' \rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}h e^{-\frac{h^2}{2}} f'(h)$ is a normalization constant chosen for convenience with prime notation used to indicate a derivative. Note that Eq. (2) reduces to a linear teacher model $y^*(\vec{\mathbf{x}}) = \vec{\mathbf{x}} \cdot \vec{\boldsymbol{\beta}}$ when $f(h) = h$.

We draw the input features for each data point independently and identically from a normal distribution with zero mean and variance σ_X^2/N_f. We consider ground truth parameters $\vec{\boldsymbol{\beta}}$ and label noise ε that are drawn independently from normal distributions with zero mean and variances σ_β^2 and σ_ε^2, respectively. Furthermore, we assume the labels are centered so that f has zero mean with respect to its argument.

B. Model Architectures (Student Models)

We consider a student model of the form

$$\hat{y}(\vec{\mathbf{x}}) = \vec{\mathbf{z}}(\vec{\mathbf{x}}) \cdot \hat{\mathbf{w}}, \tag{3}$$

where $\hat{\mathbf{w}}$ is a vector of N_p fit parameters. For the random linear features model, the vector of 'hidden" features $\vec{\mathbf{z}}(\vec{\mathbf{x}})$ takes the form

$$\vec{\mathbf{z}}(\vec{\mathbf{x}}) = W^T \vec{\mathbf{x}}, \tag{4}$$

where W is a random transformation matrix of size $N_f \times N_p$, whose elements are drawn independently from a normal distribution with zero mean and variance σ_W^2/N_p.

C. Fitting Procedure

We train each model on a training data set consisting of M data points, $\mathcal{D} = \{(y_a, \vec{\mathbf{x}}_a)\}_{a=1}^{M}$. For convenience, we organize the vectors of input features in the training set into an observation matrix X of size $M \times N_f$ and define the length-M vectors of training labels $\vec{\mathbf{y}}$, training label noise $\vec{\boldsymbol{\varepsilon}}$, and label predictions for the training set $\hat{\mathbf{y}}$. We also organize the vectors of hidden features evaluated on the input features of the training set, $\{\vec{\mathbf{z}}(\vec{\mathbf{x}}_a)\}_{a=1}^{M}$, into the rows of a hidden feature matrix Z of size $M \times N_p$.

Given a set of training data $\mathcal{D}$, we solve for the optimal values of the fit parameters $\hat{\mathbf{w}}$ by minimizing the standard ridge regression loss function composed of the mean squared label error with L_2 regularization,

$$L(\hat{\mathbf{w}}; \mathcal{D}) = \frac{1}{2}\|\Delta\vec{\mathbf{y}}\|^2 + \frac{\lambda}{2}\|\hat{\mathbf{w}}\|^2, \tag{5}$$

where the notation $\|\cdot\|$ indicates an L_2 norm, $\Delta\vec{\mathbf{y}} = \vec{\mathbf{y}} - \hat{\mathbf{y}}$ is the vector of residual label errors for the training set, and λ is the regularization parameter. The exact solution for the fit parameters resulting from this loss function are

$$\hat{\mathbf{w}} = \left[\lambda I_{N_p} + Z^T Z\right]^{-1} Z^T \vec{\mathbf{y}}. \tag{6}$$

We will often work in the "ridge-less limit" where we take the limit $\lambda \to 0$. In this limit, we refer to the matrix $Z^T Z$ in the above expression as the Hessian matrix.

D. Model Evaluation

To evaluate a model's prediction accuracy, we measure the training and test (generalization) errors. We define the training error as the mean squared residual label error of the training data,

$$\mathcal{E}_{\text{train}} = \frac{1}{M}\|\Delta\vec{\mathbf{y}}\|^2. \tag{7}$$

We define the interpolation threshold as the model complexity (number of fit parameters N_p) at which the training error becomes exactly zero (in the ridge-less limit). Analogously, we define the test error as the mean squared error evaluated on a test data set, $\mathcal{D}' = \{(y'_a, \vec{\mathbf{x}}'_a)\}_{a=1}^{M'}$, composed of M' new data points drawn independently from the same data distribution as the training set,

$$\mathcal{E}_{\text{test}} = \frac{1}{M'}\|\Delta\vec{\mathbf{y}}'\|^2, \tag{8}$$

where $\Delta\vec{\mathbf{y}}' = \vec{\mathbf{y}}' - \hat{\mathbf{y}}'$ is a length-M' vector of residual label errors between the vector of test labels $\vec{\mathbf{y}}'$ and their predicted values $\hat{\mathbf{y}}'$. Furthermore, we define the ensemble-averaged training and test errors, $\langle\mathcal{E}_{\text{train}}\rangle$ and $\langle\mathcal{E}_{\text{test}}\rangle$, respectively, by taking averages of the above definitions with respect to all sources of randomness (e.g., X, $\vec{\boldsymbol{\varepsilon}}$, $\vec{\boldsymbol{\beta}}$, etc.).

E. Bias-Variance Decomposition

The bias-variance decomposition separates test error into components stemming from three distinct sources: bias, variance, and noise. Here, we utilize the standard definitions of bias and variance [1, 2],

$$\text{Bias}[\hat{y}(\vec{\mathbf{x}})] = \text{E}_{\mathcal{D}}[\hat{y}(\vec{\mathbf{x}})] - y^*(\vec{\mathbf{x}}) \tag{9}$$

$$\text{Var}[\hat{y}(\vec{\mathbf{x}})] = \text{E}_{\mathcal{D}}\left[\hat{y}^2(\vec{\mathbf{x}})\right] - \text{E}_{\mathcal{D}}[\hat{y}(\vec{\mathbf{x}})]^2, \tag{10}$$

where $\vec{\mathbf{x}}$ is an arbitrary test data point and the subscript $\mathcal{D}$ denotes the sampling average with respect to the training set (i.e., with respect to the input features X and label noise $\vec{\boldsymbol{\varepsilon}}$, but not the ground truth parameters $\vec{\boldsymbol{\beta}}$).

In order to incorporate other sources of randomness (e.g., $\vec{\boldsymbol{\beta}}$ and W), we define the more general ensemble-averaged squared bias and variance, respectively, as

$$\langle\text{Bias}^2[\hat{y}]\rangle = \text{E}_{\vec{\boldsymbol{\beta}},W,\vec{\mathbf{x}}}\left[\text{Bias}[\hat{y}(\vec{\mathbf{x}})]^2\right] \tag{11}$$

$$\langle\text{Var}[\hat{y}]\rangle = \text{E}_{\vec{\boldsymbol{\beta}},W,\vec{\mathbf{x}}}[\text{Var}[\hat{y}(\vec{\mathbf{x}})]]. \tag{12}$$

Using these definitions, we define the ensemble-averaged bias-variance decomposition of the test error,

$$\langle\mathcal{E}_{\text{test}}\rangle = \langle\text{Bias}^2[\hat{y}]\rangle + \langle\text{Var}[\hat{y}]\rangle + \sigma_\varepsilon^2, \tag{13}$$

which we utilize throughout this work.

F. Derivation of Closed-Form Solutions

Following the derivations in Ref. 35, we utilize the zero-temperature cavity method to derive closed-form expressions for the training error, test error, bias, and variance. In this derivation, we work in the thermodynamic limit, where $N_f, M, N_p \to \infty$, but their ratios, $\alpha_f = N_f/M$ and $\alpha_p = N_p/M$, remain finite. Our results are exact in this limit. Furthermore, we utilize the procedure described in Ref. 37 to reproduce the closed-form solution for the eigenvalues spectrum of the Hessian matrix for this model. We refer the reader to the Appendix for further details on these calculations.

III. ANALYTIC EXPRESSIONS

We find that the closed-form solutions for the random linear features model are characterized by three distinct regimes, each defined by which of the following three quantities is the smallest: the number of input features N_f, the number of fit parameters (hidden features) N_p, or the size of the training set M. In terms of $\alpha_f = N_f/M$ and $\alpha_p = N_p/M$, the expressions for the ensemble-averaged training error, test error, bias, and variance are

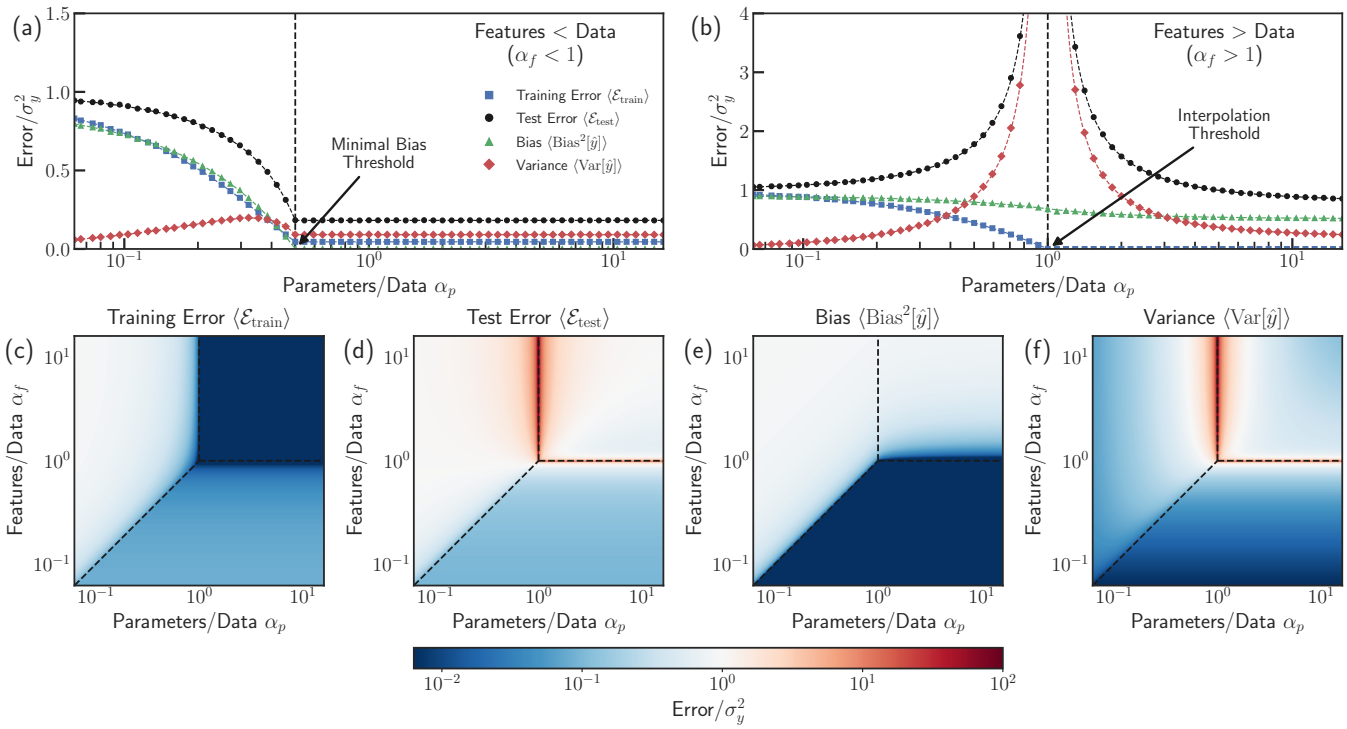

FIG. 1. **Random Linear Features Model (Two-layer Linear Neural Network).** Analytic solutions plotted as a function of $\alpha_p = N_p/M$ with fixed $\alpha_f = N_f/M$ for **(a)** less input features than training data points ($\alpha_f = 1/2$) and **(b)** more input features than training data points ($\alpha_f = 4$). Shown are the ensemble-averaged training error (blue squares), test error (black circles), squared bias (green triangles), and variance (red diamonds). Analytic solutions are indicated as dashed lines with numerical results shown as points. In (a), a black dashed vertical line marks the boundary between the large bias and minimal bias underparameterized regimes at $\alpha_p = \alpha_f$, while in (b), a similar line marks the boundary between the under- and overparameterized regimes at $\alpha_p = 1$. Analytic solutions as a function of α_p and α_f are also shown for the the ensemble-averaged **(c)** training error, **(d)** test error, **(e)** squared bias, and **(f)** variance. Results are shown for a linear teacher model $y(\vec{\mathbf{x}}) = \vec{\mathbf{x}} \cdot \vec{\boldsymbol{\beta}} + \varepsilon$, a signal-to-noise ratio of $\sigma_\beta^2 \sigma_X^2 / \sigma_\varepsilon^2 = 10$, and have been scaled by the variance of the training set labels $\sigma_y^2 = \sigma_\beta^2 \sigma_X^2 + \varepsilon^2$. In each panel, black dashed lines indicate boundaries between different regimes of the solutions depending on which is the smallest of the quantities M, N_f, or N_p. See Appendix for additional numerical details.

