Harmless interpolation in regression and classification with structured features

Andrew D. McRae
Georgia Tech

Santhosh Karnik
Michigan State University

Mark A. Davenport
Georgia Tech

Vidya Muthukumar
Georgia Tech

Abstract

Overparametrized neural networks tend to perfectly fit noisy training data yet generalize well on test data. Inspired by this empirical observation, recent work has sought to understand this phenomenon of *benign overfitting* or *harmless interpolation* in the much simpler linear model. Previous theoretical work critically assumes that either the data features are statistically independent or the input data is high-dimensional; this precludes general nonparametric settings with structured feature maps. In this paper, we present a general and flexible framework for upper bounding regression and classification risk in a reproducing kernel Hilbert space. A key contribution is that our framework describes precise sufficient conditions on the data Gram matrix under which harmless interpolation occurs. Our results recover prior independent-features results (with a much simpler analysis), but they furthermore show that harmless interpolation can occur in more general settings such as features that are a bounded orthonormal system. Furthermore, our results show an asymptotic separation between classification and regression performance in a manner that was previously only shown for Gaussian features.

1 INTRODUCTION

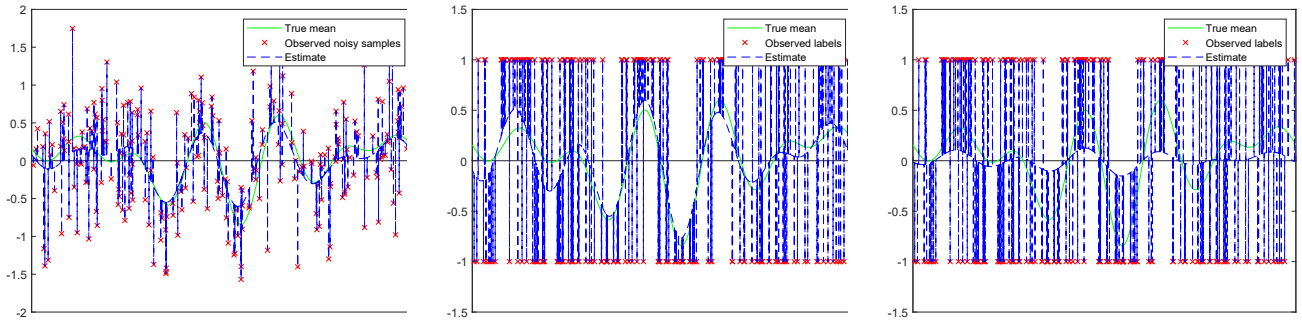
Overparametrized neural networks tend to perfectly fit, or *interpolate*, noisy training data. Somewhat surprisingly, these overparametrized networks also tend to generalize well (C. Zhang et al., 2017). More recently, this phenomenon of "harmless interpolation" was also empirically demonstrated in the much simpler model families of kernel machines (Belkin et al., 2018) and overparametrized linear models (Belkin et al., 2019). These observations have motivated a large body of research that aims to develop a mathematical understanding of the generalization properties of interpolating solutions and the impact of fitting noise (see Section 1.2 and Appendix B for more related work).

While these theoretical results represent significant progress, they come with some caveats. Most notably, harmless interpolation has only been shown under (a) strong assumptions on the feature distribution or (b) high dimension of the input data. For example, the strongest guarantees on harmless interpolation assume that the features consist of independent random variables (or are a linear transformation of such a vector). Similarly, the positive results on consistency of kernel interpolation require the dimension of the input data to grow with the size of the training set.

To see why these assumptions may not be realistic, consider the problem of simple linear regression using a Fourier series model $f(x) = \sum_\ell a_\ell e^{i2\pi\ell x}$ for a function f on the interval $[0, 1]$, where ℓ may range over all integers or, as we will later assume, a subset $\{-d, \dots, d\}$. Here the input data dimension is 1, and the features are given by $v_\ell(x) = e^{\mathbf{j}2\pi\ell x}$. If x is uniformly distributed, the features $\{v_\ell(x)\}_\ell$ (all evaluated at the same random x), though *uncorrelated*, are not *independent*. In this (and many other) examples, the input data can be low-dimensional and the features may not be independent. Whether harmless interpolation is possible with high-dimensional feature maps on such *constant*-dimensional data remains an open question. As a first effort, Muthukumar et al., 2020 show that harmless interpolation can occur with structured feature maps with uniformly spaced data, but whether this can be shown for the more realistic case of randomly-sampled data has remained open.

A second question is how the interpolation phenomenon applies to the *classification* problem. For example, Chatterji and Long, 2021; Muthukumar et

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.



(a) Gaussian-noise observations

(b) Binary labels—good regression performance

(c) Binary labels—poor regression but good classification performance

Figure 1: Interpolation in various regimes. This uses the bi-level Fourier series framework of Section 4.

al., 2021 show that the max-margin support vector machine can achieve good performance even when the corresponding regression task does not. These results require the very strong assumption of independent (sub)Gaussian features. Whether this asymptotic separation between regression and classification tasks exists in more general kernel settings is not addressed by this literature.

1.1 Our Contributions

In this paper, we provide new non-asymptotic risk bounds for both regression and classification tasks with the standard Hilbert-norm regularizer under minimal regularity assumptions. Our results apply for an arbitrarily small amount of regularization (including the interpolating regime) and are summarized below.

Harmless interpolation in kernel regression. For the regression task, we obtain new non-asymptotic risk bounds on the mean-squared-error of the Hilbert-norm regularized estimator, which includes the cases of kernel ridge regression and minimum-Hilbert-norm interpolation. In Section 2.2, we give error bounds for fixed sample locations. In Section 2.3, we give a variety of concentration results from random sampling that, when combined with our fixed-sample theorems, yield high-probability guarantees of harmless interpolation. Our results imply harmless interpolation in significantly more general settings than previous works (see Section 1.2 for a comparison to prior work). Our results recover existing independent-feature results (e.g., Bartlett et al., 2020) but also apply to other examples such as bounded orthonormal systems (BOSs). BOSs include many popular feature ensembles such as sinusoids and Chebyshev polynomials. Figure 1a shows an example of a function estimate that yields strong regression performance for the case of sinusoidal Fourier basis features.

Asymptotic separation between kernel classification and regression. We next analyze the classification error of the minimum-Hilbert-norm interpolator of binary labels. Although good regression performance implies good classification performance (see Figure 1b), the reverse is not true. In Section 3, we derive a simple bound on classification error that can be much tighter than the bound on regression error, and we present another fixed-sample error bound useful for bounding the regression risk. Then, for the case of bounded orthonormal system features, we demonstrate an asymptotic separation between the regression and classification tasks. Figure 1c illustrates how the minimum-norm label interpolator can have poor regression performance but good classification performance.

1.2 Related Work

Harmless interpolation. Recent work has shown that the "harmless interpolation" phenomenon becomes more pronounced with increased (effective) overparameterization when the minimum-Hilbert-norm interpolator is used in kernel regression or the minimum-norm interpolator is used in linear regression in a variety of models. All of the current known harmless interpolation results assume at least one of the following (see Appendix B for a more complete citation list): (a) independence of features, (b) sub-Gaussianity in the feature vector, (c) high data dimension, or (d) explicit structure in the kernel operator/feature map. For specific kernels like the Laplace kernel, statistically consistent interpolation may actually *require* growing data dimension with the number of training examples (Rakhlin & Zhai, 2019), as the data dimension fundamentally alters the eigenvalues of the Laplace kernel integral operator. In contrast, our results do not explicitly posit any of these assumptions. Our sufficient conditions for harmless interpolation are expressed purely in terms of the eigenvalues of the kernel integral operator, and do not utilize special structure either on the eigenfunctions or the integral operator itself.

Classification versus regression. General techniques from statistical learning theory (e.g., Bartlett and Mendelson, 2002; Vapnik, 2000) do not differentiate between classification and regression tasks. However, the idea that classification is easier than regression is well-known: the main idea is we do not need near-zero bias, but rather how much signal is recovered only needs to be large relative to the variance. This idea goes back to (Friedman, 1997), and has primarily been used to obtain faster non-asymptotic rates for classification relative to regression in a number of scenarios (Audibert & Tsybakov, 2007; Devroye et al., 1996; Koltchinskii & Beznosova, 2005). A separation in statistical consistency between the two tasks was shown more recently in Muthukumar et al., 2021. Similar sharp analyses for classification error have also been provided for the related high-dimensional linear discriminant analysis setting (Cao et al., 2021; Chatterji & Long, 2021; Wang & Thrampoulidis, 2021). These results all make restrictive assumptions of Gaussianity, independent sub-Gaussian features, or Gaussian mixture models; the most fine-grained analyses (Muthukumar et al., 2021; Wang & Thrampoulidis, 2021) require Gaussian design. With our more general analysis, we show that the previous restrictive assumptions can be avoided and demonstrate that the separation between classification and regression consistency is a general phenomenon. Although our error expressions are less sharp nonasymptotically than those that assume Gaussian features, the consistency implications are nearly identical.

General kernel regression. Finally, our work continues a substantial literature on general linear and RKHS regression. Space limitations prevent a comprehensive review, but we note that our analysis techniques most closely resemble the approach of Hsu et al., 2014; T. Zhang, 2005, who analyze explicitly regularized ridge regression under random design with minimal assumptions on the data distribution. Other notable works are Caponnetto and De Vito, 2007; Steinwart et al., 2009, which also use techniques based on the kernel integral operator. These works assume a power-law eigenvalue decay to get power-law regression error bounds. Our results apply to more general kernels with an arbitrary eigenvalue decay and give a more refined bias-variance decomposition of error. Significantly, none of these works analyze interpolating solutions in the presence of noise.

2 KERNEL REGRESSION

Our results are presented in terms of reproducing kernel Hilbert space regression (with traditional linear regression as a special case). We first introduce the analytical framework and then present our main results.

2.1 Kernel Regression Introduction

We first review the general theory of regression in reproducing kernel Hilbert spaces. A more thorough introduction to kernel theory can be found in many standard references, such as Schölkopf and Smola, 2002, Wendland, 2004, and Chapter 12 of Wainwright, 2019.

Let X be a set, and let $\mathcal{H}$ be a real reproducing kernel Hilbert space over X with kernel $k\colon X \times X \to \mathbf{R}$. For $f \in \mathcal{H}$ and $x \in X$, we have $f(x) = \langle f, k_x \rangle_{\mathcal{H}}$, where $k_x \coloneqq k(\cdot, x)$. Note that this implies that $k(x, y) = \langle k_x, k_y \rangle_{\mathcal{H}}$.

Suppose $f^* \in \mathcal{H}$, and we observe $y_i = f^*(x_i) + \xi_i$, $i = 1, \dots, n$, where $x_1, \dots, x_n \in X$ are sample points, and the ξ_i's represent noise or other measurement error. We use the kernel ridge regression estimate

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \alpha \|f\|_{\mathcal{H}}^2,$$

where $\alpha \geq 0$ is a regularization term. When $\alpha \to 0$, we get the minimum-Hilbert-norm interpolator,

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}} \text{ s.t. } f(x_i) = y_i \ \forall i = 1, \dots, n.$$

By the standard kernel regression formula, we have $\hat{f}(x) = \sum_{i=1}^{n} \hat{z}_i k(x, x_i)$ where the vector $\hat{z} = (\alpha I_n + K)^{-1} y$, and K is the kernel Gram matrix with $K_{ij} = \langle k_{x_i}, k_{x_j} \rangle_{\mathcal{H}} = k(x_i, x_j)$. We denote by $\mathcal{A}\colon \mathcal{H} \to \mathbf{R}^n$ the *sampling operator*, which is defined by $(\mathcal{A}(f))_i = f(x_i) = \langle f, k_{x_i} \rangle_{\mathcal{H}}$. The adjoint of the sampling operator is given by $\mathcal{A}^*(z) = \sum_{i=1}^{n} z_i k_{x_i}$ for all $z \in \mathbf{R}^n$. Then the Gram matrix is $K = \mathcal{A}\mathcal{A}^*$, and we can write the kernel regression estimate in terms of the standard ridge regression formulas:

$$\hat{f} = \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}^*)^{-1} y = (\alpha \mathcal{I}_{\mathcal{H}} + \mathcal{A}^*\mathcal{A})^{-1} \mathcal{A}^* y.$$

Note that, in general, the second expression is only well-defined if $\alpha > 0$ (since $\mathcal{A}$ is rank-deficient if $\mathcal{H}$ is infinite-dimensional).

We analyze two terms in this estimator. The first is the estimator that would be obtained in the absence of noise, which is given by

$$\hat{f}_0 \coloneqq \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}^*)^{-1} \mathcal{A} f^* = (\alpha \mathcal{I}_{\mathcal{H}} + \mathcal{A}^*\mathcal{A})^{-1} \mathcal{A}^*\mathcal{A} f^*.$$

The second is the contribution to the estimate due to noise, which we denote by the function $\epsilon(x)$. We have

$$\epsilon = \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}^*)^{-1} \xi,$$

where $\xi = (\xi_1, \dots, \xi_n)$. This leads to a standard decomposition in the error of the estimator $\hat{f}$ in terms of its bias and variance.

To characterize the test error, we need a sampling model. Let μ be a probability measure on X. We then define the kernel integral operator $\mathcal{T}$ as

$$(\mathcal{T}(f))(x) = \int_X k(x,y) f(y)\, d\mu(y)$$

with respect to the measure μ. Under mild regularity/continuity conditions (see, e.g., Steinwart and Scovel, 2012 for a thorough analysis), we have the eigenvalue decomposition

$$\mathcal{T}(f) = \sum_{\ell=1}^{\infty} \lambda_\ell \langle v_\ell, f\rangle_{L_2} v_\ell,$$

where $\{v_\ell\}_{\ell=1}^{\infty}$ is an orthonormal basis for $L_2(X,\mu)$, and $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots$ are the eigenvalues of $\mathcal{T}$ arranged in decreasing order. Furthermore, we have

$$k(x,y) = \sum_{\ell=1}^{\infty} \lambda_\ell v_\ell(x) v_\ell(y).$$

We can handle the finite-dimensional case by setting $\lambda_\ell = 0$ for $\ell > d$, where $d = \dim(\mathcal{H})$ (furthermore, the standard linear regression case can be recovered with $X = \mathbf{R}^d$ and $k(x,y) = \langle x, y\rangle_{\ell_2}$). Note that in order to interpolate an arbitrary set of samples, we need the dimension d to be at least the number of samples n (otherwise, the linear system is overdetermined).

