Covariate Shift in High-Dimensions and Overparameterized Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Covariate shift–when the input distributions in the training and test sets differ, but the conditional label distribution remains fixed–often results in the unpredictable degradation of machine learning systems. In particular, theoretical understanding of covariate shift in deep learning models is lacking. Studying the high-dimensional asymptotics of random feature regression has emerged as a paradigm for understanding the generalization properties of modern neural networks, which moves beyond the worst-case bounds of uniform convergence theory. We extend the framework of exact asymptotics to accommodate more general data generating processes–including covariate shift between train and test distributions–and characterize the limiting test error, bias, and variance of random feature predictors in this setting. This motivates a natural partial ordering over covariate shifts that provides a sufficient condition for when the test error will be harmed (or even helped) by the shift. We find the benefits of overparameterization extend to settings with covariate shift by proving several monotonicity properties of the test error, bias, and variance. Additionally, we show our framework is sufficiently rich to replicate (perhaps surprising) empirical phenomena where linear trends in generalization error are observed between models evaluated on unshifted and shifted distributions.

1 Introduction

Theoretical justification for almost all machine learning methods fundamentally depends upon the equality of the distributions from which the training and test data are drawn. Yet, in many real-world applications, this equality is demonstrably violated—naturally-occurring distribution shift between the training data and the data encountered during deployment is the norm, not the exception [18]. Even *non-adversarial* changes in distributions can uncover the surprising fragility of modern machine learning models [36, 37, 27, 15, 8, 31]. Such shifts are distinct from adversarial examples, which require explicit poisoning attacks [12]. Rather, they can result from mild corruptions, such as changes of camera angle or blur [15], to subtle, unintended changes in data recollection procedures [37]. Moreover, this fragility is limiting the application of deep learning in certain safety-critical areas [18].

Empirical studies of distribution shift have observed several intriguing phenomena, including linear trends between model performance on shifted and unshifted test distributions [37, 15, 18], dramatic degradation in calibration [31], and surprising spurious inductive biases [8]. 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 even help) a model's performance, and by how much, are not understood. Although uniform convergence theory has formed a powerful foundation for understanding generalization in classical models (see *e.g.* [42]), there is growing evidence these techniques do not explain the behavior of modern, deep learning methods (even in the

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

absence of distribution shift) [28]. Indeed, classical measures of model complexity, such as various norms of the parameters, have been found to lead to ambiguous conclusions [30].

In this paper, we bring the tools of exact, high-dimensional asymptotics to bear on the phenomenology of distribution shift. We study a specific type of distribution shift called covariate shift, where the distribution of the training covariates does not equal that of the test covariates but 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 unstructured, 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 notion of covariate shift strength and 3.) replicate patterns from empirical work [36, 37, 16] characterizing how model complexity affects in-distribution vs. out-of-distribution generalization.

1.1 Contributions

Our primary contributions are to:

1. Calculate and analyze theoretical predictions for the test error, bias, and variance of random feature regression for general multivariate Gaussian covariates under covariate shift in the high-dimensional asymptotic limit (see Section 2.2);
2. Provide a *definition* of a class of covariate shifts and a partial ordering over them that is a sufficient condition determining whether they increase or decrease test error (see Definition 1);
3. Analyze the bias and variance under covariate shift to argue that the benefits of overparameterization extend to this setting, *i.e.* larger models generalize better, and find examples of nonmonotonicity in the bias with respect to model complexity (see Section 2.4);
4. Replicate the linear trends between unshifted and shifted generalization errors observed in several models and datasets [36, 37], providing understanding for this surprising phenomenon (see Section 2.3).

1.2 Related Work

There is an extensive literature on empirical methods and algorithms for unsupervised domain adaptation (DA): where data from a labeled source domain is used to generalize to an unlabeled target domain [39, 11, 3, 44, 43, 23, 22]. The number of theoretical works studying the impact of distribution shift on generalization are far fewer. One pioneering work in this vein [5], proves VC-dimension-based error bounds for classification that are augmented by a discrepancy measure between shifted source and target domain. [7] demonstrates a similar class of uniform convergence-based results in the setting of kernel regression. More recently, [20] provides non-vacuous margin-based bounds for self-training under the condition that the source domain gradually shifts toward the target domain. In [25], assumptions based on model similarity are used to help explain why classifiers exhibit linear trends between their accuracies on shifted and unshifted test sets.

In contrast to the above lines of work, which provide *worst-case* generalization bounds, we build upon a recent line of work developing tools for understanding the exact, high-dimensional asymptotics of regression methods. A series of theoretical works have computed asymptotically exact expressions for the test error of minimum-norm interpolated least squares and ridge regression in the random design setting [4, 9, 14]. However, the purview of the linear model is limited by the fact that the model complexity or number of parameters is directly tied to the input dimension. An exact analysis of random feature regression applied to a fixed data generating process, provides a rich but tractable class of models to gain further insight into generalization phenomena [2, 26]. Random feature methods are of particular interest because of their connection to neural networks, with the number of random features serving as an analogue to the network width (or model complexity) [29, 17]. However, this class of models are also an interesting subject of study (and a practical method for data analysis) in their own right [35, 38]. 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 [33, 34, 24, 32, 10].

