The Annals of Statistics
2020, Vol. 48, No. 3, 1329–1347
https://doi.org/10.1214/19-AOS1849

JUST INTERPOLATE: KERNEL "RIDGELESS" REGRESSION CAN GENERALIZE

BY TENGYUAN LIANG[1] AND ALEXANDER RAKHLIN[2]

[1]*Econometrics and Statistics, Booth School of Business, University of Chicago, tengyuan.liang@chicagobooth.edu*
[2]*Center for Statistics & IDSS, Massachusetts Institute of Technology, rakhlin@mit.edu*

In the absence of explicit regularization, Kernel "Ridgeless" Regression with nonlinear kernels has the potential to fit the training data perfectly. It has been observed empirically, however, that such interpolated solutions can still generalize well on test data. We isolate a phenomenon of implicit regularization for minimum-norm interpolated solutions which is due to a combination of high dimensionality of the input data, curvature of the kernel function and favorable geometric properties of the data such as an eigenvalue decay of the empirical covariance and kernel matrices. In addition to deriving a data-dependent upper bound on the out-of-sample error, we present experimental evidence suggesting that the phenomenon occurs in the MNIST dataset.

1. Introduction. According to conventional wisdom, explicit regularization should be added to the least-squares objective when the Hilbert space $\mathcal{H}$ is high or infinite dimensional [1, 11, 14, 15, 24, 25, 28]:

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\big(f(x_i)-y_i\big)^2+\lambda\|f\|_{\mathcal{H}}^2. \tag{1.1}$$

The regularization term is introduced to avoid "overfitting" since kernels provide enough flexibility to fit training data exactly (i.e., interpolate it). From the theoretical point of view, the regularization parameter λ is a knob for balancing bias and variance, and should be chosen judiciously. Yet, as noted by a number of researchers in the last few years,[1] the best out-of-sample performance, empirically, is often attained by setting the regularization parameter to *zero* and finding the minimum-norm solution among those that interpolate the training data. The mechanism for good out-of-sample performance of this interpolation method has been largely unclear [5, 31].

As a concrete motivating example, consider the prediction performance of Kernel Ridge Regression for various values[2] of the regularization parameter λ on subsets of the MNIST dataset. For virtually all pairs of digits, the best out-of-sample mean squared error is achieved at $\lambda = 0$. Contrary to the standard bias-variance-tradeoffs picture we have in mind, the test error is monotonically decreasing as we decrease λ (see Figure 1 and further details in Section 6).

We isolate what appears to be a new phenomenon of *implicit regularization* for interpolated minimum-norm solutions in Kernel "Ridgeless" Regression. This regularization is due to the curvature of the kernel function and "kicks in" only for high-dimensional data and for "favorable" data geometry. We provide out-of-sample statistical guarantees in terms of spectral

Received September 2018; revised February 2019.
MSC2010 subject classifications. 68Q32, 62G08.
Key words and phrases. Minimum-norm interpolation, reproducing kernel Hilbert spaces, implicit regularization, high dimensionality, data-dependent bounds, kernel methods, spectral decay.

[1]In particular, we thank M. Belkin, B. Recht, L. Rosasco and N. Srebro for highlighting this phenomenon.
[2]We take $\lambda \in \{0, 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64, 1.28\}$.

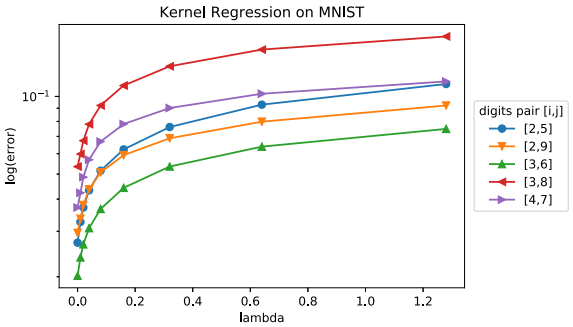

FIG. 1. *Test performance of Kernel Ridge Regression on pairs of MNIST digits for various values of regularization parameter* λ*, normalized by variance of* y *in the test set (for visualization purposes).*

decay of the empirical kernel matrix and the empirical covariance matrix, under additional technical assumptions.

Our analysis rests on the recent work in random matrix theory. In particular, we use a suitable adaptation of the argument of [13] who showed that high-dimensional random kernel matrices can be approximated in spectral norm by linear kernel matrices plus a scaled identity. While the message of [13] is often taken as "kernels do not help in high dimensions," we show that such a random matrix analysis helps in explaining the good performance of interpolation in Kernel "Ridgeless" Regression.

1.1. *Literature review.* Grace Wahba [28] pioneered the study of nonparametric regression in reproducing kernel Hilbert spaces (RKHS) from the computational and statistical perspectives. One of the key aspects in that work is the role of the decay of eigenvalues of the kernel (at the population level) in rates of convergence. The analysis relies on explicit regularization (ridge parameter λ) for the bias-variance trade-off. The parameter is either chosen to reflect the knowledge of the spectral decay at the population level [11] (typically unknown to statistician), or by the means of cross-validation [15]. Interestingly, the explicit formula of Kernel Ridge Regression has been introduced as "kriging" in the literature before, and was widely used in Bayesian statistics [9, 28].

In the learning theory community, Kernel Ridge Regression is known as a special case of Support Vector Regression [24, 26, 27]. Notions like metric entropy [10] or "effective dimension" [8] were employed to analyze the guarantees on the excess loss of Kernel Ridge Regression, even when the model is misspecified. We refer the readers to [17] for more details. Again, the analysis leans crucially on the explicit regularization, as given by a careful choice of λ, for the model complexity and approximation trade-off, and mostly focusing on the fixed dimension and large sample size setting. However, to the best of our knowledge, the literature stays relatively quiet in terms of what happens to the minimum norm interpolation rules, that is, $\lambda = 0$. As pointed out by [4, 5], the existing bounds in nonparametric statistics and learning theory do not apply to interpolated solution either in the regression or the classification setting. In this paper, we aim to answer when and why interpolation in RKHS works, as a starting point for explaining the good empirical performance of interpolation using kernels in practice [5, 31].

2. Preliminaries.

2.1. *Problem formulation.* Suppose we observe n i.i.d. pairs (x_i, y_i), $1 \le i \le n$, where x_i are the covariates with values in a compact domain $\Omega \subset \mathbb{R}^d$ and $y_i \in \mathbb{R}$ are the responses (or, labels). Suppose the n pairs are drawn from an unknown probability distribution $\mu(x, y)$. We

are interested in estimating the conditional expectation function $f_*(x) = \mathbf{E}(\mathbf{y}|\mathbf{x} = x)$, which is assumed to lie in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$. Suppose the RKHS is endowed with the norm $\|\cdot\|_{\mathcal{H}}$ and corresponding positive definite kernel $K(\cdot,\cdot) : \Omega \times \Omega \to \mathbb{R}$. The interpolation estimator studied in this paper is defined as

$$\widehat{f} = \arg\min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}}, \quad \text{s.t.} \quad f(x_i) = y_i, \quad \forall i. \tag{2.1}$$

Let $X \in \mathbb{R}^{n \times d}$ be the matrix with rows $x_1, \ldots, x_n$ and let Y be the vector of values $y_1, \ldots, y_n$. Slightly abusing the notation, we let $K(X, X) = [K(x_i, x_j)]_{ij} \in \mathbb{R}^{n \times n}$ be the kernel matrix. Extending this definition, for $x \in \Omega$ we denote by $K(x, X) \in \mathbb{R}^{1 \times n}$ the matrix of values $[K(x, x_1), \ldots, K(x, x_n)]$. When $K(X, X)$ is invertible, solution to (2.1) can be written in the closed form

$$\widehat{f}(x) = K(x, X)K(X, X)^{-1}Y. \tag{2.2}$$

In this paper, we study the case when $K(X, X)$ is full rank, taking (2.2) as the starting point. For this interpolating estimator, we provide high-probability (with respect to a draw of X) upper bounds on the integrated squared risk of the form

$$\mathbf{E}\big(\widehat{f}(\mathbf{x}) - f_*(\mathbf{x})\big)^2 \le \phi_{n,d}\big(X, f^*\big). \tag{2.3}$$

Here, the expectation is over $\mathbf{x} \sim \mu$ and $Y|X$, and $\phi_{n,d}$ is a data-dependent upper bound. We remark that upper bounds of the form (2.3) also imply prediction loss bounds for excess square loss with respect to the class $\mathcal{H}$, as $\mathbf{E}(\widehat{f}(\mathbf{x}) - f_*(\mathbf{x}))^2 = \mathbf{E}(\widehat{f}(\mathbf{x}) - \mathbf{y})^2 - \mathbf{E}(f_*(\mathbf{x}) - \mathbf{y})^2$.