We will also use the following well-known fact throughout our analysis: for any $f, g \in L_2$, we have

$$\langle f, g\rangle_{L_2} = \langle \mathcal{T}^{1/2} f, \mathcal{T}^{1/2} g\rangle_{\mathcal{H}}.$$

Hence, $\mathcal{T}^{1/2}$ is an isometry from L_2 to $\mathcal{H}$. Note that this implies that for every $f \in \mathcal{H}$,

$$\|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{\infty} \frac{\langle f, v_\ell\rangle_{L_2}^2}{\lambda_\ell}.$$

Intuitively, we expect that if f has small/bounded $\mathcal{H}$-norm, most of its energy is captured by components corresponding to relatively large eigenvalues. Therefore, it is feasible to recover an accurate (in L_2) estimate of f, even though f lies in an infinite-dimensional space.

We will assume $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \mu$, i.e., the training examples are drawn from the same measure as the test example $x \sim \mu$. Since we are evaluating a regression task, we wish to bound the squared (excess) prediction loss $\mathbf{E}(\hat{f}(x) - f^*(x))^2 = \|\hat{f} - f^*\|_{L_2}^2$. We will provide non-asymptotic upper bounds on $\|\hat{f} - f^*\|_{L_2}^2$ as a function of the number of training examples n. We will also focus on understanding scenarios for which we obtain statistical consistency, i.e., $\|\hat{f} - f^*\|_{L_2}^2 \to 0$ as $n \to \infty$.

2.2 Main Results for Deterministic Sample Locations

To state our main results, we introduce some additional notation. Here and for the rest of this section, p will be a fixed integer that we can tune in our analysis. We divide the function space $L_2(X,\mu)$ into two parts: $G = \operatorname{span}\{v_1, \ldots, v_p\}$ denotes the space spanned by the first p eigenfunctions of $\mathcal{T}$, and $G^\perp$ denotes its orthogonal complement (in both L_2 and $\mathcal{H}$). Accordingly, we split our sampling operator into two parts: $\mathcal{A}_G = \mathcal{A}|_G$ and $\mathcal{R} = \mathcal{A}|_{G^\perp}$. Intuitively, if p is chosen such that $\lambda_{p+1}, \lambda_{p+2}, \ldots$ are relatively small, we expect G to contain most of the energy in any given function $f \in \mathcal{H}$. A key fact is that the Gram matrix can be decomposed as $\mathcal{A}\mathcal{A}^* = \mathcal{A}_G \mathcal{A}_G^* + \mathcal{R}\mathcal{R}^*$. The dimension p is similar to (but more flexible than) the regularization-dependent effective dimension in Hsu et al., 2014; T. Zhang, 2005.

Since $p = \dim(G)$ is finite, we can recover a function in G from a finite number of samples. We state this quantitatively by analyzing the restricted sampling operator $\mathcal{A}_G$. To state concentration results on G in terms of the L_2 norm, we denote $\mathcal{C} = \mathcal{A}_G$, and we let $\mathcal{C}^* = \mathcal{T}_G^{-1} \mathcal{A}_G^*$ be its adjoint with respect to the L_2 inner product. Note that $\frac{1}{n}\mathbf{E}\,\mathcal{A}_G^*\mathcal{A}_G = \mathcal{T}_G$, where $\mathcal{T}_G = \mathcal{T}|_G$. Therefore,

$$\frac{1}{n}\mathbf{E}\,\mathcal{C}^*\mathcal{C} = \mathcal{I}_G,$$

where $\mathcal{I}_G$ is the identity operator on G. Provided that $n \gg p$, we expect $\frac{1}{n}\mathcal{C}^*\mathcal{C} \approx \mathcal{I}_G$. We will analyze how closely this holds later; we first state *deterministic* results that depend on the error in this approximation.

The second key approximation regards the "remainder Gram matrix" $\mathcal{R}\mathcal{R}^*$. Previous interpolation literature has assumed that this matrix is approximately a multiple of the identity I_n (or is in some sense "well-conditioned"). We will again analyze how accurately this holds later, but for now, we will state our main results assuming that $\alpha I_n + \mathcal{R}\mathcal{R}^*$ is upper and lower bounded by multiples of the identity. There is no requirement that $\alpha \geq 0$; in principle, our framework applies to negative regularization (Tsigler & Bartlett, 2020), but we do not explore this aspect in detail.

Finally, we will assume, for simplicity and brevity, that $f^* \in G$ exactly. If this did not hold, there would be another term in the "bias" error bound whose size is directly proportional to the size of $\mathcal{P}_{G^\perp}(f^*)$, which in turn is negligible provided that $f^* \in \mathcal{H}$ (i.e., f^* has bounded $\mathcal{H}$-norm). Note that kernel methods run into fundamental approximation-theoretic limitations in the absence of a bounded-$\mathcal{H}$-norm assumption (Belkin, 2018; Donhauser et al., 2021; Ghorbani et al., 2021).

Theorem 1 (Bias). *Suppose that*

1. $\alpha_L I_n \preceq \alpha I_n + \mathcal{R}\mathcal{R}^* \preceq \alpha_U I_n$ *for some numbers* $\alpha_U \geq \alpha_L > 0$, *and*
2. $\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \leq c$ *for some* $c < 1$.

Let $\bar{\alpha} = \frac{2\alpha_U\alpha_L}{\alpha_U+\alpha_L}$ *be the harmonic mean of* α_U *and* α_L. *Then, for any* $f^* \in G$, *we have*

$$\left\|\hat{f}_0 - f^*\right\|_{L_2} \lesssim \min\left\{\sqrt{\lambda_1}, \frac{1}{1-c}\frac{\bar{\alpha}}{n\sqrt{\lambda_p}}, \frac{1}{1-c}\sqrt{\frac{\bar{\alpha}}{n}}\right\} \times \left(1 + \sqrt{\frac{n\lambda_{p+1}}{\bar{\alpha}}}\right)\|f^*\|_{\mathcal{H}}.$$

Theorem 2 (Variance). *Suppose the conditions of Theorem 1 hold, and let* $\tilde{\alpha} = \frac{\alpha_U+\alpha_L}{2}$. *Furthermore, suppose the* ξ_i*'s are zero-mean and independent with variance bounded by* σ^2. *Then*

$$\mathbf{E}_\xi\|\epsilon\|_{L_2}^2 \lesssim \sigma^2\left(\frac{\alpha_U}{\alpha_L} + 1\right)^2\left(\frac{p}{n} + \frac{\mathrm{tr}_{L_2}(\mathcal{R}^*\mathcal{R})}{\tilde{\alpha}^2}\right). \quad (1)$$

Section 2.4 contains simplified proof sketches of Theorems 1 and 2; we provide complete proofs in Appendix C in the supplementary material. The reader should note that our proofs consist of relatively simple linear algebra. Compare this, for example, to Bartlett et al., 2020 or Muthukumar et al., 2021, where the analysis depends delicately on the independence (or, in the latter case, even Gaussianity) of the features via rather complicated matrix manipulations.

We could also obtain a high-probability (with respect to ξ) bound on the variance (if, e.g., the ξ_i's are sub-Gaussian), but we omit this to preserve the clarity and simplicity of the result. We outline how one could do this in Appendix C.2.

2.3 Operator Concentration Results

We now state operator concentration results on three important quantities: (a) the deviation of the residual Gram matrix $\mathcal{R}\mathcal{R}^*$ from a multiple of the identity, (b) the quantity $\mathrm{tr}_{L_2}(\mathcal{R}^*\mathcal{R})$ which appears in the variance bound, and (c) the deviation of $\frac{1}{n}\mathcal{C}^*\mathcal{C}$ from $\mathcal{I}_G$. All proofs are contained in Appendix D in the supplementary material. We begin with our most general results that apply under minimal assumptions.

2.3.1 General Residual Concentration

Let $k^R(x,y) = \sum_{\ell>p}\lambda_\ell v_\ell(x)v_\ell(y)$ be the reproducing kernel restricted to $G^\perp$.

Lemma 1 (Generic residual Gram matrix).

$$\mathbf{E}\|\mathcal{R}\mathcal{R}^* - (\mathrm{tr}\,\mathcal{T}_{G^\perp})I_n\|^2 \lesssim n^2\,\mathrm{tr}(\mathcal{T}_{G^\perp}^2) + \|k^R(\cdot,\cdot) - \mathrm{tr}\,\mathcal{T}_{G^\perp}\|_\infty^2,$$

where $\|k^R(\cdot,\cdot) - \mathrm{tr}\,\mathcal{T}_{G^\perp}\|_\infty = \sup_x\{|k^R(x,x) - \mathrm{tr}\,\mathcal{T}_{G^\perp}|\}$.

Note for this result to give $\alpha_L I_n \preceq \mathcal{R}\mathcal{R}^* \preceq \alpha_U I_n$ where α_U/α_L is bounded, we need $\mathrm{tr}\,\mathcal{T}_{G^\perp} \gtrsim n\sqrt{\mathrm{tr}(\mathcal{T}_{G^\perp}^2)}$. Even when $\{\lambda_\ell\}_{\ell>p}$ are all equal (see Section 3.1), we need $\dim\mathcal{H} = d \gtrsim n^2$. While this may seem restrictive, it is not possible to do better without additional assumptions on the features. In Appendix E, we show that in the case of Fourier features, $\lambda_{\max}(\mathcal{R}\mathcal{R}^*)/\lambda_{\min}(\mathcal{R}\mathcal{R}^*) \gtrsim \frac{n^4}{\tau^2 d^2}$ with probability at least $1 - e^{-\tau}$, and thus $d \gtrsim n^2$ is necessary. This can be significantly relaxed when the features are independent, as shown in Section 2.3.3.

To bound the variance, we will use the following expectation throughout the rest of this paper:

Lemma 2 (Generic trace bound on $\mathcal{R}^*\mathcal{R}$).

$$\mathbf{E}\,\mathrm{tr}_{L_2}(\mathcal{R}^*\mathcal{R}) = n\,\mathrm{tr}(\mathcal{T}_{G^\perp}^2) = n\sum_{\ell>p}\lambda_\ell^2.$$

Note that Lemma 2, Theorem 2, and the approximate identity $\mathcal{R}\mathcal{R}^* \approx (\mathrm{tr}\,\mathcal{T}_{G^\perp})I_n$ combine to bound the variance error as $\|\epsilon\|_{L_2}^2 \lesssim \frac{p}{n} + n\left(\sum_{\ell>p}\lambda_\ell^2\right)/\left(\sum_{\ell>p}\lambda_\ell\right)^2$. This is identical to the bound provided in Bartlett et al., 2020.

2.3.2 Bounded Orthonormal System

Our results show that harmless interpolation can occur in much more general settings than independent and/or sub-Gaussian features. An important class of features that are not independent or sub-Gaussian is a *bounded orthonormal system* (BOS).

On the subspace G defined before, the basis $v_1, \ldots, v_p$ is a BOS if it is an orthonormal basis in L_2 (as we have already assumed) and, further, we have

$$\sum_{\ell=1}^p v_\ell^2(x) \leq Cp$$

μ-almost surely in x for some constant $C \geq 1$. Equivalently, for all $f \in G$, $\|f\|_\infty^2 \leq Cp\|f\|_{L_2}^2$.

This assumption is satisfied by many popular choices of features including sinusoids (see Section 4), Chebyshev polynomials, and the standard Euclidean basis on $\mathbf{R}^d$. One can also often show that kernel eigenfunctions satisfy this property, such as when the data lie on a low-dimensional manifold (McRae et al., 2020).

It is easy to derive concentration inequalities for bounded orthonormal systems via matrix/operator concentrations results for sums of bounded independent random matrices (e.g., Tropp, 2015—see our supplementary material for details). A bound that is useful for our purposes is the following:

Lemma 3 (BOS sampling operator on G)**.** *If G is spanned by a bounded orthonormal system with constant C, then, for $t > 0$, with probability at least $1 - e^{-t}$,*

$$\frac{1}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \lesssim \sqrt{\frac{Cp(t+\log p)}{n}} + \frac{Cp(t+\log p)}{n}.$$

Thus if $n \gtrsim Cp\log p$, we can have, say, $\frac{1}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \le 1/4$ (or any other small constant) with high probability.

In general, the $Cp\log p$ sample complexity is optimal under the BOS assumption. As a simple example, consider the following basis $\{v_1, \dots, v_p\}$ on $\mathbf{R}^p$ (written as functions on $\{1, \dots, p\}$): for uniquely determined constants c_1 and c_2, set the measure to be $\mu(\{j\}) = \frac{1}{Cp}$ for $j < p$ and $\mu(\{p\}) = c_1$, and set $v_\ell = \sqrt{Cp}\delta_\ell$ for $\ell < p$ and $v_p = c_2\delta_p$. One can easily verify by a coupon collector argument that we need $O(Cp\log p)$ samples from μ merely to sample every coordinate at least once.