Recent empirical work has shown across a suite of benchmarks that standard robustness interventions provide little to no protection against naturally, occurring distribution shifts [37, 40]. Indeed, further

work has shown that the standard procedure of performing empirical risk minimization (ERM) on clean, unshifted training data often maintains the best performance on out-of-distribution (OOD) benchmarks across a variety of domains compared to competing methods (such as invariant risk minimization (IRM)). One of the most striking observations, in the context of natural distribution shifts (*i.e.* which might arise from subtle changes in data resampling) and data corruption models (i.e. snow, fog, blur, *etc*...), is that, simply put, model robustness to shifted data improves with increasing classifier accuracy on unshifted data [37, 40, 15, 27] in a variety of domains. For example, an increase in 1 accuracy point of a classifier on the unshifted CIFAR-10 test set corresponds to a 1.7 accuracy point increase on the CIFAR-10.1 dataset (a dataset subject to natural distribution shift) [37].

2 Covariate Shift in Random Feature Regression

Problem setup and notation. As in prior work studying random feature regression [14, 26, 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 further define the *overparameterization ratio* as ϕ/ψ which is the limit of $\frac{n_1}{m}$ 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 non-linear teacher neural networks (see [26, 2] for more details). With this in mind, 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, $\mathbf{x}_i \sim \mathcal{N}(0, \Sigma)$ with positive definite covariance matrix Σ, and the labels are generated by a linear function parameterized by $\beta \in \mathbb{R}^{n_0}$, whose entries are drawn independently from $\mathcal{N}(0, 1)$. In particular

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

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ 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 [35]. Here the random features are given by a single-layer neural network with random weights. Given a set of training data $X = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$ and a prospective test data point $\mathbf{x}$, the (random) featurizations of the train and test data are given as

$$F := \sigma(W_1 X/\sqrt{n_0}) \quad \text{and} \quad f := \sigma(W_1 \mathbf{x}/\sqrt{n_0})\,, \tag{2}$$

respectively, for a random weight matrix $W_1 \in \mathbb{R}^{n_1 \times n_0}$ with i.i.d. standard Gaussian entries and activation function $\sigma(\cdot) : \mathbb{R}^{n_0} \to \mathbb{R}$ applied elementwise. The corresponding kernel is

$$K(\mathbf{x}_1, \mathbf{x}_2) := \frac{1}{n_1} \sigma(W_1 \mathbf{x}_1/\sqrt{n_0})^\top \sigma(W_1 \mathbf{x}_2/\sqrt{n_0})\,. \tag{3}$$

Similarly, the model's predictions are defined as

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

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 γ is a ridge regularization constant. Our central object of study is the test loss for a datapoint $\mathbf{x} \sim \mathcal{N}(0, \Sigma^*)$ where $\Sigma^* \neq \Sigma$. The test error (without label noise on the test point) can be written as

$$E_{\Sigma^*} = \mathbb{E}_\beta \mathbb{E}_{\mathbf{x}} \mathbb{E}[(y(\mathbf{x}) - \hat{y}(\mathbf{x}))^2] - \sigma_\epsilon^2 = \mathbb{E}_{\mathbf{x}} \mathbb{E}[(\beta^\top \mathbf{x}/\sqrt{n_0} - Y K^{-1} K_{\mathbf{x}})^2]\,, \tag{5}$$

where the outer expectation over β has been suppressed since the quantity concentrates sharply around its mean (see the Appendix for more details) and the innermost expectation is computed over all the randomness from training (*i.e.* W_1, X, $\boldsymbol{\epsilon}$). We further decompose the test error into its bias and variance components by defining,

$$E_{\Sigma^*} = \underbrace{\mathbb{E}_{\mathbf{x}} \mathbb{E}[(\mathbb{E}[\hat{y}(\mathbf{x})] - y(\mathbf{x}))^2]}_{B_{\Sigma^*}} + \underbrace{\mathbb{E}_{\mathbf{x}}[\mathbb{V}[\hat{y}(\mathbf{x})]]}_{V_{\Sigma^*}}\,. \tag{6}$$

The bias and variance are similarly computed over all the randomness from training (*i.e.* W_1, X, $\boldsymbol{\epsilon}$).

Spectral assumptions. Our primary interest is in situations where the covariate distribution between the training and test distribution are *not equal*. Since Σ and Σ^* are deterministic inputs to the problem, we encode their difference and limiting behavior via their spectra, as this information is necessary for the random matrix theory calculation (see Appendix C) and to state our results. We use $\lambda_1, \ldots, \lambda_{n_0}$ and $v_1, \ldots, v_{n_0}$ to denote a choice for the eigenvalues and eigenvectors of Σ, and $\nu_1, \ldots, \nu_{n_0}$ and $u_1, \ldots, u_{n_0}$ for those of Σ^*, both ordered in decreasing magnitude. $r_i := \sum_{j=1}^{n_0} (u_j \cdot v_i)^2 \nu_j$ denotes the sequence of *overlap coefficients* associated to Σ^* with respect to Σ. We also use $\bar{\mathrm{tr}}$ to denote the dimension-normalized trace. Regularity assumptions on the spectra of Σ and Σ^* are necessary to state the limiting behavior of this system; it is not sufficient to consider the spectra of these matrices individually, they must be considered jointly 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 $(0, \infty)^2$ 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, i.e. *the problem under analysis.*[1]

Often we use (λ, r) for random variables sampled jointly from μ and denote the marginal of λ under μ with μ_{train}. Since the conditional expectation $\mathbb{E}[r|\lambda]$ is an important object in our study, we assume for simplicity that the marginal of λ under μ is either absolutely continuous or a finite sum of delta masses. We also assume that the expectations of λ and r are finite.

