ANISOTROPIC RANDOM FEATURE REGRESSION IN HIGH DIMENSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

In contrast to standard statistical wisdom, modern learning algorithms typically find their best performance in the overparameterized regime in which the model has many more parameters than needed to fit the training data. A growing number of recent works have shown that random feature models can offer a detailed theoretical explanation for this unexpected behavior, but typically these analyses have utilized isotropic distributional assumptions on the underlying data generation process, thereby failing to provide a realistic characterization of real-world models that are designed to identify and harness the structure in natural data. In this work, we examine the high-dimensional asymptotics of random feature regression in the presence of structured data, allowing for arbitrary input correlations and arbitrary alignment between the data and the weights of the target function. We define a partial order on the space of weight-data alignments and prove that generalization performance improves in response to stronger alignment. We also clarify several previous observations in the literature by distinguishing the behavior of the sample-wise and parameter-wise learning curves, finding that sample-wise multiple descent can occur at scales dictated by the eigenstructure of the data covariance, but that parameter-wise multiple descent is limited to double descent, although strong anisotropy can induce additional signatures such as wide plateaus and steep cliffs. Finally, these signatures are related to phase transitions in the spectrum of the feature kernel matrix, and unlike the double descent peak, persist even under optimal regularization.

1 INTRODUCTION

While deep learning races ahead with its impressive list of practical successes, it has often done so by building larger models and training on ever larger datasets. For example, recent large language models leverage billions (Brown et al., 2020; Raffel et al., 2019) or even trillions (Fedus et al., 2021) of parameters. As the state-of-the-art pushes the computational frontier, model development, experimentation, and exploration become increasingly costly, both in terms of the computational resources and person-hours.

In order to facilitate the development of improved models, it is becoming ever more necessary to develop theoretical models that can help guide model development in these high-dimensional scenarios. Indeed, our theoretical understanding is so limited that even linear regression, the most basic machine learning method of all, is still yielding surprises in high dimensions (Hastie et al., 2019).

While linear regression can capture many salient features of high-dimensional *data*, it is not a reasonable model for high-dimensional *models*. In particular, there is no natural data-independent notion of model size: the number of parameters in the model is intimately tied (usually equal) to the dimensionality of the datapoints. This limitation means that linear regression may not accurately capture the effect of overparameterization, which appears to be a crucial ingredient in modern deep learning models (Zhang et al., 2016).

Beyond linear regression, the next simplest model is arguably the random feature model of Rahimi et al. (2007), which turns out to be tractable in high dimensions while also allowing for a natural data-independent lever to vary model size – the number of random features. Random feature models have also gained independent interest through their relationship to neural kernels and Gaussian

processes (Neal, 1996; Lee et al., 2017; Novak et al., 2018; Lee et al., 2019; Jacot et al., 2018). While recent works have examined the asymptotic performance of random feature models, they have done so in restricted setting of isotropic covariates and target functions (Mei & Montanari, 2019; d'Ascoli et al., 2020; Adlam & Pennington, 2020a;b).

Natural data tends to exhibit strong anisotropy and, as such, it is unclear to what extent the conclusions drawn from the isotropic analysis carry over to practical settings. In this work, we study high-dimensional random feature regression for anisotropic data and anisotropic target functions, generalizing prior work in this direction that was restricted to linear regression.

1.1 Contributions

Our primary contributions are to:

1. Derive theoretical predictions for the test error, bias, and variance of random feature regression for anisotropic Gaussian covariates under general linear target functions in the high-dimensional asymptotic limit (Section 3.1);
2. Prove that overparameterization reduces the total test error, bias, and variance, even in the presence of anisotropy (Section 3.2);
3. Define a partial order on the space of weight-data alignments and prove that generalization performance improves in response to stronger alignment (Definition 2.1, Section 3.3);
4. Demonstrate that anisotropy can induce sample-wise multiple descent, but prove that parameter-wise multiple descent is limited to double descent (Section 3.4);
5. Provide a theoretical explanation for steep cliffs observed in parameter-wise learning curves that persist under optimal regularization, and relate them to phase transitions in the spectrum of the feature kernel matrix (Section 3.4).

1.2 Related work

Our analysis builds on a series of works that have studied the exact high-dimensional asymptotics of the test error for a growing class of model families and data distributions. Early work in this direction examined minimum-norm interpolated least squares and ridge regression in the high-dimensional random design setting (Belkin et al., 2019; Dobriban et al., 2018; Hastie et al., 2019), finding surprising phenomena such as double descent and a negative optimal ridge constant (Kobak et al., 2020). These analyses were generalized in (Richards et al., 2021; Wu & Xu, 2020; Hastie et al., 2019) to allow for anisotropic covariance matrices and weights that generate the targets. This anisotropic configuration was studied in further detail in (Mel & Ganguli, 2021), uncovering nuanced behavior such as sample-wise multiple descent and a sequence of phase transitions in the performance and optimal regularization.

While linear regression can provide valuable insight into generalization performance in high dimensions, it is not sufficiently complex to faithfully reproduce the features of non-linear models, such as the effects of overparameterization. Random feature models, on the other hand, can capture many such features while still maintaining analytical tractability. Recent results have developed a detailed understanding of isotropic random feature regression in the high-dimensional asymptotic limit, providing a precise characterization of peaks in the test error and the benefits of overparameterization (Mei & Montanari, 2019; d'Ascoli et al., 2020; Adlam & Pennington, 2020a). The origin of these peaks was explained by means of a fine-grained decomposition of the bias and variance in (Adlam & Pennington, 2020b; Lin & Dobriban, 2020). While most prior work examining the high-dimensional asymptotics of random feature models have leveraged the isotropy of the covariate distribution, this assumption was recently relaxed in (Liao et al., 2020) in the context of random Fourier feature models, and in (Anonymous, 2021) in the study of covariate shift. These works did not pursue a detailed investigation of the anisotropy itself, nor did they analyze the interaction with anistropic target weights as we do here.