2.2. *Notation and background on RKHS.* For an operator A, its adjoint is denoted by A^*. For real matrices, the adjoint is the transpose. For any $x \in \Omega$, let $K_x : \mathbb{R} \to \mathcal{H}$ be such that

$$f(x) = \langle K_x, f\rangle_{\mathcal{H}} = K_x^* f. \tag{2.4}$$

It follows that for any $x, z \in \Omega$

$$K(x, z) = \langle K_x, K_z\rangle_{\mathcal{H}} = K_x^* K_z. \tag{2.5}$$

Let us introduce the integral operator $\mathcal{T}_\mu : L^2_\mu \to L^2_\mu$ with respect to the marginal measure $\mu(x)$:

$$\mathcal{T}_\mu f(z) = \int K(z, x) f(x)\, d\mu(x), \tag{2.6}$$

and denote the set of eigenfunctions of this integral operator by $e(x) = \{e_1(x), e_2(x), \ldots, e_p(x)\}$, where p could be ∞. We have that

$$\mathcal{T}_\mu e_i = t_i e_i, \quad \text{and} \quad \int e_i(x) e_j(x)\, d\mu(x) = \delta_{ij}. \tag{2.7}$$

Denote $T = \operatorname{diag}(t_1, \ldots, t_p)$ as the collection of nonnegative eigenvalues. Adopting the spectral notation,

$$K(x, z) = e(x)^* T e(z).$$

Via this spectral characterization, the interpolation estimator (2.1) takes the following form:

$$\widehat{f}(x) = e(x)^* T e(X)\big[e(X)^* T e(X)\big]^{-1} Y. \tag{2.8}$$

Extending the definition of K_x, it is natural to define the operator $K_X : \mathbb{R}^n \to \mathcal{H}$. Denote the sample version of the kernel operator to be

$$\widehat{\mathcal{T}} := \frac{1}{n} K_X K_X^* \tag{2.9}$$

and the associated eigenvalues to be $\lambda_j(\widehat{T})$, indexed by j. The eigenvalues are the same as those of $\frac{1}{n}K(X,X)$. It is sometimes convenient to express $\widehat{T}$ as the linear operator under the basis of eigenfunctions, in the following matrix sense:

$$\widehat{T} = T^{1/2}\left(\frac{1}{n}e(X)e(X)^*\right)T^{1/2}.$$

We write $\mathbf{E}_\mu[\cdot]$ to denote the expectation with respect to the marginal $\mathbf{x} \sim \mu$. Furthermore, we denote by

$$\|g\|_{L^2_\mu}^2 = \int g^2 \, d\mu(x) = \mathbf{E}_\mu g^2(\mathbf{x})$$

the squared L^2 norm with respect to the marginal distribution. The expectation $\mathbf{E}_{Y|X}[\cdot]$ denotes the expectation over $y_1, \ldots, y_n$ conditionally on $x_1, \ldots, x_n$.

3. Main result. We impose the following assumptions:

(A.1) High dimensionality: there exists universal constants $c, C \in (0, \infty)$ such that $c \le d/n \le C$. Denote by $\Sigma_d = \mathbf{E}_\mu[x_i x_i^*]$ the covariance matrix, assume that the operator norm $\|\Sigma_d\|_{\mathrm{op}} \le 1$.

(A.2) $(8+m)$-moments: $z_i := \Sigma_d^{-1/2} x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, are i.i.d. random vectors. Furthermore, the entries $z_i(k), 1 \le k \le d$ are i.i.d. from a distribution with $\mathbf{E}z_i(k) = 0$, $\mathrm{Var}(z_i(k)) = 1$ and $|z_i(k)| \le C \cdot d^{\frac{2}{8+m}}$, for some $m > 0$.

(A.3) Noise condition: there exists a $\sigma > 0$ such that $\mathbf{E}[(f_*(\mathbf{x}) - \mathbf{y})^2 | \mathbf{x} = x] \le \sigma^2$ for all $x \in \Omega$.

(A.4) Nonlinear kernel: for any $x \in \Omega$, $K(x,x) \le M$. Furthermore, we consider the inner product kernels of the form

$$K(x,x') = h\left(\frac{1}{d}\langle x, x'\rangle\right) \tag{3.1}$$

for a nonlinear smooth function $h(\cdot) : \mathbb{R} \to \mathbb{R}$ in a neighborhood of 0.

While we state the main theorem for inner product kernels, the results follow under suitable modifications[3] for Radial Basis Function (RBF) kernels of the form

$$K(x,x') = h\left(\frac{1}{d}\|x - x'\|^2\right). \tag{3.2}$$

We postpone the discussion of the assumptions until after the statement of the main theorem.

Let us first define the following quantities related to curvature of h:

$$\begin{aligned} \alpha &:= h(0) + h''(0)\frac{\mathrm{Tr}(\Sigma_d^2)}{d^2}, \\ \beta &:= h'(0), \\ \gamma &:= h\left(\frac{\mathrm{Tr}(\Sigma_d)}{d}\right) - h(0) - h'(0)\frac{\mathrm{Tr}(\Sigma_d)}{d}. \end{aligned} \tag{3.3}$$

[3]We refer the readers to [13] for explicit extensions to RBF kernels.

THEOREM 1. *Define*

$$\begin{aligned}\phi_{n,d}(X, f_*) &= \mathbf{V} + \mathbf{B} \\ &:= \frac{8\sigma^2 \|\Sigma_d\|_{\mathrm{op}}}{d} \sum_j \frac{\lambda_j(\frac{XX^*}{d} + \frac{\alpha}{\beta} 11^*)}{[\frac{\gamma}{\beta} + \lambda_j(\frac{XX^*}{d} + \frac{\alpha}{\beta} 11^*)]^2} \\ &\quad + \|f_*\|_{\mathcal{H}}^2 \inf_{0 \le k \le n} \left\{ \frac{1}{n} \sum_{j>k} \lambda_j(K_X K_X^*) + 2M\sqrt{\frac{k}{n}} \right\}.\end{aligned} \tag{3.4}$$

Under the assumptions (A.1)–(A.4) *and for d large enough, with probability at least* $1 - 2\delta - d^{-2}$ (*with respect to a draw of design matrix X*), *the interpolation estimator* (2.2) *satisfies*

$$\mathbf{E}_{Y|X} \|\widehat{f} - f_*\|_{L^2_\mu}^2 \le \phi_{n,d}(X, f_*) + \epsilon(n, d). \tag{3.5}$$

Here, the remainder term $\epsilon(n, d) = O(d^{-\frac{m}{8+m}} \log^{4.1} d) + O(n^{-\frac{1}{2}} \log^{0.5}(n/\delta))$.

A few remarks are in order. First, the upper bound is *data-dependent* and can serve as a certificate (assuming that an upper bound on σ^2, $\|f_*\|_{\mathcal{H}}^2$ can be guessed) that interpolation will succeed. The bound also suggests the regimes when the interpolation method should work. The two terms in the estimate of Theorem 1 represent upper bounds on the variance and bias of the interpolation estimator, respectively. Unlike the explicit regularization analysis (e.g., [8]), the two terms are not controlled by a tunable parameter λ. Rather, the choice of the nonlinear kernel K itself leads to an *implicit control* of the two terms through curvature of the kernel function, favorable properties of the data and high dimensionality. We remark that for the linear kernel ($h(a) = a$), we have $\gamma = 0$, and the bound on the variance term can become very large in the presence of small eigenvalues. In contrast, curvature of h introduces regularization through a nonzero value of γ. We also remark that the bound "kicks in" in the high-dimensional regime: the error term decays with both d and n.

We refer to the favorable structure of eigenvalues of the data covariance matrix as *favorable geometric properties* of the data. The first term (variance) is small when the data matrix enjoys certain decay of the eigenvalues, thanks to the implicit regularization γ. The second term (bias) is small when the eigenvalues of the kernel matrix decay fast or the kernel matrix is effectively low rank. Note that the quantities α, β are constants, and γ scales with $(\mathrm{Tr}(\Sigma_d)/d)^2$. We will provide a detailed discussion on the trade-off between the bias and variance terms for concrete examples in Section 4.

We left the upper bound of Theorem 1 in a data-dependent form for two reasons. First, an explicit dependence on the data tells us whether interpolation can be statistically sound on the given dataset. Second, for general spectral decay, current random matrix theory falls short of characterizing the spectral density nonasymptotically except for special cases [7, 13].

Discussion of the assumptions.

