Overparameterization Improves Robustness to Covariate Shift in High Dimensions

Nilesh Tripuraneni[*]
U.C. Berkeley[†]
nilesh_tripuraneni@berkeley.edu

Ben Adlam[*]
Brain Team, Google Research
adlam@google.com

Jeffrey Pennington[*]
Brain Team, Google Research
jpennin@google.com

Abstract

A significant obstacle in the development of robust machine learning models is *covariate shift*, a form of distribution shift that occurs when the input distributions of the training and test sets differ while the conditional label distributions remain the same. Despite the prevalence of covariate shift in real-world applications, a theoretical understanding in the context of modern machine learning has remained lacking. In this work, we examine the exact high-dimensional asymptotics of random feature regression under covariate shift and present a precise characterization of the limiting test error, bias, and variance in this setting. Our results motivate a natural partial order over covariate shifts that provides a sufficient condition for determining when the shift will harm (or even help) test performance. We find that overparameterized models exhibit enhanced robustness to covariate shift, providing one of the first theoretical explanations for this ubiquitous empirical phenomenon. Additionally, our analysis reveals an exact linear relationship between the in-distribution and out-of-distribution generalization performance, offering an explanation for this surprising recent observation.

1 Introduction

Theoretical justification for almost all machine learning methods relies upon the equality of the distributions from which the training and test data are drawn. Nevertheless, in many real-world applications, this equality is violated—naturally-occurring distribution shift between the training data and the data encountered during deployment is the rule, not the exception [31]. Even *non-adversarial* changes in distributions can uncover the surprising fragility of modern machine learning models [55, 56, 43, 26, 12, 50]. Such shifts are distinct from adversarial examples, which require explicit poisoning attacks [22]; rather, they can result from mild corruptions, ranging from changes of camera angle or blur [26], to subtle, unintended changes in data acquisition procedures [56]. Moreover, this fragility limits the application of deep learning in certain safety-critical areas [31].

Empirical studies of distribution shift have observed several intriguing phenomena, including linear trends between model performance on shifted and unshifted test distributions [56, 26, 31], dramatic degradation in calibration [50], and surprising spurious inductive biases [12]. Theoretical understanding of why such patterns occur across a variety of real-world domains is scant. Even basic questions such as what makes a certain distribution shift likely to hurt (or help) a model's performance,

[*] Equal contribution.

[†] Work performed while the author was at Google.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

and by how much, are not understood. One reason that these phenomena have eluded theoretical understanding is that there is often a strong coupling between model and distribution, implying that the effect of a given shift cannot usually be understood in a model-agnostic way. Another reason that satisfactory explanations have remained lacking is that the go-to formalism for studying generalization in classical models, namely uniform convergence theory (see e.g. [66]), may be insufficient to explain the behavior of modern deep learning methods (even in the absence of distribution shift) [47, 68]. Indeed, classical measures of model complexity, such as various norms of the parameters, have been found to lead to ambiguous conclusions [49].

In this paper, we follow a different approach: instead of focusing on worst-case bounds for generic distributions, we study average-case behavior for narrowly specified distributions. While this change in perspective sacrifices generality, it allows us to derive more precise predictions, which we believe are necessary to fully capture the relevant phenomenology. We study a specific type of distribution shift called covariate shift, in which the distributions of the training and test covariates differ, while the conditional distribution of the labels given the covariates remains fixed. Using random matrix theory, we perform an asymptotically exact computation of the generalization error of random feature regression under covariate shift. The random feature model provides a useful testbed to (1) investigate the interplay between various factors such as model complexity, label noise, bias, variance, and covariate shift; (2) rigorously define a *model-agnostic* notion for the strength of covariate shift; and (3) provide a theoretical explanation for the linear relationships recently observed between in-distribution and out-of-distribution generalization performance [55, 56, 27].

1.1 Contributions

Our primary contributions are to:

1. Provide a *model-agnostic* partial order over covariate shifts that is sufficient to determine when a shift will increase or decrease the test error in random feature regression (see Def. 4.1);
2. Compute the test error, bias, and variance of random feature regression for general multivariate Gaussian covariates under covariate shift in the high-dimensional limit (see Sec. 5.1);
3. Prove that overparameterization enhances robustness to covariate shift, and that the error, bias, and variance are nonincreasing functions of the number of excess parameters (see Sec. 5.3);
4. Deduce an exact linear relationship between in-distribution and out-of-distribution generalization performance, offering an explanation for this surprising recent empirical observation (see Sec. 5.4).

1.2 Related work

There is extensive literature on the empirical analysis of distribution shift in all of its myriad forms, ranging from domain adaptation [62, 20, 8, 70, 69, 37, 36] to defenses against adversarial attacks [39, 60] to distributionally robust optimization [59, 15, 16], among many others. Interestingly, for naturally occurring distribution shifts [31, 26], standard robustness interventions provide little protection [56, 63]. Indeed, empirical risk minimization on clean, unshifted training data often performs better on out-of-distribution benchmarks than more sophisticated methods [31]. One of the most striking observations in the context of natural distribution shifts is that model robustness improves with the classifier's accuracy [56, 63, 26, 43]. For example, if a classifier's accuracy increases by 1.0% on the unshifted CIFAR-10 test set, this tends to increase its accuracy by 1.7% on the CIFAR-10.1 dataset (a dataset with natural distribution shift) [56]. Moreover, such linear trends between the unshifted and shifted measures of error have now been observed in several contexts [56, 63, 43, 40, 44].

The number of theoretical works studying the impact of distribution shift on generalization is far smaller. One pioneering work provides VC-dimension-based error bounds for classification that are augmented by a discrepancy measure between source and target domains [10], while another demonstrates a similar class of uniform convergence-based results in the setting of kernel regression [11]. Recent work shows that learning domain-invariant features is insufficient to guarantee generalization when the class-conditional distributions of features may shift [69]. When the source domain gradually shifts toward the target domain, non-vacuous margin-based bounds for self-training can be established [32]. In [40], assumptions based on model similarity are used to help explain why classifiers exhibit linear trends between their accuracies on shifted and unshifted test sets [56, 43].