$$\langle \mathcal{E}_{\text{train}} \rangle = \begin{cases} (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)(1-\alpha_f) & \text{if} \quad N_f < N_p, M \\ \sigma_\beta^2 \sigma_X^2 \frac{(1-\alpha_p)(\alpha_f - \alpha_p)}{\alpha_f} + (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)(1-\alpha_p) & \text{if} \quad N_p < N_f, M \\ 0 & \text{if} \quad M < N_f, N_p \end{cases} \tag{14}$$

$$\langle \mathcal{E}_{\text{test}} \rangle = \begin{cases} (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{1}{(1-\alpha_f)} & \text{if} \quad N_f < N_p, M \\ \sigma_\beta^2 \sigma_X^2 \frac{(\alpha_f - \alpha_p)}{\alpha_f (1-\alpha_p)} + (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{1}{(1-\alpha_p)} & \text{if} \quad N_p < N_f, M \\ \sigma_\beta^2 \sigma_X^2 \frac{\alpha_p (\alpha_f - 1)}{\alpha_f (\alpha_p - 1)} + (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{(\alpha_f \alpha_p - 1)}{(\alpha_f - 1)(\alpha_p - 1)} & \text{if} \quad M < N_f, N_p \end{cases} \tag{15}$$

$$\langle \text{Bias}^2[\hat{y}] \rangle = \begin{cases} \sigma_{\delta y^*}^2 & \text{if} \quad N_f < N_p, M \\ \sigma_\beta^2 \sigma_X^2 \frac{(\alpha_f - \alpha_p)}{\alpha_f} + \sigma_{\delta y^*}^2 & \text{if} \quad N_p < N_f, M \\ \sigma_\beta^2 \sigma_X^2 \frac{\alpha_p (\alpha_f - 1)^2}{\alpha_f (\alpha_f \alpha_p - 1)} + \sigma_{\delta y^*}^2 & \text{if} \quad M < N_f, N_p \end{cases} \tag{16}$$

$$\langle \text{Var}[\hat{y}] \rangle = \begin{cases} (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{\alpha_f}{(1-\alpha_f)} & \text{if} \quad N_f < N_p, M \\ \sigma_\beta^2 \sigma_X^2 \frac{\alpha_p (\alpha_f - \alpha_p)}{\alpha_f (1-\alpha_p)} + (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{\alpha_p}{(1-\alpha_p)} & \text{if} \quad N_p < N_f, M \\ \sigma_\beta^2 \sigma_X^2 \frac{\alpha_p (\alpha_f - 1)(\alpha_f - 1 + \alpha_p - 1)}{\alpha_f (\alpha_p - 1)(\alpha_f \alpha_p - 1)} + (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{(\alpha_f - 1 + \alpha_p - 1)}{(\alpha_f - 1)(\alpha_p - 1)} & \text{if} \quad M < N_f, N_p, \end{cases} \tag{17}$$

where we have taken the limit $\lambda \to 0$ (with leading order terms of order λ^2 reported in the Appendix for quantities reported here as zero). The quantity $\sigma_{\delta y^*}^2$ is the

statistical variance of the nonlinear components of the true labels, measured via their deviation from a linear teacher model, $\sigma^2_{\delta y^*} = \mathrm{E}_{\vec{\mathbf{x}}}[(y^*(\vec{\mathbf{x}}) - \vec{\mathbf{x}}\cdot\vec{\boldsymbol{\beta}})^2]$. In the thermodynamic limit, we find that this quantity evaluates to

$$\sigma^2_{\delta y^*} = \sigma^2_\beta \sigma^2_X \Delta f, \qquad \Delta f = \frac{\langle f^2\rangle - \langle f'\rangle^2}{\langle f'\rangle^2}$$
$$\langle f^2\rangle = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{d}h e^{-\frac{1}{2}h^2} f^2(h) \tag{18}$$
$$\langle f'\rangle = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{d}h e^{-\frac{1}{2}h^2} f'(h).$$

In Figs. 1(a) and (b), we plot the training error, test error, bias, and variance as a function of $\alpha_p = N_p/M$ for fixed $\alpha_f = N_f/M$ for the two cases $\alpha_f < 1$ and $\alpha_f > 1$, respectively. To gain a better grasp of the full set of solutions, we also plot all quantities in Eqs. (14)-(17) as a function of both α_p and α_f in Figs. 1(c)-(f). All solutions are shown for a linear teacher model $y^*(\vec{\mathbf{x}}) = \vec{\mathbf{x}}\cdot\vec{\boldsymbol{\beta}}$ ($\sigma^2_{\delta y^*} = 0$).

The three regimes we observe in the closed-form solutions are separated by three distinct phase transitions. Examining the training error in Fig. 1(c), we find that it goes to zero at two of the transitions, $\alpha_p = 1$ with $\alpha_f \geq 1$ and $\alpha_f = 1$ with $\alpha_p \geq 1$, giving rise to an interpolation boundary. On one side of this boundary, where $\alpha_p < 1$ or $\alpha_f < 1$ (there are less data points M than fit parameters N_p or input features N_f), the model is underparameterized, while beyond this boundary the model is overparameterized and in the "interpolation" regime. We note that the interpolation transition for this model is markedly different from that of the random nonlinear features model (nonlinear activation function) where the interpolation threshold occurs at $\alpha_p = 1$ independently of α_f (see Sec. V and Ref. 35).

We find that the test error and variance diverge along the entire interpolation boundary, while the bias remains finite. This is in contrast with previous studies which employed non-standard definitions of bias and variance and found that both variance and bias diverge at the interpolation threshold [9]. Examining Figs. 1(a) and (b), we find that the test error (and similarly, the variance) exhibits very different behavior as a function of α_p, depending on whether $\alpha_f < 1$ (less input features than training data points $N_f < M$) or $\alpha_f > 1$ (more input features than training data points $N_f > M$). When $\alpha_f > 1$, the test error diverges at $\alpha_p = 1$ and decreases monotonically in the overparameterized regime. In contrast, when $\alpha_f < 1$, the test error monotonically decreases to a small, constant value at $\alpha_p \geq \alpha_f$. Although this model does not display the full canonical double-descent behavior in either case, the test error in the overparameterized regime is always at least as small as – if not smaller than – that of the underparameterized regime for fixed α_f.

Examining the bias in Fig 1(c) reveals that there is an additional phase transition in the underparameterized regime located at the boundary $\alpha_f = \alpha_p$ for $\alpha_p \leq 1$ and $\alpha_f \leq 1$ (i.e., when the number of input features N_f equals the number of hidden features N_p, with both N_f and N_p less than the number of data points M). This transition divides the non-interpolation solutions into two pieces. When $\alpha_p < \alpha_f$ (less fit parameters than input features $N_p < N_f$), the model exhibits a large bias because there are not enough fit parameters (or hidden features) to fully express the input features in the data [see Eq. (16)]. In contrast, when $\alpha_p > \alpha_f$ (more fit parameters than input features $N_p > N_f$), the only contribution to the bias is a small constant $\sigma^2_{\delta y^*}$ stemming from the nonlinear components of the labels. For the special case of a linear teacher model shown in the figures, the bias is identically zero in this regime.

Interestingly, we also observe that $\sigma^2_{\delta y^*}$ appears as an additive component to the label noise σ^2_ε in the training error, test error, and variance, indicating that the model interprets the nonlinear components of the labels as effective noise [35].

IV. PHASE TRANSITIONS, SUSCEPTIBLITIES, AND EIGENVALUE SPECTRA

In the previous section, we found that the analytic solutions for the random linear features model are characterized by three distinct regimes separated by three different phase transitions. As a natural byproduct of our cavity derivations, we find that each of these phase transitions is marked by a diverging susceptibility. In particular, setting the gradient of the loss function in Eq. (5) equal to a small nonzero field $\vec{\boldsymbol{\eta}}$, such that $\partial L/\partial\hat{\mathbf{w}} = \vec{\boldsymbol{\eta}}$, we define the susceptibility matrix $\partial\hat{\mathbf{w}}/\partial\vec{\boldsymbol{\eta}}$. This quantity measures the sensitivity of the fit parameters to small perturbation in the gradient and can be shown to be equivalent to the inverse Hessian of the loss function. Taking the trace of this matrix, we define the scalar susceptibility

$$\nu = \frac{1}{N_p}\mathrm{Tr}\left.\frac{\partial\hat{\mathbf{w}}}{\partial\vec{\boldsymbol{\eta}}}\right|_{\vec{\boldsymbol{\eta}}=0} = \frac{1}{N_p}\mathrm{Tr}\left[\lambda I_{N_p} + Z^T Z\right]^{-1}. \tag{19}$$

In the small λ limit, we make the approximation $\nu \approx \lambda^{-1}\nu_{-1} + \nu_0$. In exact matrix form, we find that the two coefficients are

$$\nu_{-1} = 1 - \frac{1}{N_p}\mathrm{rank}(Z^T Z), \quad \nu_0 = \frac{1}{N_p}\mathrm{Tr}\left[Z^T Z\right]^+, \tag{20}$$

where $^+$ denotes a Moore-Penrose pseudoinverse.

In Figs. 2(a) and (b), we plot the analytic closed-form expressions for these two quantities in the thermodynamic limit as a function of α_f and α_p (see Appendix for expressions). We find that the first coefficient ν_{-1} counts the fraction of fit parameters that go beyond the minimum needed to attain minimal training error. In

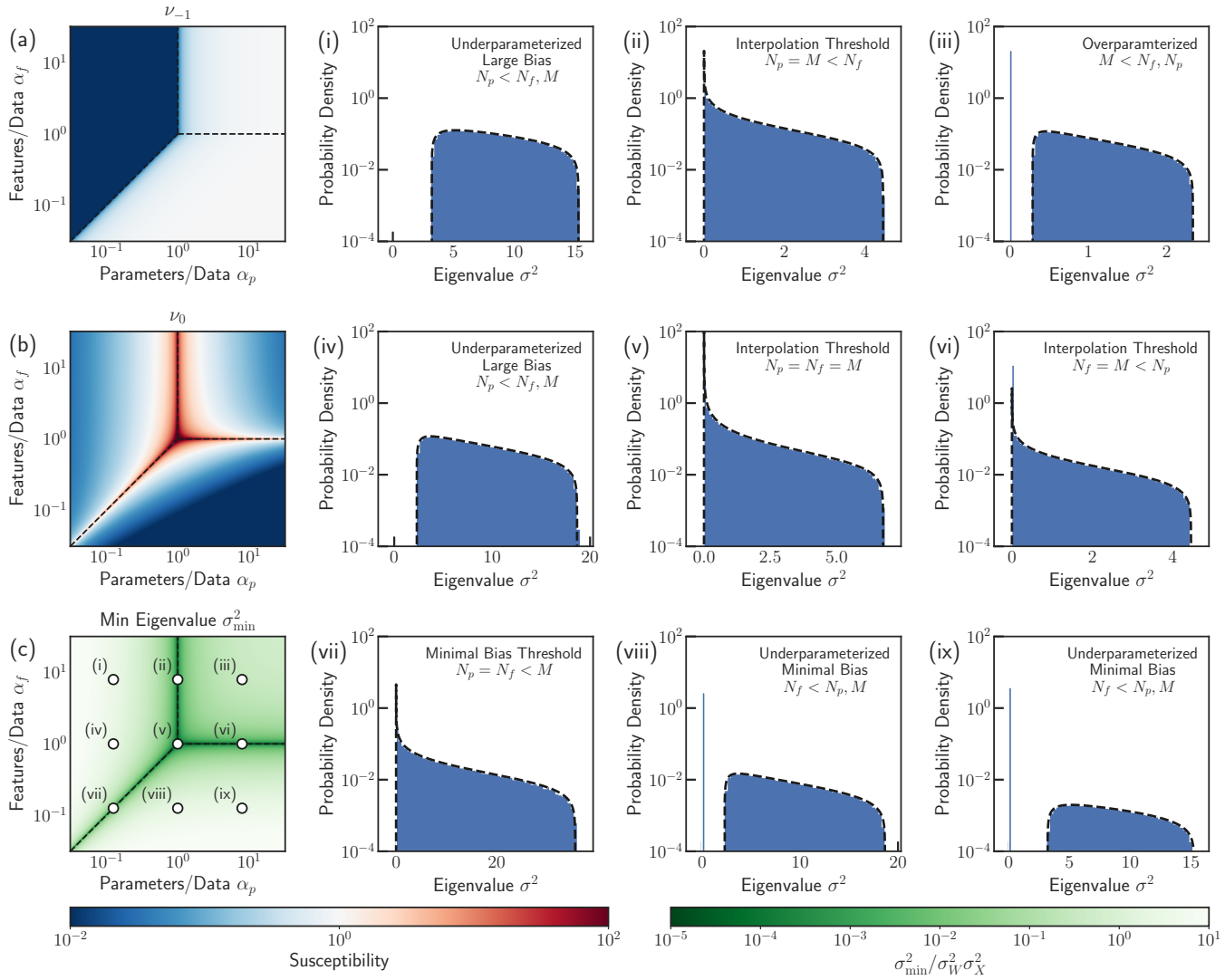

FIG. 2. **Susceptibilities and Eigenvalue Spectra (a)-(b)** Analytic solutions for the susceptibility ν which measures the sensitivity of the fit parameters to small perturbations in the gradient. In the small λ limit, $\nu \approx \lambda^{-1}\nu_{-1} + \nu_0$. **(a)** The coefficient ν_{-1} counts the fraction of unutilized fit parameters, or the fraction of parameters beyond that needed to attain minimal training error. **(b)** The coefficient ν_0 diverges at each phase transition when Z^TZ has a small eigenvalue. **(c)** Analytic solution for the minimum nonzero eigenvalue $\sigma^2_{\min}$ of the Hessian matrix Z^TZ. **(i)-(ix)** Examples of the full eigenvalue spectrum are shown for each of the corresponding points in (c). See Appendix for additional simulation details.

contrast, the second coefficient ν_0 diverges along each phase boundary. Based on the exact matrix form of ν_0 in Eq. (20), we infer that these divergences can be attributed to small eigenvalues in the Hessian matrix Z^TZ.