From the technical perspective, our analytic techniques build upon these works and a series of recent results stemming from the literature on random matrix theory and operator-valued free probability (Pennington & Worah, 2018; 2019; Adlam et al., 2019; Adlam & Pennington, 2020a; Louart et al., 2018; Péché et al., 2019; Far et al., 2006; Mingo & Speicher, 2017).

2 PRELIMINARIES

2.1 PROBLEM SETUP AND NOTATION

Following numerous prior works on random feature regression in high dimensions (Hastie et al., 2019; Mei & Montanari, 2019; Adlam & Pennington, 2020a; Adlam et al., 2019; Anonymous, 2021), we focus on the proportional asymptotics in which the number of samples m, input feature dimension n_0, and hidden layer size n_1 all tend to infinity at the same rate, with $\phi := n_0/m$ and $\psi := n_0/n_1$ held constant. We define the *overparameterization ratio* as $\phi/\psi = \frac{n_1}{m}$, which characterizes the normalized complexity of the random feature model.

We consider the task of learning an unknown function from m independent samples $(\mathbf{x}_i, y_i) \in \mathbb{R}^{n_0} \times \mathbb{R}, \ , i = 1, \ldots, m$, where the covariates are Gaussian, $\mathbf{x}_i \sim \mathcal{N}(0, \Sigma)$, and the labels are generated by a linear function,

$$y(\mathbf{x}_i) = \beta^\top \mathbf{x}_i / \sqrt{n_0} + \boldsymbol{\epsilon}_i \,, \tag{1}$$

where the additive noise $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is present on the training points only. We emphasize that assuming a linear target function sacrifices little in terms of generality – for a broad class of nonlinear target functions, there exists a linear function whose behavior is entirely equivalent in the proportional asymptotics (Mei & Montanari, 2019; Adlam & Pennington, 2020a). In contrast to prior work on high-dimensional random feature regression, we do not assume that the coefficient vector β is isotropic; instead, we allow for an arbitrary weight covariance, namely $\beta \sim \mathcal{N}(0, \Sigma_\beta)$. .

The predictive model is defined by kernel ridge regression using unstructured random feature maps (Rahimi et al., 2007) for which the random features are given by a single-layer neural network with random weights. In particular, for a training dataset $X = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$ and a test point $\mathbf{x}$, the random features are given by,

$$F := \sigma(WX/\sqrt{n_0}) \quad \text{and} \quad f := \sigma(W\mathbf{x}/\sqrt{n_0}) \,, \tag{2}$$

where $W \in \mathbb{R}^{n_1 \times n_0}$ is a random weight matrix with i.i.d. standard Gaussian entries, and $\sigma(\cdot) : \mathbb{R} \to \mathbb{R}$ is an activation function that is applied elementwise. The induced kernel is then given by,

$$K(\mathbf{x}_1, \mathbf{x}_2) := \frac{1}{n_1} \sigma(W\mathbf{x}_1/\sqrt{n_0})^\top \sigma(W\mathbf{x}_2/\sqrt{n_0}) \,, \tag{3}$$

and the model's predictions on a point x are

$$\hat{y}(\mathbf{x}) = Y K^{-1} K_{\mathbf{x}} \,, \tag{4}$$

where $Y := [y(\mathbf{x}_1), \ldots, y(\mathbf{x}_m)]$, $K := K(X, X) + \gamma I_m$, $K_{\mathbf{x}} := K(X, \mathbf{x})$, and $\gamma \geq 0$. Owing to the implicit regularization effect of the nonlinear feature maps (Bartlett et al., 2021), in low noise settings the optimal value of γ can sometimes be negative (Kobak et al., 2020; Wu & Xu, 2020). For simplicity, we nevertheless make the standard assumption that $\gamma \geq 0$, though we emphasize our techniques readily accommodate negative values.

The training error is defined as,

$$E_{\text{train}}^{\Sigma} = \mathbb{E}_\beta \mathbb{E}[(y(X) - \hat{y}(X))^2] = \mathbb{E}[(\beta^\top X/\sqrt{n_0} - Y K^{-1} K(X, X))^2] \,, \tag{5}$$

and the test error (without label noise on the test point) can be written as

$$E_\Sigma = \mathbb{E}_\beta \mathbb{E}_{\mathbf{x}} \mathbb{E}[(y(\mathbf{x}) - \hat{y}(\mathbf{x}))^2] - \sigma_\epsilon^2 = \mathbb{E}_{\mathbf{x}} \mathbb{E}[(\beta^\top \mathbf{x}/\sqrt{n_0} - Y K^{-1} K_{\mathbf{x}})^2] \,, \tag{6}$$

where the outermost expectation over β has been suppressed since the quantity concentrates sharply around its mean and the inner expectation is computed over all the randomness from training, *i.e.* W, X, and $\boldsymbol{\epsilon}$. The test error can be decomposed into its bias and variance components as

$$E_\Sigma = \underbrace{\mathbb{E}_{\mathbf{x}} \mathbb{E}[(\mathbb{E}[\hat{y}(\mathbf{x})] - y(\mathbf{x}))^2]}_{B_{\Sigma_\beta}} + \underbrace{\mathbb{E}_{\mathbf{x}}[\mathbb{V}[\hat{y}(\mathbf{x})]]}_{V_{\Sigma_\beta}} \,. \tag{7}$$

Following (Adlam & Pennington, 2020b), the bias and variance are also computed with respect to all the randomness from training, *i.e.* W, X, $\boldsymbol{\epsilon}$.