When the eigenspaces of Σ and Σ^* are aligned and $r_i = \nu_i = \Phi(\lambda_i)$ for some smooth function Φ, the support of the LJSD degenerates. Here, Assumption 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 [9, 26]. One special case of note, is when there is no shift *i.e.* Φ is the identity, which has LJSD $\mu_\emptyset$ defined as

$$(\lambda, \lambda) \text{ for } \lambda \sim \mu_{\text{train}}. \tag{8}$$

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

Assumption 2. *The activation functions $\sigma : \mathbb{R} \to \mathbb{R}$ are assumed to be differentiable almost everywhere. We assume there exists a universal constant C such that, $\|\sigma(x)\|, \|\sigma'(x)\| = O(\exp(Cx))$.*

2.1 When does covariate shift provably hurt (or help)?

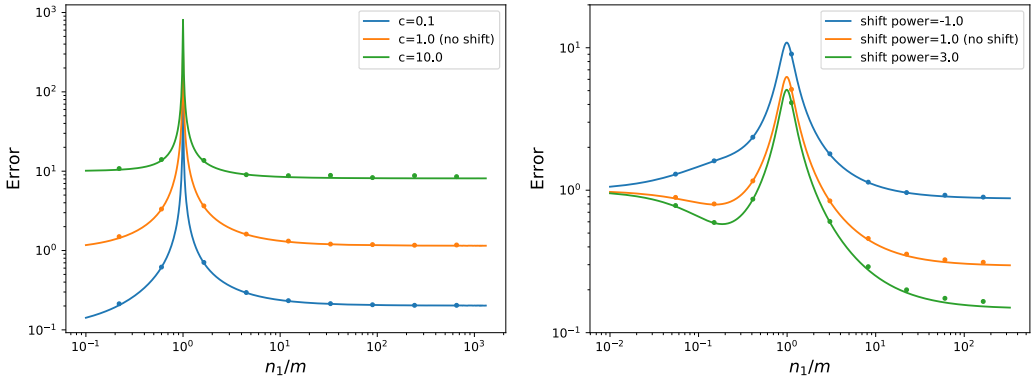
We first pause to consider (ridgeless) linear regression (LR) and use this section to illustrate some of the technical points that arise in the random feature case. The LR estimator is defined by $\hat{\beta} = (XX^\top)^{-1}XY$ under the generative model defined as before, $y_i = \mathbf{x}_i^\top \beta/\sqrt{n_0} + \epsilon$, except with fixed β. The test risk is defined as before in (5).

Proposition 1. *Fix the dimension n_0 and sample size m such that $m > n_0 + 1$. Then for LR*

$$E_{\Sigma^*} = \sigma_\epsilon^2 \frac{n_0}{m - n_0 - 1} \cdot \bar{\mathrm{tr}}(\Sigma^* \Sigma^{-1}) = \sigma_\epsilon^2 \frac{n_0/m}{1 - n_0/m - 1/m} \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{r_i}{\lambda_i}.$$

Even in this simple case, we can see how the test error depends on the interaction between the eigenspaces of Σ and Σ^* as well as their magnitudes. While Proposition 1 is correct for finite m, the limiting behavior is described similarly as an expectation over the distribution μ introduced in Assumption 1: $\sigma_\epsilon^2 \phi/(1-\phi)\mathbb{E}_\mu[r/\lambda]$. For the random feature case, our results only describe the limiting behavior of the test error, and similar functionals of μ are necessary to characterize this behavior with μ serving to define the limit of the covariate shift.

[1] 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 Appendix A.



(a) Test error for varying scales c on test covariance.

(b) Test Error for normalized *power-law* shifts.

Figure 1: Solid lines are theoretical predictions while dots are results from corresponding numerical simulations for $n_0 = 3000$. **(a)** Test error vs. n_1/m for tanh activation, $n_0/m = 2$, $\sigma_\epsilon^2 = 1$, $\lambda = 10^{-6}$ with $\mu := \delta_{(1,c)}$ for varying c. Simply changing the scale c of the test covariance relative to the training covariance can result in greater or less test error relative to $c = 1$. **(b)** Test error vs. n_1/m for ReLU activation, $n_0/m = 1/2$, $\sigma_\epsilon^2 = 1$, $\lambda = 10^{-3}$ with μ a 4-spike covariance subject to power-law shifts for different θ (see Example 1). Varying the shift-power θ for the test covariance results in an ordering of errors for all overparameterization ratios. The minimum test error is also achieved in the overparameterized regime ($n_1 \to \infty$) for all shifts.