- The assumption in (A.1) that $c \le d/n \le C$ emphasizes that we work in a high-dimensional regime where d scales on the order of n. This assumption is used in the proof of [13], and the particular dependence on c, C can be traced in that work if desired. Rather than doing so, we "folded" these constants into mild additional power of $\log d$. The same goes for the assumption on the scaling of the trace of the population covariance matrix.
- The assumption in (A.2) that $Z_i(k)$ are i.i.d. across $k = 1, \ldots, d$ is a strong assumption that is required to ensure the favorable high-dimensional effect. Relaxing this assumption is left for future work.

- The existence of $(8+m)$-moments for $|z_i(k)|$ is enough to ensure $|z_i(k)| \le C \cdot d^{\frac{2}{8+m}}$ for $1 \le i \le n$, $1 \le k \le d$ almost surely (see Lemma 2.2 in [30]). Remark that the assumption of existence of $(8+m)$-moments in (A.2) is relatively weak. In particular, for bounded or sub-Gaussian variables, $m = \infty$ and the error term $\epsilon(n,d)$ scales as $d^{-1} + n^{-1/2}$, up to log factors. See Lemma B.1 [21] for an explicit calculation in the Gaussian case.
- Finally, as already mentioned, the main result is stated for the inner product kernel, but can be extended to the RBF kernel using an adaptation of the analysis in [13].

4. Behavior of the data-dependent bound. In this section, we estimate, both numerically and theoretically, the nonasymptotic data-dependent upper bound in Theorem 1 in several regimes. To illustrate the various trade-offs, we divide the discussion into two main regimes: $n > d$ and $n < d$. Without loss of generality, we take as an illustration the nonlinearity $h(t) = \exp(2t)$ and $K(x,x') = \exp(2\langle x, x'\rangle/d)$, with the implicit regularization $\mathbf{r} := \gamma/\beta \asymp (\mathrm{Tr}(\Sigma_d)/d)^2$. In our discussion, we take both n and d large enough so that the residuals in Theorem 1 are negligible. The main theoretical results in this section, Corollaries 4.1 and 4.2, are direct consequences of the data-dependent bound in Theorem 1.

Case $n > d$. We can further bound the variance and the bias, with the choice $k = 0$, as

$$\mathbf{V} \precsim \frac{1}{d}\sum_j \frac{\lambda_j(\frac{XX^*}{d})}{[\mathbf{r} + \lambda_j(\frac{XX^*}{d})]^2} = \frac{1}{n}\sum_{j=1}^{d} \frac{\lambda_j(\frac{XX^*}{n})}{[\frac{d}{n}\mathbf{r} + \lambda_j(\frac{XX^*}{n})]^2}, \tag{4.1}$$

$$\mathbf{B} \precsim \frac{1}{n}\sum_{j=1}^{n} \lambda_j(K_X K_X^*) \asymp \mathbf{r} + \frac{1}{d}\sum_{j=1}^{d} \lambda_j\left(\frac{XX^*}{n}\right). \tag{4.2}$$

We first illustrate numerically the bias-variance trade-off by varying the geometric properties of the data in terms of the population spectral decay of $\mathbf{x}$. We shall parametrize the eigenvalues of the covariance, for $0 < \kappa < \infty$, as

$$\lambda_j(\Sigma_d) = \left(1 - \left((j-1)/d\right)^{\kappa}\right)^{1/\kappa}, \quad 1 \le j \le d.$$

The parameter κ controls approximate "low-rankness" of the data: the closer κ is to 0, the faster does the spectrum of the data decay. This is illustrated in the top row of Figure 6 on page 1344. By letting $\kappa \to 0$, $\mathbf{r}$ can be arbitrary small as

$$\frac{\mathrm{Tr}(\Sigma_d)}{d} \asymp \int_0^1 \left(1 - t^{\kappa}\right)^{1/\kappa} dt = \frac{\Gamma(1+1/\kappa)^2}{\Gamma(1+2/\kappa)} \in [0,1].$$

We will focus on three cases, $\kappa \in \{e^{-1}, e^0, e^1\}$, for the decay parameter, and values $d = 100$, $n \in \{500, 2000\}$. The data-dependent upper bounds on $\mathbf{V}$ and $\mathbf{B}$ are summarized in Table 1. More detailed plots are postponed to Figure 6 (in this figure, we plot the ordered eigenvalues and the spectral density for both the population and empirical covariances). Table 1 shows that as κ increases (a slower spectral decay), the implicit regularization parameter becomes larger, resulting in a decreasing variance and an increasing bias.

We also perform simulations to demonstrate the trade-off between bias and variance in the generalization error. The result is shown in Figure 2. For each choice of (n,d) pair, we vary the spectral decay of the kernel by changing gradually $\kappa \in [e^{-2}, e^2]$, and plot the generalization error on the log scale. We postpone the experiment details to Section 6.2, but point out important phenomena observed in Figures 2–3: (1) an extremely fast spectral decay (small κ) will generate insufficient implicit regularization that would hurt the generalization performance due to a large variance term; (2) a very slow spectral decay (large κ) will result

TABLE 1
Case $n > d$: variance bound **V** (4.1), *bias bound* **B** (4.2)

Spectral Decay	Implicit Reg	$n/d = 5$		$n/d = 20$	
		V	**B**	**V**	**B**
$\kappa = e^{-1}$	0.005418	14.2864	0.07898	9.4980	0.07891
$\kappa = e^{0}$	0.2525	0.4496	0.7535	0.1748	0.7538
$\kappa = e^{1}$	0.7501	0.1868	1.6167	0.05835	1.6165

in a large bias, which can also hurt the generalization performance; (3) certain favorable spectral decay achieves the best trade-off, resulting in the best generalization error.

We now theoretically demonstrate scalings within the $n > d$ regime when both **V** and **B** vanish. For simplicity, we consider Gaussian **x**.

COROLLARY 4.1 (General spectral decay: $n > d$). *Consider general eigenvalue decay with $\|\Sigma_d\|_{\mathrm{op}} \le 1$. Then with high probability,*

$$\mathbf{V} \precsim \frac{\mathrm{Tr}(\Sigma_d^{-1})}{n}, \qquad \mathbf{B} \precsim \mathbf{r} + \frac{\mathrm{Tr}(\Sigma_d)}{d}.$$

To illustrate the behavior of the estimates in Corollary 4.1, consider the following assumptions on the population covariance matrix.

EXAMPLE 4.1 (Low rank). Let $\Sigma_d = \mathrm{diag}(1, \dots, 1, 0, \dots, 0)$ with ϵd ones, $\epsilon \in (0, 1)$. In this case, $\mathbf{r} = \epsilon^2$ and $\lambda_j(XX^*/n) \ge (1 - \sqrt{\epsilon d/n})^2$ with high probability by standard results in random matrix theory. Then

$$\mathbf{V} \precsim \frac{\epsilon d}{n} \frac{(1 - \sqrt{\epsilon d/n})^2}{(\epsilon^2 d/n + (1 - \sqrt{\epsilon d/n})^2)^2} \asymp \frac{d}{n}\epsilon, \qquad \mathbf{B} \precsim \epsilon^2 + \epsilon.$$

Therefore, as $\epsilon \to 0$, both terms vanish for $n > d$.

EXAMPLE 4.2 (Approximately low rank). Let $\Sigma_d = \mathrm{diag}(1, \epsilon, \dots, \epsilon)$ for small $\epsilon > 0$. In this case, $\mathbf{r} = \epsilon^2$ and $\lambda_j(XX^*/n) \ge \epsilon(1 - \sqrt{d/n})^2$ with high probability. Then

$$\mathbf{V} \precsim \frac{d}{n} \frac{\epsilon(1 - \sqrt{d/n})^2}{(\epsilon^2 d/n + \epsilon(1 - \sqrt{d/n})^2)^2} \asymp \frac{d}{n}\frac{1}{\epsilon}, \qquad \mathbf{B} \precsim \epsilon^2 + \epsilon.$$

For instance, for $\epsilon \asymp (d/n)^{1/2}$, both terms vanish for $n \gg d$.

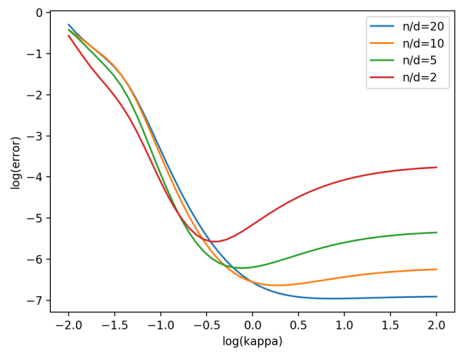

FIG. 2. *Generalization error as a function of varying spectral decay. Here, $d = 200$, $n = 400, 1000, 2000, 4000$.*

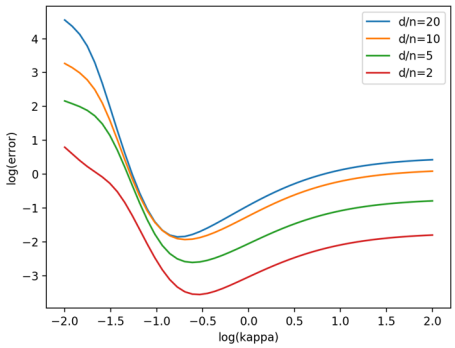

FIG. 3. *Generalization error as a function of varying spectral decay. Here*, $n = 200$, $d = 400, 1000, 2000, 4000$.