Our technical tools build on a series of works that have studied the exact high-dimensional limit of the test error for a growing class of model families and data distributions. In the context of linear models, recent work analyzes ridge regression for general covariances and a general non-isotropic source condition on the parameters which generate the targets [57], extending earlier work studying minimum-norm interpolated least squares and ridge regression in the random design setting [9, 14, 24]. The non-isotropy of source parameter effectively induces a shift on the bias term of this model, but the phenomenon is distinct from the covariate shifts we study here. Beyond linear regression, random feature models provide a rich but tractable class of models to gain further insight into generalization phenomena [2, 3, 41, 35]. These methods are of particular interest because of their connection to neural networks, with the number of random features corresponding to the network width (or model complexity) [48, 33, 29], and because they serve as a practical method for data analysis in their own right [54, 61]. In this context, a precise characterization of the gaps between uniform convergence and the (asymptotic) exact test error as a function of the sample size and number of random features can be derived [68]. Since this paper's publication, we released follow-up work considering unequal scales in the training and test distributions and optimal regularization [64]. 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 free probability [53, 52, 1, 2, 38, 51, 19, 45].

2 Preliminaries

2.1 Problem setup and notation

As in prior work studying random feature regression [24, 41, 2, 1], we compute the test error in the high-dimensional, proportional asymptotics where the dataset size 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 fixed. We refer to ϕ/ψ as the *overparameterization ratio*, which is the limit of n_1/m and characterizes the normalized complexity of the (random) feature model.

Interestingly, in this high-dimensional limit, the conditional distribution of a linear labeling function is asymptotically equivalent to a wide class of nonlinear teacher functions (see [41, 2] for more details). With this in mind, we consider the task of learning an unknown function from m i.i.d. samples $(\mathbf{x}_i, y_i) \in \mathbb{R}^{n_0} \times \mathbb{R}$ for $i \in \{1, \ldots, m\}$, where the covariates are Gaussian, $\mathbf{x}_i \sim \mathcal{N}(0, \Sigma)$ with positive definite covariance matrix Σ, and the labels are generated by a linear function parameterized by $\beta \in \mathbb{R}^{n_0}$, drawn from $\mathcal{N}(0, I_{n_0})$. In particular

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

where $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is additive label noise on the training points.

We study the class of prediction models defined by kernel ridge regression using unstructured random feature maps [54]. The random features are given by a single-layer, fully-connected neural network with random weights. Given a set of training data $X = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$ and a prospective test point $\mathbf{x}$, the random features embeddings of the training and test data are given by

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

for a random weight matrix $W \in \mathbb{R}^{n_1 \times n_0}$ with i.i.d. standard Gaussian entries and an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ applied elementwise. The induced kernel is

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

and the model's predictions are given by $\hat{y}(\mathbf{x}) = YK^{-1}K_\mathbf{x}$, 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 $\gamma \geq 0$ is a ridge regularization constant[1]. Owing to the implicit regularization effect of the nonlinear feature maps [7], in low noise settings the optimal value of γ can sometimes be negative [30]. For simplicity, we nevertheless make the standard assumption that $\gamma \geq 0$, though we emphasize our techniques readily accommodate negative values.

[1] We overload the definition of K to include the additive regularization whenever no arguments are present. Also, if $\gamma = 0$ and K is not full-rank, K^{-1} should be understood as the Moore–Penrose pseudoinverse.

Our central object of study is the expected test loss for a datapoint $\mathbf{x} \sim \mathcal{N}(0, \Sigma^*)$ where Σ^* may be different from the training covariance Σ. The test error (without label noise on the test point) is

$$
\begin{aligned}
E_{\Sigma^*} &= \mathbb{E}[(\beta^\top \mathbf{x}/\sqrt{n_0} - YK^{-1}K_{\mathbf{x}})^2] && (4)\\
&= \underbrace{\mathbb{E}_{\mathbf{x},\beta}[(\mathbb{E}[\hat{y}(\mathbf{x})] - y(\mathbf{x}))^2]}_{B_{\Sigma^*}} + \underbrace{\mathbb{E}_{\mathbf{x},\beta}[\mathbb{V}[\hat{y}(\mathbf{x})]]}_{V_{\Sigma^*}}, && (5)
\end{aligned}
$$

where the inner expectations defining the bias and variance are computed over W, X, and Y. We decompose the training and test covariance matrices into eigenbases as $\Sigma = \sum_{i=1}^{n_0} \lambda_i \mathbf{v}_i \mathbf{v}_i{}^\top$ and $\Sigma^* = \sum_{i=1}^{n_0} \lambda_i^* \mathbf{v}_i^* \mathbf{v}_i^*{}^\top$, where the eigenvalues are in nondecreasing magnitude, i.e. $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{n_0}$ and $\lambda_1^* \leq \lambda_2^* \leq \ldots \leq \lambda_{n_0}^*$. We define the *overlap coefficients*

$$
r_i := \mathbf{v}_i^\top \Sigma^* \mathbf{v}_i = \sum_{j=1}^{n_0} (\mathbf{v}_j^* \cdot \mathbf{v}_i)^2 \lambda_j^* \tag{6}
$$

to measure the alignment of Σ^* with the ith eigendirection of Σ. In particular, r_i is the induced norm of $\mathbf{v}_i$ with respect to Σ^*. We use $\bar{\mathrm{tr}}$ to denote the dimension-normalized trace: for a matrix $A \in \mathbb{R}^{n \times n}$, $\bar{\mathrm{tr}}(A) = \frac{1}{n}\mathrm{tr}(A)$. We use $\|A\|_\infty$ and $\|A\|_F$ to denote the operator norm and Frobenius norm of matrix A respectively. Finally, we use $\delta_{\mathbf{x}}$ to denote the Dirac delta function centered at $\mathbf{x}$.

2.2 Assumptions

Regularity assumptions on the spectra of Σ and Σ^* are necessary to state the limiting behavior of this system. As in [67], it is not sufficient to consider the spectra of these matrices individually; they must be considered jointly. We do this in an eigenbasis of Σ.

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

$$
\mu_{n_0} := \frac{1}{n_0} \sum_{i=1}^{n_0} \delta_{(\lambda_i, r_i)} \tag{7}
$$

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 train and test distributions[2]*. Additionally, we require that $\limsup_{n_0} \max(\|\Sigma\|_\infty, \|\Sigma^*\|_\infty) \leq C$ for a constant C.*

Often we use (λ, r) for random variables sampled jointly from μ and denote the marginal of λ under μ with μ_{train}. The conditional expectation $\mathbb{E}[r|\lambda]$ is an important object in our study. We frequently overload the notation $\mathbb{E}[r|\lambda]$ to view it as a function of λ, and we assume the following for simplicity.