To illustrate this connection between the eigenvalues of the Hessian and the susceptibility ν, we note that for this problem, the inverse Hessian is equivalent to the Green's function and can be used to extract the eigenvalue spectrum [37] (see Appendix for derivation). In Fig. 2(c), we show the analytic solution for the minimum nonzero eigenvalue $\sigma^2_{\min}$ of Z^TZ. Consistent with ν_0, we find that $\sigma^2_{\min}$ goes to zero along each phase transition. In Figs. 2(i)-(ix), we also plot the eigenvalue distributions for the points indicated in Fig. 2(c). While ν_0 captures the distribution of nonzero eigenvalues, ν_{-1} captures the weight of the delta function at zero in the overparameterized and minimal bias regimes. In each regime and along each phase boundary, these distributions are qualitatively similar to the Marchenko-Pastur distribution [38]. Along each phase transition, the gap in the distribution goes to zero, while the gap is finite away from each boundary. The presence of this eigenvalue gap was previously shown to be the root cause of the decrease in variance in the overparameterized regime [35].

To understand the source of these small eigenvalue gaps, we note that the types of random matrices we consider in this work typically exhibit infinitesimally small eigenvalues in the thermodynamic limit if they contain an

equal number of rows and columns. This fact allows us to identify which matrix is the root cause of each transition. Since Z is a product of X and W, this phenomenon arises in two forms. First, Z can exhibit a small eigenvalue if either X or W is square and the expression of its input feature space is not limited by its product with the other matrix (e.g., if X is square and $N_p \geq N_f = M$ or W is square and $M \geq N_f = N_p$). This behavior explains the interpolation transition at $\alpha_f = 1$ which arises due to small eigenvalues in X, but does not extend below $\alpha_p = 1$ when the rank of W becomes too low to preserve every direction in the space of input features encoded in X. Similarly, the minimal bias transition at $\alpha_f = \alpha_p$ occurs due to small eigenvalues in W, disappearing above $\alpha_f = 1$ when the rank of X is too low to fully express the input feature space of W. Second, Z can exhibit a small eigenvalue if it is square and full rank, giving rise to the transition at $\alpha_p = 1$, but only when $\alpha_f \geq 1$.

Interestingly, we observe that the test error only diverges at the two phase transitions to the interpolation regime, but not at the minimal bias transition, despite σ^2_{min} going to zero in all three cases. This lack of divergence is explained by the fact that the minimal bias transition arises due to small eigenvalues in W, which is used to transform both the training data and the test data. Since both data sets are transformed in the same way, predictions by the model for the test set based on the training set will not be limited by small eigenvalues in W, and the test error will not diverge when W is square ($N_f = N_p$).

Finally, we wish to emphasize that the two coefficients of ν each capture a different aspect of learning. The first coefficient ν_{-1} captures overparameterization, counting the fraction of fit parameters in excess of that needed to achieve zero training error. Meanwhile, ν_0 characterizes phase transitions, diverging at the interpolation threshold in concert with the variance and at the minimal bias transition. We refer the reader to Ref. 35 for a thorough discussion of the physical interpretations of these susceptibilities for a general nonlinear model, along with the other susceptibilities of note that appear in our derivations.

V. COMPARISON TO LINEAR REGRESSION AND THE RANDOM NONLINEAR FEATURES MODEL

One of the more surprising results of our analysis is that the phase diagram for the random *linear* features model is qualitatively different from the random *nonlinear* features model. On other hand, we find that the random linear feature model is qualitatively similar to ordinary ridge-less regression when the number of hidden features matches the input features. To better understand the similarities and differences, we have reproduced the phase diagrams for all three models in Fig. 3 (see Ref. [35] for a detailed analysis of ridge regression and the random nonlinear features model).

First, we compare the random linear features model to linear regression in which the number of hidden features matches the input features,

$$\vec{\mathbf{z}}(\vec{\mathbf{x}}) = \vec{\mathbf{x}}. \tag{21}$$

Fig. 3(a) shows the training error, test error, bias, and variance for linear regression. Since linear regression lacks basis functions, in Fig. 3(b), we plot the same quantities for the random linear features model for the special case where the number of input features equals the number of hidden features ($\alpha_f = \alpha_p$).

We observe that along this cut of the phase diagram, the random linear features model behaves qualitatively similar to linear regression, with variance first increasing as one approaches the interpolation threshold ($\alpha_p = 1$) and then decreasing monotonically beyond the threshold. Meanwhile, the bias is zero below the interpolation threshold and then increases once one crosses the interpolation threshold. As discussed in detail in Ref. [35], the underlying reason for the increase in bias for $\alpha_p > 1$ is that in this regime, the model does not have enough training data points to sample the entire input feature space. Therefore, any predictions made about these unsampled directions represent implicit assumptions of the model.

Next, we compare to the random nonlinear features model in which the hidden features take the form

$$\vec{\mathbf{z}}(\vec{\mathbf{x}}) = \frac{1}{\langle \varphi' \rangle} \frac{\sigma_W \sigma_X}{\sqrt{N_p}} \varphi\left(\frac{\sqrt{N_p}}{\sigma_W \sigma_X} W^T \vec{\mathbf{x}} \right) \tag{22}$$

where φ is a nonlinear activation function that acts separately on each element of its input and $\langle \varphi' \rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dh e^{-\frac{h^2}{2}} \varphi'(h)$ is a normalization constant. Figs. 2(c)-(f) show the training error, test error, bias, and variance for this model for the case of ReLU activation, $\varphi(h) = \max(0, h)$, as a function of both α_p and α_f, while Figs. 2(g)-(j) depict the same for the random linear features model.

We observe that while the interpolation transition boundary at $\alpha_p = 1$ for $\alpha_f \geq 1$ remains the same, the addition of a nonlinear activation function suppresses the interpolation transition at $\alpha_f = 1$ for $\alpha_p > 1$, along with the transition to a minimal bias regime at $\alpha_f = \alpha_p$ for $\alpha_p < 1$. At the same time, the interpolation transition at $\alpha_p = 1$ is extended to all values of α_f.

The two changes to the shape of the interpolation boundary can be attributed to the behavior of small eigenvalues in the Hessian. Upon the introduction of a nonlinear activation, the interpolation transition at $\alpha_p = 1$ for $\alpha_f < 1$ arises due to the creation of new small eigenvalues in Z. The nonlinear transformation promotes Z to full rank when it is square ($N_p = M$), even if the product XW is not full rank. As a result, Z exhibits small eigenvalues at this transition whether or not XW has a small eigenvalue, translating to small eigenvalues

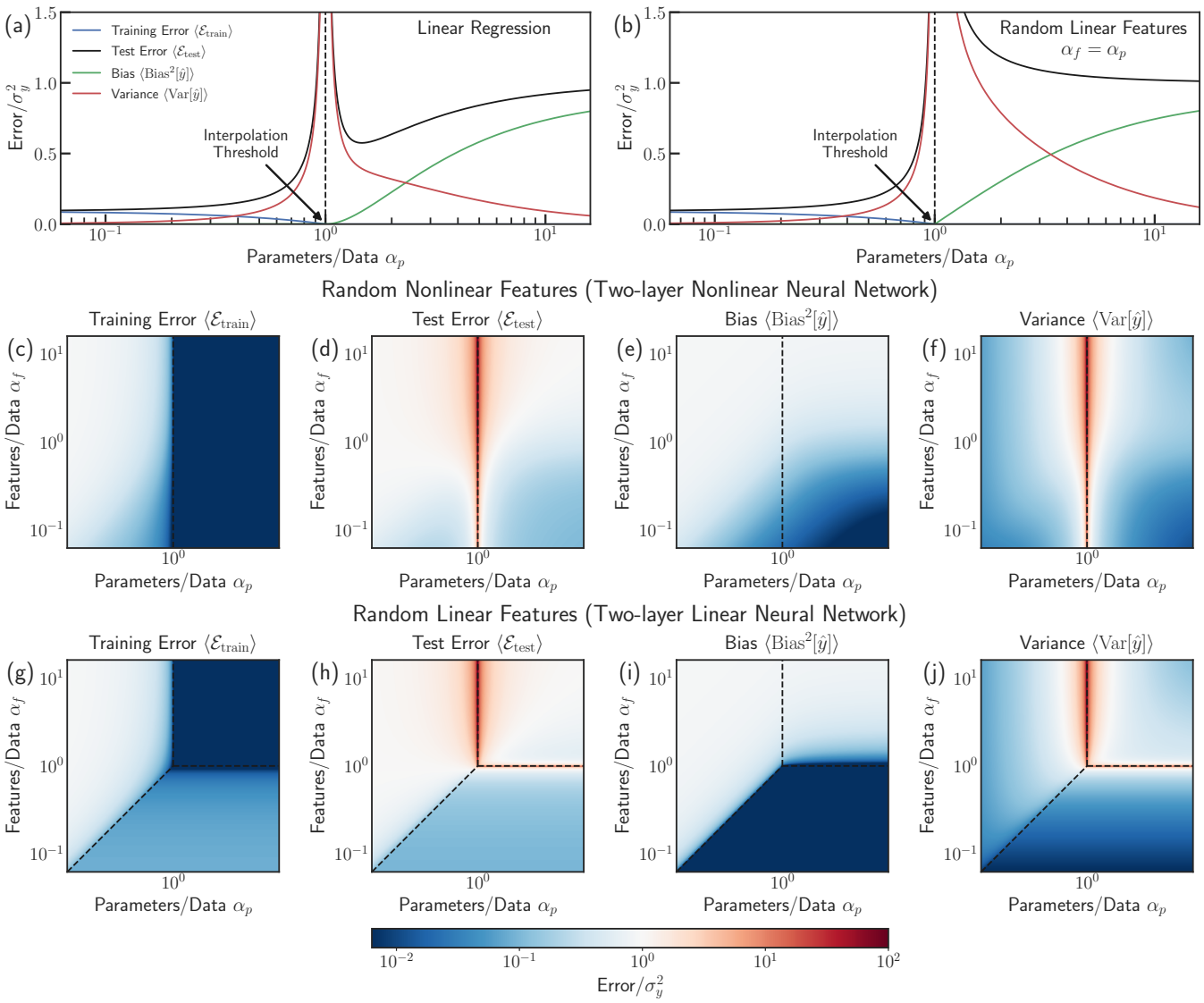

FIG. 3. **Comparison of the Random Linear Features Model to Linear Regression (No Basis Functions) and the Random Nonlinear Features Model. (a)** The training error, test error, bias, and variance as a function of α_p (or equivalently, α_f) for linear regression (no basis functions) , and **(b)** the same quantities for the random linear features model for the special case $\alpha_f = \alpha_p$. The training error, test error, bias, and variance as a function of α_p and α_f for **(c)-(f)** the random nonlinear features model with ReLU activation, $\varphi(h) = \max(0, h)$, and **(g)-(j)** the random linear features model. In all panels, black dashed lines indicate phase transitions. Results for linear regression and the random nonlinear features model are reproduced from Ref. 35.

in the Hessian and a divergence in the test error. In contrast, our random matrix theory analysis suggests that the presence of a nonlinear activation function suppresses the transition at $\alpha_f = 1$ with $\alpha_p > 1$ by serving as an implicit regularizer of small eigenvalues in the Hessian (this behavior was previously observed in Ref. 15). In particular, the use of nonlinear basis functions masks divergences arising from small eigenvalues that arise when the design matrix X is square ($N_f = M$).

Finally, to account for the removal of the minimal bias transition when $\alpha_f = \alpha_p$ with $\alpha_p < 1$, we note that the training data is generated using a teacher model that depends directly on the input features, while the nonlinear model first applies a nonlinear transformation. This nonlinear basis masks properties of the underlying input feature space like its dimension, introducing additional bias. Therefore, the bias does not approach a minimal value at $\alpha_p = \alpha_f$, even if there are in principle enough hidden features to fully encode the full space of input features. Instead, we observe that the bias only reaches a minimum in the limit $\alpha_p \to \infty$ for fixed α_f.

VI. CONCLUSIONS

A central question in machine learning is understanding why complicated models with many more parameters than training data points can still make accurate predictions. Here, we have tackled this problem by analyzing one of the simplest examples of non-trivial supervised learning: regression with random linear features. Despite the simplicity of the model, it exhibits remarkably rich behavior with multiple phase transitions.