EXAMPLE 4.3 (Nonparametric slow decay). Consider $\lambda_j(\Sigma_d) = j^{-\alpha}$ for $0 < \alpha < 1$. Then $\mathbf{r} \asymp d^{-2\alpha}$. One can bound w.h.p. (see (B.4))

$$\mathbf{V} \asymp \frac{1}{n}\int_0^d t^\alpha \, dt \asymp \frac{d^{\alpha+1}}{n}, \qquad \mathbf{B} \precsim d^{-2\alpha} + d^{-\alpha}.$$

Balancing the two terms, one obtains a nonparametric upper bound $n^{-\frac{\alpha}{2\alpha+1}}$. A similar analysis can be carried out for $\alpha \geq 1$.

Case $d > n$. In this case, we can further bound the variance and the bias, with the choice $k = 0$, as

$$\mathbf{V} \precsim \frac{1}{d}\sum_{j=1}^{n} \frac{\lambda_j(\frac{XX^*}{d})}{[\mathbf{r} + \lambda_j(\frac{XX^*}{d})]^2}, \tag{4.3}$$

$$\mathbf{B} \precsim \frac{1}{n}\sum_{j=1}^{n} \lambda_j(K_X K_X^*) \asymp \mathbf{r} + \frac{1}{n}\sum_{j=1}^{n} \lambda_j\left(\frac{XX^*}{d}\right). \tag{4.4}$$

We first numerically illustrate the trade-off between the variance and the bias upper bounds. We consider three cases $\kappa \in \{e^{-1}, e^0, e^1\}$, and $d = 2000$, $n \in \{400, 100\}$. As before, we find a trade-off between $\mathbf{V}$ and $\mathbf{B}$ with varying κ; the results are summarized in Table 2. Additionally, Figure 7 provides a plot of the ordered eigenvalues, as well as spectral density for both the population and empirical covariances. As one can see, for a general eigenvalue decay, the spectral density of the population and the empirical covariance can be quite distinct. We again plot the generalization error in Figure 3 as a function of κ.

We now theoretically showcase an example in the $d \gg n$ regime where both $\mathbf{V}$ and $\mathbf{B}$ vanish. Again consider $\mathbf{x}$ being Gaussian for simplicity.

TABLE 2
Case $d > n$: variance bound $\mathbf{V}$ (4.3), *bias bound* $\mathbf{B}$ (4.4)

		$d/n = 5$		$d/n = 20$	
Spectral Decay	Implicit Reg	$\mathbf{V}$	$\mathbf{B}$	$\mathbf{V}$	$\mathbf{B}$
$\kappa = e^{-1}$	0.005028	3.9801	0.07603	0.7073	0.07591
$\kappa = e^0$	0.2503	0.1746	0.7513	0.04438	0.7502
$\kappa = e^1$	0.7466	0.06329	1.6106	0.01646	1.6102

COROLLARY 4.2 (General spectral decay: $d > n$). *With high probability, it holds that*

$$\mathbf{V} \precsim \frac{n}{d}\frac{1}{4\mathbf{r}}, \qquad \mathbf{B} \precsim \mathbf{r} + \frac{\operatorname{Tr}(\Sigma_d)}{d}.$$

The variance bound follows from the fact that $\frac{t}{(\mathbf{r}+t)^2} \le \frac{1}{4\mathbf{r}}$ *for all* t.

EXAMPLE 4.4 (Favorable spectral decay for $d \gg n$). Recall $\operatorname{Tr}(\Sigma_d)/d = \mathbf{r}^{1/2}$. With the choice $\mathbf{r} = (n/d)^{2/3}$, both terms vanish for $d \gg n$ as

$$\mathbf{V} \precsim \frac{n}{d}\frac{1}{4\mathbf{r}}, \qquad \mathbf{B} \precsim \mathbf{r}^{1/2}.$$

In this case, the spectrum satisfies $\operatorname{Tr}(\Sigma_d)/d = O((n/d)^{1/3})$.

5. Proofs. To prove Theorem 1, we decompose the mean square error into the bias and variance terms (Lemma 5.1), and provide data-dependent bound for each (Sections 5.2 and 5.3).

5.1. *Bias-variance decomposition*. The following is a standard bias-variance decomposition for an estimator. We remark that it is an equality, and both terms have to be small to ensure the desired convergence.

LEMMA 5.1. *The following decomposition for the interpolation estimator* (2.2) *holds*:

$$\mathbf{E}_{Y|X}\|\widehat{f} - f_*\|_{L^2_\mu}^2 = \mathbf{V} + \mathbf{B}, \tag{5.1}$$

where

$$\mathbf{V} := \int \mathbf{E}_{Y|X}\left|K_x^* K_X (K_X^* K_X)^{-1}(Y - \mathbf{E}[Y|X])\right|^2 d\mu(x), \tag{5.2}$$

$$\mathbf{B} := \int \left|K_x^*\left[K_X (K_X^* K_X)^{-1} K_X^* - I\right] f_*\right|^2 d\mu(x). \tag{5.3}$$

PROOF OF LEMMA 5.1. Recall the closed form solution of the interpolation estimator:

$$\widehat{f}(x) = K_x^* K_X (K_X^* K_X)^{-1} Y = K(x, X) K(X, X)^{-1} Y.$$

Define $E = Y - \mathbf{E}[Y|X] = Y - f_*(X)$. Since $\mathbf{E}_{Y|X} E = 0$, we have

$$\begin{aligned}
\widehat{f}(x) - f_*(x) &= K_x^* K_X (K_X^* K_X)^{-1} E + K_x^*\left[K_X (K_X^* K_X)^{-1} K_X^* - I\right] f_*, \\
\mathbf{E}_{Y|X}\left(\widehat{f}(x) - f_*(x)\right)^2 &= \mathbf{E}_{Y|X}\left(K_x^* K_X (K_X^* K_X)^{-1} E\right)^2 \\
&\quad + \left|K_x^*\left[K_X (K_X^* K_X)^{-1} K_X^* - I\right] f_*\right|^2.
\end{aligned}$$

Using Fubini's theorem,

$$\begin{aligned}
\mathbf{E}_{Y|X}\|\widehat{f} - f_*\|_{L^2_\mu}^2 &= \int \mathbf{E}_{Y|X}\left(\widehat{f}(x) - f_*(x)\right)^2 d\mu(x) \\
&= \int \mathbf{E}_{Y|X}\left|K_x^* K_X (K_X^* K_X)^{-1} E\right|^2 d\mu(x) \\
&\quad + \int \left|K_x^*\left[K_X (K_X^* K_X)^{-1} K_X^* - I\right] f_*\right|^2 d\mu(x).
\end{aligned}$$

□

5.2. *Variance.* In this section, we provide upper estimates on the variance part $\mathbf{V}$ in (5.2).

THEOREM 2 (Variance). *Let $\delta \in (0,1)$. Under the assumptions* (A.1)–(A.4), *with probability at least $1-\delta-d^{-2}$ with respect to a draw of X,*

$$\mathbf{V} \le \frac{8\sigma^2\|\Sigma_d\|}{d}\sum_j \frac{\lambda_j(\frac{XX^*}{d}+\frac{\alpha}{\beta}11^*)}{[\frac{\gamma}{\beta}+\lambda_j(\frac{XX^*}{d}+\frac{\alpha}{\beta}11^*)]^2} + \frac{8\sigma^2}{\gamma^2}d^{-(4\theta-1)}\log^{4.1} d, \tag{5.4}$$

for $\theta = \frac{1}{2}-\frac{2}{8+m}$ and for d large enough.

REMARK 5.1. Let us discuss the first term in equation (5.4) and its role in implicit regularization induced by the curvature of the kernel, eigenvalue decay and high dimensionality. In practice, the data matrix X is typically centered, so $1^*X = 0$. Therefore, the first term is effectively

$$\sum_j f_r\left(\lambda_j\left(\frac{XX^*}{d}\right)\right), \quad \text{where } f_r(t) := \frac{t}{(r+t)^2} \le \frac{1}{4r}.$$

This formula explains the effect of implicit regularization, and captures the "effective rank" of the training data X. We would like to emphasize that this measure of complexity is distinct from the classical notion of effective rank for regularized kernel regression [8], where the "effective rank" takes the form $\sum_j g_r(t_j)$ with $g_r(t) = t/(r+t)$, with t_j is the eigenvalue of the population integral operator $\mathcal{T}$.