As noted by (Hastie et al., 2019; Wu & Xu, 2020; Mel & Ganguli, 2021), the test error of linear regression depends on the geometry of (Σ, Σ_β) (or (Σ, β) in the case of nonrandom β), and we will

find the same to be true for random feature regression. As such, we decompose the training and β covariance matrices into eigenbases as $\Sigma = \sum_{i=1}^{n_0} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ and $\Sigma_\beta = \sum_{i=1}^{n_0} \lambda_i^\beta \mathbf{v}_i^\beta \mathbf{v}_i^{\beta\top}$, where the eigenvalues are in nondecreasing magnitude, *i.e.* $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{n_0}$ and $\lambda_1^\beta \leq \lambda_2^\beta \leq \ldots \leq \lambda_{n_0}^\beta$. We then define the *overlap coefficients* as

$$q_i := \mathbf{v}_i^\top \Sigma_\beta \mathbf{v}_i = \sum_{j=1}^{n_0} (\mathbf{v}_j^\beta \cdot \mathbf{v}_i)^2 \lambda_j^\beta \,, \tag{8}$$

to measure the alignment of Σ_β with the ith eigendirection of Σ. In particular, q_i is the induced norm of $\mathbf{v}_i$ with respect to Σ_β.

Finally, for a matrix $A \in \mathbb{R}^{n \times n}$, we use $\bar{\mathrm{tr}}$ to denote the dimension-normalized trace, $\bar{\mathrm{tr}}(A) = \frac{1}{n}\mathrm{tr}(A)$, and $\|A\|_\infty$ to denote its spectral norm. We write $\delta_{\mathbf{x}}$ to denote the Dirac delta function centered at $\mathbf{x}$.

2.2 Assumptions

In order to define the limiting behavior of the test error in Eq. (6), it is necessary to impose some regularity conditions on Σ and Σ_β. As in (Wu & Xu, 2020; Anonymous, 2021), it is not sufficient to consider the limiting spectra of these matrices individually; because their eigenspaces can align, they must be considered jointly. It is convenient to do this in an eigenbasis of Σ.

Assumption 1. *We define the joint spectral distribution (JSD) as*

$$\mu_{n_0} := \frac{1}{n_0} \sum_{i=1}^{n_0} \delta_{(\lambda_i, q_i)} \tag{9}$$

and assume it converges in distribution to some μ, a distribution on $\mathbb{R}_+^2$ as $n_0 \to \infty$. We refer to μ as the limiting joint spectral distribution (LJSD), and emphasize that this defines the relevant limiting properties of the covariate and weight distributions[1]*. Additionally, we require that $\limsup_{n_0} \max(\|\Sigma\|_\infty, \|\Sigma_\beta\|_\infty) \leq C$ for a constant C.*

Often we use (λ, q) for random variables sampled jointly from μ and denote the marginal of λ under μ with μ_{data}. Since the conditional expectation $\mathbb{E}[q|\lambda]$ is an important object in our study, we assume the following for simplicity.

Assumption 2. *μ is either absolutely continuous or a finite sum of delta masses and the expectations of λ and q are finite. Moreover, $\mathbb{E}_\mu[\lambda q] = 1$.*

The normalization condition $\mathbb{E}_\mu[\lambda q] = 1$ is not necessary for our analysis, but, as in (Mel & Ganguli, 2021), it enforces consistent signal-to-noise ratios across different target functions and thereby facilitates meaningful comparisons of the test error.

When the eigenspaces of Σ and Σ_β are aligned and $q_i = \lambda_i^\beta = \Phi(\lambda_i)$ for some smooth function Φ, the support of the LJSD degenerates. Here, Assumption 1 is essentially equivalent to assuming the spectral distribution of Σ converges in distribution to some μ_{data}, which is a standard assumption in the regression literature (Dobriban et al., 2018; Mei & Montanari, 2019). One important case is when $\Sigma_\beta = I_{n_0}$, *i.e.* $\Phi : \lambda \to 1$, in which case the LJSD degenerates to $\mu_\emptyset$ defined by

$$\mu_\emptyset(\lambda, q) := \mu_{\text{data}}(\lambda)\delta_1(q) \,. \tag{10}$$

As our analysis will eventually take place in the high-dimensional limit, we further define the asymptotic covariance scale as $s = \lim_{n_0 \to \infty} \bar{\mathrm{tr}}(\Sigma) = \mathbb{E}_\mu[\lambda]$ under the limiting behavior specified in Assumption 1.

Throughout this paper, we also enforce the following standard regularity assumptions on the activation functions to ensure the existence of the moments and derivatives we compute.

Assumption 3. *The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is assumed to be differentiable almost everywhere. We assume there exists a universal constant C such that, $|\sigma(x)|, |\sigma'(x)| = O(\exp(Cx))$.*

[1] The JSD depends not only on Σ and Σ_β but also on a choice of eigendecomposition for Σ when it has repeated eigenvalues; however, as in (Anonymous, 2021), all possible choices lead to the same formulas and conclusions.

2.3 A SIMPLE FAMILY OF ANISOTROPIC DISTRIBUTIONS

Because the above assumptions allow for such a general class of covariance structures, it will prove useful to consider our results in the context of a simple family of distributions that readily admits straightforward interpretation.