Assumption 2. *μ is either absolutely continuous or a finite sum of delta masses. Moreover, the expectations of λ and r are finite.*

When the eigenspaces of Σ and Σ^* are aligned and $r_i = \lambda_i^* = \Phi(\lambda_i)$ for some smooth function Φ, the support of the LJSD degenerates. Here, Assump. 1 is essentially equivalent to assuming the empirical spectral distribution of Σ converges in distribution to some μ_{train}, which is a standard assumption in the regression literature [14, 41]. One special case of note is when there is no shift, i.e. Φ is the identity, in which case the LJSD degenerates to $\mu_\emptyset$ defined by

$$
\mu_\emptyset(\lambda, r) := \mu_{\text{train}}(\lambda)\delta_\lambda(r), \quad \textit{i.e.} \quad (\lambda, \lambda) \sim \mu_\emptyset \ \text{ for } \ \lambda \sim \mu_{\text{train}}. \tag{8}
$$

As our analysis will eventually take place in the high-dimensional limit, we further define the asymptotic scales of the training and test covariances as $s := \lim_{n_0 \to \infty} \bar{\mathrm{tr}}(\Sigma) = \mathbb{E}_\mu[\lambda]$ and $s_* := \lim_{n_0 \to \infty} \bar{\mathrm{tr}}(\Sigma^*) = \mathbb{E}_\mu[r]$ under the limiting behavior specified in Assump. 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 that, $|\sigma(x)|, |\sigma'(x)| \leq c_0 \exp(c_1 x)$ for constants c_0, c_1.*

[2] Note that the EJSD depends not only on Σ and Σ^* but also on a choice of eigendecomposition for Σ when it has repeated eigenvalues. However, all possible choices for the EJSD form an equivalence class, and the ambiguity does not affect later definitions and conclusions. See Sec. A3.

2.3 A simple family of diatomic distributions

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

Definition 2.1. *For $\alpha \geq 1$ and $\theta \in \mathbb{R}$, we define the family of (α, θ)-diatomic LJSDs with θ-power-law shifts as*

$$\mu^{\textit{diatomic}}_{\alpha,\theta} := \frac{1}{\alpha+1}\delta_{(\alpha, C\alpha^{\theta})} + \frac{\alpha}{\alpha+1}\delta_{(\alpha^{-1}, C\alpha^{-\theta})}, \tag{9}$$

where C is a normalization constant chosen so that $\mathbb{E}_{\mu^{\textit{diatomic}}_{\alpha,\theta}}[r] = 1$. Note that $\mu^{\textit{diatomic}}_{\alpha,\theta}$ is the limit of

$$\Sigma_{ij} := \begin{cases} \alpha & \textit{if } i = j \textit{ and } i \leq \lfloor \frac{n_0}{1+\alpha} \rfloor \\ \alpha^{-1} & \textit{if } i = j \textit{ and } i > \lfloor \frac{n_0}{1+\alpha} \rfloor \\ 0 & \textit{if } i \neq j \end{cases} \qquad \textit{and} \quad \Sigma^* := \frac{1}{\bar{\operatorname{tr}}(\Sigma^{\theta})}\Sigma^{\theta} \,. \tag{10}$$

This simple two-parameter family of distributions captures the fast eigenvalue decay observed in many datasets in machine learning, for which the covariance spectra are often dominated by several large eigenvalues and exhibit a long tail of many small eigenvalues [34]. Note that the trivial case of $\alpha = 1$ yields an identity covariance with no shift. For the nontrivial setting $\alpha > 1$, the exponent θ parameterizes the strength of the shift in an intuitive way: when $\theta = 1$, there is no shift; when $\theta < 1$, $\alpha^{\theta} < \alpha$, so the large eigendirections of the training distribution are suppressed in the test distribution, suggesting that the shift makes learning harder; when $\theta > 1$, $\alpha^{\theta} > \alpha$, so the large eigendirections of the training distribution are further emphasized in the test distribution, suggesting that the shift makes learning easier. We will return to the notion of shift strength in Secs. 3 and 4.

3 Motivating example: linear regression

We first consider the relatively simple case of ridgeless linear regression (LR), which will help build some intuition for the more general analysis of random feature regression in Sec. 5.1. Assuming the labels are generated by the linear model defined above, i.e. $y_i = \beta^{\top}\mathbf{x}_i/\sqrt{n_0} + \varepsilon_i$, the estimator is given by $\hat{\beta} = (XX^{\top})^{-1}XY$, and the test risk (see Eq. (4)) has the following simple form.

Proposition 3.1. *For fixed dimension n_0 and sample size $m > n_0 + 1$, the test error of LR is given by*

$$E^{\textit{LR}}_{\Sigma^*} = \sigma^2_{\epsilon}\frac{n_0}{m - n_0 - 1}\bar{\operatorname{tr}}(\Sigma^*\Sigma^{-1}) = \sigma^2_{\epsilon}\frac{n_0}{m - n_0 - 1}\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{r_i}{\lambda_i}. \tag{11}$$

Under Assump. 1, as $n_0, m \to \infty$ with $\phi = n_0/m$ fixed, $E^{\textit{LR}}_{\Sigma^*} \to E^{\textit{LR}}_{\mu} = \sigma^2_{\epsilon}\phi/(1-\phi)\mathbb{E}_{\mu}[r/\lambda]$.

One immediate question is whether a given shift will increase or decrease the test error relative to E^{LR}_{Σ}. While the precise answer is of course determined by the value of $\bar{\operatorname{tr}}(\Sigma^*\Sigma^{-1})$, it useful for the subsequent analysis to develop an understanding of the individual contributions to this term. Similar decompositions of the test error into eigenspaces have proved useful in a variety of other contexts, e.g. [42, 4]. We begin with a specific example in the setting of the finite-dimensional analog of Def. 2.1.