PROOF OF THEOREM 2. From the definition of $\mathbf{V}$ and $E[Y|X] = f_*(X)$,

$$\begin{aligned}\mathbf{V} &= \int \mathbf{E}_{Y|X}\operatorname{Tr}\big(K_x^*K_X(K_X^*K_X)^{-1}(Y-f_*(X))(Y-f_*(X))^*\\ &\qquad \times (K_X^*K_X)^{-1}K_X^*K_x\big)\,d\mu(x)\\ &\le \int \big\|(K_X^*K_X)^{-1}K_X^*K_x\big\|^2\big\|\mathbf{E}_{Y|X}\big[(Y-f_*(X))(Y-f_*(X))^*\big]\big\|\,d\mu(x).\end{aligned}$$

Due to the fact that $\mathbf{E}_{Y|X}[(Y_i - f_*(X_i))(Y_j - f_*(X_j))] = 0$ for $i \ne j$, and $\mathbf{E}_{Y|X}[(Y_i - f_*(X_i))^2] \le \sigma^2$, we have that $\|\mathbf{E}_{Y|X}[(Y-f_*(X))(Y-f_*(X))^*]\| \le \sigma^2$, and thus

$$\mathbf{V} \le \sigma^2\int \big\|(K_X^*K_X)^{-1}K_X^*K_x\big\|^2\,d\mu(x) = \sigma^2\mathbf{E}_\mu\big\|K(X,X)^{-1}K(X,\mathbf{x})\big\|^2.$$

Let us introduce two quantities for the ease of derivation. For α, β, γ defined in (3.3), let

$$K^{\mathrm{lin}}(X,X) := \gamma I + \alpha 11^T + \beta\frac{XX^*}{d} \in \mathbb{R}^{n\times n}, \tag{5.5}$$

$$K^{\mathrm{lin}}(X,x) := \beta\frac{Xx^*}{d} \in \mathbb{R}^{n\times 1}, \tag{5.6}$$

and $K^{\mathrm{lin}}(x,X)$ being the transpose of $K^{\mathrm{lin}}(X,x)$. By Proposition A.2, with probability at least $1-\delta-d^{-2}$, for $\theta = \frac{1}{2}-\frac{2}{8+m}$ the following holds:

$$\big\|K(X,X) - K^{\mathrm{lin}}(X,X)\big\| \le d^{-\theta}\big(\delta^{-1/2}+\log^{0.51} d\big).$$

As a direct consequence, one can see that

$$\big\|K(X,X)^{-1}\big\| \le \frac{1}{\gamma - d^{-\theta}(\delta^{-1/2}+\log^{0.51} d)} \le \frac{2}{\gamma}, \tag{5.7}$$

$$\begin{aligned}\|K(X,X)^{-1}K^{\mathrm{lin}}(X,X)\| &\le 1+\|K(X,X)^{-1}\|\\ &\quad\cdot\|K(X,X)-K^{\mathrm{lin}}(X,X)\|\\ &\le \frac{\gamma}{\gamma-d^{-\theta}(\delta^{-1/2}+\log^{0.51}d)}\le 2,\end{aligned}\tag{5.8}$$

provided d is large enough, in the sense that

$$d^{-\theta}\big(\delta^{-1/2}+\log^{0.51}d\big)\le\gamma/2.$$

By Lemma B.2 (for Gaussian case, Lemma B.1),

$$\mathbf{E}_\mu\|K(\mathbf{x},X)-K^{\mathrm{lin}}(\mathbf{x},X)\|^2\le d^{-(4\theta-1)}\log^{4.1}d.\tag{5.9}$$

Let us proceed with the bound

$$\begin{aligned}\mathbf{V}&\le\sigma^2\mathbf{E}_\mu\|K(X,X)^{-1}K(X,\mathbf{x})\|^2\\ &\le 2\sigma^2\mathbf{E}_\mu\|K(X,X)^{-1}K^{\mathrm{lin}}(X,\mathbf{x})\|^2\\ &\quad+2\sigma^2\|K(X,X)^{-1}\|^2\cdot\mathbf{E}_\mu\|K(X,\mathbf{x})-K^{\mathrm{lin}}(X,\mathbf{x})\|^2\\ &\le 2\sigma^2\|K(X,X)^{-1}K^{\mathrm{lin}}(X,X)\|^2\mathbf{E}_\mu\|K^{\mathrm{lin}}(X,X)^{-1}K^{\mathrm{lin}}(X,\mathbf{x})\|^2\\ &\quad+\frac{8\sigma^2}{\gamma^2}d^{-(4\theta-1)}\log^{4.1}d\\ &\le 8\sigma^2\mathbf{E}_\mu\|K^{\mathrm{lin}}(X,X)^{-1}K^{\mathrm{lin}}(X,\mathbf{x})\|^2+\frac{8\sigma^2}{\gamma^2}d^{-(4\theta-1)}\log^{4.1}d,\end{aligned}$$

where the the third inequality relies on (5.9) and (5.7), and the fourth inequality follows from (5.8).

One can further show that

$$\begin{aligned}&\mathbf{E}_\mu\|K^{\mathrm{lin}}(X,X)^{-1}K^{\mathrm{lin}}(X,\mathbf{x})\|^2\\ &\quad=\mathbf{E}_\mu\operatorname{Tr}\Bigg(\bigg[\gamma I+\alpha 11^*+\beta\frac{XX^*}{d}\bigg]^{-1}\\ &\qquad\times\beta\frac{X\mathbf{x}}{d}\beta\frac{\mathbf{x}^*X^*}{d}\bigg[\gamma I+\alpha 11^*+\beta\frac{XX^*}{d}\bigg]^{-1}\Bigg)\\ &\quad=\operatorname{Tr}\Bigg(\bigg[\gamma I+\alpha 11^*+\beta\frac{XX^*}{d}\bigg]^{-1}\\ &\qquad\times\beta^2\frac{X\Sigma_dX^*}{d^2}\bigg[\gamma I+\alpha 11^*+\beta\frac{XX^*}{d}\bigg]^{-1}\Bigg)\\ &\quad\le\frac{1}{d}\|\Sigma_d\|\operatorname{Tr}\Bigg(\bigg[\gamma I+\alpha 11^*+\beta\frac{X^*X}{d}\bigg]^{-1}\\ &\qquad\times\beta^2\frac{X^*X}{d}\bigg[\gamma I+\alpha 11^*+\beta\frac{X^*X}{d}\bigg]^{-1}\Bigg)\\ &\quad\le\frac{1}{d}\|\Sigma_d\|\operatorname{Tr}\Bigg(\bigg[\gamma I+\alpha 11^*+\beta\frac{X^*X}{d}\bigg]^{-1}\end{aligned}$$

$$\times \left[\beta^2 \frac{X^*X}{d} + \alpha\beta 11^*\right]\left[\gamma I + \alpha 11^* + \beta\frac{X^*X}{d}\right]^{-1}\Big)$$

$$= \frac{1}{d}\|\Sigma_d\| \sum_j \frac{\lambda_j(\frac{XX^*}{d} + \frac{\alpha}{\beta}11^*)}{[\frac{\gamma}{\beta} + \lambda_j(\frac{XX^*}{d} + \frac{\alpha}{\beta}11^*)]^2}.$$

We conclude that with probability at least $1 - \delta - d^{-2}$,

$$\mathbf{V} \le 8\sigma^2 \mathbf{E}_\mu \left\| K^{\text{lin}}(X, X)^{-1} K^{\text{lin}}(X, \mathbf{x}) \right\|^2 + \frac{8\sigma^2}{\gamma^2} d^{-(4\theta-1)} \log^{4.1} d \tag{5.10}$$

$$\le \frac{8\sigma^2\|\Sigma_d\|}{d} \sum_j \frac{\lambda_j(\frac{XX^*}{d} + \frac{\alpha}{\beta}11^*)}{[\frac{\gamma}{\beta} + \lambda_j(\frac{XX^*}{d} + \frac{\alpha}{\beta}11^*)]^2} + \frac{8\sigma^2}{\gamma^2} d^{-(4\theta-1)} \log^{4.1} d \tag{5.11}$$

for d large enough. □

5.3. *Bias*.