We found that the phase diagram of the model has three distinct phases: (i) an underparameterized regime with finite training error and large bias, (ii) a second underparameterized regime with minimal bias, and (iii) an overparameterized, or interpolation, regime with zero training error. We also showed that at the transition to the interpolation regime, the variance but not the bias diverges. For this reason, while the classical bias-variance trade-off captures much of the behavior of the model in the underparameterized regime, it fails to describe the interpolation regime where the variance decreases with increasing model complexity.

We showed that the divergence of the variance is due to the presence of small eigenvalues in the Hessian matrix. This is consistent with the general picture advocated in Refs. 8 and 35 that large test errors are associated with the closing of a spectral gap near the interpolation transition. On both sides of the transition, when the spectral gap is large, it is easy to distinguish noise from poorly sampled directions in feature space. However, when the gap closes this is no longer possible, accounting for the large variance.

Our work suggests that many of the fundamental features of double-descent can be understood even when considering simple convex models. An important question is how to generalize the intuitions developed here to more complex settings. In contrast with the random linear features model, modern deep learning methods are often non-convex and attempt to learn meaningful features directly from data. In the future, it will be interesting to see how this changes the understanding of the bias-variance trade-off developed here [39–41].

ACKNOWLEDGMENTS

This work was supported by NIH NIGMS grant 1R35GM119461 and a Simons Investigator in the Mathematical Modeling of Living Systems (MMLS) award to PM. The authors also acknowledge support from the Shared Computing Cluster administered by Boston University Research Computing Services.

[1] Stuart Geman, Elie Bienenstock, and René Doursat, "Neural Networks and the Bias/Variance Dilemma," Neural Computation **4**, 1–58 (1992).
[2] Christopher M. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006).
[3] Pankaj Mehta, Marin Bukov, Ching Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab, "A high-bias, low-variance introduction to Machine Learning for physicists," Physics Reports **810**, 1–124 (2019).
[4] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, "Understanding Deep Learning Requires Re-thinking Generalization," International Conference on Learning Representations (ICLR) (2017).
[5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," Proceedings of the National Academy of Sciences **116**, 15849–15854 (2019).
[6] Marco Loog, Tom Viering, Alexander Mey, Jesse H. Krijthe, and David M. J. Tax, "A brief prehistory of double descent," Proceedings of the National Academy of Sciences **117**, 10625–10626 (2020).
[7] Ben Adlam and Jeffrey Pennington, "Understanding double descent requires a fine-grained bias-variance decomposition," Advances in Neural Information Processing Systems (NeurIPS) **33**, 11022–11032 (2020).
[8] Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky, "High-dimensional dynamics of generalization error in neural networks," Neural Networks **132**, 428–446 (2020).
[9] Jimmy Ba, Murat Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang, "Generalization of Two-layer Neural Networks: An Asymptotic Viewpoint," International Conference on Learning Representations (ICLR) (2020).
[10] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová, "Optimal errors and phase transitions in high-dimensional generalized linear models," Proceedings of the National Academy of Sciences **116**, 5451–5460 (2019).
[11] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler, "Benign overfitting in linear regression," Proceedings of the National Academy of Sciences **117**, 30063–30070 (2020).
[12] Mikhail Belkin, Daniel Hsu, and Ji Xu, "Two Models of Double Descent for Weak Features," SIAM Journal on Mathematics of Data Science **2**, 1167–1180 (2020).
[13] Koby Bibas, Yaniv Fogel, and Meir Feder, "A New Look at an Old Problem: A Universal Learning Approach to Linear Regression," IEEE International Symposium on Information Theory (ISIT) , 2304–2308 (2019).
[14] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis, "A Model of Double Descent for High-Dimensional Logistic Regression," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) , 4267–4271 (2020).
[15] Stéphane D'Ascoli, Levent Sagun, and Giulio Biroli, "Triple descent and the two kinds of overfitting: Where & why do they appear?" Advances in Neural Information Processing Systems (NeurIPS) **33** (2020).

[16] Stéphane D'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala, "Double Trouble in Double Descent: Bias and Variance(s) in the Lazy Regime," Proceedings of the 37th International Conference on Machine Learning (ICML) PMLR, **119**, 2280–2290 (2020).

[17] Michał Dereziński, Feynman Liang, and Michael W. Mahoney, "Exact expressions for double descent and implicit regularization via surrogate random design," Advances in Neural Information Processing Systems (NeurIPS) **33**, 5152–5164 (2020).

[18] Oussama Dhifallah and Yue M. Lu, "A Precise Performance Analysis of Learning with Random Features," (2020), arXiv:2008.11904.

[19] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová, "Generalisation error in learning with random features and the hidden manifold model," Proceedings of the 37th International Conference on Machine Learning (ICML) PMLR, **119**, 3452–3462 (2020).

[20] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani, "Surprises in High-Dimensional Ridgeless Least Squares Interpolation," (2019), arXiv:1903.08560.

[21] Arthur Jacot, Berfin Şimşek, Francesco Spadaro, Clément Hongler, and Franck Gabriel, "Implicit regularization of random feature models," Proceedings of the 37th International Conference on Machine Learning (ICML) PMLR, **119**, 4631–4640 (2020).

[22] Ganesh Ramachandra Kini and Christos Thrampoulidis, "Analytic Study of Double Descent in Binary Classification: The Impact of Loss," IEEE International Symposium on Information Theory (ISIT) , 2527–2532 (2020).

[23] Andrew K. Lampinen and Surya Ganguli, "An analytic theory of generalization dynamics and transfer learning in deep linear networks," International Conference on Learning Representations (ICLR) (2019).

[24] Zhu Li, Weijie J. Su, and Dino Sejdinovic, "Benign overfitting and noisy features," (2020), arXiv:2008.02901.

[25] Tengyuan Liang and Alexander Rakhlin, "Just interpolate: Kernel "Ridgeless" regression can generalize," Annals of Statistics **48**, 1329–1347 (2020).

[26] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai, "On the Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels," Proceedings of Thirty Third Conference on Learning Theory PMLR, **125**, 2683–2711 (2020).

[27] Zhenyu Liao, Romain Couillet, and Michael W Mahoney, "A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent," Advances in Neural Information Processing Systems (NeurIPS) **33** (2020).

[28] Licong Lin and Edgar Dobriban, "What causes the test error? Going beyond bias-variance via ANOVA," Journal of Machine Learning Research **22**, 1–82 (2021).

[29] Partha P Mitra, "Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation," (2019), arXiv:1906.03667.

[30] Song Mei and Andrea Montanari, "The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve," Communications on Pure and Applied Mathematics (2021), 10.1002/cpa.22008.

[31] Vidya Muthukumar, Kailas Vodrahalli, and Anant Sahai, "Harmless interpolation of noisy data in regression," IEEE International Symposium on Information Theory (ISIT) , 2299–2303 (2019).

[32] Preetum Nakkiran, "More Data Can Hurt for Linear Regression: Sample-wise Double Descent," (2019), arXiv:1912.07242.

[33] Ji Xu and Daniel Hsu, "On the number of variables to use in principal component regression," Advances in Neural Information Processing Systems (NeurIPS) **32** (2019).

[34] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma, "Rethinking Bias-Variance Trade-off for Generalization of Neural Networks," Proceedings of the 37th International Conference on Machine Learning (ICML) PMLR, **119**, 10767–10777 (2020).

[35] Jason W. Rocks and Pankaj Mehta, "Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models," Physical Review Research **4**, 013201 (2022).

[36] Thomas Dupic and Isaac Pérez Castillo, "Spectral density of products of Wishart dilute random matrices. Part I: the dense case," (2014), arXiv:1401.7802.

[37] Wenping Cui, Jason W. Rocks, and Pankaj Mehta, "The perturbative resolvent method: Spectral densities of random matrix ensembles via perturbation theory," (2020), arXiv:2012.00663.

[38] Vladimir Alexandrovich Marčenko and Leonid Andreevich Pastur, "Distribution of eigenvalues for some sets of random matrices," Mathematics of the USSR-Sbornik **1**, 457–483 (1967).

[39] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina, "Entropy-SGD: biasing gradient descent into wide valleys," Journal of Statistical Mechanics: Theory and Experiment **2019**, 124018 (2019).

[40] Carlo Baldassi, Enrico M. Malatesta, Matteo Negri, and Riccardo Zecchina, "Wide flat minima and optimal generalization in classifying high-dimensional Gaussian mixtures," Journal of Statistical Mechanics: Theory and Experiment **2020**, 124012 (2020).

[41] Fabrizio Pittorino, Carlo Lucibello, Christoph Feinauer, Enrico M. Malatesta, Gabriele Perugini, Carlo Baldassi, Matteo Negri, Elizaveta Demyanenko, and Riccardo Zecchina, "Entropic gradient descent algorithms and wide flat minima," International Conference on Learning Representations (ICLR) (2021).

Appendix A: Cavity Derivations

In this section, we provide detailed derivations of all closed-form solutions for the random linear features model. These calculations follow the general procedure laid out in Ref. 35.

1. Notational 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}).

2. Nonlinear Label Decomposition

In order to calculate the statistical properties of the nonlinear teacher model, we first decompose the labels into their linear and nonlinear components as follows:

$$y(\vec{\mathbf{x}}) = \vec{\mathbf{x}} \cdot \vec{\boldsymbol{\beta}} + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}) + \varepsilon, \qquad \vec{\boldsymbol{\beta}} \equiv \Sigma^{-1}_{\vec{\mathbf{x}}} \mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}, y^*(\vec{\mathbf{x}})]. \tag{A1}$$

The first term in this decomposition captures the linear correlations between the labels y^* and the input features $\vec{\mathbf{x}}$ via the ground truth parameters $\vec{\boldsymbol{\beta}}$. The second term, defined as $\delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}) \equiv y^*(\vec{\mathbf{x}}) - \vec{\mathbf{x}} \cdot \vec{\boldsymbol{\beta}}$, represents the remaining nonlinear component of the labels. By defining the ground truth parameters as shown above, where $\Sigma_{\vec{\mathbf{x}}} = \mathrm{Cov}_{\vec{\mathbf{x}}}[\vec{\mathbf{x}}, \vec{\mathbf{x}}^T]$ is the covariance matrix of the input features (assumed to be invertible), it can be proven that the linear and nonlinear components are statistically independent. Using the definition of the nonlinear component of the labels $\delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}})$, it is straightforward to show that their mean and covariance in the thermodynamic limit evaluate to

$$\mathrm{E}[\delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_a)] = 0, \qquad \mathrm{Cov}[\delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_a), \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_b)] = \sigma^2_{\delta y^*} \delta_{ab}, \tag{A2}$$

where $\vec{\mathbf{x}}_a$ and $\vec{\mathbf{x}}_b$ are to independent data points and we have defined the variance $\sigma^2_{\delta y^*}$ of the nonlinear components as

$$\begin{aligned} \sigma^2_{\delta y^*} &= \sigma^2_\beta \sigma^2_X \Delta f, \qquad \Delta f = \frac{\langle f^2 \rangle - \langle f' \rangle^2}{\langle f' \rangle^2} \\ \langle f^2 \rangle &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dh e^{-\frac{h^2}{2}} f^2(h), \qquad \langle f' \rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dh e^{-\frac{h^2}{2}} f'(h). \end{aligned} \tag{A3}$$

The decomposition in Eq. (A1) and its statistical properties are derived in detail in Ref. 35.

3. General Solutions

Next, we derive some useful formulas for the ensemble-averaged quantities we wish to calculate. First, we express the ensemble-averaged training error as

$$\langle \mathcal{E}_{\mathrm{train}} \rangle = \langle \Delta y^2 \rangle, \qquad \langle \Delta y^2 \rangle = \mathrm{E}\left[\frac{1}{M} \sum_b \Delta y_b^2 \right] \tag{A4}$$

where we have defined $\langle \Delta y^2 \rangle$ as the mean squared label error for the training data.

Next, we evaluate the average over the test data set in the ensemble average of the test error,

$$\langle \mathcal{E}_{\mathrm{test}} \rangle = \sigma^2_X \langle \Delta \beta^2 \rangle + \sigma^2_{\delta y^*} + \sigma^2_\varepsilon. \tag{A5}$$

To obtain this expression, we have defined the set of ground truth parameters estimated by the model as $\hat{\boldsymbol{\beta}} \equiv W\hat{\mathbf{w}}$ and the corresponding residual parameter error $\Delta\vec{\boldsymbol{\beta}} \equiv \vec{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}$. The quantity $\langle\Delta\beta^2\rangle$ is then the mean squared residual parameter error,

$$\langle\Delta\beta^2\rangle = \mathrm{E}\left[\frac{1}{N_f}\sum_k \Delta\beta_k^2\right]. \tag{A6}$$

To correctly calculate the ensemble average of the squared bias, we make use of the following trick: we reinterpret the square of the average over $\mathcal{D}$ as two separate averages over uncorrelated training data sets,

$$\begin{aligned}\mathrm{Bias}^2[\hat{y}(\vec{\mathbf{x}})] &= (\mathrm{E}_{\mathcal{D}}[\hat{y}(\vec{\mathbf{x}})] - y^*(\vec{\mathbf{x}}))^2 \\ &= \mathrm{E}_{\mathcal{D}_1,\mathcal{D}_2}[(y(\vec{\mathbf{x}}) - \hat{y}_1(\vec{\mathbf{x}}))(y(\vec{\mathbf{x}}) - \hat{y}_2(\vec{\mathbf{x}}))].\end{aligned} \tag{A7}$$

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.