Example 3.1. *For the finite form of the (α, θ)-diatomic density defined in Eq. (10), the test error of LR is given by*

$$E^{\textit{LR}}_{\Sigma^*} = \sigma^2_{\epsilon}\frac{n_0}{m - n_0 - 1}\frac{\alpha + w(\alpha^{2\theta-1} - \alpha)}{1 + w(\alpha^{2\theta} - 1)} \quad \textit{for} \quad w = \frac{1}{n_0}\left\lfloor \frac{n_0}{1+\alpha} \right\rfloor, \tag{12}$$

and so $\frac{\partial}{\partial\theta}E^{\textit{LR}}_{\Sigma^*} = \sigma^2_{\epsilon}\frac{n_0}{m-n_0-1}\frac{2(1-w)w\alpha^{2\theta-1}(1-\alpha^2)\log(\alpha)}{(1+w(\alpha^{2\theta}-1))^2} \leq 0$, which implies $E^{\textit{LR}}_{\Sigma^*_1} \leq E^{\textit{LR}}_{\Sigma^*_2}$ whenever $\theta_1 \geq \theta_2$, in accordance with the discussion in Sec. 2.3. It follows from Eq. (10) that the condition $\theta_1 \geq \theta_2$ not only implies $\bar{\operatorname{tr}}(\Sigma^*_1\Sigma^{-1}) \leq \bar{\operatorname{tr}}(\Sigma^*_2\Sigma^{-1})$, but also that the ratios of overlap coefficients $r_{i,1}/r_{i,2}$ form a nondecreasing sequence. It is this condition involving all the eigendirections that will generalize to the nonlinear random feature setting in Sec. 4.

The following proposition captures the essence of these considerations in the context of linear regression. See Sec. A4 for the proof.

Proposition 3.2. *Let $r_{i,1}$ and $r_{i,2}$ denote the overlap coefficients*[3] *of Σ_1^* and Σ_2^* relative to Σ. If $\mathrm{tr}(\Sigma_2^*) \geq \mathrm{tr}(\Sigma_1^*)$ and the ratios $r_{i,1}/r_{i,2}$ form a nondecreasing sequence, then in the setting of Prop. 3.1, $E^{LR}_{\Sigma_2^*} \geq E^{LR}_{\Sigma_1^*}$.*

Whereas the (α, θ)-diatomic LJSDs explicitly enforce the trace normalizations $\mathrm{tr}(\Sigma) = \mathrm{tr}(\Sigma_1^*) = \mathrm{tr}(\Sigma_2^*) = 1$, Prop. 3.2 provides sufficient conditions for the ordering of test errors for non-unit traces. The fact that $E^{\mathrm{LR}}_{\Sigma^*}$ scales linearly with the overall scale of Σ^* is a unique feature of linear regression and does not generalize to the nonlinear random feature setting. We return to this issue in Sec. 4.

4 Definition of shift strength

Deriving conditions on whether a shift will hurt or help a model's performance is crucial to building an understanding of covariate shift. Motivated in part by the above results for linear regression, and in part by the results for random feature regression that we present in Sec. 5.3, we introduce the following definition of shift strength, which is a direct generalization of the conditions of Prop. 3.2:

Definition 4.1. *Let μ_1 and μ_2 be LJSDs with the same marginal distribution of λ, denoted μ_{train}. If the asymptotic overlap coefficients are such that $\mathbb{E}_{\mu_1}[r|\lambda]/\mathbb{E}_{\mu_2}[r|\lambda]$ is nondecreasing as a function of λ on the support of μ_{train} and $\mathbb{E}_{\mu_1}[r] \leq \mathbb{E}_{\mu_2}[r]$, we say μ_1 is* easier *than μ_2 (or μ_2 is* harder *than μ_1), and write $\mu_1 \leq \mu_2$. Comparing against the case of no shift $\mu_\emptyset$, we say μ_1 is* easy *when $\mu_1 \leq \mu_\emptyset$ and* hard *when $\mu_1 \geq \mu_\emptyset$.*

A priori, there is little reason to hope that such a *model-independent* definition of shift strength would adequately characterize a shift's impact on the total error, bias, or variance of a given model. Even for the relatively simple case of random feature kernel regression, the nonlinear feature maps of Eq. (2) would seem to inextricably couple the covariance distribution to the model.

Nevertheless, as we show in Sec. 5.1, the coupling between model and shift simplifies considerably in the high-dimensional proportional asymptotics. It is characterized by a handful of constants that depend solely on the overall covariance scale $\mathbb{E}_\mu[r]$ and a collection of functionals of μ, whose magnitudes can be bounded in terms of the ratio $\mathbb{E}_\mu[r|\lambda]/\mathbb{E}_{\mu_\emptyset}[r|\lambda]$. The conditions that Def. 4.1 places on $\mathbb{E}_\mu[r]$ and $\mathbb{E}_\mu[r|\lambda]/\mathbb{E}_{\mu_\emptyset}[r|\lambda]$ can be augmented by various constraints on the model to derive bounds on how the total error and bias will respond to a shift of a given strength. This perspective introduces considerable complexity and we present the details of this analysis elsewhere.

In this work, we focus on a simpler, surprising result: by merely normalizing the scales of the covariate distributions (i.e. enforcing $s = s_*$), Def. 4.1 provides a *model-independent* definition of shift strength that determines how random feature models respond to shifts of different strength. This observation motivates the following assumption.

Assumption 4. *The training and test covariance scales are equal, $\mathbb{E}_\mu[\lambda] = \mathbb{E}_\mu[r]$, i.e. $s = s_*$.*

We emphasize that Assump. 4 reflects common practice for many models and data modalities, as preprocessing techniques such as standardization are ubiquitous and many architectural components such as layer- or batch-normalization achieve a similar effect [21, 28, 46].

5 Covariate shift in random feature kernel regression

5.1 Main results

Our main results characterize the high-dimensional limits of the test error, bias, and variance of the nonlinear random feature model of Sec. 2. Before stating them, we first introduce some additional constants that capture the effect of the nonlinearity σ. For $z \sim \mathcal{N}(0, s)$, define

$$\eta := \mathbb{V}[\sigma(z)]\,, \quad \rho := \left(\tfrac{1}{s}\mathbb{E}[z\sigma(z)]\right)^2\,, \quad \zeta := s\rho\,, \quad \text{and} \quad \omega := s(\eta/\zeta - 1)\,. \tag{13}$$

Our results also depend on the covariance spectra through two sets of functionals of μ,

$$\mathcal{I}_{a,b}(x) := \phi\, \mathbb{E}_\mu\left(\lambda^a\,(\phi + x\lambda)^{-b}\right) \quad \text{and} \quad \mathcal{I}^*_{a,b}(x) := \phi\, \mathbb{E}_\mu\left(r\lambda^{a-1}\,(\phi + x\lambda)^{-b}\right). \tag{14}$$

[3] Recall the definition of the overlap coefficients in Eq. (6).