THEOREM 3 (Bias). *Let $\delta \in (0, 1)$. The bias, under the only assumptions that $K(x, x) \le M$ for $x \in \Omega$, and X_i's are i.i.d. random vectors, is upper bounded as*

$$\mathbf{B} \le \|f_*\|_{\mathcal{H}}^2 \cdot \inf_{0\le k\le n} \left\{ \frac{1}{n}\sum_{j>k} \lambda_j\big(K(X, X)\big) + 2\sqrt{\frac{k}{n}}\sqrt{\frac{\sum_{i=1}^n K(x_i, x_i)^2}{n}} \right\} + 3M\sqrt{\frac{\log 2n/\delta}{2n}}, \tag{5.12}$$

with probability at least $1 - \delta$.

PROOF OF THEOREM 3. In this proof, when there is no confusion, we use $f(x) = \sum_{i=1}^p e_i(x) f_i$ where f_i denotes the coefficients of f under the basis $e_i(x)$. Adopting this notation, we can write $f(x) = e(x)^* f$ where $f = [f_1, f_2, \ldots, f_p]^T$ also denotes a possibly infinite vector. For the bias, it is easier to work in the frequency domain using the spectral decomposition. Recalling the spectral characterization in the preliminary section,

$$\begin{aligned}
\mathbf{B} &= \int \left|e^*(x)T^{1/2}\big[T^{1/2}e(X)\big(e(X)^*Te(X)\big)^{-1}e(X)^*T^{1/2} - I\big]T^{-1/2}f_*\right|^2 d\mu(x) \\
&\le \int \left\|\big[T^{1/2}e(X)\big(e(X)^*Te(X)\big)^{-1}e(X)^*T^{1/2} - I\big]T^{1/2}e(x)\right\|^2 d\mu(x) \\
&\quad \cdot \left\|T^{-1/2}f_*\right\|^2 \\
&= \|f_*\|_{\mathcal{H}}^2 \int \left\|\big[T^{1/2}e(X)\big(e(X)^*Te(X)\big)^{-1}e(X)^*T^{1/2} - I\big]T^{1/2}e(x)\right\|^2 d\mu(x).
\end{aligned}$$

Here, we use the fact that $T^{-1/2}f_* = \sum_i t_i^{-1/2} f_{*,i} e_i$ and $\|T^{-1/2}f_*\|^2 = \sum_i f_{*,i}^2/t_i = \|f_*\|_{\mathcal{H}}^2$. Next, recall the empirical Kernel operator with its spectral decomposition $\widehat{T} = \widehat{U}\widehat{\Lambda}\widehat{U}^*$, with $\widehat{\Lambda}_{jj} = \frac{1}{n}\lambda_j(K(X, X))$. Denote the top k columns of $\widehat{U}$ to be $\widehat{U}_k$, and $P_{\widehat{U}_k}^{\perp}$ to be projection to the eigenspace orthogonal to $\widehat{U}_k$. By observing that $T^{1/2}e(X)(e(X)^*Te(X))^{-1}e(X)^*T^{1/2}$ is a projection matrix, it is clear that for all $k \le n$,

$$\begin{aligned}
\mathbf{B} &\le \|f_*\|_{\mathcal{H}}^2 \int \left\|P_{\widehat{U}}^{\perp}\big(T^{1/2}e(x)\big)\right\|^2 d\mu(x) \\
&\le \|f_*\|_{\mathcal{H}}^2 \int \left\|P_{\widehat{U}_k}^{\perp}\big(T^{1/2}e(x)\big)\right\|^2 d\mu(x).
\end{aligned} \tag{5.13}$$

We continue the study of the last quantity using techniques inspired by [24]. Denote the function g indexed by any rank-k projection U_k as

$$g_{U_k}(x) := \|P_{U_k}(T^{1/2}e(x))\|^2 = \operatorname{Tr}(e^*(x)T^{1/2}U_kU_k^TT^{1/2}e(x)). \tag{5.14}$$

Clearly, $\|U_kU_k^T\|_F = \sqrt{k}$. Define the function class

$$\mathcal{G}_k := \{g_{U_k}(x) : U_k^TU_k = I_k\}.$$

It is clear that $g_{\widehat{U}_k} \in \mathcal{G}_k$. Observe that $g_{\widehat{U}_k}$ is a random function that depends on the data X, and we will bound the bias term using the empirical process theory. It is straightforward to verify that

$$\begin{aligned}
\mathbf{E}_{\mathbf{x}\sim\mu}\|P_{\widehat{U}_k}^{\perp}(T^{1/2}e(\mathbf{x}))\|^2 &= \int \|P_{\widehat{U}_k}^{\perp}(T^{1/2}e(x))\|^2\,d\mu(x),\\
\widehat{\mathbf{E}}_n\|P_{\widehat{U}_k}^{\perp}(T^{1/2}e(\mathbf{x}))\|^2 &= \frac{1}{n}\sum_{i=1}^n\|P_{\widehat{U}_k}^{\perp}(T^{1/2}e(x_i))\|^2\\
&= \operatorname{Tr}(P_{\widehat{U}_k}^{\perp}\widehat{T}P_{\widehat{U}_k}^{\perp}) = \sum_{j>k}\widehat{\Lambda}_{jj} = \frac{1}{n}\sum_{j>k}\lambda_j(K(X,X)).
\end{aligned}$$

Using symmetrization Lemma B.4 with $M = \sup_{x\in\Omega}K(x,x)$, with probability at least $1-2\delta$,

$$\begin{aligned}
&\int \|P_{\widehat{U}_k}^{\perp}(T^{1/2}e(x))\|^2\,d\mu(x) - \frac{1}{n}\sum_{j>k}\lambda_j(K(X,X))\\
&= \mathbf{E}_{\mu}\|P_{\widehat{U}_k}^{\perp}(T^{1/2}e(\mathbf{x}))\|^2 - \widehat{\mathbf{E}}_n\|P_{\widehat{U}_k}^{\perp}(T^{1/2}e(\mathbf{x}))\|^2\\
&\le \sup_{U_k:U_k^TU_k=I_k}(\mathbf{E}-\widehat{\mathbf{E}}_n)\|P_{U_k}^{\perp}(T^{1/2}e(\mathbf{x}))\|^2\\
&\le 2\mathbf{E}_{\epsilon}\sup_{U_k:U_k^TU_k=I_k}\frac{1}{n}\sum_{i=1}^n\epsilon_i\big(\|T^{1/2}e(x_i)\|^2 - \|P_{U_k}(T^{1/2}e(x_i))\|^2\big)\\
&\quad + 3M\sqrt{\frac{\log 1/\delta}{2n}}
\end{aligned}$$

by the Pythagorean theorem. Since ϵ_i's are symmetric and zero-mean and $\|T^{1/2}e(x_i)\|^2$ does not depend on U_k, the last expression is equal to

$$2\mathbf{E}_{\epsilon}\sup_{g\in\mathcal{G}_k}\frac{1}{n}\sum_{i=1}^n\epsilon_ig(x_i) + 3M\sqrt{\frac{\log 1/\delta}{2n}}.$$

We further bound the Rademacher complexity of the set $\mathcal{G}_k$,

$$\begin{aligned}
\mathbf{E}_{\epsilon}\sup_{g\in\mathcal{G}_k}\frac{1}{n}\sum_{i=1}^n\epsilon_ig(x_i) &= \mathbf{E}_{\epsilon}\sup_{U_k}\frac{1}{n}\sum_{i=1}^n\epsilon_ig_{U_k}(x_i)\\
&= \mathbf{E}_{\epsilon}\frac{1}{n}\sup_{U_k}\Big\langle U_kU_k^T, \sum_{i=1}^n\epsilon_iT^{1/2}e(x_i)e^*(x_i)T^{1/2}\Big\rangle\\
&\le \frac{\sqrt{k}}{n}\mathbf{E}_{\epsilon}\Big\|\sum_{i=1}^n\epsilon_iT^{1/2}e(x_i)e^*(x_i)T^{1/2}\Big\|_F
\end{aligned}$$

by the Cauchy–Schwarz inequality and the fact that $\|U_k U_k^T\|_F \le \sqrt{k}$. The last expression is can be further evaluated by the independence of ϵ_i's,

$$\begin{aligned}
&\frac{\sqrt{k}}{n}\left\{\mathbf{E}_\epsilon\left\|\sum_{i=1}^n \epsilon_i T^{1/2} e(x_i)e^*(x_i)T^{1/2}\right\|_F^2\right\}^{1/2} \\
&\quad = \frac{\sqrt{k}}{n}\left\{\sum_{i=1}^n \|T^{1/2}e(x_i)e^*(x_i)T^{1/2}\|_F^2\right\}^{1/2} \\
&\quad = \sqrt{\frac{k}{n}}\sqrt{\frac{\sum_{i=1}^n K(x_i,x_i)^2}{n}}.
\end{aligned}$$