One immediate question is whether a shift increases or decreases the test error relative to E_Σ. It is immediately clear that simply rescaling, as in $\Sigma^* = c\Sigma$ for a constant $c > 0$, rescales E_{Σ^*} by the same factor. This phenomenology can also be observed in the random feature model in Fig. 1a. Setting aside this trivial behavior[2], Proposition 1 motivates a partial ordering over shifts that settles the errors' ordering in certain nontrivial cases. Let $r_{i,1}, r_{i,2}$ be the sequence of overlap coefficients of Σ_1^* and Σ_2^* relative to Σ respectively, with the normalization $\mathrm{tr}(\Sigma_1^*) = \mathrm{tr}(\Sigma_2^*)$. In the non-asymptotic setting, if a pair Σ_1^* and Σ_2^* of covariances have overlap coefficients such that $r_{i,2}/r_{i,1}$ is a non-decreasing (or resp. non-increasing) sequence, a simple argument (see Appendix B) shows $E_{\Sigma_2^*} \geq E_{\Sigma_1^*}$ (or resp. $E_{\Sigma_2^*} \leq E_{\Sigma_1^*}$). Note one can set $\Sigma_1^* = \Sigma$ to compare to E_Σ. This condition and argument generalize naturally to the limiting expression.

Definition 1. Let μ_1 and μ_2 be LJSDs such that the marginal of λ is the same under μ_1 and μ_2 and that $\mathbb{E}_{\mu_1} r = \mathbb{E}_{\mu_2} r$. If the asymptotic overlap coefficients are such that $\mathbb{E}_{\mu_1}[r|\lambda]/\mathbb{E}_{\mu_2}[r|\lambda]$ is non-decreasing (or resp. non-increasing) in λ, we say the shift μ_2 is *harder* (or resp. *easier*)[3] than μ_1. Moreover, if μ_1 has no shift ($\mu_1 = \mu_\emptyset$) then we say that μ_2 is *hard* (or resp. *easy*) when the above condition is satisfied.

Note that when we refer to the relative *hardness* of two limiting measures (or say one measure is *hard*) we *always* implicitly enforce the normalizations $\mathbb{E}_{\mu_1} r = \mathbb{E}_{\mu_2} r$ and equality of their marginals in λ to ensure their comparison is meaningful. As we will see in Section 2.3, this definition is also revealing in the random feature model. To probe the behavior arising from Definition 1, we introduce the following example that captures much of the phenomena we explore.

Example 1. For $\alpha \geq 1$ and $\theta \in \mathbb{R}$, we define the family of α-*spike* distributions with *power-law* distributional shifts as

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

where C is such that $\mathbb{E}_\mu[r] = 1$. For Eq. (9), $\mathbb{E}_\mu[r|\lambda] = C\lambda^\theta$. Note that this is the limit of

$$\Sigma_{ij} := \begin{cases} \alpha & \text{if } i = j \text{ and } i \leq \frac{n_0}{1+\alpha} \\ \alpha^{-1} & \text{if } i = j \text{ and } i > \frac{n_0}{1+\alpha} \\ 0 & \text{if } i \neq j \end{cases} \tag{10}$$

[2]Indeed, pre-processing techniques such as standardization are designed to enforce $\bar{\mathrm{tr}}(\Sigma) = \bar{\mathrm{tr}}(\Sigma^*) = 1$.

[3]The finite condition ($r_{i,2}/r_{i,1}$ is non-decreasing/non-increasing) implies this definition; set $\mu_i := \sum_j \delta_{(\lambda_{j,i}, r_{j,i})}/n_0$.

and $\Sigma^* := C\Sigma^\theta$, with the standard basis chosen for the eigendecomposition of Σ.

Despite its simplicity, the training distribution is a 1-parameter family that captures the fast eigenvalue decay of many datasets in machine learning. Such spectra are often dominated by several large eigenvalues and a long tail of many small eigenvalues [21]. The power θ parameterizes the strength of the shift (when $\theta = 1$, there is no shift). When $\theta < 1$ the directions in the training distribution that are explored more (*i.e.* $\alpha > 1$) are explored less in the test distribution since $\alpha^\theta < \alpha$, suggesting the shift makes learning harder. Similarly for $\theta > 1$ the converse is true: those directions which are explored more in the training distribution are explored yet further in the test distribution, suggesting the shift makes learning easier. Indeed, it is easy to check for two shifts indexed by $\theta_1 > \theta_2$, the latter shift is *harder* than the former in the sense of Definition 1.

2.2 Main Result: Exact Asymptotics for Random Feature Regression under Covariate Shift

Our main result characterizes the limits of the test error, bias, and variance of our model as $m \to \infty$ in terms of the solution to a self-consistent equation. This equation is complicated but can be solved numerically to make theoretical predictions. Let $s := \mathbb{E}_\mu[\lambda]$ and $s_* := \mathbb{E}_\mu[r]$, then define

$$\eta := \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(\sqrt{s}z)^2] \;-\; [\mathbb{E}_{z\sim\mathcal{N}(0,1)}\sigma(\sqrt{s}z)]^2, \quad \zeta := s\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma'(\sqrt{s}z)]^2$$
$$\eta_* := \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(\sqrt{s_*}z)^2] \;-\; [\mathbb{E}_{z\sim\mathcal{N}(0,1)}\,\sigma(\sqrt{s_*}z)]^2, \quad \text{and} \quad \zeta_* := s_*\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma'(\sqrt{s_*}z)]^2,$$

which capture the effect of the activation nonlinearity, and $\xi := \sqrt{\zeta_* s/\zeta s_*}$, which measures the mismatch in the overall scale between the training and test covariances. As for linear regression, we introduce some functionals of μ. Define