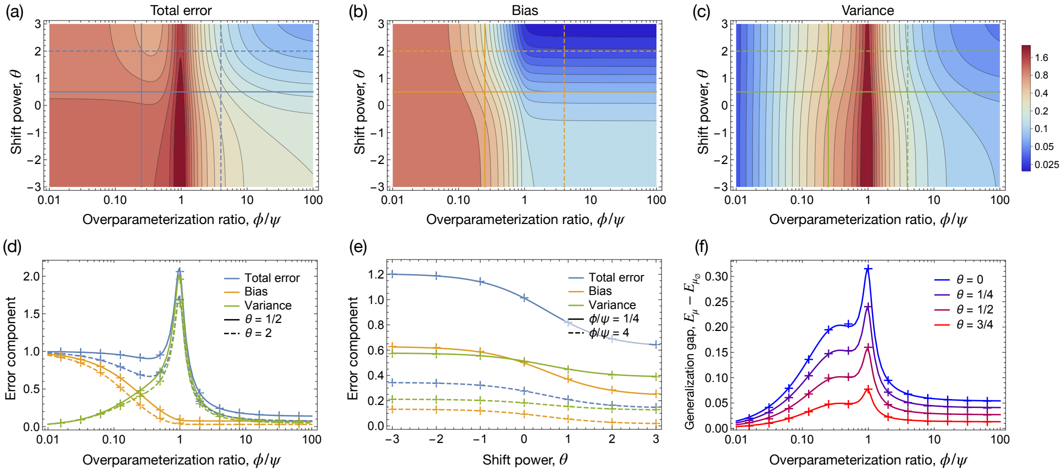

Figure 1: The asymptotic predictions of Thm. 5.1 as a function of the overparameterization ratio ($\phi/\psi = n_1/m$) and the shift power (θ) for the $(2,\theta)$-diatomic LJSD (Eq. (9)) with $\phi = n_0/m = 0.5$, $\sigma =$ ReLU, $\gamma = 0.001$, and $\sigma_\varepsilon^2 = 0.1$. (**a**) The test 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 Prop. 5.2. (**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 Prop. 5.3. In (**a**,**b**), the total error and bias are nonincreasing functions of θ, as in Prop. 5.1. (**d**) 1D horizontal slices of (**a**,**b**,**c**) demonstrate the monotonicity in ϕ/ψ predicted by Props. 5.2 and 5.3. (**e**) 1D vertical slices of (**a**,**b**,**c**) demonstrate the monotonicity in θ predicted by Prop. 5.1 (the variance also appears monotonic, but it need not be in general). (**f**) The generalization gap between the error on shifted and unshifted distributions is a nonincreasing function of ϕ/ψ in the overparameterized regime, as in Prop. 5.4. Markers in (**d**,**e**,**f**) show simulations for $n_0 = 512$ and agree well with the asymptotic predictions.

Theorem 5.1. *Under Assumps. 1, 2, 3 and 4, as $n_0, n_1, m \to \infty$ the test error E_{Σ^*} converges to $E_\mu = B_\mu + V_\mu$, with the bias B_μ and variance V_μ given by*

$$B_\mu = \phi \mathcal{I}_{1,2}^* \tag{15}$$

$$V_\mu = -\rho \frac{\psi}{\phi}\frac{\partial x}{\partial \gamma}\Big(\mathcal{I}_{1,1}(\omega + \phi\mathcal{I}_{1,2})(\omega + \mathcal{I}_{1,1}^*) + \frac{\phi^2}{\psi}\gamma\bar{\tau}\mathcal{I}_{1,2}\mathcal{I}_{2,2}^* + \gamma\tau\mathcal{I}_{2,2}(\omega + \phi\mathcal{I}_{1,2}^*) + \sigma_\varepsilon^2\Big((\omega + \phi\mathcal{I}_{1,2})(\omega + \mathcal{I}_{1,1}^*) + \frac{\phi}{\psi}\gamma\bar{\tau}\mathcal{I}_{2,2}^*\Big)\Big), \tag{16}$$

where x is the unique nonnegative real root of $x = \frac{1-\gamma\tau}{\omega+\mathcal{I}_{1,1}}$, $\frac{\partial x}{\partial\gamma} = -\frac{x}{\gamma + \rho\gamma(\tau\psi/\phi+\bar{\tau})(\omega+\phi\mathcal{I}_{1,2})}$, and

$$\tau = \frac{\sqrt{(\psi-\phi)^2 + 4x\psi\phi\gamma/\rho} + \psi - \phi}{2\psi\gamma} \quad \text{and} \quad \bar{\tau} = \frac{1}{\gamma} + \frac{\psi}{\phi}\big(\tau - \frac{1}{\gamma}\big)\,. \tag{17}$$

Numerical predictions from Thm. 5.1 can be obtained by first solving the self-consistent equation for x by fixed-point iteration, $x \mapsto \frac{1-\gamma\tau}{\omega+\mathcal{I}_{1,1}}$, and then plugging the result into the remaining terms. Fig. 1 shows excellent agreement between these asymptotic predictions and finite-size simulations.

At times we will find it convenient to consider the ridgeless limit of Thm. 5.1. By carefully expanding x and τ for small γ, it is straightforward to obtain the following corollary.

Corollary 5.1. *In the setting of Thm. 5.1, as the ridge regularization constant $\gamma \to 0$, $E_\mu = B_\mu + V_\mu$ with B_μ given in Eq. (15) and V_μ given by*

$$V_\mu = \frac{\psi}{|\phi-\psi|}x(\sigma_\varepsilon^2 + \mathcal{I}_{1,1})(\omega + \mathcal{I}_{1,1}^*) + \begin{cases} x\Big(1 - \frac{x(\omega-\sigma_\varepsilon^2)}{1-x^2\mathcal{I}_{2,2}}\Big)\mathcal{I}_{2,2}^* & \phi \geq \psi \\ \frac{x^2\psi\mathcal{I}_{2,2}}{\phi - x^2\psi\mathcal{I}_{2,2}}\big(\omega + \phi\mathcal{I}_{1,2}^*\big) & \phi < \psi \end{cases}, \tag{18}$$

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

Taking $\sigma(x) = x$ and $\psi \to 0$ in Cor. 5.1 yields an expression for the test error of ridgeless linear regression that agrees with [24] and with the asymptotic form of Prop. 3.1 (see Sec. A4.2).