Therefore, for all $k \le n$, with probability at least $1 - 2n\delta$,

$$\begin{aligned}
\mathbf{B} \le \|f_*\|_{\mathcal{H}}^2 \cdot \inf_{0\le k\le n}\Bigg\{&\frac{1}{n}\sum_{j>k}\lambda_j\big(K(X,X)\big) + 2\sqrt{\frac{k}{n}}\sqrt{\frac{\sum_{i=1}^n K(x_i,x_i)^2}{n}} \\
&+ 3M\sqrt{\frac{\log 1/\delta}{2n}}\Bigg\}.
\end{aligned}$$

□

REMARK 5.2. Let us compare the bounds obtained in this paper to those one can obtain for classification with a margin. For classification, Theorem 21 in [2] shows that the misclassification error is upper bounded with probability at least $1 - \delta$ as

$$\mathbf{E}\mathbf{1}\big(\mathbf{y}\widehat{f}(\mathbf{x}) < 0\big) \le \mathbf{E}\phi_\gamma\big(\mathbf{y}\widehat{f}(\mathbf{x})\big) \le \widehat{\mathbf{E}}_n\phi_\gamma\big(\mathbf{y}\widehat{f}(\mathbf{x})\big) + \frac{C_\delta}{\gamma\sqrt{n}}\sqrt{\frac{\sum_{i=1}^n K(x_i,x_i)}{n}},$$

where $\phi_\gamma(t) := \max(0, 1 - t/\gamma) \wedge 1$ is the margin loss surrogate for the indicator loss $\mathbf{1}(t < 0)$. By tuning the margin γ, one obtains a family of upper bounds.

Now consider the noiseless regression scenario (i.e., $\sigma = 0$ in (A.1)). In this case, the variance contribution to the risk is zero, and

$$\begin{aligned}
\mathbf{E}_{Y|X}\|\widehat{f} - \mathbf{y}\|_{L_\mu^2}^2 &= \mathbf{E}_{Y|X}\|\widehat{f} - f_*\|_{L_\mu^2}^2 \\
&= \mathbf{E}\big[P_n^\perp f_*\big]^2 \\
&\le \mathbf{E}\big[P_k^\perp f_*\big]^2 \\
&\le \widehat{\mathbf{E}}_n\big[P_k^\perp f_*\big]^2 + C'_\delta\sqrt{\frac{k}{n}}\sqrt{\frac{\sum_{i=1}^n K(x_i,x_i)^2}{n}},
\end{aligned}$$

where P_k is the best-rank k projection (based on X) and $P_k^\perp$ denotes its orthogonal projection. By tuning the parameter k (similar as the $1/\gamma$ in classification), one can balance the RHS to obtain the optimal trade-off.

However, classification is easier than regression in the following sense: $\widehat{f}$ can present a nonvanishing bias in estimating f_*, but as long as the bias is below the empirical margin level, it plays no effect in the margin loss $\phi_\gamma(\cdot)$. In fact, for classification, under certain conditions, one can prove exponential convergence for the generalization error [18].

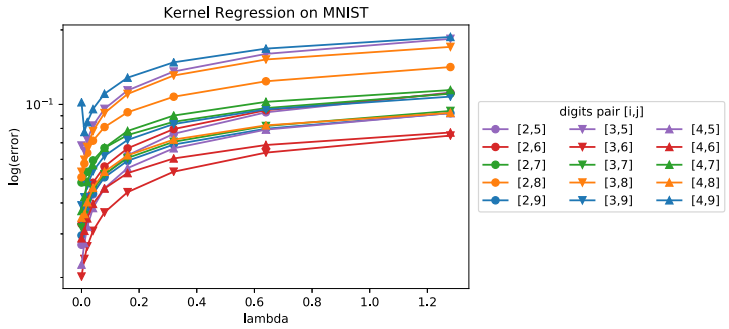

FIG. 4. *Test error, normalized as in* (6.1). *The y-axis is on the log scale.*

6. Experiments.

6.1. *MNIST.* In this section, we provide full details of the experiments on MNIST [19]. Our first experiment considers the following problem: for each pair of distinct digits (i, j), $i, j \in \{0, 1, \ldots, 9\}$, label one digit as 1 and the other as -1, then fit the Kernel Ridge Regression with Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2/d)$, where $d = 784$ is the dimension as analyzed in our theory (also the default choice in the Scikit-learn package [23]). For each of the $\binom{10}{2} = 45$ pairs of experiments, we chose $\lambda = 0$ (no regularization, interpolation estimator), $\lambda = 0.1$ and $\lambda = 1$. We evaluated the performance on the *out-of-sample* test dataset, with the error metric

$$\frac{\sum_i (\widehat{f}(x_i) - y_i)^2}{\sum_i (\bar{y} - y_i)^2}. \tag{6.1}$$

Remarkably, among all 45 experiments, no-regularization performs the best. We refer to the table in Section B for a complete list of numerical results. For each experiment, the sample size is roughly $n \approx 10{,}000$.

The second experiment is to perform the similar task on a finer grid of regularization parameter $\lambda \in \{0, 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64, 1.28\}$. Again, in all but one pair, the interpolation estimator performs the best in out-of-sample prediction. We refer to Figure 4 for details.

To conclude this experiment, we plot the eigenvalue decay of the empirical kernel matrix and the sample covariance matrix for the 5 experiments shown in the Introduction. The two plots are shown in Figure 5. Both plots exhibit a fast decay of eigenvalues, supporting the theoretical finding that interpolation performs well on a test set in such situations.

On the other hand, it is easy to construct examples where the eigenvalues do not decay and interpolation performs poorly. This is the case, for instance, if X_i are i.i.d. from spherical

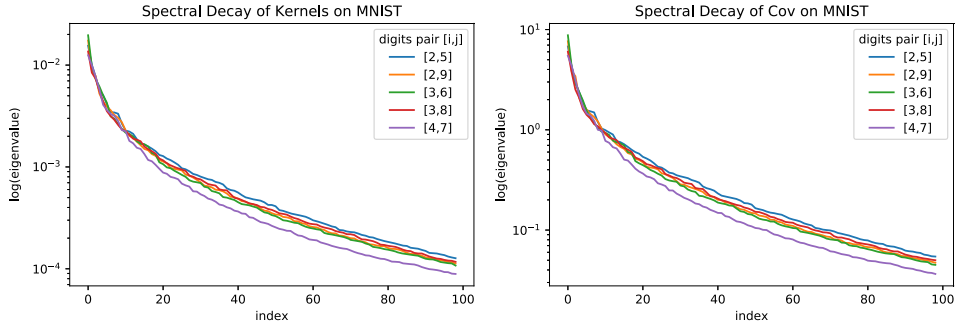

FIG. 5. *Spectral decay. The y-axis is on the log scale.*

Gaussian. One can show that in the high-dimensional regime, the variance term itself (and not just the upper bound on it) is large. Since the bias-variance decomposition is an equality, it is not possible to establish good L^2_μ convergence.

6.2. *A synthetic example.* In this section, we provide the details of the synthetic experiments mentioned in Section 4 for Tables 1–2 and Figures 2–3. We choose the RBF kernel as the nonlinearity with $h(t) = \exp(-t)$. Again, we consider a family of eigenvalue decays for the covariance matrix parametrized by κ, with the small κ describing fast spectral decay

$$\lambda_j(\Sigma_{d,\kappa}) = \left(1 - \left((j-1)/d\right)^\kappa\right)^{1/\kappa}, \quad 1 \le j \le d.$$

We set a target nonlinear function f_* in the RKHS with kernel $K(x, x') = h(\|x - x'\|^2/d)$ as

$$f_*(x) = \sum_{l=1}^{100} K(x, \theta_l), \quad \theta_l \overset{\text{i.i.d.}}{\sim} N(0, I_d).$$

For each parameter triplet (n, d, κ), we generate data in the following way:

$$x_i \sim N(0, \Sigma_{d,\kappa}), \qquad y_i = f_*(x_i) + \epsilon_i$$

for $1 \le i \le n$ where $\epsilon_i \sim N(0, \sigma^2)$ is independent noise, with $\sigma = 0.1$ (Figures 2–3) and $\sigma = 0.5$ (Figures 8). Figures 6–7 contrast the difference between the population and empirical eigenvalues for various parameter triplets (n, d, κ).

We now explain Figures 2–3, which illustrate the true generalization error in this synthetic example, by varying the spectral decay κ, for a particular case of high dimensionality ratio d/n. Here, we plot the *out-of-sample* test error for the interpolated min-norm estimator $\widehat{f}$ on fresh new test data (x_t, y_t) from the same data generating process, with the error metric

$$\text{error} = \frac{\sum_t (\widehat{f}(x_t) - f_*(x_t))}{\sum_t (y_t - \bar{y})^2}.$$

The error plots are shown in Figure 2 (for $n > d$) and 3 (for $d > n$), and Figure 8 for the high noise case. On the x-axis, we plot the $\log(\kappa)$, and on the y-axis the log(error). Each curve corresponds to the generalization error behavior (and the bias and variance trade-off) as we vary spectral decay from fast to slow (as κ increases) for a particular choice of d/n or n/d ratio. Clearly, for a general pair of high dimensionality ratio d/n, there is a "sweet spot" of κ (favorable geometric structure) such that the trade-off is optimized.