2.3.3 Independent Features

To compare to prior work, we list independent-feature concentration results that can be plugged into our Theorem 1. Suppose that for $x \sim \mu$, the features $\{v_\ell(x)\}$ are independent random variables. The key benefit this gives us is that we can now write the residual Gram matrix $\mathcal{R}\mathcal{R}^*$ as a sum of *independent* random rank-1 matrices. To see this, define the vectors

$$w_\ell = (v_\ell(x_1), v_\ell(x_2), \dots, v_\ell(x_n)) \in \mathbf{R}^n.$$

We have already been assuming that the entries of each w_ℓ are independent (since they only depend on the independent variables x_i), but an independent features assumption implies that the entire set of random vectors $\{w_\ell\}_{\ell\ge 1}$ is independent. We can then write

$$\mathcal{R}\mathcal{R}^* = \sum_{\ell > p} \lambda_\ell w_\ell \otimes w_\ell.$$

We state a formal result for sub-Gaussian independent features. We expect similar results hold for much weaker tail conditions.

Lemma 4 (Independent features residual Gram matrix)**.** *Suppose the features $\{v_\ell(x)\}_{\ell\ge 1}$ are zero-mean, independent, and sub-Gaussian. Then, for $t > 0$, with probability at least $1 - e^{-t}$,*

$$\|\mathcal{R}\mathcal{R}^* - (\operatorname{tr}\mathcal{T}_{G^\perp})I_n\| \lesssim \sqrt{(n+t)\operatorname{tr}(\mathcal{T}_{G^\perp}^2)} + (n+t)\lambda_{p+1}.$$

The zero-mean assumption is for simplicity and can easily be relaxed at the cost of a more complicated theorem statement. Note that this is stronger than Lemma 1 in two ways: first, the bound holds with exponentially high probability as opposed to being merely in expectation. Second, we have effectively replaced the n^2 in Lemma 1 by n, greatly reducing the amount of overparametrization we need.

Note for this result to give $\alpha_L I_n \preceq \mathcal{R}\mathcal{R}^* \preceq \alpha_U I_n$ where α_U/α_L is bounded, we need

$$n \lesssim \frac{\operatorname{tr}\mathcal{T}_{G^\perp}}{\lambda_{p+1}} = \frac{1}{\lambda_{p+1}}\sum_{\ell > p}\lambda_\ell$$

(this also gives us $\sqrt{n\operatorname{tr}(\mathcal{T}_{G^\perp}^2)} \lesssim \operatorname{tr}\mathcal{T}_{G^\perp}$ by Cauchy-Schwartz). This is identical to the requirement that $r_{k^*}(\Sigma) \ge bn$ in Bartlett et al., 2020.

We can also obtain slightly improved results (vs. the BOS assumption) for concentration of $\mathcal{C}^*\mathcal{C}$:

Lemma 5 (Sampling operator on G under independent features)**.** *With probability at least $1 - e^{-t}$,*

$$\left\|\frac{1}{n}\mathcal{C}^*\mathcal{C} - \mathcal{I}_G\right\|_{L_2} \lesssim \sqrt{\frac{p+t}{n}} + \frac{p+t}{n}.$$

Thus we only require $n \gtrsim p$ to obtain the required concentration. For a proof, see, for example, Vershynin, 2018, Section 4.6.

2.4 Informal Proof Sketch (Deterministic)

Here we outline the basic proof structure of Theorems 1 and 2. We will perform the analysis as though $\alpha I_n + \mathcal{R}\mathcal{R}^* = \bar{\alpha} I_n$ and $\mathcal{C}^*\mathcal{C} = n\mathcal{I}_G$ (equivalently, $\mathcal{A}_G^*\mathcal{A} = n\mathcal{T}_G$), and we will write "$\approx$" where we make these substitutions. The main additional steps we need are to quantify the error due to these approximations.

2.4.1 Bias Term (from Signal)

Note that since $f^* \in G$, we have

$$\begin{aligned}
\hat{f}_0 &= \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}_G f^* \\
&\approx \mathcal{A}^*(\bar{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}_G f^* \\
&= \mathcal{A}^*\mathcal{A}_G(\bar{\alpha}\mathcal{I}_G + \mathcal{A}_G^*\mathcal{A}_G)^{-1} f^* \\
&\approx \begin{bmatrix}\mathcal{A}_G^* \\ \mathcal{R}^*\end{bmatrix}\mathcal{A}_G(\bar{\alpha}\mathcal{I}_G + n\mathcal{T}_G)^{-1} f^*.
\end{aligned}$$

From this we obtain

$$\begin{aligned}
\mathcal{P}_G(\hat{f}_0) &= \mathcal{A}_G^*\mathcal{A}_G(\bar{\alpha}\mathcal{I}_G + n\mathcal{T}_G)^{-1} f^* \\
&\approx n\mathcal{T}_G(\bar{\alpha}\mathcal{I}_G + n\mathcal{T}_G)^{-1} f^* \\
&= \overline{\mathcal{S}} f^*,
\end{aligned}$$

where $\overline{\mathcal{S}} \coloneqq n\mathcal{T}_G(\bar{\alpha}\mathcal{I}_G + n\mathcal{T}_G)^{-1}$ is the idealized "survival" operator, representing the extent to which the original signal f^* is preserved. We then have $f^* - \mathcal{P}_G(\hat{f}_0) \approx \overline{\mathcal{B}} f^*$, where $\overline{\mathcal{B}} \coloneqq \mathcal{I}_G - \overline{\mathcal{S}} = \bar{\alpha}(\bar{\alpha}\mathcal{I}_G + n\mathcal{T}_G)^{-1}$ is the idealized "bias" operator. One can verify that

$$\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2} \lesssim \min\left\{\sqrt{\lambda_1}, \frac{\bar{\alpha}}{n\sqrt{\lambda_p}}, \sqrt{\frac{\bar{\alpha}}{n}}\right\}.$$

This bounds $\|\mathcal{P}_G(\hat{f}_0) - f^*\|_{L_2}$ for Theorem 1; the formal theorem has an extra factor of $1/(1-c)$ which comes from the approximation errors (recall that c is determined by how accurate our idealizing approximation are—see the statement of Theorem 1 for the precise definition).

Next, note that $\mathcal{P}_{G^\perp}(\hat{f}_0) \approx \mathcal{R}^*\mathcal{C}\frac{\overline{\mathcal{B}}}{\bar{\alpha}} f^*$, where we have substituted $\mathcal{C}$ for $\mathcal{A}_G$. Because

$$\|\mathcal{R}^*\|_{\ell_2\to L_2} \le \|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2}\|\mathcal{R}^*\|_{\ell_2\to\mathcal{H}} \lesssim \sqrt{\lambda_{p+1}\bar{\alpha}},$$

we have

$$\begin{aligned}\left\|\mathcal{R}^*\mathcal{C}\frac{\overline{\mathcal{B}}}{\bar{\alpha}}\right\|_{\mathcal{H}\to L_2} &\le \|\mathcal{R}^*\|_{\ell_2\to L_2}\|\mathcal{C}\|_{L_2\to\ell_2}\frac{\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2}}{\bar{\alpha}} \\ &\lesssim \sqrt{\frac{n\lambda_{p+1}}{\bar{\alpha}}}\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2},\end{aligned}$$

which allows us to bound $\|\mathcal{P}_{G^\perp}(\hat{f}_0)\|_{L_2}$.

2.4.2 Variance Term (from Noise)

Making similar approximations as above (with $\tilde{\alpha}$ instead of $\bar{\alpha}$ – the distinction comes from the approximation arguments we use in the formal proof), we have

$$\begin{aligned}\epsilon &= \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}*)^{-1}\xi \\ &\approx \begin{bmatrix}\mathcal{A}_G^* \\ \mathcal{R}^*\end{bmatrix}(\tilde{\alpha}I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\xi \\ &= \begin{bmatrix}(\tilde{\alpha}\mathcal{I}_G + \mathcal{A}_G^*\mathcal{A}_G)^{-1}\mathcal{A}_G^*\xi \\ \mathcal{R}^*(\tilde{\alpha}I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\xi\end{bmatrix} \\ &\approx \begin{bmatrix}(\tilde{\alpha}\mathcal{T}_G^{-1} + n\mathcal{I}_G)^{-1}\mathcal{C}^*\xi \\ \mathcal{R}^*(\tilde{\alpha}I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\xi\end{bmatrix}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbf{E}_\xi\|\epsilon\|_{L_2}^2 &\approx \sigma^2\left\|\begin{bmatrix}(\tilde{\alpha}\mathcal{T}_G^{-1} + n\mathcal{I}_G)^{-1}\mathcal{C}^* \\ \mathcal{R}^*(\tilde{\alpha}I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\end{bmatrix}\right\|_{HS,\ell_2\to L_2}^2 \\ &\lesssim \sigma^2\left(\frac{1}{n^2}\operatorname{tr}_{L_2}(\mathcal{C}^*\mathcal{C}) + \frac{1}{\tilde{\alpha}^2}\operatorname{tr}_{L_2}(\mathcal{R}^*\mathcal{R})\right) \\ &\approx \sigma^2\left(\frac{1}{n^2}\operatorname{tr}_{L_2}(n\mathcal{I}_G) + \frac{1}{\tilde{\alpha}^2}\operatorname{tr}_{L_2}(\mathcal{R}^*\mathcal{R})\right) \\ &= \sigma^2\left(\frac{p}{n} + \frac{1}{\tilde{\alpha}^2}\operatorname{tr}_{L_2}(\mathcal{R}^*\mathcal{R})\right).\end{aligned}$$

The factor of α_U/α_L comes from the approximation arguments.

3 KERNEL CLASSIFICATION

We now consider the case of classification, in which the observation y is a (noisy) binary label in $\{-1, 1\}$ with a distribution depending on x. Our approach is to perform ordinary linear/kernel regression on the binary labels y_i with the squared loss function. Although this seems counter-intuitive, recent results (e.g., Hui and Belkin, 2021) have shown that training with the squared-loss is highly competitive with the more common cross-entropy loss function in practice. Separately, recent results have also shown that regression on binary labels is, in some interesting overparametrized cases, equivalent to the maximum-margin SVM (e.g., Hsu et al., 2021; Muthukumar et al., 2021). Inspired by these findings, we study the minimum-ℓ_2-norm interpolator of the binary labels $\{y_i\}_{i=1}^n$ and its ensuing *classification* error.

Through the lens of regression, our target function f^* is now replaced by

$$\eta^*(x) \coloneqq \mathbf{E}(y \mid x) = 2\,\mathbf{P}(y = 1 \mid x) - 1.$$

The label noise is $\xi = y - \eta^*(x)$. Note that $\mathbf{E}[\xi|x] = 0$ by definition, and $\operatorname{var}(\xi \mid x) = 1 - (\eta^*)^2(x)$. Our assumption on the label noise model is that $\eta^* \in G$.

The regression procedure yields an estimator $\hat{\eta}$ of η^*. Then, our classification rule is given by $\hat{y} = \operatorname{sign}(\hat{\eta})$. Given a probability distribution μ over x, the *excess risk* of the classification rule with respect to the Bayes-optimal classifier is given by

$$\mathcal{E} \coloneqq \mathbf{P}(\hat{y} \ne y) - \mathbf{P}(y \ne \operatorname{sign}(\eta^*)).$$

Standard calculations (see Friedman, 1997) give

$$\mathcal{E} = \int |\eta^*(x)|\mathbf{1}_{\{\operatorname{sign}(\hat{\eta}(x))\ne\operatorname{sign}(\eta^*(x))\}}d\mu(x).$$

Thus the excess risk is the average of the sign error of $\hat{\eta}$ versus η^* modulated by how distinguishable the two classes are (which is represented by $|\eta^*|$).

To bound $\mathcal{E}$, we decompose our estimate $\hat{\eta}$ as

$$\hat{\eta} = s\eta^* + \hat{\eta}_r, \tag{2}$$

where s is a parameter that we can tune in our analysis, and $\hat{\eta}_r$ is the residual. If $s > 0$, we have

$$\{\operatorname{sign}(\hat{\eta}) \ne \operatorname{sign}(\eta^*)\} \subseteq \{|\hat{\eta}_r| \ge s|\eta^*|\},$$

so

$$\mathcal{E} \le \frac{1}{s}\int |\hat{\eta}_\perp(x)|\, d\mu(x) = \frac{\|\hat{\eta}_r\|_{L_1}}{s} \le \frac{\|\hat{\eta}_r\|_{L_2}}{s}, \tag{3}$$

where the norm inequality is due to the fact that μ is a probability measure. For reasons that will shortly

become clear, we will call s the *survival factor* and $\hat{\eta}_r$ the *residual*.

A first possible choice for the quantities in (2) is $s = 1$ and $\hat{\eta}_r = \hat{\eta} - \eta^*$. This choice yields $\mathcal{E} \le \|\hat{\eta} - \eta\|_{L_1}$; therefore, small regression error implies small excess classification risk. However, we are interested in cases in which the regression error is not small but the classification error is. To use the bound (3), we would need to show that we can have the ratio $\|\hat{\eta}_r\|_{L_2}/s$ be very small with a different choice of $s \ll 1$.

We now show how this can work. We recall the idealized "survival" and "bias" operators $\overline{\mathcal{S}}$ and $\overline{\mathcal{B}}$ from Section 2.4. Note that to bound the regression error we show that $\hat{\eta} \approx \overline{\mathcal{S}}(\eta^*)$ and that $\|\eta^* - \overline{\mathcal{S}}\eta^*\|_{L_2} = \|\overline{\mathcal{B}}\eta^*\|_{L_2}$ is small. For the classification problem, an interesting new possibility arises. As a simple example, suppose all the first p eigenvalues $\lambda_1, \dots, \lambda_p$ are identically 1. Then $\overline{\mathcal{S}} = \frac{n}{\bar{\alpha}+n}\mathcal{I}_G$. If $\bar{\alpha} \ll n$, then the ideal bias $\overline{\mathcal{B}} = \frac{\bar{\alpha}}{\bar{\alpha}+n}\mathcal{I}_G$ will be small. However, what if $\alpha \gtrsim n$, in which case the bias is not small? We cannot get small regression error, but for classification, we can apply (3) while choosing $s = \frac{n}{\bar{\alpha}+n}$. Then, as long as

$$\|\hat{\eta} - \overline{\mathcal{S}}\eta^*\|_{L_2} \ll \frac{n}{\bar{\alpha}+n},$$

we will have small excess classification risk. In Section 3.1, we use this observation to provide sufficient conditions for classification consistency, and demonstrate that these conditions are significantly weaker than the ones needed to be regression-consistent. This apporach is qualitatively similar to that of Muthukumar et al., 2021, which provides a slightly sharper bound but relies on a special form of η^* and Gaussianity of the features. Their techniques do not easily extend to a more general setting.