5.2 Harder shifts increase the bias and test error

The bias and variance in Thm. 5.1 depend on the covariate shift exclusively through $\mathcal{I}^*_{a,b}$ — all other terms such as x, τ, $\bar{\tau}$, and $\mathcal{I}_{a,b}$ only depend on the marginal of λ under μ. The functionals $\mathcal{I}^*_{a,b}$ generalize the simple ratio of overlap coefficients to eigenvalues, $\mathbb{E}_\mu[r/\lambda]$, that characterizes the error for linear regression (indeed, $\mathcal{I}^*_{0,0} = \phi\mathbb{E}_\mu[r/\lambda]$). In contrast, the error in the random feature setting is a combination of multiple such terms. Nevertheless, Def. 4.1 enables comparisons of the individual $\mathcal{I}^*_{a,b}$ functionals, which provide sufficient conditions to order the error and bias.

Proposition 5.1. *Consider two LJSDs such that $\mu_1 \leq \mu_2$ (see Def. 4.1). Then, in the setting of Thm. 5.1, $B_{\mu_1} \leq B_{\mu_2}$ and, if $\sigma_\varepsilon^2 \leq \omega$, $E_{\mu_1} \leq E_{\mu_2}$.*

Prop. 5.1 shows that Def. 4.1 provides an essentially model-independent condition to determine the impact of covariate shift on the test error[4]. Interestingly, both the bias (which arises from both regularization and model misspecification) and the total error (which has additional variance contributions from the randomness induced by W, X, and ϵ) respond to shifts in tandem in the regime of small label noise. (For large label noise, the variance can dominate the error and cause violations of monotonicity; see Sec. 5.5.) Prop. 5.1 is illustrated for the (α, θ)-diatomic LJSD in Fig. 1: following the vertical lines upward in (a), (b), or the x-axis rightward in (e) yields easier shifts and a corresponding decrease in the bias and total error.

5.3 The benefit of overparameterization

While Prop. 5.1 shows that harder shifts increase the error, it is natural to wonder whether this increase can be mitigated by judicious model selection. In practice, empirical investigations have shown that the performance of large, overparameterized models tends to deteriorate less under distribution shift than their smaller counterparts [26]. We obtain a number of theoretical results that formally prove the benefit of overparameterization in our random feature setting.

First, we show that the bias decreases (or stays constant) when additional random features are added, which increases the model capacity and accords with the intuition of the bias as a measure of the model's ability to fit the data.

Proposition 5.2. *In the setting of Thm. 5.1, the bias B_μ is a nonincreasing function of the overparameterization ratio ϕ/ψ.*

In contrast to the bias, which is monotonic for all overparameterization ratios, the variance can exhibit nonmonotonic behavior in the underparameterized regime. On the other hand, the following proposition shows that in the overparameterized regime, the variance is also nonincreasing. Note that our proof requires the setting of ridgeless regression ($\gamma = 0$), but numerical investigation suggests this condition may not be necessary (see Fig. 1).

Proposition 5.3. *In the setting of Cor. 5.1 and in the overparameterized regime (i.e. $\psi < \phi$), the variance V_μ is a nonincreasing function of the overparameterization ratio ϕ/ψ.*

The explosion of variance at the interpolation threshold and then its subsequent decay have been demonstrated in previous exact asymptotic studies of random feature regression in the absence of covariate shift, in stark contrast to what classical theory would suggest [2, 41]. Prop. 5.3 confirms the existence of analogous behavior under covariate shift.

Taken together, Props. 5.2 and 5.3 imply that some of the benefits of overparameterization extend to models evaluated out-of-distribution. An additional benefit is that overparameterized models are more robust: the difference in error between unshifted and shifted test distributions is smaller for larger models. A formal statement of this enhanced robustness is given in the following result.

Proposition 5.4. *Consider two LJSDs such that $\mu_1 \leq \mu_2$ (see Def. 4.1). Then, in the setting of Cor. 5.1 and in the overparameterized regime (i.e. $\psi < \phi$), the generalization gap $E_{\mu_2} - E_{\mu_1}$ is a nonincreasing function of the overparameterization ratio ϕ/ψ.*

[4] The weak model-dependence arising from the condition $\sigma_\varepsilon^2 < \omega$ can instead be regarded as a small label-noise condition that can always be satisfied for $\omega > 0$, which is true for any nonlinear σ (see Sec. A1).

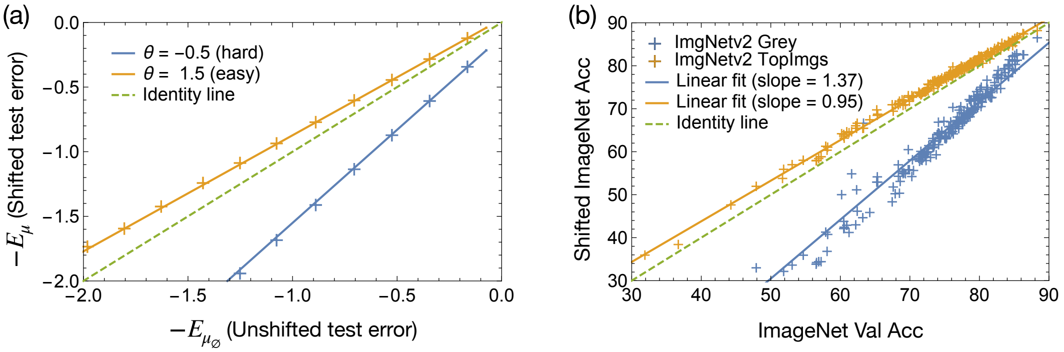

Figure 2: Linear relationship between in-distribution and out-of-distribution generalization error. **(a)** Asymptotic predictions for shifted versus unshifted error for models with varying degrees of overparameterization $\phi/\psi > 1$, obtained via Cor. 5.1 for the $(3, \theta)$-diatomic LJSD (Eq. (9)) with $\phi = n_0/m = 0.5$, $\sigma = \text{ReLU}$, $\sigma_\varepsilon^2 = 0.01$ and two different values of the shift-power θ. Markers represent simulations for $n_0 = 512$. The negated errors are plotted so that performance improves left to right and bottom to top, in order to match the behavior of the accuracy metric. **(b)** Reproduction of the empirical results of [56, 63], showing the relationship between the classification accuracy of various models on the original ImageNet test set and two shifted ImageNet datasets: a "hard" dataset with greyscale corruptions, Grey, and an "easy" dataset with high inter-annotator agreement, TopImgs. In both (**a**) and (**b**), the slope is greater than one for the hard shift and less than one for the easy shift, in accordance with Prop. 5.5.