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

Theorem 2. *Fix $z \in \mathbb{C}^+$, then $x(z)$ satisfying*

$$x = \frac{2\zeta\psi}{\frac{\zeta}{x}(\psi-\phi) + \frac{\sqrt{\zeta(\zeta(\phi-\psi)^2 - (4x\psi z\phi)s)}}{x} + 2\psi((\eta-\zeta)s + \zeta\mathcal{I}_{1,1}(x))} \tag{11}$$

has a unique solution in $\mathbb{C}^+$ that defines x. Define $\tau_1(z)$, $\bar{\tau}_1(z)$, and $\tau_2(z)$ via the equations

$$\tau_1 = -\frac{\sqrt{(\phi-\psi)^2 - \frac{4sx\psi z\phi}{\zeta}} + \psi - \phi}{2\psi z}, \quad x = -z\frac{\zeta}{s}\tau_1\bar{\tau}_1, \quad \textit{and} \quad \tau_2 = \tau_1\mathcal{I}_{1,1}(x).$$

Let $\Re(z) = -\gamma$ and assume a LJSD of μ as $n_0, n_1, m \to \infty$. Then $E_{\Sigma^*} \to E_\mu$ equal to

$$\lim_{\Im(x)\to 0+}\Bigg\{s_* - 2\xi\frac{x}{\phi}\mathcal{I}^*_{1,1}(x) + -\xi^2\sigma_\varepsilon^2\frac{\partial}{\partial z}\frac{x}{\tau_1}\big((\eta_*/\zeta_* - 1)s_* + \mathcal{I}^*_{0,1}(x)\big) + (\eta_t - \zeta_t)\frac{\partial}{\partial z}\big(\frac{\psi}{\phi}z\tau_2\big)$$
$$+ \xi^2\Big(s_* - \big(\eta - \zeta - \frac{1}{z\bar{\tau}_1^2}\big)\frac{\partial}{\partial z}\Big(z\bar{\tau}_1\mathcal{I}^*_{0,1}(x)\Big) - \big(\frac{1}{z\bar{\tau}_1} + 2\big)\mathcal{I}^*_{0,1}(x)\Big)\Bigg\}. \tag{12}$$

Moreover, the random feature model's limiting bias and variance are characterized as

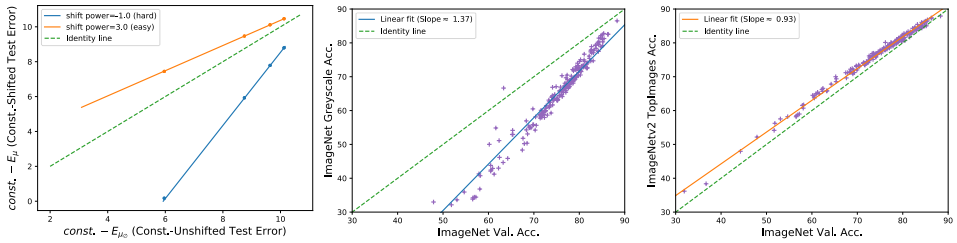
$$B_{\Sigma^*} \to \lim_{\Im(x)\to 0+}(1-\xi)^2 s_* + 2(1-\xi)\xi\mathcal{I}^*_{0,1}(x) + \xi^2\phi\mathcal{I}^*_{0,2}(x), \quad \textit{with} \quad V_{\Sigma^*} = E_{\Sigma^*} - B_{\Sigma^*}. \tag{13}$$

The power of Theorem 2 lies in its generality. However, the result simplifies considerably in the ridgeless limit, which is an interesting object of study in its own right.

Corollary 1. *In the setting of Theorem 2, as $\gamma \to 0$, the limiting test error E_{Σ^*} converges to*

$$(1-\xi)^2 s_* - \xi(2-\xi)\omega_* + \frac{x}{|\phi-\psi|}\big(\sigma_\varepsilon^2\xi^2\psi - \xi^2\omega\psi + \frac{\xi^2\phi}{x} + \frac{2\xi(1-\xi)|\phi-\psi|}{x}\big)\big(\omega_* + \mathcal{I}^*_{0,1}(x)\big)$$
$$+ \begin{cases} \xi^2(\sigma_\varepsilon^2 - \omega)\frac{x^2\mathcal{I}^*_{1,2}(x)}{1-x^2\mathcal{I}_{2,2}(x)} & \phi \geq \psi \\ \xi^2\phi\Big(\frac{1}{x\psi} + \frac{\omega_* - \omega + \phi(\mathcal{I}^*_{0,2}(x) - \mathcal{I}_{0,2}(x))}{\phi - x^2\psi\mathcal{I}_{2,2}(x)}\Big) & \phi < \psi \end{cases}.$$

where $x = \frac{\min(1,\phi/\psi)}{\omega + \mathcal{I}_{1,1}(x)}$, $\omega = (\eta/\zeta - 1)s$, $\omega_* \equiv (\eta_*/\zeta_* - 1)s_*$ and $B_{\Sigma^*}, V_{\Sigma^*}$ defined as before.