To illustrate the effects of multiple scales in the input data, we define the following d-scale LJSDs: for $d \in \mathbb{N}$ and real $0 < \alpha \leq 1$ and θ,

$$\mu_{\alpha,\theta}^{d\text{-scale}} := \sum_{j=0}^{d-1} \frac{1}{d} \delta_{(C\alpha^j, D\alpha^{\theta j})}, \tag{11}$$

where $C := \left(\frac{1}{d}\frac{1-\alpha^d}{1-\alpha}\right)^{-1}$ and $D := \frac{1}{C}\left(\frac{1}{d}\frac{1-\alpha^{(\theta+1)d}}{1-\alpha^{(\theta+1)}}\right)^{-1}$ enforce the normalization conditions $s = \mathbb{E}_{\mu_{\alpha,\theta}^{d\text{-scale}}}[\lambda] = 1$ and $\mathbb{E}_{\mu_{\alpha,\theta}^{d\text{-scale}}}[q\lambda] = 1$.

These simple covariance models capture the hierarchy of scales that often characterize the structure of natural datasets. Note that by setting $\alpha = 1$ and $s = 1$, the distribution reduces to the isotropic case with identity covariance. For the nontrivial setting in which $\alpha < 1$, the data exhibits a sequence of d distinct scales where each scale is a factor α smaller than the previous one. The exponent θ parameterizes the strength of the weight-data alignment in an intuitive way. When $\theta = 0$, there is no alignment, and Σ_β is proportional to the identity. When $\theta > 0$, there is strong alignment, as the large eigendirections of the training distribution correspond to large eigendirections of Σ_β. When $\theta < 0$, there is anti-alignment, as the large eigendirections of the training distribution correspond to small eigendirections of Σ_β.

2.4 DEFINITION OF THE STRENGTH OF WEIGHT-DATA ALIGNMENT

As the preceding discussion has suggested and as we will see explicitly in the following section, the LJSD μ captures all of the information about the pair of covariance matrices (Σ, Σ_β) that is relevant for describing the asymptotic test error, bias, and variance. As the alignment between Σ and Σ_β can vary in strength among the different eigendirections, the concept of alignment is inherently multi-dimensional. Nevertheless, by requiring strong alignment along the larger eigendirections, it is possible to define the following natural partial order on the space of possible weight-data alignments.

Definition 2.1. *Let μ_1 and μ_2 be LJSDs with the same marginal distribution of λ. If the asymptotic overlap coefficients are such that $\mathbb{E}_{\mu_1}[\lambda q|\lambda] / \mathbb{E}_{\mu_2}[\lambda q|\lambda]$ is nondecreasing in λ, we say that μ_1 is more strongly aligned than μ_2 and write $\mu_1 \leq \mu_2$. Comparing against the case of isotropic weight distribution, $\mu_\emptyset$, we say μ_1 is* aligned *when $\mu_1 \leq \mu_\emptyset$ and* anti-aligned *when $\mu_1 \geq \mu_\emptyset$.*

The parameter θ of the d-scale model in Eq. (11) provides a quantitative measure of alignment under Definition 2.1, formalizing the intuition given in Section 2.3. As we will see in Section 3.3, the asymptotic generalization performance of random feature regression is strictly ordered under this definition of alignment strength.

3 MAIN RESULTS

Our main result characterizes the high-dimensional asymptotic limits of the test error, bias, and variance of the nonlinear random feature model of Section 2. Before stating the result, we introduce some additional notation that captures the effect of the nonlinearity σ,

$$\eta := \mathbb{V}_{z\sim\mathcal{N}(0,s)}[\sigma(z)]\,, \quad \rho := \left(\tfrac{1}{s}\mathbb{E}_{z\sim\mathcal{N}(0,s)}[z\sigma(z)]\right)^2, \quad \zeta := s\rho\,, \quad \omega := s(\eta/\zeta - 1)\,. \tag{12}$$

where s is defined above as $= \lim_{n_0\to\infty} \bar{\mathrm{tr}}(\Sigma) = \mathbb{E}_\mu[\lambda]$. The constant $\omega \geq 0$ is a measure of the degree of nonlinearity, with $\omega = 0$ corresponding to linear activation functions (see Lemma A.1). We also introduce two functionals of μ, which capture all the relevant spectral information needed for describing the test loss,

$$\mathcal{I}_{a,b} := \phi\, \mathbb{E}_\mu\left(\lambda^a (\phi + x\lambda)^{-b}\right) \quad \text{and} \quad \mathcal{I}_{a,b}^\beta := \phi\, \mathbb{E}_\mu\left(q\lambda^a (\phi + x\lambda)^{-b}\right). \tag{13}$$