To combine this framework with our previous interpolation results, note that, under our new notation, $\hat{\eta} = \hat{\eta}_0 + \epsilon$, where $\hat{\eta}_0 = \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}\eta^*$ and $\epsilon = \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\xi$. We will show that $\hat{\eta}_0 \approx \overline{\mathcal{S}}\eta^*$ and ϵ is small. For the first objective, we present here a more refined version of Theorem 1 that bounds the error to $\overline{\mathcal{S}}\eta^*$ rather than η^* itself. This will be used to characterize the classification error in Section 3.1.

Lemma 6 (More refined bias estimate)**.** *Under the conditions of Theorem 1 (assuming c is bounded away from 1 so that $(1-c)^{-1}$ is subsumed into the constants),*

$$\|\hat{\eta}_0 - \overline{\mathcal{S}}\eta^*\|_{L_2} \lesssim \left(c + \sqrt{\frac{n\lambda_{p+1}}{\bar{\alpha}}}\right) \times \min\left\{\lambda_1, \frac{\bar{\alpha}}{n\sqrt{\lambda_p}}, \sqrt{\frac{\bar{\alpha}}{n}}\right\}\|\eta^*\|_{\mathcal{H}}.$$

The proof of Lemma 6 is an easy modification of the proof of Theorem 1 (see Appendix C.1).

3.1 Bi-level Ensemble Asymptotic Analysis

We now examine the implications of this refined classification analysis in a bounded orthonormal system (BOS). In particular, suppose that the eigenfunctions are all bounded (e.g., a Fourier series for periodic functions on an interval). For the eigenvalues, we consider the bi-level ensemble as defined in Muthukumar et al., 2021 with non-negative parameters n, β, q, r (where $\beta > 1$ and $r < 1$). This ensemble contains $d = n^\beta$ features, of which $p = n^r$ have "large" eigenvalues, and the remaining $d - p$ eigenvalues are small and their relative magnitude depends on the parameter q. Specifically, we set

$$\lambda_\ell = \begin{cases} 1, & 1 \le \ell \le p \\ n^{-(\beta-r-q)}, & p < \ell \le d. \end{cases} \tag{4}$$

We require $q < \beta - r$ to ensure that the "small" eigenvalues are actually smaller than 1.

Corollary 1. *Consider the bi-level ensemble with parameters n, β, q, r, and suppose that the eigenfunctions are all bounded by an absolute constant. Further, suppose that $\beta > 2$ and $r < 1$, and $\eta^* \in G$. Then we obtain the following asymptotic results as $n \to \infty$:*

1. *If $q < 1 - r$, as $n \to \infty$, $\|\hat{\eta} - \eta^*\|_{L_2} \to 0$ in probability, and therefore both regression and classification are consistent.*
2. *If $q > 1 - r$, $\|\hat{\eta}\|_{L_2} \to 0$ in probability, and therefore regression is inconsistent for nonzero η^*.*
3. *If $q < \frac{3}{2}(1-r)$ and $\beta > 2(r+q)$, excess classification risk $\mathcal{E} \to 0$ in probability, that is, classification is consistent.*

Corollary 1 is proved in Appendix F. Note that we have an asymptotic separation between classification and regression when $1 - r < q < \frac{3}{2}(1-r)$. This is comparable to the results of Muthukumar et al., 2021, which allow slightly larger q and smaller β but require much stronger feature assumptions.

We use the bi-level ensemble model in (4) for simplicity; however, we can obtain non-asymptotic bounds on the classification risk under more general assumptions. For a fixed value of $p := n^r$, our analysis allows the tail eigenvalues corresponding to indices $p < \ell \le d$ to be non-uniform. The requirement that the top p eigenvalues are the same is somewhat more stringent; in general, when the eigenvalues are different, the survival operator $\overline{\mathcal{S}}$ is not a multiple of the identity. This

could lead to qualitatively different behavior, as now $\hat{\eta}$ may be distorted from η^* due to differences in the eigenvalues of $\mathcal{T}_G$. This problem disappears in the case that either (a) η^* is proportional to a single eigenfunction or (b) the first p eigenvalues of $\mathcal{T}$ are identical (both of which hold in Muthukumar et al., 2021). We analyze further the extent to which we can bound the classification risk when *neither* of these assumptions holds in Appendix G. Our analysis method requires λ_p to be close to λ_1 to obtain significant gains for classification over regression.

4 NUMERICAL EXPERIMENTS

We now perform numerical experiments to demonstrate how the parameters β, r, and q of the bi-level ensemble model affect regression and classification performance. We consider the case of Fourier features $v_\ell(x) = e^{\mathbf{j}2\pi\ell x}$ for $\ell = -d, \dots, d$ over $x \in [0,1]$ with the uniform sampling measure, and the bi-level ensemble as defined in (4). The corresponding kernel function is

$$\begin{aligned} k(x,y) &= \sum_{\ell=-d}^{d} \lambda_\ell v_\ell(x)\overline{v_\ell(y)} \\ &= (1 - n^{-(\beta-r-q)})D_p(x-y) \\ &\quad + n^{-(\beta-r-q)}D_d(x-y), \end{aligned}$$

where $D_m(t) = \frac{\sin[(2m+1)\pi t]}{\sin(\pi t)}$ is the Dirichlet sinc function. We consider three cases for the bi-level ensemble parameters: $(\beta, r, q) = (2.6, 0.3, 0.3)$, $(2.6, 1/3, 5/6)$, and $(2.6, 0.8, 0.45)$. We sweep over several values of n between 10 and 3162. For each n, we generate an $\eta^* \in \text{span}\{v_\ell\}_{\ell=-p}^{p}$, scaled such that $\max_{x\in[0,1]} |\eta^*(x)| = 1$.

We first attempt to reconstruct $\eta^*(x)$ from noisy samples $y_i^{\text{reg}} = \eta^*(x_i) + \xi_i$ for $i = 1, \dots, n$ where ξ_i are i.i.d. $\mathcal{N}(0,1)$. We use the kernel ridge regression estimator $\widehat{\eta^*}^{\text{reg}} = \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}^*)^{-1} y^{\text{reg}}$ with a regularization parameter of $\alpha = 10^{-3}$. We then measure the relative L_2 error of the estimate, i.e., $\mathcal{E}^{\text{reg}} = \|\eta^* - \widehat{\eta}^{\text{reg}}\|_{L_2}^2 / \|\eta^*\|_{L_2}^2$.

We also attempt to reconstruct $\eta^*(x)$ from binary observations $y_i^{\text{class}} = +1$ with probability $(1 + \eta^*(x_i))/2$ and -1 with probability $(1 - \eta^*(x_i))/2$ for $i = 1, \dots, n$. We use the estimator $\widehat{\eta}^{\text{class}} = \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}^*)^{-1} y^{\text{class}}$ with a regularization parameter of $\alpha = 10^{-3}$. We then measure the relative excess risk, i.e.,

$$\mathcal{E}^{\text{class}} = \frac{\int |\eta^*(x)| \mathbf{1}_{\{\text{sign}(\hat{\eta}^{\text{class}}(x)) \neq \text{sign}(\eta^*(x))\}}\, dx}{\int |\eta^*(x)|\, dx}.$$

The above procedure is repeated over 100 trials. In Figure 2, we plot the relative L_2 error (averaged over 100 trials) versus n and the relative excess risk (averaged over 100 trials) versus n for each of the three sets of values for β, r, q. In the first case where $\beta = 2.6$, $r = 0.3$, and $q = 0.3$, we have $r + q < 1$ and both $\mathcal{E}^{\text{reg}}$ and $\mathcal{E}^{\text{class}}$ decrease as n increases. In the second case where $\beta = 2.6$, $r = 1/3$, and $q = 5/6$, we have $1 - r < q < \frac{3}{2}(1-r)$ and $\beta > 2r + 2q$, and $\mathcal{E}^{\text{class}}$ decreases as n increases, but $\mathcal{E}^{\text{reg}}$ does not decrease as n increases. In the third case where $\beta = 2.6$, $r = 0.8$, and $q = 0.45$, we have that $r + q > 1$, and $q > \frac{3}{2}(1-r)$, and both $\mathcal{E}^{\text{reg}}$ and $\mathcal{E}^{\text{class}}$ do not decrease as n increases.

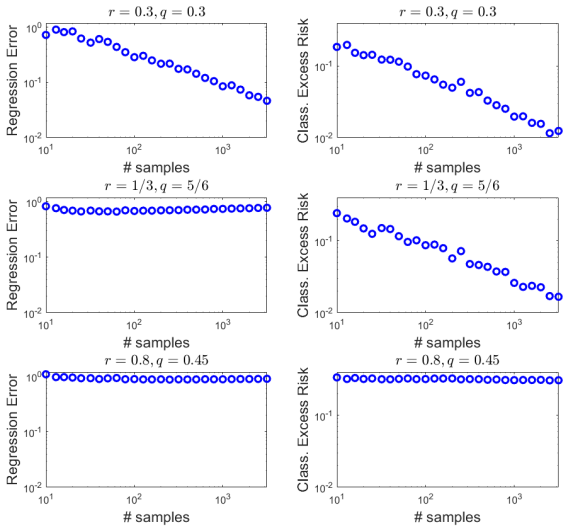


Figure 2: Relative L_2 errors versus n and the relative classification excess risks versus n for each of the three sets of bi-level ensemble parameters (averaged over 100 trials).

5 DISCUSSION

In this paper we showed that under minimal assumptions on the data and feature map (a) harmless interpolation of noise in data is possible, and (b) we can be classification-consistent in high-dimensional regimes where we are not regression-consistent. Important future directions include considering more general function models (e.g., any $f^* \in \mathcal{H}$ or even $f^* \notin \mathcal{H}$), better understanding the implications of distortion among the top eigenfunctions in classification error, and improving the non-asymptotic rates for classification risk from Section 3.1. Another intriguing question is whether there is an equivalence between interpolating binary labels and the max-margin SVM (as shown in Hsu et al., 2021; Muthukumar et al., 2021) in the more general settings considered in this paper. Finally, it would be very interesting to study whether our upper bounds (particularly for classification) can be matched by non-asymptotic lower bounds.

Acknowledgments

We thank Fanny Yang for useful discussion. This work was supported, in part, by National Science Foundation grant CCF-2107455.

References

Adlam, B., & Pennington, J. (2020). Understanding double descent requires a fine-grained bias-variance decomposition. *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*.

Audibert, J.-Y., & Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Stat.*, *35*(2), 608–633.

Ba, J., Erdogdu, M., Suzuki, T., Wu, D., & Zhang, T. (2020). Generalization of two-layer neural networks: An asymptotic viewpoint. *Proc. Int. Conf. Learn. Representations (ICLR)*.

Bartlett, P. L., Long, P. M., Lugosi, G., & Tsigler, A. (2020). Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. U.S.A.*, *117*(48), 30063–30070.

Bartlett, P. L., & Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, *3*, 463–482.

Bartlett, P. L., Montanari, A., & Rakhlin, A. (2021). Deep learning: A statistical viewpoint. *Acta Numer.*, *30*, 87–201.

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.*, *116*(32), 15849–15854.

Belkin, M. (2018). Approximation beats concentration? an approximation view on inference with smooth radial kernels. *Proc. Conf. Learn. Theory (COLT)*.

Belkin, M. (2021). Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numer.*, *30*, 203–248.

Belkin, M., Hsu, D., & Xu, J. (2020). Two models of double descent for weak features. *SIAM J. Math. Data Sci.*, *2*(4), 1167–1180.

Belkin, M., Ma, S., & Mandal, S. (2018). To understand deep learning we need to understand kernel learning. *Proc. Int. Conf. Mach. Learn. (ICML)*, *35*.

Cao, Y., Gu, Q., & Belkin, M. (2021). Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures.

Caponnetto, A., & De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, *7*(3), 331–368.

Chatterji, N. S., & Long, P. M. (2021). Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *J. Mach. Learn. Res.*, *22*.

Chinot, G., & Lerasle, M. (2020). On the robustness of the minimum ℓ_2 interpolator.

Dar, Y., Muthukumar, V., & Baraniuk, R. G. (2021). A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning.

D'Ascoli, S., Refinetti, M., Biroli, G., & Krzakala, F. (2020). Double trouble in double descent: Bias and variance(s) in the lazy regime. *Proc. Int. Conf. Mach. Learn. (ICML)*.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer.

Dhifallah, O., & Lu, Y. M. (2020). A precise performance analysis of learning with random features.

Donhauser, K., Wu, M., & Yang, F. (2021). How rotational invariance of common kernels prevents generalization in high dimensions. *Proc. Int. Conf. Mach. Learn. (ICML)*, 2804–2814.

Friedman, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Min. Knowl. Discov.*, *1*, 55–77.

Gerace, F., Loureiro, B., Krzakala, F., Mezard, M., & Zdeborova, L. (2020). Generalisation error in learning with random features and the hidden manifold model. *Proc. Int. Conf. Mach. Learn. (ICML)*.

Ghorbani, B., Mei, S., Misiakiewicz, T., & Montanari, A. (2021). Linearized two-layers neural networks in high dimension. *Ann. Stat.*, *49*(2), 1029–1054.

Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation.

Horn, R. A., & Johnson, C. R. (1985). *Matrix analysis*. Cambridge.

Hsu, D., Kakade, S. M., & Zhang, T. (2014). Random design analysis of ridge regression. *Found. Comput. Math.*, *14*, 569–600.