(a) Predictions from Proposition 4 (b) Empirical "Hard" Shift (c) Empirical "Easy" Shift

Figure 2: **(a)** Solid line shows the conclusion of Prop. 4 by plotting $c - E_\Sigma$ *vs.* $c - E_{\Sigma^*}$ for a constant c with the erf activation function, $n_0/m = 1/2$, $\sigma_\epsilon^2 = 1$ in the 4-spike model and power-law shifts. Dots represent simulations for $n_0 = 3000$. Note Proposition 4 predicts a slope $\geq / \leq 1$ for hard/easy shifts respectively, matching the empirically observed behavior in **(b)** and **(c)** plots. **(b)** Encapsulates the conclusions of [37, 40], by showing the relationship between the classification accuracy of several models (indexed by purple markers) on the original ImageNet test set and their performance on a corrupted (by greyscale) version of the ImageNet test set. **(c)** Example from [37] displaying the "easy" TopImages test set for which models perform *better* than on the original ImageNet test set. In this case the slope of the linear trend in Fig. 2c is ≤ 1 as predicted by Prop. 4.

The (relative) simplicity of Corollary 1 will allow for more detailed analytic understanding of the interplay between the number of random features and the covariate shift. By carefully taking the linear regression limit of Corollary 1 we can also recover formulas for the test risk in [14] and in Proposition 1 (see Appendix C.5).

2.3 Test Error of Random Features under Covariate Shift

The structural form of Theorem 2 implies the generalization error under covariate shift exclusively factors through the $\mathcal{I}^*$ (all other terms such as $x, \tau_1, \bar{\tau}_1$ only depend on the marginal of λ under μ and the overall scale $\mathbb{E}_\mu[r]$). As such, the intuitions arising from the setting of linear regression in the study of the $\mathrm{tr}(\Sigma^*\Sigma^{-1})$ term, can also be applied to $\mathcal{I}^*$ functionals arising in the random feature test errors using a more elaborate argument.

Proposition 3. *Consider two LJSD's μ_1 and μ_2 such that μ_2 is harder than μ_1 (recall Definition 1). Denote the limit of $E_{\Sigma_i^*}$ as E_{μ_i} as in Theorem 2. Then, if[4] $\xi \leq 1$, in the setting of Corollary 1,*

$$E_{\mu_2} \geq E_{\mu_1}.$$

After removing the impact of overall scales, Proposition 3 provides a sufficient condition to order test errors, corroborating the intuitions obtained from linear regression. By setting $\mu_1 = \mu_\emptyset$, Proposition 3 characterizes which shifts hurt (or even help) the test error relative to the unshifted case. An error ordering at finite regularization in the setting of Example 1 can be seen in Fig. 1b.

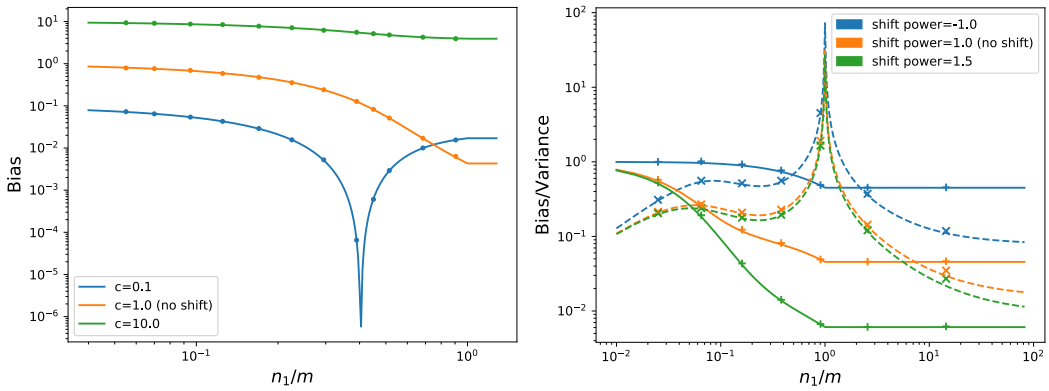
Random feature regression is a richer model than that of linear regression. In particular, the number of random features n_1 (or ϕ/ψ, when considered relative to the dataset size) serves as a surrogate for the width of the corresponding neural network. Effectively, it describes the complexity of the regression model. This allows us to make a remarkable connection to the recent empirical observations of [36, 37, 27] – which characterize generalization performance of models of varying complexity on clean test data *vs.* shifted test data.

Proposition 4. *Denote the limits of E_{Σ^*} as E_μ and E_Σ as $E_{\mu_\emptyset}$ as in Theorem 2. In the setting of Corollary 1, when $\psi \leq \phi$ (*i.e. *in the overparameterized regime)*

$$E_\mu(\psi) = E_0 + \underbrace{\xi^2\left(\frac{\omega_* + \mathcal{I}_{0,1}^*}{\omega + \mathcal{I}_{1,1}}\right)}_{\text{SLOPE}} E_{\mu_\emptyset}(\psi)\,, \tag{14}$$

where E_0 is independent of ψ. When μ is hard, SLOPE ≥ 1, *whereas when μ is easy,* SLOPE ≤ 1.

[4] Note that under an equal trace normalization between train and test covariate distributions (i.e. $s = s_*$) $\xi = 1$. The condition $\xi \leq 1$ is necessary to infer a partial ordering over test errors (see Appendix D.1.1).



(a) Bias for varying scales c on test covariance (b) Bias-variance decomposition for power-law shifts.