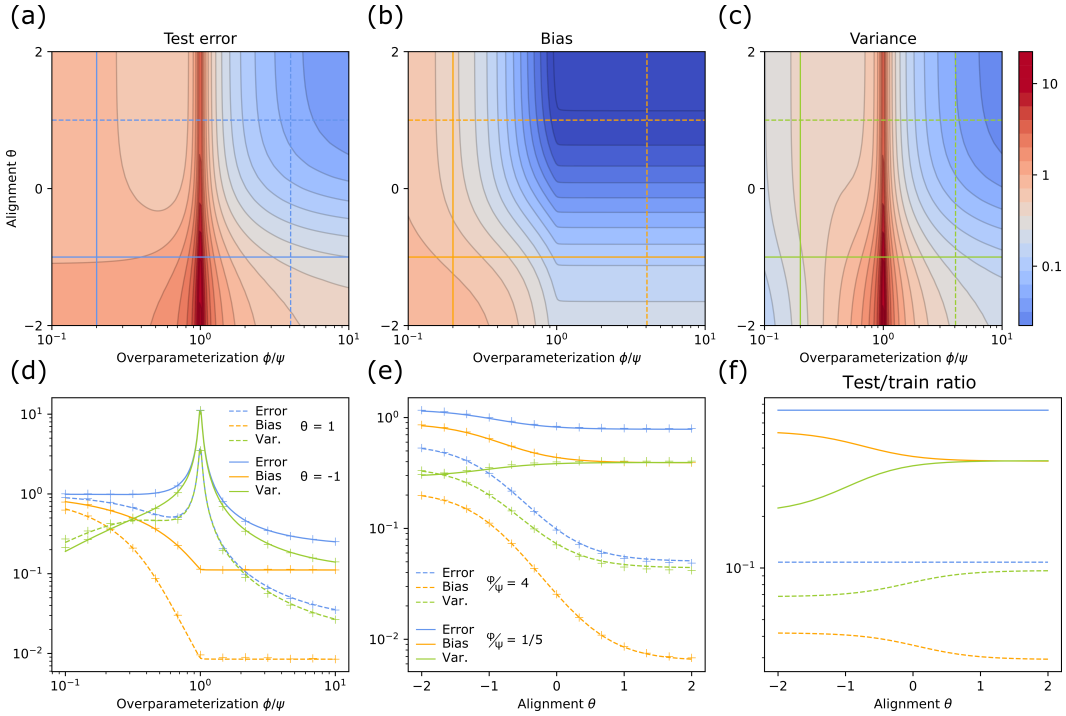

Figure 1: Test error, bias, and variance as a function of the overparameterization ratio ($\phi/\psi = n_1/m$) and the alignment θ for the 2-scale LJSD (Eq. (11)) with $\phi = n_0/m = 10/9, \sigma = \tanh, \gamma = 10^{-8}$ and $\sigma_\varepsilon^2 = 1/100$. (**a**) The total test error exhibits the characteristic double descent behavior for all shift powers. (**b**) The bias is a nonincreasing function of ϕ/ψ for all shift powers, as in Proposition 3.1. (**c**) The variance is the source of the double-descent peak, and is a nonincreasing function of ϕ/ψ for all shift powers in the overparameterized regime, as in Proposition 3.2. Blue, orange, and green lines in (a,b,c) indicate locations of vertical and horizontal slices shown in (d,e,f). (**d**) 1D horizontal slices of (a,b,c) more clearly demonstrate the monotonicity in ϕ/ψ predicted by Propositions 3.1 and 3.2. (**e**) 1D vertical slices more clearly demonstrate the monotonicity in θ of the test error and the bias predicted by Proposition 3.3, and illustrate that the variance may not decrease in response to stronger alignment (solid green trace). (**f**) Alignment only affects performance through the training error. Same legend as (e). Traces show the test error, bias, and variance normalized by the training error (ie. $\frac{E_\mu + \sigma_\varepsilon^2}{E_{\text{train}}/E_{\text{train}}(\theta=0)}$, and similarly for B_μ, V_μ). Constancy of the blue traces illustrates how the train and test error respond identically to alignment. The relative contributions of bias and variance, however, do depend on alignment. Simulations for $m = 4000$ in (d,e; crosses) agree well with asymptotic predictions.

3.1 Formulas for asymptotic bias, variance, and test error

Theorem 3.1. *Under Assumptions 1-3, as $n_0, n_1, m \to \infty$ with $\phi = n_0/m$ and $\psi = n_0/n_1$ held constant, the training error E^{Σ}_{train} tends toward the value of E^{μ}_{train} where*

$$E^{\mu}_{train} = -\gamma^2 \left(\partial_\gamma (\tau_1 \mathcal{I}^{\beta}_{1,1}) + \sigma_\varepsilon^2 \partial_\gamma \tau_1 \right) , \tag{14}$$

and the bias, variance, and test error $B_{\Sigma_\beta}, V_{\Sigma_\beta}, E_{\Sigma_\beta}$ tend toward B_μ, V_μ, E_μ with

$$E_\mu = \frac{E^{\mu}_{train}}{\gamma^2 \tau_1^2} - \sigma_\varepsilon^2 , \quad B_\mu = \phi \mathcal{I}^{\beta}_{1,2} , \quad \text{and} \quad V_\mu = E_\mu - B_\mu , \tag{15}$$

where ρ, ω are defined in (12), $\mathcal{I}_{a,b}, \mathcal{I}^{\beta}_{a,b}$ are defined in (13), and x is the unique nonnegative real root of $x = \frac{1-\gamma\tau_1}{\omega + \mathcal{I}_{1,1}}$ with $\tau_1 = \frac{\sqrt{(\psi-\phi)^2 + 4x\psi\phi\gamma/\rho} + \psi - \phi}{2\psi\gamma}$.