Hsu, D., Muthukumar, V., & Xu, J. (2021). On the proliferation of support vectors in high dimensions. *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*.

Hu, H., & Lu, Y. M. (2020). Universality laws for high-dimensional learning with random features.

Hui, L., & Belkin, M. (2021). Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *Proc. Int. Conf. Learn. Representations (ICLR)*.

Koltchinskii, V., & Beznosova, O. (2005). Exponential convergence rates in classification. *Proc. Conf. Learn. Theory (COLT)*, 295–307.

Li, Z., Zhou, Z.-H., & Gretton, A. (2021). Towards an understanding of benign overfitting in neural networks.

Liang, T., & Rakhlin, A. (2020). Just interpolate: Kernel "ridgeless" regression can generalize. *Ann. Stat.*, *48*(3), 1329–1347.

Liang, T., Rakhlin, A., & Zhai, X. (2020). On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. *Proc. Conf. Learn. Theory (COLT)*, 2683–2711.

Liao, Z., Couillet, R., & Mahoney, M. W. (2020). A random matrix analysis of random Fourier features: Beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent. *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*.

Lin, L., & Dobriban, E. (2021). What causes the test error? going beyond bias-variance via ANOVA. *J. Mach. Learn. Res.*, *22*.

McRae, A. D., Romberg, J., & Davenport, M. A. (2020). Sample complexity and effective dimension for regression on manifolds. *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*.

Mei, S., Misiakiewicz, T., & Montanari, A. (2021). Generalization error of random features and kernel methods: Hypercontractivity and kernel matrix concentration.

Mei, S., & Montanari, A. (2021). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Commun. Pure Appl. Math.*

Muthukumar, V., Narang, A., Subramanian, V., Belkin, M., Hsu, D., & Sahai, A. (2021). Classification vs. regression in overparameterized regimes: Does the loss function matter? *J. Mach. Learn. Res.*

Muthukumar, V., Vodrahalli, K., Subramanian, V., & Sahai, A. (2020). Harmless interpolation of noisy data in regression. *IEEE J. Sel. Areas Inf. Theory*, *1*(1), 67–83.

Rakhlin, A., & Zhai, X. (2019). Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. *Proc. Conf. Learn. Theory (COLT)*, 2595–2623.

Rudelson, M., & Vershynin, R. (2013). Hanson-Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.*, *18*.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

Steinwart, I., Hush, D., & Scovel, C. (2009). Optimal rates for regularized least squares regression. *Proc. Conf. Learn. Theory (COLT)*.

Steinwart, I., & Scovel, C. (2012). Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constr. Approx.*, *35*, 363–417.

Tropp, J. (2015). An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, *8*(1-2), 1–230.

Tsigler, A., & Bartlett, P. L. (2020). Benign overfitting in ridge regression.

Vapnik, V. N. (2000). *The nature of statistical learning theory*. Springer.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Cambridge.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge.

Wang, K., & Thrampoulidis, C. (2021). Benign overfitting in binary classification of gaussian mixtures. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*.

Wendland, H. (2004). *Scattered data approximation*. Cambridge.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. *Proc. Int. Conf. Learn. Representations (ICLR)*.

Zhang, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.*, *17*, 2077–2098.

Supplementary Material: Harmless interpolation in regression and classification with structured features

A NOTATION

For convenience in reading, we collect all notation that is used for the proofs in Table 1.

In addition, we will use many different norms. For a function $f\colon X \to \mathbf{R}$, $\|f\|_{L_p} \coloneqq \left((\mathbf{E}_{x\sim\mu}|f(x)|^p)^{1/p}\right)$. For $f \in \mathcal{H}$, $\|f\|_{\mathcal{H}} = \|\mathcal{T}^{-1/2}f\|_{L_2}$ is the RKHS norm. For $u \in \mathbf{R}^n$, $\|u\|_{\ell_2}$ is the standard Euclidean norm. We denote the L_2, $\mathcal{H}$, and ℓ_2 inner products by $\langle\cdot,\cdot\rangle_{L_2}$, $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, and $\langle\cdot,\cdot\rangle_{\ell_2}$, respectively.

$\|\cdot\|_{L_2}$, $\|\cdot\|_{\mathcal{H}}$, and $\|\cdot\|_{\ell_2}$ also denote operator norms when applied to operators from the corresponding Hilbert space to itself. We will write the operator norm of an operator $T\colon H_1 \to H_2$ (for any Hilbert spaces H_1 and H_2) with respect to the H_1 and H_2 norms as $\|T\|_{H_1\to H_2}$. Similarly, $\|T\|_{HS,H_1\to H_2}$ refers to the Hilbert-Schmidt norm of T with respect to the H_1 and H_2 inner products.

Table 1: Notation

Symbol(s)	Definition(s)	Description
k_x	$k_x = k(\cdot, x)$	Kernel function centered at x
$\mathcal{T}$	$\mathcal{T}(f) = \int f(x)k_x \, d\mu(x)$	Integral operator of kernel k
$\{(\lambda_\ell, v_\ell)\}_{\ell=1}^{\infty}$	$\mathcal{T}(f) = \sum_{\ell=1}^{\infty} \lambda_\ell \langle f, v_\ell\rangle_{L_2} v_\ell, \ \lambda_1 \geq \lambda_2 \geq \cdots$	Eigenvalue decomposition of $\mathcal{T}$
$\mathcal{A}$	$\mathcal{A}(f) = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$	Sampling operator from $\mathcal{H}$ to $\mathbf{R}^n$
$\mathcal{A}^*$	$\mathcal{A}^*(z) = \sum_{i=1}^n z_i k_{x_i}$	Adjoint of $\mathcal{A}$ w.r.t $\mathcal{H}$ and ℓ_2 inner products
G, $G^\perp$	$G = \operatorname{span}\{v_1, \ldots, v_p\}$	Span of first p eigenfunctions of $\mathcal{T}$ (and its complement)
$\mathcal{I}$ ($\mathcal{I}_G$)		Identity operator (restricted to G)
$\mathcal{T}_G, \mathcal{T}_{G^\perp}$	$\mathcal{T}_G = \mathcal{T}\mathcal{P}_G$, $\mathcal{T}_{G^\perp} = \mathcal{T}\mathcal{P}_{G^\perp}$	$\mathcal{T}$ restricted to G and $G^\perp$
$\mathcal{A}_G$, $\mathcal{R}$	$\mathcal{A}_G = \mathcal{A}\mathcal{P}_G, \mathcal{R} = \mathcal{A}\mathcal{P}_{G^\perp}$	Restrictions of sampling operator to G, $G^\perp$
$\mathcal{C}, \mathcal{C}^*$	$\mathcal{C} = \mathcal{A}_G$, $\mathcal{C}^* = \mathcal{T}_G^{-1}\mathcal{A}_G^*$	Sampling operator and its adjoint on G w.r.t. L_2 inner product on G
α		Explicit regularization parameter
α_L, α_U	$\alpha_L I_n \preceq \alpha I_n + \mathcal{R}\mathcal{R}^* \preceq \alpha_U I_n$	Lower and upper bounds on explicit+implicit regularization
$\bar{\alpha}, \tilde{\alpha}$	$\bar{\alpha} = \frac{2\alpha_U\alpha_L}{\alpha_U+\alpha_L}, \tilde{\alpha} = \frac{\alpha_U+\alpha_L}{2}$	Harmonic and arithmetic means of α_U, α_L
$\mathcal{B}$	$\mathcal{B} = (\mathcal{I}_G + \mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}$	Bias operator on G
$\mathcal{S}$	$\mathcal{S} = \mathcal{I}_G - \mathcal{B}$	Kernel regression operator ("survival") on G
$\overline{\mathcal{B}}$	$\overline{\mathcal{B}} = \left(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\right)^{-1}$	Idealized approximation to bias $\mathcal{B}$
$\overline{\mathcal{S}}$	$\overline{\mathcal{S}} = \mathcal{I}_G - \overline{\mathcal{B}} = \frac{n}{\bar{\alpha}}\mathcal{T}_G\left(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\right)^{-1}$	Idealized approximation to survival $\mathcal{S}$

B DETAILED INTERPOLATION LITERATURE SURVEY

Recent work has shown that the "harmless interpolation" phenomenon becomes more pronounced with increased (effective) overparameterization when the minimum-Hilbert-norm interpolator is used in kernel regression (Liang & Rakhlin, 2020; Liang et al., 2020) or the minimum-norm interpolator is used in linear regression (Adlam & Pennington, 2020; Ba et al., 2020; Bartlett et al., 2020; Belkin et al., 2020; D'Ascoli et al., 2020; Dhifallah & Lu, 2020; Gerace et al., 2020; Hastie et al., 2019; Hu & Lu, 2020; Li et al., 2021; Liao et al., 2020; Lin & Dobriban, 2021; Mei et al., 2021; Mei & Montanari, 2021; Muthukumar et al., 2020; Tsigler & Bartlett, 2020) in a variety of models. See Bartlett et al., 2021; Belkin, 2021; Dar et al., 2021 for recent surveys of this line of work.

All of these models make at least one of the following assumptions: (a) independence of features (Bartlett et al., 2020; Chinot & Lerasle, 2020; Hastie et al., 2019; Muthukumar et al., 2020), (b) sub-Gaussianity in the feature vector (Bartlett et al., 2020; Tsigler & Bartlett, 2020), (c) high data dimension (Adlam & Pennington, 2020; Ba et al., 2020; D'Ascoli et al., 2020; Dhifallah & Lu, 2020; Gerace et al., 2020; Hastie et al., 2019; Hu & Lu, 2020; Li et al., 2021; Liang & Rakhlin, 2020; Liang et al., 2020; Liao et al., 2020; Mei et al., 2021; Mei & Montanari, 2021), or (d) explicit structure in the kernel operator/feature map (Adlam & Pennington, 2020; Ba et al., 2020; Belkin et al., 2020; D'Ascoli et al., 2020; Dhifallah & Lu, 2020; Gerace et al., 2020; Hu & Lu, 2020; Li et al., 2021; Liang & Rakhlin, 2020; Liang et al., 2020; Liao et al., 2020; Lin & Dobriban, 2021; Mei et al., 2021; Mei & Montanari, 2021; Muthukumar et al., 2020). For specific kernels like the Laplace kernel, statistically consistent interpolation may actually *require* growing data dimension with the number of training examples (Rakhlin & Zhai, 2019), as the data dimension fundamentally alters the eigenvalues of the Laplace kernel integral operator. In contrast, our results do not explicitly posit any of these assumptions. Our sufficient conditions for harmless interpolation are expressed purely in terms of the eigenvalues of the kernel integral operator and do not require special structure either on the eigenfunctions or the integral operator itself.

C PROOFS OF DETERMINISTIC-SAMPLE RESULTS

We begin with the proofs of the deterministic-sample results (Theorems 1 and 2).

In this section, we will often abbreviate scaled identity operators such as aI_n, $a\mathcal{I}$, $a\mathcal{I}_G$, $a\mathcal{I}_G{}^{\perp}$ by the number a. The meaning should be clear from context.

C.1 Bias

The main technical challenge for proving Theorem 1 is bounding the approximation error between the "ideal" bias operator $\overline{\mathcal{B}} = \left(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\right)^{-1}$ (discussed in Section 2.4) and the actual bias, which turns out to be $\mathcal{B} \coloneqq (\mathcal{I}_G + \mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}$ (derived in the proof of Theorem 1 below). The following result quantifies this error.

Lemma 7. *Under the conditions of Theorem 1,*

$$\|\mathcal{B} - \overline{\mathcal{B}}\|_{\mathcal{H}\to L_2} \le \frac{c}{1-c}\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2},$$

where $c < 1$ is an upper bound on the quantity $\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2}$ (as defined in Theorem 1).

Proof. Recall that we have assumed

$$\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \le c < 1.$$

A standard perturbation argument (e.g., Horn and Johnson, 1985, p. 335) gives

$$\begin{aligned}\mathcal{B} - \overline{\mathcal{B}} &= \sum_{i=1}^{\infty}(-1)^i\left[\overline{\mathcal{B}}\left(\mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G - \frac{n}{\bar{\alpha}}\mathcal{T}_G\right)\right]^k\overline{\mathcal{B}} \\ &= \left(\sum_{i=1}^{\infty}(-1)^i\left[\left(\mathcal{T}_G^{-1} + \frac{n}{\bar{\alpha}}\mathcal{I}_G\right)^{-1}\left(\mathcal{C}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{C} - \frac{n}{\bar{\alpha}}\mathcal{I}_G\right)\right]^k\right)\overline{\mathcal{B}}\end{aligned}$$

as long as the operator norm (in any space) of the bracketed operator $\left(\mathcal{T}_G^{-1} + \frac{n}{\bar{\alpha}}\mathcal{I}_G\right)^{-1}\left(\mathcal{C}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{C} - \frac{n}{\bar{\alpha}}\mathcal{I}_G\right)$ is strictly less than 1.