Next, we apply the ensemble average and explicitly average over the test data point $\vec{\mathbf{x}}$ to obtain

$$\langle\mathrm{Bias}^2[\hat{y}(\vec{\mathbf{x}})]\rangle = \sigma_X^2\,\langle\Delta\beta_1\Delta\beta_2\rangle\,, \qquad \langle\Delta\beta_1\Delta\beta_2\rangle = \mathrm{E}\left[\frac{1}{N_f}\sum_k \Delta\beta_{1,k}\Delta_{2,k}\right] \tag{A8}$$

where we have defined $\langle\Delta\beta_1\Delta\beta_2\rangle$ as the covariance of the residual label errors between the two models trained on data sets $\mathcal{D}_1$ and $\mathcal{D}_2$.

Finally, we find an expression for the variance by subtracting the bias and noise (σ_ε^2) from the test error,

$$\langle\mathrm{Var}[\hat{y}(\vec{\mathbf{x}})]\rangle = \sigma_X^2\big(\langle\Delta\beta^2\rangle - \langle\Delta\beta_1\Delta\beta_2\rangle\big). \tag{A9}$$

Based on these expressions, we find that the training error, test error, bias, and variance depend on three key ensemble-averaged quantities: $\langle\Delta y^2\rangle$, $\langle\Delta\beta^2\rangle$, and $\langle\Delta\beta_1\Delta\beta_2\rangle$. We aim to calculate these quantities in the remainder of this derivation.

4. Linear System of Equations

In this section, we derive a linear system of equations to which we will apply the cavity method. To do this, we first evaluate the gradient of the loss function in Eq. (5) with respect to the fit parameters,

$$0 = \frac{\partial L(\hat{\mathbf{w}})}{\partial \hat{w}_J} = -\sum_b \Delta y_b Z_{bJ} + \lambda\hat{w}_J. \tag{A10}$$

In addition to this gradient equation, we will also need the equations for the residual label errors for the training set,

$$\Delta y_a = y^*(\vec{\mathbf{x}}_a) + \varepsilon_a - \sum_K \hat{w}_K Z_{aJ}. \tag{A11}$$

Next, we decompose these two sets of equations such that they are linear in the random matrices W and X, resulting in four different sets of equations,

$$\begin{aligned}\lambda\hat{w}_J &= \sum_k \hat{u}_k W_{kJ} + \eta_J \\ \hat{u}_j &= \sum_b \Delta y_b X_{bj} + \psi_j \\ \Delta y_a &= \sum_k \Delta\beta_k X_{ak} + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_a) + \varepsilon_a + \xi_a \\ \Delta\beta_j &= \beta_j - \sum_K \hat{w}_K W_{jK} + \zeta_j,\end{aligned} \tag{A12}$$

where we have also utilized Eq. (A1) to decompose the training labels into their linear and nonlinear components. We have also added a small auxiliary field, η_J, ψ_j, ξ_a, or ζ_j, to each equation. We will use these extra fields to define perturbations about the solutions to these equations with the intent of setting the fields to zero at the end of the derivation. The quantities $\hat{u}_j$ can be interpreted as representations of the fit parameters in the space of input features.

5. 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. 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

$$\begin{aligned}
\lambda\hat{w}_J &= \sum_k \hat{u}_k W_{kJ} + \eta_J + \hat{u}_0 W_{0J} \\
\hat{u}_j &= \sum_b \Delta y_b X_{bj} + \psi_j + \Delta y_0 X_{0j} \\
\Delta y_a &= \sum_k \Delta\beta_k X_{ak} + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_a) + \varepsilon_a + \xi_a + X_{a0}\Delta\beta_0 \\
\Delta\beta_j &= \beta_j - \sum_K \hat{w}_K W_{jK} + \zeta_j - \hat{w}_0 W_{j0},
\end{aligned} \tag{A13}$$

with each new variable described by a new equation,

$$\begin{aligned}
\lambda\hat{w}_0 &= \sum_k \hat{u}_k W_{k0} + \eta_0 + \hat{u}_0 W_{00} \\
\hat{u}_0 &= \sum_b \Delta y_b X_{b0} + \psi_0 + \Delta y_0 X_{00} \\
\Delta y_0 &= \sum_k \Delta\beta_k X_{0k} + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_0) + \varepsilon_0 + \xi_0 + \Delta\beta_0 X_{00} \\
\Delta\beta_0 &= \beta_0 - \sum_K \hat{w}_K W_{0K} + \zeta_0 - \hat{w}_0 W_{00}.
\end{aligned} \tag{A14}$$

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. (A13) as small perturbations to the auxiliary fields,

$$\delta\eta_J = \hat{u}_0 W_{0J}, \qquad \delta\psi_j = \Delta y_0 X_{0j}, \qquad \delta\xi_a = \Delta\beta_0 X_{a0}, \qquad \delta\zeta_j = -\hat{w}_0 W_{j0}, \tag{A15}$$

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

$$\begin{aligned}
\hat{w}_J &\approx \hat{w}_{J\backslash 0} + \sum_K \nu^{\hat{w}}_{JK}\delta\eta_K + \sum_k \phi^{\hat{w}}_{Jk}\delta\psi_k + \sum_b \chi^{\hat{w}}_{Jb}\delta\xi_b + \sum_k \omega^{\hat{w}}_{Jk}\delta\zeta_k \\
\hat{u}_j &\approx \hat{u}_{j\backslash 0} + \sum_K \nu^{\hat{u}}_{jK}\delta\eta_K + \sum_k \phi^{\hat{u}}_{jk}\delta\psi_k + \sum_b \chi^{\hat{u}}_{jb}\delta\xi_b + \sum_k \omega^{\hat{u}}_{jk}\delta\zeta_k \\
\Delta y_a &\approx \Delta y_{a\backslash 0} + \sum_K \nu^{\Delta y}_{aK}\delta\eta_K + \sum_k \phi^{\Delta y}_{ak}\delta\psi_k + \sum_b \chi^{\Delta y}_{ab}\delta\xi_b + \sum_k \omega^{\Delta y}_{ak}\delta\zeta_k \\
\Delta\beta_j &\approx \Delta\beta_{j\backslash 0} + \sum_K \nu^{\Delta\beta}_{jK}\delta\eta_K + \sum_k \phi^{\Delta\beta}_{jk}\delta\psi_k + \sum_b \chi^{\Delta\beta}_{jb}\delta\xi_b + \sum_k \omega^{\Delta\beta}_{jk}\delta\zeta_k.
\end{aligned} \tag{A16}$$

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

$$\begin{aligned}
\nu^{\hat{w}}_{JK} &= \frac{\partial\hat{w}_J}{\partial\eta_K}, & \phi^{\hat{w}}_{Jk} &= \frac{\partial\hat{w}_J}{\partial\psi_k}, & \chi^{\hat{w}}_{Jb} &= \frac{\partial\hat{w}_J}{\partial\xi_b}, & \omega^{\hat{w}}_{Jk} &= \frac{\partial\hat{w}_J}{\partial\zeta_k}, \\
\nu^{\hat{u}}_{jK} &= \frac{\partial\hat{u}_j}{\partial\eta_K}, & \phi^{\hat{u}}_{jk} &= \frac{\partial\hat{u}_j}{\partial\psi_k}, & \chi^{\hat{u}}_{jb} &= \frac{\partial\hat{u}_j}{\partial\xi_b}, & \omega^{\hat{u}}_{jk} &= \frac{\partial\hat{u}_j}{\partial\zeta_k}, \\
\nu^{\Delta y}_{aK} &= \frac{\partial\Delta y_a}{\partial\eta_K}, & \phi^{\Delta y}_{ak} &= \frac{\partial\Delta y_a}{\partial\psi_k}, & \chi^{\Delta y}_{ab} &= \frac{\partial\Delta y_a}{\partial\xi_b}, & \omega^{\Delta y}_{ak} &= \frac{\partial\Delta y_a}{\partial\zeta_k}, \\
\nu^{\Delta\beta}_{jK} &= \frac{\partial\Delta\beta_j}{\partial\eta_K}, & \phi^{\Delta\beta}_{jk} &= \frac{\partial\Delta\beta_j}{\partial\psi_k}, & \chi^{\Delta\beta}_{jb} &= \frac{\partial\Delta\beta_j}{\partial\xi_b}, & \omega^{\Delta\beta}_{jk} &= \frac{\partial\Delta\beta_j}{\partial\zeta_k}.
\end{aligned} \tag{A17}$$

a. Central Limit Theorem

Substituting the expansions in Eq. (A16) into the 0-indexed equations in Eq. (A14), we find that each of the resulting sums 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 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 $\Delta\beta_{j\backslash 0}$. The unperturbed quantities in each of these sums are statistically independent of all elements of both X and W with a 0-valued index. Using this fact, we find

$$\begin{aligned}
\sum_k \hat{u}_{k\backslash 0} W_{k0} &\approx \sigma_{\hat{w}} z_{\hat{w}}, & \sigma_{\hat{w}}^2 &= \sigma_W^2 \frac{\alpha_f}{\alpha_p} \langle \hat{u}^2 \rangle, & \langle \hat{u}^2 \rangle &= \frac{1}{N_f} \sum_k \hat{u}_{k\backslash 0}^2 \\
\sum_b \Delta y_{b\backslash 0} X_{b0} &\approx \sigma_{\hat{u}} z_{\hat{u}}, & \sigma_{\hat{u}}^2 &= \sigma_X^2 \alpha_f^{-1} \langle \Delta y^2 \rangle, & \langle \Delta y^2 \rangle &= \frac{1}{M} \sum_b \Delta y_{b\backslash 0}^2 \\
\sum_k \Delta\beta_{k\backslash 0} X_{0k} &\approx \sigma_{\Delta y} z_{\Delta y}, & \sigma_{\Delta y}^2 &= \sigma_X^2 \langle \Delta\beta^2 \rangle, & \langle \Delta\beta^2 \rangle &= \frac{1}{N_f} \sum_k \Delta\beta_{k\backslash 0}^2 \\
\sum_K \hat{w}_{K\backslash 0} W_{0K} &\approx \sigma_{\Delta\beta} z_{\Delta\beta}, & \sigma_{\Delta\beta}^2 &= \sigma_W^2 \langle \hat{w}^2 \rangle, & \langle \hat{w}^2 \rangle &= \frac{1}{N_p} \sum_K \hat{w}_{K\backslash 0}^2,
\end{aligned} \tag{A18}$$

where $z_{\hat{w}}$, $z_{\hat{u}}$, $z_{\Delta y}$, and $z_{\Delta\beta}$ are all random variables with zero mean and unit variance and can easily be shown to be statistically independent from one another.

Note that we have used the same notation, $\langle \Delta y^2 \rangle$ and $\langle \Delta\beta^2 \rangle$, for the two averages defined previously in Sec. A 3 even though they each lack an ensemble average. In doing so, we have employed 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 each of the sums containing one of the square susceptibility matrices. Using the fact that all of the susceptibility matrices are statistically independent of all elements of both X and W with a 0-valued index, we find that each of these sums is dominated by its mean with its variance going to zero in the thermodynamic limit,

$$\begin{aligned}
\sum_{jk} \omega_{jk}^{\hat{u}} W_{j0} W_{k0} &\approx \sigma_W^2 \frac{\alpha_f}{\alpha_p} \omega, & \omega &= \frac{1}{N_f} \sum_k \omega_{kk}^{\hat{u}} \\
\sum_{ab} \chi_{ab}^{\Delta y} X_{a0} X_{b0} &\approx \sigma_X^2 \alpha_f^{-1} \chi, & \chi &= \frac{1}{M} \sum_b \chi_{bb}^{\Delta y} \\
\sum_{jk} \phi_{jk}^{\Delta\beta} X_{0j} X_{0k} &\approx \sigma_X^2 \phi, & \phi &= \frac{1}{N_f} \sum_k \phi_{kk}^{\Delta\beta} \\
\sum_{JK} \nu_{JK}^{\hat{w}} W_{0J} W_{0K} &\approx \sigma_W^2 \nu, & \nu &= \frac{1}{N_p} \sum_K \nu_{KK}^{\hat{w}}
\end{aligned} \tag{A19}$$

where ω, χ, ϕ, and ν can be interpreted as a set of scalar susceptibilities.