Remark 3.1. *As noted in (Adlam & Pennington, 2020a; Hastie et al., 2019) for isotropic covariates, the test error is simply related to the training error through the generalized cross-validation (GCV)*

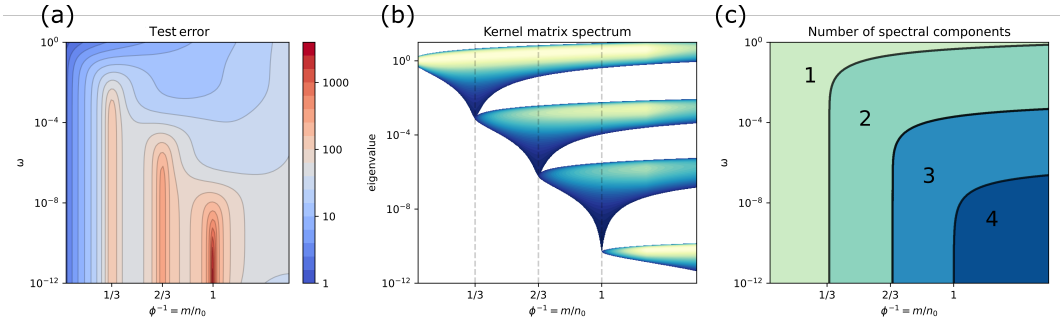

Figure 2: Sample-wise multiple descent for anisotropic data as a function of nonlinearity strength ω. All panels refer to a 3-scale covariance model with $\alpha = 10^3$, and $\psi = \frac{1}{2}, s = \rho = 1, \gamma = 10^{-13}$ and $\sigma_\varepsilon^2 = 10$. (**a**) Error exhibits multiple peaks as a function of the the sampling density ϕ^{-1}. As ω is increased, peaks are attenuated and eventually disappear in sequence starting from those due to the weakest data scales (at $m = n_0$) to those due to the strongest data scales (at $m = n_0/3$). (**b**) Limiting spectral density of the kernel matrix $K(X, X) = \frac{1}{n_1} F^T F$. Vertical slices correspond to the spectral density at the corresponding value of ϕ^{-1} for $\omega = 10^{-10}$. Peaks in the error curves in (a) occur at critical values $m = \frac{1}{3}n_0, \frac{2}{3}n_0, n_0$ where a new spectral component first appears. (**c**) Consistent with this, sample-wise learning curves exhibiting multiple descent at low ω pass through several transitions in the number of spectral components, while those at higher ω with fewer peaks pass through fewer component transitions.

formula (Golub et al., 1979). Curiously, because τ_1 is independent of β, it follows that the test error depends on β entirely through its effect on the training error; however, the bias-variance decomposition retains explicit β dependence, even conditional on the training error.

Numerical predictions from Theorem 3.1 can be obtained by first solving the self-consistent equation for x by fixed-point iteration, $x \to \frac{1-\gamma\tau_1}{\omega+\mathcal{I}_{1,1}}$, and then plugging the result into the remaining terms. For the partial derivative in the training error in (14), $\tau_1, \mathcal{I}_{1,1}^\beta$ depend implicitly on γ through the self-consistent equation for x and the definiton of $\mathcal{I}^\beta$ in (13).

As a consistency check, taking $\sigma(x) = x$ and $\psi \to 0$ in Theorem 3.1 yields an expression for the test error, bias, and variance that agrees with the results of (Wu & Xu, 2020; Mel & Ganguli, 2021) (see Section D).

At times we will find it convenient to consider the ridgeless limit of Theorem 3.1. By carefully expanding x and τ_1 for small γ, it is straightforward to obtain the following Corollary.

Corollary 3.1. *In the setting of Theorem 3.1, as the ridge regularization constant $\gamma \to 0$, $E_\mu = B_\mu + V_\mu$ with $B_\mu = \phi\mathcal{I}_{1,2}^\beta$ and V_μ given by*

$$V_\mu \overset{\gamma\to 0}{\longrightarrow} \frac{\min(\phi,\psi)}{|\phi-\psi|}(\sigma_\varepsilon^2 + \mathcal{I}_{1,1}^\beta) + \begin{cases} x\mathcal{I}_{2,2}^\beta & \text{if } \phi < \psi \\ \frac{x\mathcal{I}_{2,2}}{\omega+\phi\mathcal{I}_{1,2}}(\sigma_\varepsilon^2 + \mathcal{I}_{1,2}^\beta) & \text{otherwise} \end{cases}, \tag{16}$$

where x is the unique positive real root of $x = \frac{\min(1,\phi/\psi)}{\omega+\mathcal{I}_{1,1}}$.

3.2 The benefits of overparameterization

While there is abundant empirical evidence that overparameterization can improve the generalization performance of practical models (see e.g. (Zhang et al., 2016)), prior theoretical work has only offered explanations for this behavior in the unrealistic setting of isotropic covariates and target functions. It is therefore natural to wonder whether the benefits of overparmeterization persist in the presence of anisotropy. The following two results prove that indeed this is the case.

First, we show that the bias decreases (or stays constant) in response to an increase in the number of random features, regardless of any anisotropy in the covariates or target function weights.

Proposition 3.1. *In the setting of Theorem 3.1, the bias B_μ is a nonincreasing function of overparameterization ratio ϕ/ψ.*

The same is true for the variance in the overparameterized regime. Note that our proof requires the ridgeless setting ($\gamma = 0$), but numerical simulations suggest that the result may hold without this assumption (see Fig. 1).

Proposition 3.2. *In the setting of Corollary 3.1 and in the overparameterized regime ($\psi < \phi$), the variance V_μ is a nonincreasing function of overparameterization ratio ϕ/ψ.*

In Fig. 1, Propositions 3.1 and 3.2 are illustrated. Moving left to right in panels (a)-(d) leads to models with more parameters. The monotonicity of the bias across the whole range of parameterization is evident, as is the monotonicity of the variance in the overparmaeterized regime.