We now show that this is the case. We have

$$\begin{aligned}
\left\|\mathcal{C}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{C} - \frac{n}{\bar{\alpha}}\mathcal{I}_G\right\|_{L_2} &\leq \left\|\mathcal{C}^*\left((\alpha + \mathcal{R}\mathcal{R}^*)^{-1} - \frac{1}{\bar{\alpha}}\right)\mathcal{C}\right\|_{L_2} + \frac{1}{\bar{\alpha}}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \\
&\leq \|\mathcal{C}\|_{L_2\to\ell_2}^2 \max\left\{\left|\frac{1}{\alpha_L} - \frac{1}{\bar{\alpha}}\right|, \left|\frac{1}{\alpha_U} - \frac{1}{\bar{\alpha}}\right|\right\} + \frac{1}{\bar{\alpha}}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \\
&\leq \frac{\alpha_U - \alpha_L}{2\alpha_U\alpha_L}\left(n + \|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2}\right) + \frac{1}{\bar{\alpha}}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2},
\end{aligned}$$

where the first and third inequalities use the triangle inequality, and the second inequality uses $\alpha_L \preceq \alpha + \mathcal{R}\mathcal{R}^* \preceq \alpha_U$. Then, since

$$\left\|\left(\mathcal{T}_G^{-1} + \frac{n}{\bar{\alpha}}\mathcal{I}_G\right)^{-1}\right\|_{L_2} \leq \frac{\bar{\alpha}}{n},$$

we have

$$\begin{aligned}
&\left\|\left(\mathcal{T}_G^{-1} + \frac{n}{\bar{\alpha}}\mathcal{I}_G\right)^{-1}\left(\mathcal{C}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{C} - \frac{n}{\bar{\alpha}}\mathcal{I}_G\right)\right\|_{L_2} \\
&\quad\leq \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \left(1 + \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\right)\cdot\frac{1}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \\
&\quad\leq \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \\
&\quad\leq c.
\end{aligned}$$

Since $c < 1$, the rest of the bound follows via the expression for the infinite sum of a geometric series. □

We are now ready to prove our main deterministic bias result (Theorem 1).

Proof of Theorem 1. Since $f^* \in G$, the full expression for the noiseless regression estimate is

$$\hat{f}_0 = \begin{bmatrix}\mathcal{A}_G^* \\ \mathcal{R}^*\end{bmatrix}(\alpha + \mathcal{A}_G\mathcal{A}_G^* + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G f^*.$$

The pushthrough identity gives

$$\begin{aligned}
(\alpha + \mathcal{A}_G\mathcal{A}_G^* + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G &= (\alpha + \mathcal{R}\mathcal{R}^*)^{-1}(I_n + \mathcal{A}_G\mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1})^{-1}\mathcal{A}_G \\
&= (\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G(\mathcal{I}_G + \mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}.
\end{aligned}$$

This gives

$$\begin{aligned}
\mathcal{P}_G(\hat{f}_0) &= \mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G(\mathcal{I}_G + \mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}f^* \\
&= (\mathcal{I}_G - (\mathcal{I}_G + \mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1})f^*
\end{aligned}$$

and

$$\mathcal{P}_{G^\perp}(\hat{f}_0) = \mathcal{R}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G(\mathcal{I}_G + \mathcal{A}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}^*)^{-1}f^*.$$

We denote the actual bias and survival operators on G as

$$\begin{aligned}
&\mathcal{B} = (\mathcal{I}_G + \mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}, \text{ and} \\
&\mathcal{S} = \mathcal{I}_G - \mathcal{B}.
\end{aligned}$$

We then have

$$\mathcal{P}_G(\hat{f}_0) = \mathcal{S}f^*,$$

and

$$\mathcal{P}_{G^\perp}(\hat{f}_0) = \mathcal{R}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G\mathcal{B}f^*.$$

Clearly, $\|f^* - \mathcal{P}_G(\hat{f}_0)\|_{L_2} = \|\mathcal{B}f^*\|_{L_2} \le \|\mathcal{B}\|_{\mathcal{H}\to L_2}\|f^*\|_{\mathcal{H}}$. To bound $\|\mathcal{P}_{G^\perp}(\hat{f}_0)\|_{L_2}$, note that (recalling $\mathcal{C} = \mathcal{A}_G$)

$$\left\|\mathcal{R}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G\right\|_{L_2} \le \|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2} \cdot \|\mathcal{R}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\|_{\ell_2\to\mathcal{H}} \cdot \|\mathcal{C}\|_{L_2\to\ell_2}.$$

Note that $\|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2} = \|\mathcal{T}_{G^\perp}^{1/2}\|_{\mathcal{H}} = \sqrt{\lambda_{p+1}}$, and

$$\|\mathcal{C}\|_{L_2\to\ell_2}^2 = \|\mathcal{C}^*\mathcal{C}\|_{L_2} \approx \|n\mathcal{I}_G\|_{L_2} = n.$$

Furthermore, note that the singular values (from ℓ_2 to $\mathcal{H}$) of the operator $\mathcal{R}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}$ are

$$\frac{\sqrt{\lambda_k(\mathcal{R}\mathcal{R}^*)}}{\alpha + \lambda_k(\mathcal{R}\mathcal{R}^*)} \le \frac{1}{\sqrt{\alpha + \lambda_k(\mathcal{R}\mathcal{R}^*)}} \le \frac{1}{\sqrt{\alpha_L}},\ k = 1, \dots, n,$$

where $\lambda_k(S)$ denotes the kth eigenvalue of a symmetric matrix S. Therefore,

$$\left\|\mathcal{R}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G\right\|_{L_2} \lesssim \sqrt{\lambda_{p+1}} \cdot \frac{1}{\sqrt{\alpha_L}} \cdot \sqrt{n} = \sqrt{\frac{n\lambda_{p+1}}{\alpha_L}}.$$

Noting that $\bar{\alpha} \le 2\alpha_L$, we have

$$\|\hat{f}_0 - f^*\|_{L_2} \lesssim \left(1 + \sqrt{\frac{n\lambda_{p+1}}{\bar{\alpha}}}\right)\|\mathcal{B}\|_{\mathcal{H}\to L_2}\|f^*\|_{\mathcal{H}}.$$

Lemma 7 gives

$$\|\mathcal{B}\|_{\mathcal{H}\to L_2} \le \|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2} + \|\mathcal{B} - \overline{\mathcal{B}}\|_{\mathcal{H}\to L_2} \le \frac{1}{1-c}\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2}.$$

Also, one can also easily check that $\|B\|_{\mathcal{H}} \le 1$, and therefore $\|\mathcal{B}\|_{\mathcal{H}\to L_2} \le \sqrt{\lambda_1}$. Thus

$$\|\mathcal{B}\|_{\mathcal{H}\to L_2} \le \min\left\{\sqrt{\lambda_1}, \frac{1}{1-c}\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2}\right\}$$

Using the fact that $\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2} \lesssim \min\left\{\sqrt{\frac{\bar{\alpha}}{n}}, \frac{\bar{\alpha}}{n\sqrt{\lambda_p}}\right\}$ completes the proof.

□

With the proof of Theorem 1 complete, recall that we introduced a more refined expression for the estimation error due to bias in Lemma 6 for the purpose of bounding classification error. Note that the proof of Lemma 6 is a very simple modification of the preceding proof. The error in $G^\perp$ is bounded the same way. For the error in G, we bound the norm of $(\mathcal{S} - \overline{\mathcal{S}})f^* = (\overline{\mathcal{B}} - \mathcal{B})f^*$ instead of $f^* - \mathcal{S}f^* = \mathcal{B}f^*$, and therefore we replace $\|\mathcal{B}\|_{\mathcal{H}\to L_2}$ by $\|\overline{\mathcal{B}} - \mathcal{B}\|_{\mathcal{H}\to L_2}$ in the bound.

C.2 Variance

Recall that $\alpha_L \preceq \alpha + \mathcal{R}\mathcal{R}^* \preceq \alpha_U$ and $\tilde{\alpha} = \frac{\alpha_U + \alpha_L}{2}$. Also recall the formula

$$\epsilon = \mathcal{A}^*(\alpha + \mathcal{A}\mathcal{A}^*)^{-1}\xi.$$

To allow us to replace $\alpha + \mathcal{A}\mathcal{A}^*$ with $\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*$, we need the following result:

Lemma 8.

$$\|(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)(\alpha + \mathcal{A}\mathcal{A}^*)^{-1}\|_{\ell_2} \le \frac{1}{2}\left(\frac{\alpha_U}{\alpha_L} + 1\right).$$

Proof. Since $(\alpha + \mathcal{A}\mathcal{A}^*) - (\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*) = \alpha + \mathcal{R}\mathcal{R}^* - \tilde{\alpha}$, another perturbation expansion (see Appendix C.1) gives

$$(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1} - (\alpha + \mathcal{A}\mathcal{A}^*)^{-1} = \sum_{k=1}^{\infty}(-1)^{k+1}(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\left[(\alpha + \mathcal{R}\mathcal{R}^* - \tilde{\alpha})(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\right]^k,$$

which is valid since $\alpha_L \preceq \alpha + \mathcal{R}\mathcal{R}^* \preceq \alpha_U$ implies

$$\left\|(\alpha + \mathcal{R}\mathcal{R}^* - \tilde{\alpha})(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\right\|_{\ell_2} \le \frac{1}{\tilde{\alpha}} \cdot \frac{\alpha_U - \alpha_L}{2} = \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} < 1.$$

Then

$$I_n - (\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)(\alpha + \mathcal{A}\mathcal{A}^*)^{-1} = \sum_{k=1}^{\infty}(-1)^{k+1}\left[(\alpha + \mathcal{R}\mathcal{R}^* - \tilde{\alpha})(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\right]^k.$$

We apply the triangle inequality to get

$$\begin{aligned}
\|(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)(\alpha + \mathcal{A}\mathcal{A}^*)^{-1}\|_{\ell_2} &\le \|I_n\|_{\ell_2} + \sum_{k=1}^{\infty}\left\|(\alpha + \mathcal{R}\mathcal{R}^* - \tilde{\alpha})(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\right\|_{\ell_2}^k \\
&\le 1 + \sum_{i=1}^{\infty}\left(\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\right)^i \\
&= \frac{1}{2}\left(\frac{\alpha_U}{\alpha_L} + 1\right).
\end{aligned}$$

□

We can now prove the main "variance" error bound:

Proof of Theorem 2. Since $\mathrm{var}(\xi_i) \le \sigma^2$ for each i, we have

$$\begin{aligned}
\mathbf{E}_\xi\|\epsilon\|_{L_2}^2 &\le \sigma^2\|\mathcal{A}^*(\alpha + \mathcal{A}\mathcal{A}^*)^{-1}\|_{HS,\ell_2\to L_2}^2 \\
&= \sigma^2\|\mathcal{A}^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)(\alpha + \mathcal{A}\mathcal{A}^*)^{-1}\|_{HS,\ell_2\to L_2}^2 \\
&\le \frac{\sigma^2}{4}\left(\frac{\alpha_U}{\alpha_L} + 1\right)^2\|\mathcal{A}^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\|_{HS,\ell_2\to L_2}^2,
\end{aligned}$$

where the last inequality substitutes Lemma 8. Furthermore, we have

$$\begin{aligned}
\|\mathcal{A}^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\|_{HS,\ell_2\to L_2}^2 &= \|\mathcal{A}_G^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\|_{HS,\ell_2\to L_2}^2 + \|\mathcal{R}^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\|_{HS,\ell_2\to L_2}^2 \\
&\le \|(\tilde{\alpha} + \mathcal{A}_G^*\mathcal{A}_G)^{-1}\mathcal{A}_G^*\|_{HS,\ell_2\to L_2}^2 + \frac{\mathrm{tr}_{L_2}(\mathcal{R}^*\mathcal{R})}{\tilde{\alpha}^2} \\
&= \|(\tilde{\alpha}\mathcal{T}_G^{-1} + \mathcal{C}^*\mathcal{C})^{-1}\mathcal{C}^*\|_{HS,\ell_2\to L_2}^2 + \frac{\mathrm{tr}_{L_2}(\mathcal{R}^*\mathcal{R})}{\tilde{\alpha}^2} \\
&\lesssim \frac{p}{n} + \frac{\mathrm{tr}_{L_2}(\mathcal{R}^*\mathcal{R})}{\tilde{\alpha}^2},
\end{aligned}$$

where the last inequality is due to the fact that $\mathcal{C}$ is an $n \times p$-dimensional operator, all of whose singular values are close to $\sqrt{n}$.

Therefore,

$$\mathbf{E}_\xi\|\epsilon\|_{L_2}^2 \lesssim \sigma^2\left(\frac{\alpha_U}{\alpha_L} + 1\right)^2\left(\frac{p}{n} + \frac{\mathrm{tr}_{L_2}(\mathcal{R}^*\mathcal{R})}{\tilde{\alpha}^2}\right).$$

□

C.2.1 High-probability Noise Bounds

If the ξ_i's are sub-Gaussian, we could use the Hanson-Wright inequality for sub-Gaussian random vectors (see, e.g., Rudelson and Vershynin, 2013) to get a high-probability bound in Theorem 2,

Note that we can write

$$\|\epsilon\|_{L_2}^2 = \langle Z\xi, \xi\rangle_{\ell_2},$$

where

$$Z = (\alpha + \mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}\mathcal{T}\mathcal{A}^*(\alpha + \mathcal{A}\mathcal{A}^*)^{-1}.$$

We have already calculated an upper bound on the expectation of this quadratic form. To use the Hanson-Wright inequality to bound the upper tail, we need to bound both $\|Z\|_{\ell_2}$ and $\|Z\|_{\text{HS}}$ (where $\|Z\|_{\text{HS}}$ is the Hilbert-Schmidt norm with respect to the Euclidean inner product, also known as the Frobenius norm). By a similar argument as before, we have

$$\|Z\|_{\ell_2} \le \frac{1}{4}\left(\frac{\alpha_U}{\alpha_L} + 1\right)^2 \|\widetilde{Z}\|_{\ell_2},$$

and

$$\|Z\|_{\text{HS}} \le \frac{1}{4}\left(\frac{\alpha_U}{\alpha_L} + 1\right)^2 \|\widetilde{Z}\|_{\text{HS}},$$

where

$$\begin{aligned}\widetilde{Z} &= (\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}\mathcal{T}\mathcal{A}^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1} \\ &= \underbrace{(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}_G\mathcal{T}_G\mathcal{A}_G^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}}_{\widetilde{Z}_G} + \underbrace{(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{R}\mathcal{T}_{G^\perp}\mathcal{R}^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}}_{\widetilde{Z}_{G^\perp}}.\end{aligned}$$

Note that

$$\widetilde{Z}_G = \mathcal{A}_G(\tilde{\alpha} + \mathcal{A}_G^*\mathcal{A}_G)^{-1}\mathcal{T}_G(\tilde{\alpha} + \mathcal{A}_G^*\mathcal{A}_G)^{-1}\mathcal{A}_G^* = \mathcal{C}(\tilde{\alpha}\mathcal{T}_G^{-1} + \mathcal{C}^*\mathcal{C})^{-2}\mathcal{C}^*.$$

By a similar argument as before (in which we were effectively calculating the trace of $\widetilde{Z}_G$), we have $\|\widetilde{Z}_G\|_{\ell_2} \lesssim \frac{1}{n}$ and $\|\widetilde{Z}_G\|_{\text{HS}} \lesssim \frac{\sqrt{p}}{n}$.