Finally, it is straightforward to show that both the mean and variance each of the sums containing a rectangular susceptibility matrix goes to zero in the thermodynamic limit and can therefore be neglected.

b. Self-consistency Equations

Applying the approximations from the previous section, we find a set of self-consistent equations for $\hat{w}_0$, $\hat{u}_0$, Δy_0, and $\Delta\beta_0$,

$$\begin{aligned}
\lambda \hat{w}_0 &\approx \sigma_{\hat{w}} z_{\hat{w}} - \hat{w}_0 \sigma_W^2 \frac{\alpha_f}{\alpha_p} \omega + \eta_0 \\
\hat{u}_0 &\approx \sigma_{\hat{u}} z_{\hat{u}} + \Delta\beta_0 \sigma_X^2 \alpha_f^{-1} \chi + \psi_0 \\
\Delta y_0 &\approx \sigma_{\Delta y} z_{\Delta y} + \Delta y_0 \sigma_X^2 \phi + \delta y_{\mathrm{NL}}^*(\vec{\mathbf{x}}_0) + \varepsilon_0 + \xi_0 \\
\Delta\beta_0 &\approx \beta_0 - \sigma_{\Delta\beta} z_{\Delta\beta} - \hat{u}_0 \sigma_W^2 \nu + \zeta_0,
\end{aligned} \tag{A20}$$

where 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

$$\begin{aligned}
\hat{w}_0 &= \frac{\sigma_{\hat{w}} z_{\hat{w}} + \eta_0}{\lambda + \sigma_W^2 \frac{\alpha_f}{\alpha_p}\omega} \\
\hat{u}_0 &= \frac{\sigma_{\hat{u}} z_{\hat{u}} + \psi_0 + \sigma_X^2 \alpha_f^{-1}\chi(\beta_0 - \sigma_{\Delta\beta} z_{\Delta\beta} + \zeta_0)}{1 + \sigma_W^2 \sigma_X^2 \alpha_f^{-1}\chi\nu} \\
\Delta y_0 &= \frac{\sigma_{\Delta y} z_{\Delta y} + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_0) + \varepsilon_0 + \xi_0}{1 - \sigma_X^2 \phi} \\
\Delta\beta_0 &= \frac{\beta_0 - \sigma_{\Delta\beta} z_{\Delta\beta} + \zeta_0 - \sigma_W^2 \nu^2(\sigma_{\hat{u}} z_{\hat{u}} + \psi_0)}{1 + \sigma_W^2 \sigma_X^2 \alpha_f^{-1}\chi\nu}.
\end{aligned} \tag{A21}$$

Next, we derive a set of self-consistent equations for the scalar susceptibilities by taking appropriate derivatives of these variables with respect to the auxiliary fields,

$$\begin{aligned}
\nu &= \frac{1}{N_p}\sum_K \nu_{KK}^{\hat{w}} \approx \mathrm{E}\left[\nu_{00}^{\hat{w}}\right] = \mathrm{E}\left[\frac{\partial \hat{w}_0}{\partial \eta_0}\right] = \frac{1}{\lambda + \sigma_W^2 \frac{\alpha_f}{\alpha_p}\omega} \\
\omega &= \frac{1}{N_f}\sum_k \omega_{kk}^{\hat{u}} \approx \mathrm{E}\left[\omega_{00}^{\hat{u}}\right] = \mathrm{E}\left[\frac{\partial \hat{u}_0}{\partial \zeta_0}\right] = \frac{\sigma_X^2 \alpha_f^{-1}\chi}{1 + \sigma_W^2 \sigma_X^2 \alpha_f^{-1}\chi\nu} \\
\chi &= \frac{1}{M}\sum_b \chi_{bb}^{\Delta y} \approx \mathrm{E}\left[\chi_{00}^{\Delta y}\right] = \mathrm{E}\left[\frac{\partial \Delta y_0}{\partial \xi_0}\right] = \frac{1}{1 - \sigma_X^2 \phi} \\
\phi &= \frac{1}{N_f}\sum_k \phi_{kk}^{\Delta\beta} \approx \mathrm{E}\left[\phi_{00}^{\Delta\beta}\right] = \mathrm{E}\left[\frac{\partial \Delta\beta_0}{\partial \psi_0}\right] = -\frac{\sigma_W^2 \nu}{1 + \sigma_W^2 \sigma_X^2 \alpha_f^{-1}\chi\nu}.
\end{aligned} \tag{A22}$$

Furthermore, we note that there are two additional derivatives that have not yet appeared in the calculation up to this point, $\partial \hat{u}_j / \partial \psi_j$ and $\partial \Delta\beta_j / \partial \zeta_j$. It is clear to see from the equations for $\hat{u}_0$ and $\Delta\beta_0$ that these two additional derivatives are equivalent. Evaluating these derivatives, we define a fifth scalar susceptibility,

$$\kappa = \frac{1}{N_f}\sum_k \phi_{kk}^{\hat{u}} = \frac{1}{N_f}\sum_k \omega_{kk}^{\Delta\beta} \approx \mathrm{E}\left[\frac{\partial \hat{u}_0}{\partial \psi_0}\right] = \mathrm{E}\left[\frac{\partial \Delta\beta_0}{\partial \zeta_0}\right] = \frac{1}{1 + \sigma_W^2 \sigma_X^2 \alpha_f^{-1}\chi\nu}. \tag{A23}$$

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

$$\omega = \sigma_X^2 \alpha_f^{-1}\chi\kappa, \qquad \phi = -\sigma_W^2 \nu\kappa, \qquad \chi = \frac{1}{1 + \sigma_W^2 \sigma_X^2 \nu\kappa}, \qquad \nu = \frac{1}{\lambda + \sigma_W^2 \sigma_X^2 \alpha_p^{-1}\chi\kappa}. \tag{A24}$$

Finally, we square and average each of the expressions in Eq. (A21) to find self-consistent equations for the four mean squared averages (setting the auxiliary fields to zero),

$$\begin{aligned}
\langle \hat{w}^2 \rangle &= \frac{1}{N_p}\sum_K \hat{w}_{K\setminus 0}^2 \approx \mathrm{E}\left[\hat{w}_0^2\right] = \nu^2 \sigma_W^2 \frac{\alpha_f}{\alpha_p}\langle \hat{u}^2 \rangle \\
\langle \hat{u}^2 \rangle &= \frac{1}{N_f}\sum_k \hat{u}_{k\setminus 0}^2 \approx \mathrm{E}\left[\hat{u}_0^2\right] = \kappa^2 \sigma_X^2 \alpha_f^{-1}\langle \Delta y^2 \rangle + \omega^2\left(\sigma_\beta^2 + \sigma_W^2 \langle \hat{w}^2 \rangle\right) \\
\langle \Delta y^2 \rangle &= \frac{1}{M}\sum_b \Delta y_{b\setminus 0}^2 \approx \mathrm{E}\left[\Delta y_0^2\right] = \chi^2\left(\sigma_X^2 \langle \Delta\beta^2 \rangle + \sigma_{\delta y^*}^2 + \sigma_\varepsilon^2\right) \\
\langle \Delta\beta^2 \rangle &= \frac{1}{N_f}\sum_k \Delta\beta_{k\setminus 0}^2 \approx \mathrm{E}\left[\Delta\beta_0^2\right] = \kappa^2\left(\sigma_\beta^2 + \sigma_W^2 \langle \hat{w}^2 \rangle\right) + \phi^2 \sigma_X^2 \alpha_f^{-1}\langle \Delta y^2 \rangle.
\end{aligned} \tag{A25}$$

c. *Solution with Finite Regularization ($\lambda \sim 1$)*

Next, we derive the solutions when the regularization parameter λ is finite. To do this, we combine the self-consistency equations for the susceptibilities in Eqs. (A23) and (A24) to derive a cubic equation for χ,

$$0 = \chi^3 + (\alpha_f + \alpha_p - 2)\chi^2 + \left[(\alpha_f - 1)(\alpha_p - 1) + \alpha_f \alpha_p \bar{\lambda}\right]\chi - \alpha_f \alpha_p \bar{\lambda}, \tag{A26}$$

where we have defined the dimensionless regularization parameter

$$\bar{\lambda} = \frac{\lambda}{\sigma_W^2 \sigma_X^2}. \tag{A27}$$

This cubic equation indicates that we should expect three different solutions for χ. Using these solutions, we can derive the associated solutions for the rest of the susceptibilities. Furthermore, we solve Eq. (A25) to find

$$\begin{pmatrix} \langle \hat{w}^2 \rangle \\ \langle \hat{u}^2 \rangle \\ \langle \Delta y^2 \rangle \\ \langle \Delta \beta^2 \rangle \end{pmatrix} = \begin{pmatrix} 1 & -\sigma_W^2 \frac{\alpha_f}{\alpha_p} \nu^2 & 0 & 0 \\ -\sigma_W^2 \omega^2 & 1 & -\sigma_X^2 \alpha_f^{-1} \kappa^2 & 0 \\ 0 & 0 & 1 & -\sigma_X^2 \chi^2 \\ -\sigma_W^2 \kappa^2 & 0 & -\sigma_X^2 \alpha_f^{-1} \phi^2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \sigma_\beta^2 \omega^2 \\ (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \chi^2 \\ \sigma_\beta^2 \kappa^2 \end{pmatrix}. \tag{A28}$$

In combination with the solutions for the five scalar susceptibilities, these solutions are exact in the thermodynamic limit.

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

Next, we take the ridge-less limit in which $\lambda \to 0$. Based on the cubic equation for χ in Eq. (A26), we make the ansatz that the lowest order contribution to χ is $\mathcal{O}(1)$ in small $\bar{\lambda}$,

$$\chi \approx \chi_0 + \bar{\lambda} \chi_1. \tag{A29}$$

We then expand Eq. (A26) in orders of λ to find solutions for χ_0 and χ_1. Using these solutions, we solve for the following coefficients for the remaining susceptibilities.

$$\begin{aligned} \nu &\approx \frac{1}{\bar{\lambda}} \nu_{-1} + \nu_0 \\ \kappa &\approx \kappa_0 + \bar{\lambda} \kappa_1 \\ \phi &\approx \frac{1}{\bar{\lambda}} \phi_{-1} + \phi_0 \\ \omega &\approx \omega_0 + \bar{\lambda} \omega_1, \end{aligned} \tag{A30}$$

Finally, we expand the mean squared averages in small λ as

$$\begin{aligned} \langle \hat{w}^2 \rangle &\approx \langle \hat{w}^2 \rangle_0 + \bar{\lambda}^2 \langle \hat{w}^2 \rangle_2 \\ \langle \hat{u}^2 \rangle &\approx \langle \hat{u}^2 \rangle_0 + \bar{\lambda}^2 \langle \hat{u}^2 \rangle_2 \\ \langle \Delta y^2 \rangle &\approx \langle \Delta y^2 \rangle_0 + \bar{\lambda}^2 \langle \Delta y^2 \rangle_2 \\ \langle \Delta \beta^2 \rangle &\approx \langle \Delta \beta^2 \rangle_0 + \bar{\lambda}^2 \langle \Delta \beta^2 \rangle_2, \end{aligned} \tag{A31}$$

and then use the solutions for the susceptibilities to solve Eq. (A25) for these coefficients.

We find three sets of solutions for all quantities, corresponding to the three regimes of the random linear features model. To determine when each solution applies, we use the fact that each of the ensemble-averaged quantities $\langle \hat{w}^2 \rangle$, $\langle \hat{u}^2 \rangle$, $\langle \Delta y^2 \rangle$, and $\langle \Delta \beta^2 \rangle$ must be positive. All together, we find the solutions for the ensemble-averaged squared quantities in the $\lambda \to 0$ limit to be

$$\langle \hat{w}^2 \rangle = \begin{cases} \frac{\sigma_\beta^2}{\sigma_W^2} \frac{\alpha_f}{(\alpha_p - \alpha_f)} + \frac{(\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)}{\sigma_W^2 \sigma_X^2} \frac{\alpha_f^2}{(1-\alpha_f)(\alpha_p - \alpha_f)} & \text{if} \quad N_f < N_p, M \\ \frac{\sigma_\beta^2}{\sigma_W^2} \frac{(1-\alpha_p+\alpha_f-\alpha_p)}{(1-\alpha_p)(\alpha_f-\alpha_p)} + \frac{(\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)}{\sigma_W^2 \sigma_X^2} \frac{\alpha_f \alpha_p}{(1-\alpha_p)(\alpha_f - \alpha_p)} & \text{if} \quad N_p < N_f, M \\ \frac{\sigma_\beta^2}{\sigma_W^2} \frac{1}{(\alpha_p - 1)} + \frac{(\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)}{\sigma_W^2 \sigma_X^2} \frac{\alpha_f}{(\alpha_f - 1)(\alpha_p - 1)} & \text{if} \quad M < N_f, N_p \end{cases} \tag{A32}$$