In Fig. 1, Props. 5.2, 5.3 and 5.4 are illustrated. Following the horizontal lines rightward in (a), (b), and (c) or the x-axis rightward in (d) and (f) leads to models with more parameters. The monotonicity of the bias across the whole range of parameterization is evident, as is the necessity of considering the monotonicity of the variance and generalization gap only when $\phi > \psi$.

5.4 Linear trends between in-distribution and out-of-distribution generalization

We have discussed how overparameterization yields improvements on both unshifted and shifted test distributions, which hints that these two quantities are positively correlated. Indeed, recent work has suggested increasing model size as a path to increased robustness [26]. Additional empirical studies have further refined this observation by discovering a linear relationship between the performance of models of varying complexity on unshifted and shifted data [56, 63]. In the context of ridgeless random feature regression, we provide a formal proof of this linear relationship.

Proposition 5.5. *In the setting of Cor. 5.1 and in the overparameterized regime (i.e. $\psi < \phi$),*

$$E_\mu = E_0 + \underbrace{\left(\frac{\omega + \mathcal{I}_{1,1}^*}{\omega + \mathcal{I}_{1,1}}\right)}_{\text{SLOPE}} E_{\mu_\emptyset} \,, \tag{19}$$

parametrically in the overparameterization ratio ϕ/ψ, where E_0 and SLOPE *are constants independent of ϕ/ψ, and $E_{\mu_\emptyset}$ is the error on the unshifted distribution. Moreover,* SLOPE ≥ 1 *when μ is hard and* SLOPE ≤ 1 *when μ is easy.*

Eq. (19) implies a parametrically linear relationship between E_μ and $E_{\mu_\emptyset}$ by varying ϕ/ψ. Prop. 5.5 also makes the nontrivial prediction that an improvement on the unshifted distribution leads to a relatively greater improvement on the shifted distribution when the shift is hard, and to a relatively smaller improvement when the shift is easy. This prediction is corroborated qualitatively in the data from [55, 56, 43]. We plot this linear behavior in Fig. 2, where (a) shows the random feature model and (b) shows an example of data from [56, 63]. The striking similarity in these plots is evident.

5.5 Importance of assumptions

Props. 5.1, 5.2, 5.3, 5.4 and 5.5 rely on a number of assumptions and conditions; here we show the necessity of some of these prerequisites for our results.

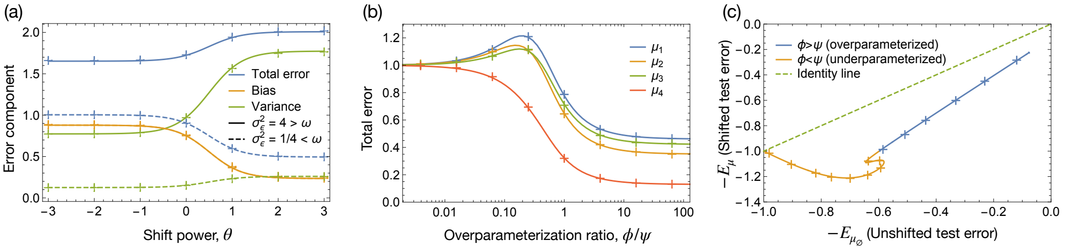

Figure 3: Relaxing the assumptions and conditions can lead to counterexamples to the propositions. (**a**) Asymptotic predictions (solid lines) and simulations with $n_0 = 4096$ (markers) for the total error, bias, and variance for the $(4, \theta)$-diatomic LJSD with $\phi = 4$, $\psi = 0.25$, $\gamma = 10^{-4}$, and $\sigma = \text{ReLU}$ (implying $\omega = 1 - \frac{2}{\pi} \approx 0.36$) as a function of increasing shift power θ. When $\sigma_\varepsilon^2 < \omega$ (dashed curves), the bias and total error are nonincreasing, as predicted by Prop. 5.1, though the variance is not. When $\sigma_\varepsilon^2 > \omega$ (solid curves), the total error is no longer nonincreasing. (**b**) Asymptotic predictions (solid lines) and simulations with $n_0 = 256$ (markers) for the total error with $\phi = 0.5$, $\gamma = 0.1$, $\sigma = \text{ReLU}$ and $\sigma_\varepsilon^2 = 0.01$ as a function of the overparameterization ratio ϕ/ψ for four different LJSDs $\mu_1, \ldots, \mu_4$, chosen such that the only comparable pairs of LJSDs under the partial order in Def. 4.1 are $\mu_1 \geq \mu_4$ and $\mu_2 \geq \mu_4$, and the strict ordering of the error for those pairs is seen for all values of ϕ/ψ. The orange (μ_2) and green (μ_3) curves cross one another, illustrating how nonmonotonicity of overlap ratios in Def. 4.1 can induce model-dependence in the ordering of the error. (**c**) Asymptotic predictions (solid lines) and simulations for $n_0 = 512$ (markers) for shifted versus unshifted error for models with varying values of the overparameterization ratio ϕ/ψ, obtained via Thm. 5.1 for the $(3, -1/2)$-diatomic LJSD with $\phi = 0.5$, $\sigma = \text{ReLU}$, $\sigma_\varepsilon^2 = 0.01$, and $\gamma = 0.005$. While the relationship is nearly linear in the overparameterized regime, it is markedly nonlinear in the underparameterized regime, highlighting the importance of overparameterization in Prop. 5.5.

Prop. 5.1 relies on a small label-noise condition ($\sigma_\varepsilon^2 \leq \omega$) to ensure that the total error is ordered with respect to shift strength. The reason this condition is necessary is that the variance can actually increase as shifts become easier. While Fig. 1 presented a configuration for which the bias, variance, and error all decrease for easier shifts, Fig. 3(a) shows that a decrease is not guaranteed for the variance, and that it can increase even under the small label-noise condition. Moreover, while the bias continues to decrease for large label noise ($\sigma_\epsilon^2 > \omega$), the variance can become so large that bias can no longer offset it, causing the error itself to increase, as seen in Fig. 3(a).

The strict ordering of overlap coefficients in Def. 4.1 are also necessary to guarantee a complete decoupling of the model and the shift strength. In the absence of these conditions, Fig. 3(b) shows how even a single out-of-order overlap coefficient induces a violation of the monotonicity with respect to shift strength suggested by Prop. 5.1 (this example is detailed further in Sec. A8).