Similarly, $\|\widetilde{Z}_{G^\perp}\|_{\ell_2} \le \frac{1}{\tilde{\alpha}^2}\|\mathcal{R}\mathcal{T}_{G^\perp}\mathcal{R}^*\|_{\ell_2}$, and $\|\widetilde{Z}_{G^\perp}\|_{\text{HS}} \le \frac{1}{\tilde{\alpha}^2}\|\mathcal{R}\mathcal{T}_{G^\perp}\mathcal{R}^*\|_{\text{HS}}$.

D PROOFS OF OPERATOR CONCENTRATION RESULTS

Proof of Lemma 1. Let $\text{diag}(Z)$ denote the projection of Z onto the space of diagonal matrices, and let $\text{diag}^\perp(Z)$ denote the orthogonal projection (i.e., onto the space of matrices with zero diagonal). Note that

$$\begin{aligned}\|\mathcal{R}\mathcal{R}^* - (\operatorname{tr}\mathcal{T}_{G^\perp})I_n\| &\le \|\text{diag}^\perp(\mathcal{R}\mathcal{R}^*)\| + \|\text{diag}(\mathcal{R}\mathcal{R}^*) - (\operatorname{tr}\mathcal{T}_{G^\perp})I_n\| \\ &\le \|\text{diag}^\perp(\mathcal{R}\mathcal{R}^*)\|_{\text{HS}} + \max_i\ |k^R(x_i, x_i) - \operatorname{tr}\mathcal{T}_{G^\perp}| \\ &\le \sqrt{\sum_{i\neq j}(k^R(x_i, x_j))^2} + \|k^R(\cdot,\cdot) - \operatorname{tr}\mathcal{T}_{G^\perp}\|_\infty.\end{aligned}$$

Squaring, taking an expectation, and noting that

$$\mathbf{E}_{\substack{\text{i.i.d.}\\ x,y\sim\mu}}(k^R(x, y))^2 = \operatorname{tr}(\mathcal{T}_{G^\perp}^2)$$

completes the proof. □

Proof of Lemma 2. We have

$$\begin{aligned}\operatorname{tr}_{L_2}(\mathcal{R}^*\mathcal{R}) &= \sum_{i=1}^n \operatorname{tr}_{L_2}(k_{x_i}^R \otimes k_{x_i}^R) \\ &= \sum_{i=1}^n \|k_{x_i}^R\|_{L_2}^2 \\ &= \sum_{i=1}^n \sum_{\ell>p} \lambda_\ell^2 v_\ell^2(x_i).\end{aligned}$$

Taking an expectation completes the proof. □

Proof of Lemma 3. We can write the operator $\mathcal{C}^*\mathcal{C}$ as a sum of independent random operators:

$$\mathcal{C}^*\mathcal{C} = \sum_{i=1}^n z(x_i) \otimes z(x_i),$$

where

$$z(x) \coloneqq \sum_{\ell=1}^p v_\ell(x) v_\ell.$$

Note that the BOS condition implies $\|z(x)\|_{L_2}^2 \le Cp$ almost surely in x. We also have $\mathbf{E}\, z(x) \otimes z(x) = \mathcal{I}_G$ for $x \sim \mu$.

We use a matrix Bernstein inequality (Tropp, 2015, Theorem 6.6.1) to analyze the zero-mean sum

$$\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G = \sum_{i=1}^p (z(x_i) \otimes z(x_i) - \mathbf{E}\, z(x_i) \otimes z(x_i)).$$

Writing $X_i = z(x_i) \otimes z(x_i) - \mathbf{E}\, z(x_i) \otimes z(x_i)$, we have $\|X_i\|_{L_2} \le Cp$ almost surely, and

$$\mathbf{E}\, X_i^2 \preceq \mathbf{E}(z(x_i) \otimes z(x_i))^2 = \mathbf{E}\|z(x_i)\|_{L_2}^2 z(x_i) \otimes z(x_i) \preceq Cp\, \mathbf{E}\, z(x_i) \otimes z(x_i) = Cp\mathcal{I}_G.$$

The Bernstein inequality then gives that for any $t > 0$, with probability at least $1 - e^{-t}$,

$$\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} = \left\| \sum_{i=1}^n X_i \right\|_{L_2} \lesssim \sqrt{Cpn(t + \log p)} + Cp(t + \log p).$$

□

Proof of Lemma 4. For $z \in \mathbf{R}^n$, we have

$$\langle \mathcal{R}\mathcal{R}^* z, z\rangle = \sum_{\ell>p} \lambda_\ell \langle w_\ell, z\rangle^2.$$

By our assumptions, this is the sum of independent random variables.

If $\|z\|_{\ell_2} = 1$, then, for each ℓ, $\langle w_\ell, z\rangle^2$ is sub-exponential (as the square of a sub-Gaussian variable; since the sub-Gaussian norm is bounded, so is the sub-exponential norm), $\mathbf{E}\langle w_\ell, z\rangle^2 = 1$, and $\mathbf{E}\langle w_\ell, z\rangle^4 \lesssim 1$.

Note that in this case, $\mathbf{E}\langle \mathcal{R}\mathcal{R}^* z, z\rangle = \operatorname{tr} \mathcal{T}_{G^\perp} = \langle (\operatorname{tr} \mathcal{T}_{G^\perp}) I_n z, z\rangle$, and

$$\begin{aligned}\mathbf{E}(\langle \mathcal{R}\mathcal{R}^* z, z\rangle - \mathbf{E}\langle \mathcal{R}\mathcal{R}^* z, z\rangle)^2 &= \sum_{\ell>p} \lambda_\ell^2\, \mathbf{E}(\langle w_\ell, z\rangle^2 - \mathbf{E}\langle w_\ell, z\rangle^2)^2 \\ &\lesssim \sum_{\ell>p} \lambda_\ell^2.\end{aligned}$$

A Bernstein inequality then implies that for $t > 0$, with probability at least $1 - e^{-t}$, we have

$$|\langle \mathcal{R}\mathcal{R}^* z, z\rangle - \operatorname{tr} \mathcal{T}_{G^\perp}| \lesssim \sqrt{\left(\sum_{\ell > p} \lambda_\ell^2\right) t} + \lambda_{p+1} t.$$

By a standard covering argument (e.g., Vershynin, 2018, Exercise 4.4.3), we then obtain, with probability at least $1 - e^{-t}$,

$$\max_{z \in S^{n-1}} |\langle \mathcal{R}\mathcal{R}^* z, z\rangle - \operatorname{tr} \mathcal{T}_{G^\perp}| \lesssim \sqrt{\left(\sum_{\ell > p} \lambda_\ell^2\right)(n+t)} + \lambda_{p+1}(n+t),$$

where S^{n-1} is the unit sphere in $\mathbf{R}^n$. □

E TIGHTNESS OF GENERAL FEATURE RESULTS

With no independence assumptions on the features $\{v_\ell(x)\}_\ell$, our general results require $d \gtrsim n^2$ in order to upper and lower bound the residual Gram matrix $\mathcal{R}\mathcal{R}^*$ by constant multiples of the identity. The following theorem shows that for Fourier features, $d \gtrsim n^2$ is in fact a necessary condition, i.e. if $d = o(n^2)$, then the condition number of $\mathcal{R}\mathcal{R}^*$ grows as $n \to \infty$.

Theorem 3. *Consider the case of Fourier features with bi-level eigenvalues, i.e. $v_\ell \in L_2([0,1])$ for $\ell = -d, \dots, d$, which are defined by $v_\ell(x) = e^{\mathbf{j}2\pi\ell x}$ for $x \in [0,1]$, and $\lambda_\ell = 1$ for $|\ell| \le p$, $\lambda_\ell = \gamma \in (0,1)$ for $p < |\ell| \le d$. Then, for any constant $\tau > 0$, the residual Gram matrix $\mathcal{R}\mathcal{R}^*$ satisfies*

$$\frac{\lambda_{max}(\mathcal{R}\mathcal{R}^*)}{\lambda_{min}(\mathcal{R}\mathcal{R}^*)} \gtrsim \frac{n^4}{\tau^2 d^2}$$

with probability at least $1 - e^{-\tau}$.

Intuitively, if there exist distinct indices $i, i' = 1, \dots, n$ such that x_i and $x_{i'}$ are very close together, then the i-th and i'-th columns (and rows) of $\mathcal{R}\mathcal{R}^*$ are nearly identical, and thus, $\mathcal{R}\mathcal{R}^*$ is nearly rank-deficient. We now make this argument rigorous.

Proof. First, pick any two indices $i, i' \in \{1, \dots, n\}$ with $i \neq i'$ and consider the 2×2 submatrix of $\mathcal{R}\mathcal{R}^*$ formed by the i-th and i'-th rows and columns, i.e,

$$(\mathcal{R}\mathcal{R}^*)_{\text{sub}} := \begin{bmatrix} k^R(x_i, x_i) & k^R(x_i, x_{i'}) \\ k^R(x_{i'}, x_i) & k^R(x_{i'}, x_{i'}) \end{bmatrix}.$$

The kernel restricted to $G^\perp$ is given by

$$\begin{aligned} k^R(x,y) &= \sum_{p < |\ell| \le d} \lambda_\ell v_\ell(x) \overline{v_\ell(y)} \\ &= \sum_{p < |\ell| \le d} \gamma e^{\mathbf{j}2\pi\ell(x-y)} \\ &= \gamma \frac{\sin[(2d+1)\pi(x-y)] - \sin[(2p+1)\pi(x-y)]}{\sin[\pi(x-y)]}. \end{aligned}$$

Hence, $k^R(x_i, x_i) = k^R(x_{i'}, x_{i'}) = 2(d-p)\gamma$.

Furthermore, using the inequality $2\cos\theta \geq 2 - \theta^2$ for $\theta \in \mathbf{R}$, we have

$$\begin{aligned}
\frac{\sin[(2d+1)\pi t] - \sin[(2p+1)\pi t]}{\sin[\pi t]} &= \sum_{p<|\ell|\leq d} e^{\mathrm{j}2\pi\ell t} \\
&= \sum_{\ell=p+1}^{d} 2\cos(2\pi\ell t) \\
&\geq \sum_{\ell=p+1}^{d} \left[2 - (2\pi\ell t)^2\right] \\
&= 2(d-p) - 4\pi^2 \left(\sum_{\ell=p+1}^{d} \ell^2\right) t^2 \\
&\geq 2(d-p) - 4\pi^2 d^2 (d-p) t^2
\end{aligned}$$

for all $t \in \mathbf{R}$, and thus,

$$\begin{aligned}
k^R(x_i, x_{i'}) = k^R(x_{i'}, x_i) &= \gamma \frac{\sin[(2d+1)\pi(x_i - x_{i'})] - \sin[(2p+1)\pi(x_i - x_{i'})]}{\sin[\pi(x_i - x_{i'})]} \\
&\geq 2(d-p)\gamma - 4\pi^2 d^2 (d-p)\gamma (x_i - x_{i'})^2.
\end{aligned}$$

We can then bound the smallest eigenvalue of $\mathcal{R}\mathcal{R}^*$ by

$$\lambda_{\min}(\mathcal{R}\mathcal{R}^*) \leq \lambda_{\min}((\mathcal{R}\mathcal{R}^*)_{\mathrm{sub}}) = k^R(x_i, x_i) - k^R(x_i, x_{i'}) \leq 4\pi^2 d^2 (d-p)\gamma (x_i - x_{i'})^2.$$

Then, by using the trivial bound $\lambda_{\max}(\mathcal{R}\mathcal{R}^*) \geq \frac{1}{n}\operatorname{tr}(\mathcal{R}\mathcal{R}^*) = \frac{1}{n} \cdot 2(d-p)\gamma n = 2(d-p)\gamma$, we have

$$\frac{\lambda_{\max}(\mathcal{R}\mathcal{R}^*)}{\lambda_{\min}(\mathcal{R}\mathcal{R}^*)} \geq \frac{2(d-p)\gamma}{4\pi^2 d^2 (d-p)\gamma (x_i - x_{i'})^2} = \frac{1}{2\pi^2 d^2 (x_i - x_{i'})^2}.$$

This bound holds for any distinct indices $i \neq i'$. A relatively straightforward calculation[1] shows that if $x_1, \ldots, x_n$ are i.i.d. Uniform$[0,1]$, then for any $\delta \in (0, \frac{1}{n-1})$,

$$\begin{aligned}
\mathbf{P}\left\{|x_i - x_{i'}| \geq \delta \text{ for all } i \neq i'\right\} &= n!\,\mathbf{P}\left\{x_{i-1} + \delta \leq x_i \text{ for all } i = 2, \ldots, n\right\} \\
&= n! \int \cdots \int_{\{0 \leq x_1,\ x_{i-1}+\delta \leq x_i \text{ for } i=2,\ldots,n,\ x_n \leq 1\}} dx_1 \cdots dx_n \\
&= n! \int \cdots \int_{\{y_i \geq 0 \text{ for } i=1,\ldots,n,\ y_1 + \cdots + y_n \leq 1-(n-1)\delta\}} dy_1 \cdots dy_n \\
&= n! \cdot \frac{1}{n!}(1-(n-1)\delta)^n \\
&= (1-(n-1)\delta)^n
\end{aligned}$$

where we made the change of variable $y_1 = x_1$ and $y_i = x_i - x_{i-1} - \delta$ for $i = 2, \ldots, n$, and we used the fact that the volume of the standard n-simplex is $\frac{1}{n!}$. Hence, if $0 < \delta < \frac{1}{n-1}$, the probability that $|x_i - x_{i'}| \leq \delta$ for some indices $i \neq i'$ is $1 - (1-(n-1)\delta)^n$.