$$\langle \hat{u}^2 \rangle = \begin{cases} \frac{\lambda^2}{\sigma_X^4 \sigma_W^4} \left[\sigma_\beta^2 \sigma_X^4 \frac{\alpha_p^3}{(\alpha_p - \alpha_f)^3} + \sigma_X^2 (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{\alpha_f \alpha_p^3}{(1-\alpha_f)(\alpha_p - \alpha_f)^3} \right] & \text{if} \quad N_f < N_p, M \\ \sigma_\beta^2 \sigma_X^4 \frac{(1-\alpha_p)(\alpha_f - \alpha_p)(1-\alpha_p+\alpha_f-\alpha_p)}{\alpha_f^3} + \sigma_X^2 (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{(1-\alpha_p)(\alpha_f - \alpha_p)}{\alpha_f^2} & \text{if} \quad N_p < N_f, M \\ \frac{\lambda^2}{\sigma_X^4 \sigma_W^4} \left[\sigma_\beta^2 \sigma_X^4 \frac{\alpha_p^3}{\alpha_f (\alpha_p - 1)^3} + \sigma_X^2 (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{\alpha_p^3}{(\alpha_f - 1)(\alpha_p - 1)^3} \right] & \text{if} \quad M < N_f, N_p \end{cases} \tag{A33}$$

$$\langle \Delta y^2 \rangle = \begin{cases} (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)(1-\alpha_f) & \text{if} \quad N_f < N_p, M \\ \sigma_\beta^2 \sigma_X^2 \frac{(1-\alpha_p)(\alpha_f-\alpha_p)}{\alpha_f} + (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)(1-\alpha_p) & \text{if} \quad N_p < N_f, M \\ \frac{\lambda^2}{\sigma_X^4 \sigma_W^4} \left[\sigma_\beta^2 \sigma_X^2 \frac{\alpha_f \alpha_p^3}{(\alpha_f-1)(\alpha_p-1)^3} + (\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2) \frac{\alpha_f^2 \alpha_p^2 (\alpha_f-1+\alpha_p-1)}{(\alpha_f-1)^3(\alpha_p-1)^3} \right] & \text{if} \quad M < N_f, N_p \end{cases} \tag{A34}$$

$$\langle \Delta \beta^2 \rangle = \begin{cases} \frac{(\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)}{\sigma_X^2} \frac{\alpha_f}{(1-\alpha_f)} & \text{if} \quad N_f < N_p, M \\ \sigma_\beta^2 \frac{(\alpha_f-\alpha_p)}{\alpha_f(1-\alpha_p)} + \frac{(\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)}{\sigma_X^2} \frac{\alpha_p}{(1-\alpha_p)} & \text{if} \quad N_p < N_f, M \\ \sigma_\beta^2 \frac{\alpha_p(\alpha_f-1)}{\alpha_f(\alpha_p-1)} + \frac{(\sigma_\varepsilon^2 + \sigma_{\delta y^*}^2)}{\sigma_X^2} \frac{(\alpha_f-1+\alpha_p-1)}{(\alpha_f-1)(\alpha_p-1)} & \text{if} \quad M < N_f, N_p. \end{cases} \tag{A35}$$

In addition, to lowest order in small λ, the five scalar susceptibilities are

$$\chi = \begin{cases} 1-\alpha_f & \text{if} \quad N_f < N_p, M \\ 1-\alpha_p & \text{if} \quad N_p < N_f, M \\ \frac{\lambda}{\sigma_W^2 \sigma_X^2} \frac{\alpha_f \alpha_p}{(1-\alpha_f)(1-\alpha_p)} & \text{if} \quad M < N_f, N_p \end{cases} \tag{A36}$$

$$\nu = \begin{cases} \frac{1}{\lambda} \frac{(\alpha_p-\alpha_f)}{\alpha_p} + \frac{1}{\sigma_W^2 \sigma_X^2} \frac{\alpha_f^2}{(1-\alpha_f)(\alpha_p-\alpha_f)} & \text{if} \quad N_f < N_p, M \\ \frac{1}{\sigma_W^2 \sigma_X^2} \frac{\alpha_f \alpha_p}{(1-\alpha_p)(\alpha_f-\alpha_p)} & \text{if} \quad N_p < N_f, M \\ \frac{1}{\lambda} \frac{(\alpha_p-1)}{\alpha_p} + \frac{1}{\sigma_W^2 \sigma_X^2} \frac{\alpha_f}{(\alpha_f-1)(\alpha_p-1)} & \text{if} \quad M < N_f, N_p \end{cases} \tag{A37}$$

$$\kappa = \begin{cases} \frac{\lambda}{\sigma_W^2 \sigma_X^2} \frac{\alpha_f \alpha_p}{(1-\alpha_f)(\alpha_p-\alpha_f)} & \text{if} \quad N_f < N_p, M \\ \frac{(\alpha_f-\alpha_p)}{\alpha_f} & \text{if} \quad N_p < N_f, M \\ \frac{(\alpha_f-1)}{\alpha_f} & \text{if} \quad M < N_f, N_p \end{cases} \tag{A38}$$

$$\omega = \begin{cases} \frac{\lambda}{\sigma_W^2} \frac{\alpha_p}{(\alpha_p-\alpha_f)} & \text{if} \quad N_f < N_p, M \\ \sigma_X^2 \frac{(1-\alpha_p)(\alpha_f-\alpha_p)}{\alpha_f^2} & \text{if} \quad N_p < N_f, M \\ \frac{\lambda}{\sigma_W^2} \frac{\alpha_p}{\alpha_f(\alpha_p-1)} & \text{if} \quad M < N_f, N_p \end{cases} \tag{A39}$$

$$\phi = \begin{cases} -\frac{1}{\sigma_X^2} \frac{\alpha_f}{(1-\alpha_f)} & \text{if} \quad N_f < N_p, M \\ -\frac{1}{\sigma_X^2} \frac{\alpha_p}{(1-\alpha_p)} & \text{if} \quad N_p < N_f, M \\ -\frac{\sigma_W^2}{\lambda} \frac{(\alpha_f-1)(\alpha_p-1)}{\alpha_f \alpha_p} & \text{if} \quad M < N_f, N_p. \end{cases} \tag{A40}$$

We use the quantities $\langle \Delta y^2 \rangle$ and $\langle \Delta \beta^2 \rangle$ above in combination with formulas for the training and test error in Sec. A 3 to obtain the expressions in Eqs. (14) and (15).

e. *Bias-Variance Decomposition*

Next, we derive the bias and variance. According to the general solutions in Eqs. (A8) and (A9), we require the quantity $\langle \Delta\beta_1 \Delta\beta_2 \rangle$. To calculate $\langle \Delta\beta_1 \Delta\beta_2 \rangle$, we apply the self-consistent equations, Eq. (A21), to two models each trained separately on one of two independent training sets, with all other random variables held in common. We specify which quantities depend on each of the training sets using a subscript 1 or 2 for data sets $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. For training set $\mathcal{D}_1$, these equations are

$$\begin{aligned} \hat{w}_{1,0} &= \nu \sigma_{\hat{w}} z_{\hat{w}_1} \\ \hat{u}_{1,0} &= \kappa \sigma_{\hat{u}} z_{\hat{u}_1} + \omega(\beta_0 - \sigma_{\Delta\beta} z_{\Delta\beta_1}) \\ \Delta y_{1,0} &= \chi(\sigma_{\Delta y} z_{\Delta y_1} + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_{1,0}) + \varepsilon_{1,0}) \\ \Delta\beta_{1,0} &= \kappa(\beta_0 - \sigma_{\Delta\beta} z_{\Delta\beta_1}) + \phi \sigma_{\hat{u}} z_{\hat{u}_1}, \end{aligned} \tag{A41}$$

while for training set $\mathcal{D}_2$, they are

$$\begin{aligned} \hat{w}_{2,0} &= \nu \sigma_{\hat{w}} z_{\hat{w}_2} \\ \hat{u}_{2,0} &= \kappa \sigma_{\hat{u}} z_{\hat{u}_2} + \omega(\beta_0 - \sigma_{\Delta\beta} z_{\Delta\beta_2}) \\ \Delta y_{2,0} &= \chi(\sigma_{\Delta y} z_{\Delta y_2} + \delta y^*_{\mathrm{NL}}(\vec{\mathbf{x}}_{2,0}) + \varepsilon_{2,0}) \\ \Delta\beta_{2,0} &= \kappa(\beta_0 - \sigma_{\Delta\beta} z_{\Delta\beta_2}) + \phi \sigma_{\hat{u}} z_{\hat{u}_2}. \end{aligned} \tag{A42}$$

Multiplying these equations and making the self-averaging approximation, we find equations for the covariance of each of the unknown variables,

$$\begin{aligned}
\langle \hat{w}_1 \hat{w}_2 \rangle &= \frac{1}{N_p} \sum_K \hat{w}_{1,K} \hat{w}_{2,K} \approx \mathrm{E}[\hat{w}_{1,0} \hat{w}_{2,0}] = \nu^2 \mathrm{E}\big[\sigma^2_{\hat{w}} z_{\hat{w}_1} z_{\hat{w}_2}\big] \\
\langle \hat{u}_1 \hat{u}_2 \rangle &= \frac{1}{N_f} \sum_k \hat{u}_{1,k} \hat{u}_{2,k} \approx \mathrm{E}[\hat{u}_{1,0} \hat{u}_{2,0}] = \kappa^2 \mathrm{E}\big[\sigma^2_{\hat{u}} z_{\hat{u}_1} z_{\hat{u}_2}\big] + \omega^2 \big(\sigma^2_\beta + \mathrm{E}\big[\sigma^2_{\Delta\beta} z_{\Delta\beta_1} z_{\Delta\beta_2}\big]\big) \\
\langle \Delta y_1 \Delta y_2 \rangle &= \frac{1}{M} \sum_b \Delta y_{1,b} \Delta y_{2,b} \approx \mathrm{E}[\Delta y_{1,0} \Delta y_{2,0}] = \chi^2 \mathrm{E}\big[\sigma^2_{\Delta y} z_{\Delta y_1} z_{\Delta y_2}\big] \\
\langle \Delta\beta_1 \Delta\beta_2 \rangle &= \frac{1}{N_f} \sum_k \Delta\beta_{1,k} \Delta\beta_{2,k} \approx \mathrm{E}[\Delta\beta_{1,0} \Delta\beta_{2,0}] = \kappa^2 \big(\sigma^2_\beta + \mathrm{E}\big[\sigma^2_{\Delta\beta} z_{\Delta\beta_1} z_{\Delta\beta_2}\big]\big) + \phi^2 \mathrm{E}\big[\sigma^2_{\hat{u}} z_{\hat{u}_1} z_{\hat{u}_2}\big].
\end{aligned} \tag{A43}$$

Next, we calculate each of the four resulting expectation values of products of random variables. Converting each of the random variables $z_{\hat{w}_1}$, $z_{\Delta\beta_1}$, etc., back into their forms as sums, we use the independence of elements of the random matrices and the other variables to find

$$\begin{aligned}
\mathrm{E}\big[\sigma^2_{\hat{w}} z_{\hat{w}_1} z_{\hat{w}_2}\big] &\approx \mathrm{E}\left[\sum_{jk} \hat{u}_{1,j\setminus 0} \hat{u}_{2,k\setminus 0} W_{j0} W_{k0}\right] = \sigma^2_W \frac{\alpha_f}{\alpha_p} \langle \hat{u}_1 \hat{u}_2 \rangle \\
\mathrm{E}\big[\sigma^2_{\Delta\beta} z_{\Delta\beta_1} z_{\Delta\beta_2}\big] &\approx \mathrm{E}\left[\sum_{JK} \hat{w}_{1,J\setminus 0} \hat{w}_{2,K\setminus 0} W_{0J} W_{0K}\right] = \sigma^2_W \langle \hat{w}_1 \hat{w}_2 \rangle \\
\mathrm{E}\big[\sigma^2_{\hat{u}} z_{\hat{u}_1} z_{\hat{u}_2}\big] &\approx \mathrm{E}\left[\sum_{ab} \Delta y_{1,a\setminus 0} \Delta y_{2,b\setminus 0} X_{1,a0} X_{2,b0}\right] = 0 \\
\mathrm{E}\big[\sigma^2_{\Delta y} z_{\Delta y_1} z_{\Delta y_2}\big] &\approx \mathrm{E}\left[\sum_{jk} \Delta\beta_{1,j\setminus 0} \Delta\beta_{2,k\setminus 0} X_{1,0j} X_{2,0k}\right] = 0.
\end{aligned} \tag{A44}$$

Substituting these results back into Eq. (A43), we find the self-consistent equations

$$\begin{aligned}
\langle \hat{w}_1 \hat{w}_2 \rangle &= \nu^2 \sigma^2_W \frac{\alpha_f}{\alpha_p} \langle \hat{u}_1 \hat{u}_2 \rangle \\
\langle \hat{u}_1 \hat{u}_2 \rangle &= \omega^2 \big(\sigma^2_\beta + \sigma^2_W \langle \hat{w}_1 \hat{w}_2 \rangle\big) \\
\langle \Delta y_1 \Delta y_2 \rangle &= 0 \\
\langle \Delta\beta_1 \Delta\beta_2 \rangle &= \kappa^2 \big(\sigma^2_\beta + \sigma^2_W \langle \hat{w}_1 \hat{w}_2 \rangle\big).
\end{aligned} \tag{A45}$$