Figure 3: Solid lines are theoretical predictions while dots are results from corresponding numerical simulations for $n_0 = 5000$. **(a)** Bias *vs.* n_1/m for tanh activation, $n_0/m = 1/2$, $\sigma_\epsilon^2 = 1/10$, $\lambda = 0$ with $\mu = \delta_{(1,c)}$ for varying c (which induces varying ξ). Note, the bias is nonmonotonic for $c = 0.1$ (where $\xi > 1$) in the underparameterized regime. **(b)** Bias-variance decomposition *vs.* n_1/m for ReLU activation, $n_0/m = 1/2$, $\sigma_\epsilon^2 = 1/10$, $\lambda = 10^{-6}$ with μ a 9-spike distribution with varying power-law shifts. Solid lines/dashed lines denote bias/variance respectively. Increasing the number of random features decreases the bias, and decreases the variance in the overparameterized regime. The variance can display non-monotonic behavior in the underparameterized regime (where $n_1 < m$).

Proposition 4 makes two non-trivial predictions that accord with two (perhaps surprising) empirical phenomena. First, the collective works [36, 37, 27] exhibit several domains where varying the complexities of classifiers results in approximately linear trends between their generalization error across clean and corrupted test sets. Proposition 4 shows in the overparameterized regime[5] varying the number of random features (a surrogate for model complexity) results in an exactly linear relationship between the test error on unshifted *vs.* shifted test data (see Fig. 2a). Second, [36, 37, 27] observes that for shifts that decrease generalization performance, models performing 1 point better on clean data generically perform roughly > 1 point better on shifted test data. Proposition 4 also qualitatively matches this behavior since for our definition of *hard* shifts, the slope of the line between the mean-squared errors is always ≥ 1. We note that for general $\gamma \neq 0$, in the underparameterized regime, we can observe nonlinear relationships between $E_\mu(\psi)$ and $E_{\mu_\emptyset}(\psi)$ (see Appendix D.1.2).

2.4 Bias, Variance, Covariate Shifts and Overparameterization

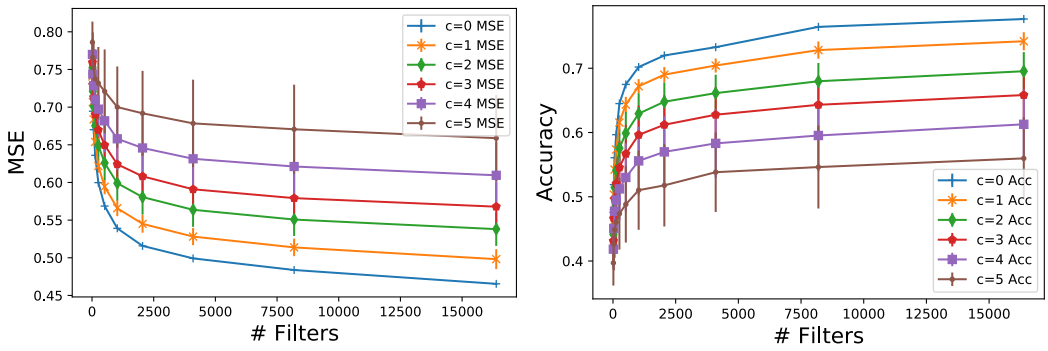
The bias-variance decomposition in Theorem 2 provides a powerful tool to understand covariate shift. We first characterize the behavior of the bias and variance as we vary the number of random features for general covariate shifts.

Proposition 5. *Denote the limit of B_{Σ^*} as B_μ and V_{Σ^*} as V_μ. Then, in the setting of Corollary 1:*

1. *the bias B_μ is a non-increasing function of the overparameterization ratio ϕ/ψ when $\xi \leq 1$;*
2. *the variance V_μ, in the overparameterized regime (i.e. for $\psi < \phi$), is a decreasing function of the overparameterization ratio ϕ/ψ.*

Proposition 5 formally establishes two generic phenomena that we also numerically observe in the ridged setting in Fig. 3b. Namely that increasing model complexity (*i.e.* increasing n_1/m) decreases the model bias under the trace normalization condition. Interestingly, in the regime $\xi \leq 1$, examples exist where the bias is non-monotonic in the underparameterized regime (see Fig. 3a). The second conclusion is that increasing the model complexity *decreases* the model variance in the overparameterized regime (also see Fig. 3b). This surprising phenomena has been demonstrated in previous exact asymptotic studies of random feature regression in the absence of covariate shift, in stark contrast to what what classical theory would suggest [2, 26]. Proposition 5 and Fig. 3b show this

[5] This tends to be the regime of interest in practice as it is often where the lowest test error is obtained.



(a) Multi-class mean-squared error vs. number of filters.

(b) Accuracy vs number of filters.

Figure 4: Error vs number of filters in convolutional layer (*i.e.* width) in random feature classifier. Averaged test performance (± 0.5 standard deviation) computed for classifier trained on CIFAR-10 and tested on corrupted datasets in CIFAR-10-C (note $c = 0$ corresponds to no corruption).

also holds under covariate shift. Together, our observations provide theoretical support for one nearly universally observed empirical phenomenon: *bigger models generalize better even under distribution shift* [37, 40, 15]. Finally, we note that increasing the hardness of the covariate shift increases bias.