If $0 < \tau < n$, we can apply this result for $\delta = \frac{\tau}{n(n-1)}$, to obtain that with probability $1 - (1 - \frac{\tau}{n})^n \geq 1 - e^{-\tau}$ there exist $i \neq i'$ such that $|x_i - x_{i'}| \leq \frac{\tau}{n(n-1)}$, and thus,

$$\frac{\lambda_{\max}(\mathcal{R}\mathcal{R}^*)}{\lambda_{\min}(\mathcal{R}\mathcal{R}^*)} \geq \frac{1}{2\pi^2 d^2 (x_i - x_{i'})^2} \geq \frac{n^2(n-1)^2}{2\pi^2 d^2 \tau^2} \gtrsim \frac{n^4}{\tau^2 d^2}.$$

If $\tau \geq n$, then it is guaranteed that there exist two indices $i \neq i'$ which satisfy $|x_i - x_{i'}| \leq \frac{1}{n-1} \leq \frac{\tau}{n(n-1)}$, and the same bound holds. □

[1] Thanks to Hans's answer at https://mathoverflow.net/questions/1294

F PROOF OF BI-LEVEL ENSEMBLE ASYMPTOTIC RESULTS

If $\beta > 2$ and $r < 1$, the concentration results Lemmas 1 and 3 will hold as n becomes large, since we will have $n \gg p \log p$ and $d - p \approx n^\beta \gg n^2$. We now apply Lemma 6 and Theorem 2 to the bi-level ensemble. Since we are in the interpolating regime, we take $\alpha = 0$; then, $\alpha_L = \lambda_{\min}(\mathcal{R}\mathcal{R}^*)$ and $\alpha_U = \lambda_{\max}(\mathcal{R}\mathcal{R}^*)$ are the smallest and largest eigenvalues of $\mathcal{R}\mathcal{R}^*$. As long as α_L and α_U are close together (which we will analyze next), we will have

$$\tilde{\alpha} \approx \bar{\alpha} \approx \sum_{\ell > p} \lambda_\ell \approx n^\beta \cdot n^{-(\beta - r - q)} = n^{r+q}.$$

Furthermore,

$$\sum_{\ell > p} \lambda_\ell^2 \approx n^\beta n^{-2(\beta - q - r)} = n^{2q + 2r - \beta}.$$

Applying these scalings to Theorem 2 gives us

$$\mathbf{E}_\xi \|\epsilon\|_{L_2}^2 \lesssim n^{r-1} + \frac{n}{n^{2(r+q)}} n^{2q+2r-\beta} = n^{r-1} + n^{1-\beta}.$$

To bound the bias, note that combining the above calculations with Lemma 1 gives

$$\begin{aligned} \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} &\lesssim \frac{1}{\bar{\alpha}} \sqrt{n^2 \sum_{\ell > p} \lambda_\ell^2} \\ &\approx \frac{1}{n^{r+q}} \sqrt{n^2 n^{2q+2r-\beta}} \\ &= n^{1-\beta/2}. \end{aligned}$$

Combining this with Lemma 3, the quantity c in Theorem 1 and Lemma 6 can be bounded as

$$c \lesssim n^{1-\beta/2} + n^{(r-1)/2} \sqrt{\log n}.$$

Then Lemma 6 gives

$$\begin{aligned} \frac{\|\hat{\eta}_0 - \overline{\mathcal{S}} \eta^*\|_{L_2}}{\|\eta^*\|_{\mathcal{H}}} &\lesssim \left(n^{1-\beta/2} + n^{(r-1)/2} \sqrt{\log n} + \sqrt{\frac{n^{-(\beta - r - q)} n}{n^{r+q}}} \right) \cdot \min\left\{1, n^{r+q-1}, n^{(r+q-1)/2}\right\} \\ &\lesssim \left(n^{1-\beta/2} + n^{(r-1)/2} \sqrt{\log n} \right) \min\{1, n^{r+q-1}\}. \end{aligned}$$

Recall from (3) that excess classification risk has upper bound $\mathcal{E} \le \frac{\|\hat{\eta}_r\|_{L_2}}{s}$ for any decomposition $\hat{\eta}_0 = s\eta^* + \hat{\eta}_r$ with an $s > 0$ that we can choose. We will now characterize the terms s and $\|\hat{\eta}_r\|_{L_2}$, beginning with the factor s. The ideal survival operator is given by

$$\overline{\mathcal{S}} = \mathcal{I}_G - \left(\mathcal{I}_G + \frac{n}{\bar{\alpha}} \mathcal{T}_G \right)^{-1} = \frac{1}{1 + \bar{\alpha} n} \mathcal{I}_G \approx \frac{1}{1 + n^{r+q-1}} \mathcal{I}_G.$$

Then, we can decompose

$$\hat{\eta} = \overline{\mathcal{S}} \eta^* + \hat{\eta}_r \approx \frac{1}{1 + n^{r+q-1}} \eta^* + \hat{\eta}_r,$$

where $\hat{\eta}_r = \epsilon + \hat{\eta}_0 - \overline{\mathcal{S}} \eta^*$. This gives us $s \approx \frac{1}{1 + n^{r+q-1}}$.

Next, we bound $\|\hat{\eta}_r\|_{L_2}$. We have

$$\|\hat{\eta}_r\|_{L_2} \lesssim n^{(r-1)/2} + n^{(1-\beta)/2} + \left(n^{1-\beta/2} + n^{(r-1)/2} \sqrt{\log n} \right) \cdot \min\{1, n^{r+q-1}\} \|\eta^*\|_{L_2}.$$

Above, we used the fact that $\|\eta^*\|_{L_2} = \|\eta^*\|_{\mathcal{H}}$.

There are several cases to consider (recall that we are already assuming $\beta > 2$ and $r < 1$):

1. $q < 1-r$: In this case, $s \approx \frac{1}{1+n^{r+q-1}} \to 1$, and $\|\hat{\eta}_r\|_{L_2} \to 0$. Thus both the excess regression and classification risk converge to 0 as $n \to \infty$.

2. If $q > 1-r$, we have $s \to 0$ and $\|\hat{\eta}_r\|_{L_2} \to 0$, so $\|\hat{\eta}\|_{L_2} \to 0$. Therefore will will *not* get regression consistency (for nonzero η^*).

3. If $1-r < q < \frac{3}{2}(1-r)$ and $\beta > 2r+2q$, then $s \to 0$, but $\frac{\|\hat{\eta}_r\|_{L_2}}{s} \approx \|\hat{\eta}_r\|_{L_2} \cdot (1+n^{r+q-1}) \to 0$, so the excess classification risk converges to zero as $n \to \infty$ even though the regression risk does not.

4. If $1-r < q < \frac{3}{2}(1-r)$ but $\beta < 2r+2q$ or if $q > \frac{3}{2}(1-r)$, our analysis does not yield any convergence results. It is an interesting and important direction for future work to characterize precisely what relations between the parameters q, r, β are both sufficient and necessary for classification risk to go to 0 as $n \to \infty$.

G DISTORTION ANALYSIS

In this section, we analyze more carefully the regularization-induced distortion. In particular, we consider how different the (deterministic) ideal survival operator $\overline{\mathcal{S}}$ is from a multiple of the identity. Recall that

$$\overline{\mathcal{S}} = \mathcal{I}_G - \Big(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\Big)^{-1} = \frac{n}{\bar{\alpha}}\mathcal{T}_G\Big(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\Big)^{-1}.$$

We want to solve

$$\begin{aligned}\arg\min_{s>0} \ \big\|s\mathcal{I}_G - \overline{\mathcal{S}}\big\|_{\mathcal{H}\to L_2} &= \arg\min_{s>0} \ \Big\|s\mathcal{T}_G^{1/2} - \mathcal{T}_G^{1/2}\overline{\mathcal{S}}\Big\|_{L_2} \\ &= \arg\min_{s>0} \ \max_{1\le\ell\le p} \ \sqrt{\lambda_\ell}\left|s - \frac{\lambda_\ell}{\lambda_\ell + \frac{\bar{\alpha}}{n}}\right|.\end{aligned}$$

We abbreviate $b \coloneqq \frac{\bar{\alpha}}{n}$. The objective function in s is convex as the maximum of convex functions. Some convex analysis tell us that there must be (at least) two distinct $i, j \in \{1, \dots, p\}$ such that, for s at its optimal value s^*, both i and j achieve the maximum over ℓ, and the arguments to the absolute value have different signs. Assuming, without loss of generality, that $\lambda_j > \lambda_i$, this implies

$$\begin{aligned}\arg\min_{s>0} \ \max_{1\le\ell\le p} \ \sqrt{\lambda_\ell}\left|s - \frac{\lambda_\ell}{\lambda_\ell + \frac{\bar{\alpha}}{n}}\right| &= \sqrt{\lambda_i}\left(s^* - \frac{\lambda_i}{\lambda_i + b}\right) \\ &= \sqrt{\lambda_j}\left(\frac{\lambda_j}{\lambda_j + b} - s^*\right).\end{aligned}$$

Note that the last expression is increasing in λ_j, so we can take $j = 1$. Solving for s^* gives

$$s^* = \frac{\lambda_i\lambda_1 + b(\lambda_i + \lambda_1 - \sqrt{\lambda_i\lambda_1})}{(b+\lambda_i)(b+\lambda_1)}.$$

Plugging this into the objective function gives

$$\|s^*\mathcal{I}_G - \overline{\mathcal{S}}\|_{\mathcal{H}\to L_2} = \max_i \ \frac{b\sqrt{\lambda_i\lambda_1}(\sqrt{\lambda_1} - \sqrt{\lambda_i})}{(b+\lambda_1)(b+\lambda_i)}.$$

One can check that if $\lambda_p \ge \frac{\lambda_1}{\left(1+\sqrt{1+\frac{\lambda_1}{b}}\right)^2}$, this minimum is achieved for $i = p$. Otherwise, we can find an upper bound by optimizing over continuous λ:

$$\begin{aligned}\max_i \ \frac{b\sqrt{\lambda_i\lambda_1}(\sqrt{\lambda_1} - \sqrt{\lambda_i})}{(b+\lambda_1)(b+\lambda_i)} &\le \max_{\lambda\ge 0} \ \frac{b\sqrt{\lambda\lambda_1}(\sqrt{\lambda_1} - \sqrt{\lambda})}{(b+\lambda_1)(b+\lambda)} \\ &= \frac{b\lambda_1^{3/2}}{2(b+\lambda_1)(b+\sqrt{b(b+\lambda_1)})},\end{aligned}$$

where the minimum is achieved at $\lambda = \frac{\lambda_1}{\left(1+\sqrt{1+\frac{\lambda_1}{b}}\right)^2}$.

Whatever value of λ we use, we then have, for the corresponding choice of s,

$$\frac{\|s\mathcal{I}_G - \overline{\mathcal{S}}\|_{\mathcal{H}\to L_2}}{s} = \frac{b\sqrt{\lambda_1\lambda}(\sqrt{\lambda_1}-\sqrt{\lambda})}{\lambda_1\lambda + b(\lambda_1+\lambda-\sqrt{\lambda_1\lambda})}.$$

For $\lambda = \frac{\lambda_1}{\left(1+\sqrt{1+\frac{\lambda_1}{b}}\right)^2}$, we get

$$\frac{\|s\mathcal{I}_G - \overline{\mathcal{S}}\|_{\mathcal{H}\to L_2}}{s} \le \frac{\sqrt{b\lambda_1(b+\lambda_1)}}{2b+2\lambda_1+\sqrt{b(b+\lambda_1)}}.$$

If $\frac{\bar{\alpha}}{n} = b \gtrsim \lambda_1$, then this last bound is approximately $\sqrt{\lambda_1} \approx \|\mathcal{I}_G\|_{\mathcal{H}\to L_2}$, so there appears to be little hope of getting small classification error from this bound.

Alternatively, if $\frac{\bar{\alpha}}{n} \ll \lambda_1$, we get

$$\frac{\|s\mathcal{I}_G - \overline{\mathcal{S}}\|_{\mathcal{H}\to L_2}}{s} \lesssim \sqrt{\frac{\bar{\alpha}}{n}}.$$

However, recall that the *regression* error is of the same order, so this analysis does not significantly improve our classification risk.

Therefore, the only regime in which we gain anything over the regression analysis is when $\lambda_p > \frac{\lambda_1}{\left(1+\sqrt{1+\frac{\lambda_1}{b}}\right)^2}$.

If $b \gtrsim \lambda_1$, then this constraint implies that λ_p/λ_1 is not very small. Furthermore,

$$\frac{\|s^*\mathcal{I}_G - \overline{\mathcal{S}}\|_{\mathcal{H}\to L_2}}{s^*} \approx \sqrt{\frac{\lambda_p}{\lambda_1}}(\sqrt{\lambda_1}-\sqrt{\lambda_p}).$$

Since λ_p is not too small, this ratio is only small when λ_1 and λ_p are very close together.

If $b \lesssim \lambda_1$, the constraint implies $\lambda_p \gtrsim b$. Then

$$\frac{\|s^*\mathcal{I}_G - \overline{\mathcal{S}}\|_{\mathcal{H}\to L_2}}{s^*} \approx \frac{b}{\sqrt{\lambda_1\lambda_p}}(\sqrt{\lambda_1}-\sqrt{\lambda_p}).$$

This is better than the previous case when b is small, and it improves over the regression error bound when λ_1 and λ_p are close. However, note that in this case we get $c \gtrsim 1$, so unless λ_p is very close to λ_1, there is no significant improvement over regression error.