Next, we make the ansatz that the ensemble-averaged covariances are $\mathcal{O}(1)$ in small $\bar{\lambda}$ with the next order terms at $\mathcal{O}(\bar{\lambda}^2)$,

$$\begin{aligned}
\langle \hat{w}_1 \hat{w}_2 \rangle &\approx \langle \hat{w}_1 \hat{w}_2 \rangle_0 + \bar{\lambda}^2 \langle \hat{w}_1 \hat{w}_2 \rangle_2 \\
\langle \hat{u}_1 \hat{u}_2 \rangle &\approx \langle \hat{u}_1 \hat{u}_2 \rangle_0 + \bar{\lambda}^2 \langle \hat{u}_1 \hat{u}_2 \rangle_2 \\
\langle \Delta\beta_1 \Delta\beta_2 \rangle &\approx \langle \Delta\beta_1 \Delta\beta_2 \rangle_0 + \bar{\lambda}^2 \langle \Delta\beta_1 \Delta\beta_2 \rangle_2 .
\end{aligned} \tag{A46}$$

All together, the covariances in the limit $\lambda \to 0$ are

$$\langle \hat{w}_1 \hat{w}_2 \rangle = \begin{cases} \frac{\sigma^2_\beta}{\sigma^2_W} \frac{\alpha_f}{(\alpha_p - \alpha_f)} & \text{if} \quad N_f < N_p, M \\ \frac{\sigma^2_\beta}{\sigma^2_W} \frac{\alpha_p}{(\alpha_f - \alpha_p)} & \text{if} \quad N_p < N_f, M \\ \frac{\sigma^2_\beta}{\sigma^2_W} \frac{1}{(\alpha_f \alpha_p - 1)} & \text{if} \quad M < N_f, N_p \end{cases} \tag{A47}$$

$$\langle \hat{u}_1 \hat{u}_2 \rangle = \begin{cases} \frac{\lambda^2}{\sigma^4_X \sigma^4_W} \sigma^2_\beta \sigma^4_X \frac{\alpha_p^3}{(\alpha_p - \alpha_f)^3} & \text{if} \quad N_f < N_p, M \\ \sigma^2_\beta \sigma^4_X \frac{(1-\alpha_p)^2 (\alpha_f - \alpha_p)}{\alpha_f^3} & \text{if} \quad N_p < N_f, M \\ \frac{\lambda^2}{\sigma^4_X \sigma^4_W} \sigma^2_\beta \sigma^4_X \frac{\alpha_p^3}{\alpha_f (\alpha_p - 1)^2 (\alpha_f \alpha_p - 1)} & \text{if} \quad M < N_f, N_p \end{cases} \tag{A48}$$

$$\langle \Delta y_1 \Delta y_2 \rangle = 0 \tag{A49}$$

$$\langle \Delta\beta_1 \Delta\beta_2 \rangle = \begin{cases} \frac{\lambda^2}{\sigma_X^4 \sigma_W^4} \sigma_\beta^2 \frac{\alpha_f^2 \alpha_p^3}{(1-\alpha_f)^2(\alpha_p-\alpha_f)^3} & \text{if} \quad N_f < N_p, M \\ \sigma_\beta^2 \frac{(\alpha_f-\alpha_p)}{\alpha_f} & \text{if} \quad N_p < N_f, M \\ \sigma_\beta^2 \frac{\alpha_p(\alpha_f-1)^2}{\alpha_f(\alpha_f\alpha_p-1)} & \text{if} \quad M < N_f, N_p. \end{cases} \tag{A50}$$

Finally, we use the solution for $\langle \Delta\beta_1 \Delta\beta_2 \rangle$ to find the bias and variance according to Eqs. (A8) and (A9), resulting in Eqs. (16) and (17).

Appendix B: Spectral Densities of Kernel Matrices

Here, we derive the spectral densities for the kernel matrix $Z^T Z$ using the technique laid out in Ref. 37. According to this formalism, the spectral density of the kernel can be written in terms of the scalar susceptibility ν, defined in the previous section, using the formula

$$\rho(x) = -\frac{1}{\pi} \lim_{\varepsilon \to 0^+} \operatorname{Im} \nu(-x + i\varepsilon). \tag{B1}$$

In addition, we expect there to be delta function of eigenvalues located at zero. Although the above formula can in principle be used to obtain the fraction of eigenvalues at zero, for convenience, we instead use the scalar susceptibility χ, which can be shown to be exactly

$$\chi = \frac{1}{M} \operatorname{Tr} \frac{\partial \Delta \vec{\mathbf{y}}}{\partial \vec{\boldsymbol{\xi}}} = \frac{1}{M} \operatorname{Tr} \left[I_M - Z Z^+ \right] = 1 - \frac{1}{M} \operatorname{rank}(Z^T Z). \tag{B2}$$

The fraction of eigenvalues at zero is then

$$f_{\text{zero}} = 1 - \frac{1}{N_p} \operatorname{rank}(Z^T Z) = \frac{\chi + \alpha_p - 1}{\alpha_p}. \tag{B3}$$

Next, we define dimensional versions of ν and λ,

$$\bar{\nu} = \sigma_W^2 \sigma_X^2 \nu, \qquad \bar{\lambda} = \frac{\lambda}{\sigma_W^2 \sigma_X^2}. \tag{B4}$$

Using the self-consistent equations for the scalar susceptibilities in Eqs. (A23) and (A24), we find a cubic equation for $\bar{\nu}$,

$$0 = (\alpha_p \bar{\lambda} \bar{\nu})^3 + [1 - \alpha_p + \alpha_f - \alpha_p] (\alpha_p \bar{\lambda} \bar{\nu})^2 + \left[(1-\alpha_p)(\alpha_f - \alpha_p) + \alpha_f \alpha_p \bar{\lambda} \right] (\alpha_p \bar{\lambda} \bar{\nu}) - \alpha_f \alpha_p^2 \bar{\lambda}. \tag{B5}$$

Solving this cubic equation analytically is very involved, so we refer to the solution in Ref. 36. Instead, we solve this equation numerically for the negative imaginary roots of $\nu(\lambda)$ with $\lambda = -x$, according to Eq. (B1). However, we also need to find the interval over which the eigenvalue spectrum is positive. To do this, we rewrite the equation in general form for $\alpha_p \bar{\lambda} \bar{\nu}$,

$$(\alpha_p \bar{\lambda} \bar{\nu})^3 + a_2 (\alpha_p \bar{\lambda} \bar{\nu})^2 + a_1 (\alpha_p \bar{\lambda} \bar{\nu}) + a_0 = 0, \tag{B6}$$

where the coefficients are

$$\begin{aligned} a_0 &= -\alpha_f \alpha_p^2 \bar{\lambda} \\ a_1 &= (1-\alpha_p)(\alpha_f - \alpha_p) + \alpha_f \alpha_p \bar{\lambda} \\ a_2 &= 1 - \alpha_p + \alpha_f - \alpha_p. \end{aligned} \tag{B7}$$

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

$$D(\lambda) = R^2 - Q^3 \tag{B8}$$

with

$$\begin{aligned} Q &= \frac{1}{9}\left(a_2^2 - 3a_1\right) \\ R &= \frac{1}{54}\left(9 a_2 a_1 - 27 a_0 - 2 a_2^3\right). \end{aligned} \tag{B9}$$

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 the solution for χ that we found previously, giving us

$$f_{\text{zero}} = \begin{cases} 1 - \frac{\alpha_f}{\alpha_p} & \text{if} \quad N_f < N_p, M \\ 0 & \text{if} \quad N_p < N_f, M \\ 1 - \alpha_p^{-1} & \text{if} \quad M < N_f, N_p \end{cases} = \max\left(0, 1 - \frac{\alpha_f}{\alpha_p}, 1 - \alpha_p^{-1}\right). \tag{B10}$$

Appendix C: Numerical Simulation Details

In this section, we explain our procedures for generating numerical results. Fig. 4 provides comparisons to numerical results for the training error, test error, bias, and variance.

1. 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 use a linear teacher model $y^*(\vec{\mathbf{x}}) = \vec{\mathbf{x}} \cdot \vec{\boldsymbol{\beta}}$ ($\sigma_{\delta y^*}^2 = 0$) for all plots.

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

$$\left[\lambda I_{N_p} + Z^T Z\right] \hat{\mathbf{w}} = Z^T \vec{\mathbf{y}} \tag{C1}$$

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. (6) in the main text.

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

$$\left[\lambda I_M + Z Z^T\right] \hat{\mathbf{a}} = \vec{\mathbf{y}}, \tag{C2}$$

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 $\hat{\mathbf{w}} = Z^T \hat{\mathbf{a}}$.

2. Bias-Variance Decompositions

To efficiently calculate the ensemble-averaged bias and variance, we take inspiration from Eq. (A7). 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 $\vec{\mathbf{y}}^{*\prime}$ We then record the dot 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.

3. 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.

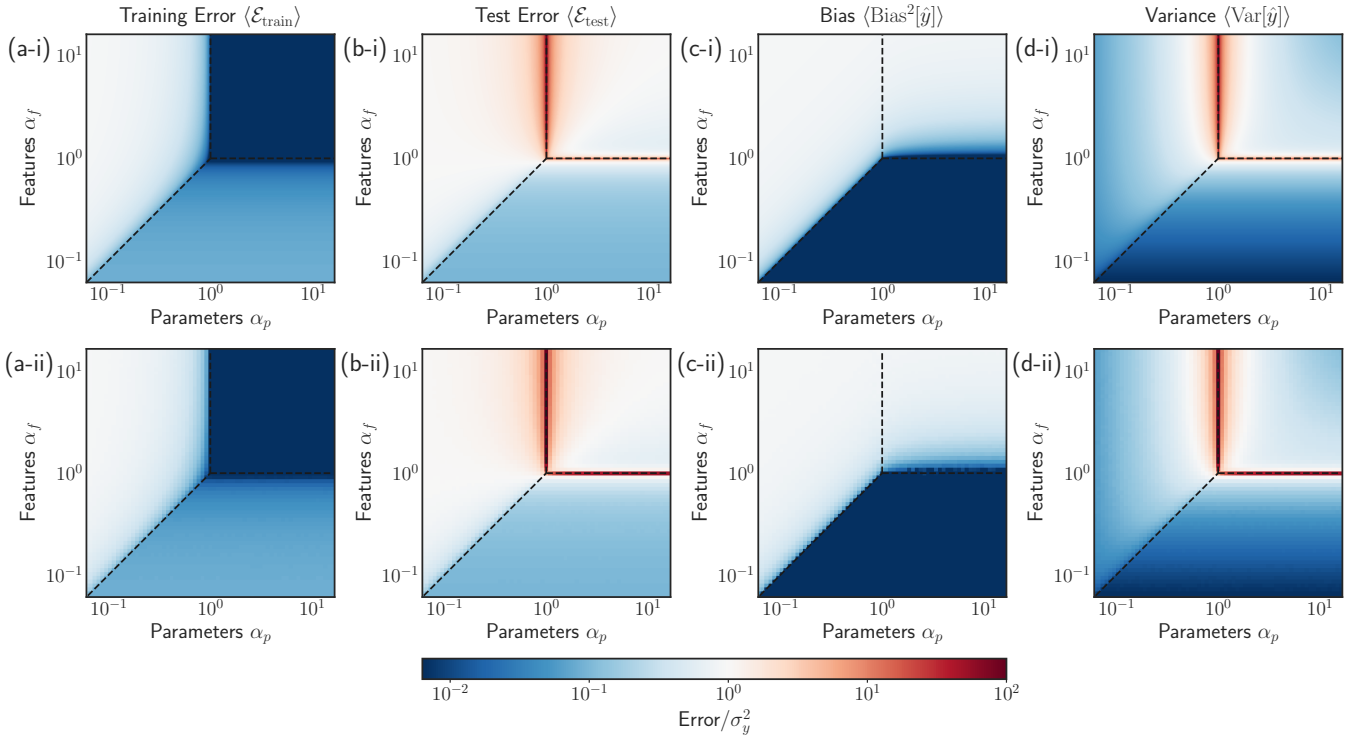

FIG. 4. **Comparison of analytic and numerical results for the random linear features model: Training error and bias-variance decomposition.** (Top Row) Analytic solutions and (Bottom Row) numerical results are shown as a function of $\alpha_p = N_p/M$ and $\alpha_f = N_f/M$. Plotted are the ensemble-averaged (a) training error, (b) test error, (c) squared bias, and (d) variance. In each panel, black dashed lines show boundaries between different regimes of solutions depending on which is smallest of the quantities M, N_f, or N_p. The vertical and horizontal lines bound the interpolation, or overparameterized, regime, located at $\alpha_p > 1$ and $\alpha_f > 1$, while the diagonal line marks the boundary between the large bias and minimal bias underparameterized regimes for a linear teacher model. All solutions have been scaled by the variance of the training set labels $\sigma_y^2 = \sigma_\beta^2 \sigma_X^2 + \sigma_\varepsilon^2$.