Proposition 6. *Denote the limit of $B_{\Sigma_i^*}$ as B_{μ_i} as in Theorem 2. Consider two LJSD's μ_1 and μ_2 such that μ_2 is harder than μ_1 (recall Definition 1). Then, if $\xi \leq 1$, in the setting of Theorem 2*

$$B_{\mu_2} \geq B_{\mu_1}.$$

In situations where the label noise is small, the bias arising from model misspecification is often the dominant source of error.

3 Experiments

We use a family of shifted test sets alongside the clean CIFAR-10 test set: the CIFAR-10-C datasets from [16] consisting of 19 families of image corruptions (*e.g.* Gaussian blur, fog, *etc*...) at 5 increasing levels of strength, to the original CIFAR-10 test set. We show the key insights from our theoretical analysis are qualitatively observed under naturally occurring distribution shifts. For our experiments, we use a (random) convolutional kernel (least-squares) classifier, similar to the Coates-Ng feature map described in [6, 37]. We study the effects of varying the number of output channels in the convolution layer (*i.e.* increasing the number of random features) on classifiers in the overparameterized regime.

In Fig. 4 we can see increasing the corruption strength results in an *ordered* drop/increase in MSE/Accuracy of the classifier in correspondence to the difficulty of the shift. Similarly, increasing the overparameterization ratio improves the classifier accuracy/error on both the clean and corrupted test sets. Finally, we find that higher mean-squared error on the test distribution correlates well with the classifier's accuracy on average. Overall, these phenomena accord well with our theoretical predictions from the random feature model. Additional experiments and details are included in Appendix D.2, including results regarding the empirical bias-variance decomposition.

4 Conclusion

We have presented an exact, asymptotic calculation of the test error and bias/variance decomposition for random feature kernel regression in the presence of covariate shift. After carefully defining a partial ordering over covariate shifts (motivated by the setting of linear regression) we have shown how our analysis can qualitatively recover many empirically observed phenomena: such as the fact that overparameterization is still advantageous even under covariate shift and why [37] observe linear trends in generalization error. Extending our results to hold in the non-asymptotic regime is an important future direction, as is accommodating more general neural network models where features are learned (*i.e.* not random). Similarly, investigating the impact of covariate shift for other loss functions (*i.e.* the hinge, logistic, or $0/1$ losses) also warrants further study and is not addressed in our current work.

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 *International Conference on Machine Learning*, pages 74–84. PMLR, 2020.

[3] C. J. Becker, C. M. Christoudias, and P. Fua. Non-linear domain adaptation with boosting. In *Neural Information Processing Systems (NIPS)*, 2013.

[4] 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.

[5] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

[6] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 215–223, Fort Lauderdale, FL, USA, 11–13 Apr 2011. JMLR Workshop and Conference Proceedings.

[7] C. Cortes and M. Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

[8] 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.

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

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

[11] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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

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

[15] 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, 2020.

[16] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[17] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[18] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. WILDS: A benchmark of in-the-wild distribution shifts. *arXiv*, 2020.

[19] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

[20] A. Kumar, T. Ma, and P. Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pages 5468–5479. PMLR, 2020.

[21] J. Lee, S. S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. *arXiv preprint arXiv:2007.15801*, 2020.

[22] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. *arXiv preprint arXiv:1602.04433*, 2016.

[23] 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.

[24] 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.

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

[26] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[27] 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.

[28] V. Nagarajan and J. Z. Kolter. Uniform convergence may be unable to explain generalization in deep learning. *arXiv preprint arXiv:1902.04742*, 2019.

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

[30] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. *arXiv preprint arXiv:1706.08947*, 2017.

[31] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019.

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

[33] J. Pennington and P. Worah. The spectrum of the fisher information matrix of a single-hidden-layer neural network. In *NeurIPS*, pages 5415–5424, 2018.

[34] J. Pennington and P. Worah. Nonlinear random matrix theory for deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124005, 2019.

[35] A. Rahimi, B. Recht, et al. Random features for large-scale kernel machines. In *NIPS*, volume 3, page 5. Citeseer, 2007.

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

[37] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet?, 2019.

[38] 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.

[39] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

[40] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[41] D. von Rosen. Moments for the inverted wishart distribution. *Scandinavian Journal of Statistics*, pages 97–109, 1988.

[42] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[43] H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532. PMLR, 2019.

[44] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31:8559–8570, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We believe the all main claims in our contributions section in Section 1.1 are directly supported by theoretical results later established in the paper.
   (b) Did you describe the limitations of your work? [Yes] In our conclusion we discuss directions for extending our work (and the implicit limitations herein).
   (c) Did you discuss any potential negative societal impacts of your work? [No] While it is hard to envision all future applications of this research, we do not believe this theoretical work will raise any ethical concerns or will generate any adverse future societal consequences. We do hope that obtaining a more rigorous understanding of covariate shift will ultimately lead to safer and more reliable machine learning methods.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We state our assumptions in Section 2.
   (b) Did you include complete proofs of all theoretical results? [Yes] Proofs are contained in the Appendices.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We included code/instructions in the Appendices.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Experimental details are included in the Appendices.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We included error bars in our real data experiment in Section 3.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We included our compute set-up in the Appendices.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We believe the CIFAR-10/CIFAR-10C datasets do not contain either of these.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]