Finally, we note that the exact linear relationship between in-distribution and out-of-distribution generalization characterized by Prop. 5.5 in Eq. (19) relies crucially on the overparameterization condition, $\psi < \phi$, as evidenced in Fig. 3(c), which shows marked nonlinearity in the underparameterized regime. This observation is perhaps unsurprising, as severely underparameterized models tend towards chance predictions, which produce comparable errors on shifted and unshifted data. Indeed, similar nonlinear behavior is seen in the low-accuracy regime for realistic models [56, Figure 17].

6 Conclusion

We have presented an exact, asymptotic calculation of the test error, bias, and variance for random feature kernel regression in the presence of covariate shift. After defining a partial order over covariate shifts (motivated by the setting of linear regression), we have proved that harder shifts imply increased error. Our results capture many empirical phenomena such as the fact that overparameterization is beneficial even under covariate shift and that a linear relationship exists between the generalization error on shifted and unshifted data. 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 covariate shift for other loss functions.

Acknowledgments and Disclosure of Funding

The authors would like to thank Rodolphe Jenatton, Horia Mania, Ludwig Schmidt, D. Sculley, Vaishaal Shankar, Lechao Xiao, and Steve Yadlowsky for valuable discussions.

This work was performed at and funded by Google. No third party funding was used.

References

[1] B. Adlam, J. Levinson, and J. Pennington. A random matrix perspective on mixtures of nonlinearities for deep learning. *arXiv preprint arXiv:1912.00827*, 2019.

[2] B. Adlam and J. Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 74–84. PMLR, 2020.

[3] B. Adlam and J. Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11022–11032. Curran Associates, Inc., 2020.

[4] M. S. Advani, A. M. Saxe, and H. Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

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

[6] M. Banna, J. Najim, and J. Yao. A CLT for linear spectral statistics of large random information-plus-noise matrices. *Stochastic Processes and their Applications*, 130(4):2250–2281, 2020.

[7] P. L. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: a statistical viewpoint, 2021.

[8] C. J. Becker, C. M. Christoudias, and P. Fua. Non-linear domain adaptation with boosting. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

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

[10] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.

[11] C. Cortes and M. Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theor. Comput. Sci.*, 519:103–126, 2014.

[12] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

[13] Y. Deshpande and A. Montanari. Sparse pca via covariance thresholding. *Journal of Machine Learning Research*, 17(141):1–41, 2016.

[14] E. Dobriban and S. Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247 – 279, 2018.

[15] J. Duchi, T. Hashimoto, and H. Namkoong. Distributionally robust losses for latent covariate mixtures, 2020.

[16] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378 – 1406, 2021.

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

[18] L. Erdős, H.-T. Yau, and J. Yin. Bulk universality for generalized wigner matrices. *Probability Theory and Related Fields*, 154(1-2):341–407, 2012.

[19] R. R. Far, T. Oraby, W. Bryc, and R. Speicher. Spectra of large block matrices. *arXiv preprint cs/0610045*, 2006.

[20] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 513–520, Madison, WI, USA, 2011. Omnipress.

[21] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[22] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[23] G. Grimmett. Percolation. In *Percolation*, pages 1–31. Springer, 1999.

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

[25] J. W. Helton, T. Mai, and R. Speicher. Applications of realizations (aka linearizations) to free probability. *Journal of Functional Analysis*, 274(1):1–79, 2018.

[26] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, October 2021.

[27] D. Hendrycks and T. G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[28] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org, 2015.

[29] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8580–8589, Red Hook, NY, USA, 2018. Curran Associates Inc.

[30] D. Kobak, J. Lomond, and B. Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020.

[31] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. Wilds: A benchmark of in-the-wild distribution shifts. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021.

[32] A. Kumar, T. Ma, and P. Liang. Understanding self-training for gradual domain adaptation. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5468–5479. PMLR, 13–18 Jul 2020.

[33] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as gaussian processes. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[34] J. Lee, S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15156–15172. Curran Associates, Inc., 2020.

[35] Z. Liao, R. Couillet, and M. W. Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13939–13950. Curran Associates, Inc., 2020.

[36] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 136–144, Red Hook, NY, USA, 2016. Curran Associates Inc.

[37] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2208–2217, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

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

[39] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[40] H. Mania and S. Sra. Why do classifier accuracies show linear trends under distribution shift?, 2021.

[41] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 06 2021.

[42] G. Mel and S. 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 M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7578–7587. PMLR, 18–24 Jul 2021.

[43] J. Miller, K. Krauth, B. Recht, and L. Schmidt. The effect of natural distribution shift on question answering models. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6905–6916. PMLR, 13–18 Jul 2020.

[44] J. P. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7721–7735. PMLR, 18–24 Jul 2021.

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

[46] Z. Nado, S. Padhy, D. Sculley, A. D'Amour, B. Lakshminarayanan, and J. Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

[47] V. Nagarajan and J. Z. Kolter. Uniform convergence may be unable to explain generalization in deep learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[48] R. M. Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.

[49] B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro. Exploring generalization in deep learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[50] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your models uncertainty? evaluating predictive uncertainty under dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

[52] J. Pennington and P. Worah. Nonlinear random matrix theory for deep learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[53] J. Pennington and P. Worah. The spectrum of the fisher information matrix of a single-hidden-layer neural network. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[54] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

[55] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

[56] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 2019.

[57] D. Richards, J. Mourtada, and L. Rosasco. Asymptotics of ridge(less) regression under general source condition. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3889–3897. PMLR, 13–15 Apr 2021.

[58] D. V. Rosen. On moments of the inverted wishart distribution. *Statistics*, 30(3):259–278, 1997.

[59] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[60] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry. Adversarially robust generalization requires more data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[61] V. Shankar, A. Fang, W. Guo, S. Fridovich-Keil, J. Ragan-Kelley, L. Schmidt, and B. Recht. Neural kernels without tangents. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8614–8623. PMLR, 13–18 Jul 2020.

[62] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005, 2007.

[63] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18583–18599. Curran Associates, Inc., 2020.

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

[65] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[66] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

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

[68] Z. Yang, Y. Bai, and S. Mei. Exact gap between generalization error and uniform convergence in random feature models. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 11704–11715. PMLR, 2021.

[69] H. Zhao, R. T. D. Combes, K. Zhang, and G. Gordon. On learning invariant representations for domain adaptation. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7523–7532. PMLR, 09–15 Jun 2019.

[70] H. Zhao, S. Zhang, G. Wu, J. M. F. Moura, J. P. Costeira, and G. J. Gordon. Adversarial multiple source domain adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.