3.3 Weight-data alignment reduces the bias and test error

In the context of high-dimensional linear regression, a number of recent works formalized the intuition that generalization performance should improve if the weights of the target function align with the large eigendirections of the training covariance (Hastie et al., 2019; Mel & Ganguli, 2021). It is natural to wonder whether this behavior is restricted to linear models of the data, or whether a similar response to alignment might persists more generally. The following result proves this to be the case not only for the total test error, but for the bias and the bias-to-variance ratio as well.

Proposition 3.3. *Let μ_1, μ_2 be two LJSDs such that $\mu_1 \leq \mu_2$ (see Definition 2.1). Then $B_{\mu_1} \leq B_{\mu_2}$, $E_{\mu_1} \leq E_{\mu_2}$, and $B_{\mu_1}/V_{\mu_1} \leq B_{\mu_2}/V_{\mu_2}$.*

Proposition 3.3 shows that Definition 2.1 provides an essentially model-independent condition for comparing the relative difficulty of learning two target functions. We illustrate this monotonicity in Fig. 1 for the 2-scale LJSD of Eq. (11). Following vertical lines upward in panels (a-c) or the x-axis left-to-right in (e,f) yields stronger alignment and a corresponding decrease in the bias and test error. Panel (f) shows the alignment-dependence of the bias, variance, and test error when normalized by the training error. As suggested by Remark 3.1, the normalized test error is independent of the alignment strength, and as intuition would suggest and as predicted by Proposition 3.3, the normalized bias decreases in response to stronger alignment.

3.4 Anisotropy induces structured learning curves

Strong anisotropy has been observed to induce structured learning curves in the context of linear regression (Wu & Xu (2020); Mel & Ganguli (2021)). It is natural to wonder whether these effects generalize to the case of nonlinear random features. Figure 2 shows that this is indeed the case for sample-wise learning curves, where $\phi^{-1} = m/n_0$ is varied while $\psi = n_0/n_1$ is held fixed. Horizontal slices through Fig. 2 (a) exhibit sample-wise multiple descent. Fig. 2 (b) and (c) show that these peaks are associated with phase transitions at critical values of ϕ^{-1} where the spectrum of the kernel matrix $\frac{1}{n_1}F^T F$ acquires a new component. Increased nonlinearity ω attenuates this effect.

In contrast, the following proposition shows that parameter-wise multiple descent does not occur even in the presence of strong anisotropy, as long as μ is *aligned*.

Proposition 3.4. *If μ is* aligned *(see Definition 2.1), then, in the setting of Corollary 3.1, the test error has at most two interior critical points as a function of the overparameterization ratio ϕ/ψ.*

Remark 3.2. *Since k-fold descent requires at least k critical points, we conclude that multiple ($k > 2$) descent does not occur, even in the presence of covariate anisotropy, so long as μ is aligned.*

Fig. 3(a,b) shows parameter-wise learning curves for various values of θ in the 2- and 3-scale models. As predicted by Proposition 3.4, the curves for aligned LJSDs ($\theta \geq 0$) have at most two critical points (not shown). In fact, even for anti-aligned LJSDs ($\theta < 0$), most curves also have at most two critical points; however, for small enough θ, an additional critical point emerges, and indeed the upper-most traces in Fig. 3(b) showcase a second peak in the learning curves.

Even though parameter-wise multiple descent is not possible for aligned LJSDs, strong anisotropy can nevertheless induce other signatures, including wide plateaus and steep cliffs. To understand the

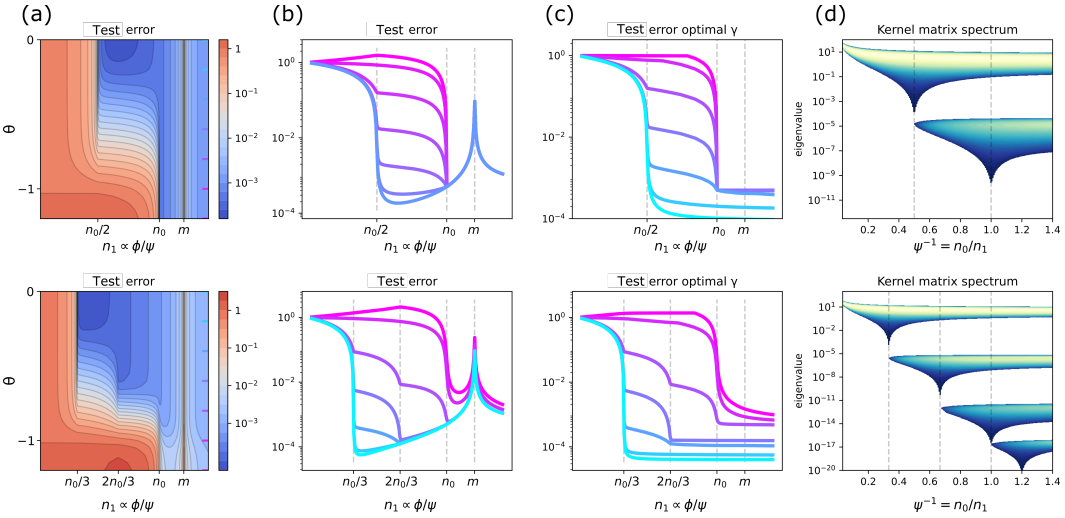

Figure 3: Strong anisotropy causes steep cliffs in the test error as a function of the overparameterization ratio $\phi/\psi = n_1/m$. $\phi = 5/6, \lambda = 10^{-22}, s = \rho = 1, \omega = 10^{-16}$ and $\sigma_\varepsilon^2 = 10^{-4}$. Top row: $d = 2$-scale model with $\alpha = 10^{-5}$; bottom row: 3-scale model with $\alpha = 10^{-6}$. (**a**,**b**) Test error exhibits the characteristic peak at the interpolation threshold, but additionally displays steep cliffs at critical values of n_1 corresponding to multiples of n_0/d (b, dashed vertical lines; m also shown). The θ value associated to each trace is indicated by the colored ticks in (a). (**c**) Unlike the peak at the interpolation threshold, cliffs in the error traces in (b) persist under optimal regularization. (**d**) The spectrum of the kernel matrix $K(X, X) = \frac{1}{n_1} F^\top F$ undergoes a phase transition at each of the critical values of n_1 (dashed vertical lines).

origin of these effects, consider the ridgeless limit, for which the self-consistent equation for x can be written as,

$$\omega \frac{x}{\phi} + \mathbb{E}_\mu \frac{\lambda}{\lambda + \frac{\phi}{x}} = \frac{1}{\phi} \min\left(1, \frac{\phi}{\psi}\right) . \tag{17}$$

For small ω, the left side becomes relatively insensitive to changes in x whenever $\lambda_- \ll \phi/x \ll \lambda_+$ for some wide spectral gap with boundary (λ_-, λ_+). It follows that, in the underparameterized regime ($\phi < \psi$), the derivative $\partial(\phi/\psi)/\partial x$ is small, or, equivalently, $\partial x/\partial(\phi/\psi)$ is large. This strong sensitivity of x to changes in the overparameterization ratio ϕ/ψ induces similarly strong changes to the error because $\partial E_\mu/\partial x$ is bounded. As a result, the parameter-wise learning curves exhibit sharp downward cliffs in the underparameterized regime for sufficiently small ω and sufficiently large spectral gaps λ_+/λ_-. We make this argument more precise in Appendix E.

Fig. 3(a) illustrates these cliffs in the context of the $d = 2$- and 3-scale LJSD. Along horizontal slices, the error turns sharply downward as a function of $n_1 \propto \phi/\psi$ at integer multiples of n_0/d. As can be seen in Fig. 3(b), when the alignment is increased and β overlaps more with large λs, the first error cliff, corresponding to the largest scales in the data, strengthens relative to others (cliff at $\phi/\psi = 1/2$ in the 2-scale model and $\phi/\psi = 1/3$ in the 3-scale model). Fig. 3(c) shows that, unlike the peak at the interpolation threshold, cliffs persist under optimal regularization. Finally, Fig. 3(d) illustrates how steep decreases in the error are associated with the appearance of a new component in the spectrum of the kernel matrix at critical values of ϕ/ψ.

4 Conclusions

We presented an asymptotically exact calculation of the test error, bias, and variance for random feature kernel regression in the presence of anisotropic covariates and target function weights. After defining a partial order over weight-data alignments, we proved that stronger alignment decreases the bias and test error. We also proved that the benefits of overparameterization persists in the anisotropic setting, and that weight-data alignment guarantees that parameter-wise multiple descent is limited to double descent. In contrast, we demonstrated that anisotropy can induce sample-wise

multiple descent, and parameter-wise cliffs, and that their structure is dictated by the eigenstructure of the data covariance. Future directions include extending our results to the non-asymptotic regime, accommodating feature learning and more general neural network models, and investigating the impact of anisotropy for other loss functions.

REFERENCES

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

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *arXiv preprint arXiv:2011.03321*, 2020b.

Ben Adlam, Jake Levinson, and Jeffrey Pennington. A random matrix perspective on mixtures of nonlinearities for deep learning. *arXiv preprint arXiv:1912.00827*, 2019.

Anonymous. Details withheld to preserve blind review. 2021.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*, 2021.

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(32):15849–15854, 2019.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Stéphane d'Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: Where & why do they appear? *arXiv preprint arXiv:2006.03509*, 2020.

Edgar Dobriban, Stefan Wager, et al. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Reza Rashidi Far, Tamer Oraby, Wlodzimierz Bryc, and Roland Speicher. Spectra of large block matrices. *arXiv preprint cs/0610045*, 2006.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*, 2021.

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

Geoffrey Grimmett. Percolation. In *Percolation*, pp. 1–31. Springer, 1999.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

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

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *J. Mach. Learn. Res.*, 21:169–1, 2020.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32:8572–8583, 2019.

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. *arXiv preprint arXiv:2006.05013*, 2020.

Licong Lin and Edgar Dobriban. What causes the test error? going beyond bias-variance via anova. *arXiv preprint arXiv:2010.05170*, 2020.

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

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Gabriel Mel and Surya Ganguli. A theory of high dimensional regression with arbitrary correlations between input features and target functions: sample complexity, multiple descent curves and a hierarchy of phase transitions. In *International Conference on Machine Learning*, pp. 7578–7587. PMLR, 2021.

James A Mingo and Roland Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pp. 29–53. Springer, 1996.

Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. *arXiv preprint arXiv:1810.05148*, 2018.

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

Jeffrey Pennington and Pratik Worah. The spectrum of the fisher information matrix of a single-hidden-layer neural network. In *NeurIPS*, pp. 5415–5424, 2018.

Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124005, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Ali Rahimi, Benjamin Recht, et al. Random features for large-scale kernel machines. In *NIPS*, volume 3, pp. 5. Citeseer, 2007.

Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge(less) regression under general source condition. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3889–3897. PMLR, 13–15 Apr 2021. URL http://proceedings.mlr.press/v130/richards21b.html.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *arXiv preprint arXiv:2006.05800*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.