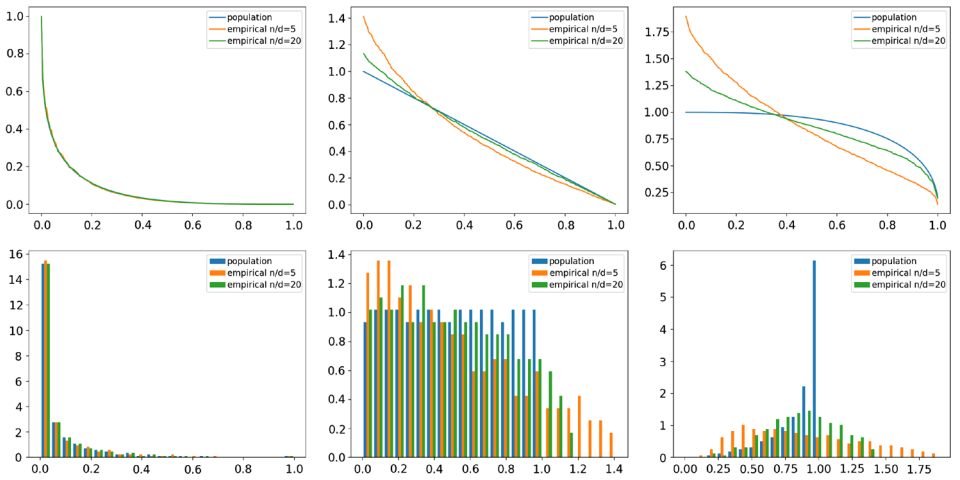

FIG. 6. *Varying spectral decay*: *case* $n > d$. *Columns from left to right*: $\kappa = e^{-1}, e^0, e^1$. *Rows from top to bottom*: *ordered eigenvalues, and the histogram of eigenvalues. Here, we plot the population eigenvalues for* Σ_d, *and the empirical eigenvalues for* X^*X/n. *In this simulation,* $d = 100$, $n = 500, 2000$.

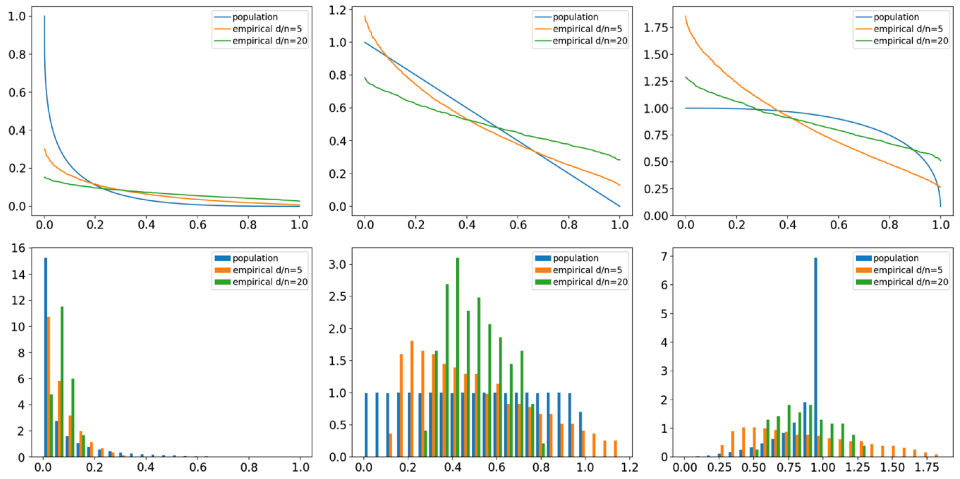

FIG. 7. *Varying spectral decay*: *case* $d > n$. *Columns from left to right*: $\kappa = e^{-1}, e^{0}, e^{1}$. *Rows from top to bottom*: *ordered eigenvalues*, *and the histogram of eigenvalues*. *Here*, *we plot the population eigenvalues for* Σ_d, *and the empirical eigenvalues for* XX^*/d. *In this simulation*, $d = 2000$, $n = 400, 100$.

7. Further discussion. This paper is motivated by the work of [5] and [31], who, among others, observed the good out-of-sample performance of interpolating rules. This paper continues the line of work in [3, 4, 6] on understanding theoretical mechanisms for the good out-of-sample performance of interpolation. We leave further investigations on the connection between kernel ridgeless regression and two-layer neural networks as a future work [12].

From an algorithmic point of view, the minimum-norm interpolating solution can be found either by inverting the kernel matrix, or by performing gradient descent on the least-squares objective (starting from 0). Our analysis can then be viewed in the light of recent work on implicit regularization of optimization procedures [16, 20, 22, 29].

The paper also highlights a novel type of implicit regularization. In addition, we discover that once we parametrize the geometric properties—the spectral decay—we discover the familiar picture of the bias-variance trade-off, controlled by the implicit regularization that adapts to the favorable geometric property of the data. Moreover, if one explicitly parametrizes the choice of the kernel by, say, the bandwidth, we are likely to see the familiar picture of the bias-variance trade-off, despite the fact that the estimator is always interpolating. Whether one can achieve optimal rates of estimation (under appropriate assumptions) for the right choice of the bandwidth appears to be an interesting and difficult statistical question. Another open question is whether one can characterize situations when the interpolating

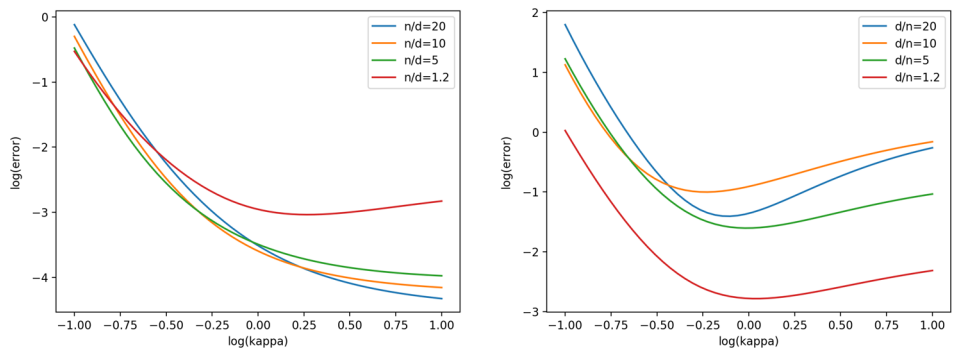

FIG. 8. *Varying spectral decay*: *generalization error for high noise case*. *Left*: $d = 200$, $n = 4000, 2000, 1000, 240$. *Right*: $n = 200$, $d = 4000, 2000, 1000, 240$.

minimum-norm solution is dominating the regularized solution in terms of expected performance.

Acknowledgments. The first author was supported by the George C. Tiao faculty fellowship.

The second author was supported in part by the NSF under Grant no. CDS&E-MSS 1521529, the DARPA Lagrange program, and by the MIT-Sensetime Alliance.

SUPPLEMENTARY MATERIAL

Supplement to: "Just interpolate: Kernel "Ridgeless" regression can generalize" (DOI: 10.1214/19-AOS1849SUPP; .pdf). Due to space constraints, we have relegated remaining proofs to the Supplement.

REFERENCES

[1] ALVAREZ, M. A., ROSASCO, L. and LAWRENCE, N. D. (2012). Kernels for vector-valued functions: A review. *Found. Trends Mach. Learn.* **4** 195–266.

[2] BARTLETT, P. L. and MENDELSON, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** 463–482. MR1984026 https://doi.org/10.1162/153244303321897690

[3] BELKIN, M. (2018). Approximation beats concentration? an approximation view on inference with smooth radial kernels. ArXiv preprint. Available at arXiv:1801.03437.

[4] BELKIN, M., HSU, D. and MITRA, P. (2018). Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. ArXiv preprint. Available at arXiv:1806.05161.

[5] BELKIN, M., MA, S. and MANDAL, S. (2018). To understand deep learning we need to understand kernel learning. ArXiv preprint. Available at arXiv:1802.01396.

[6] BELKIN, M., RAKHLIN, A. and TSYBAKOV, A. B. (2018). Does data interpolation contradict statistical optimality? ArXiv preprint. Available at arXiv:1806.09471.

[7] BOSE, A., CHATTERJEE, S. and GANGOPADHYAY, S. (2003). Limiting spectral distributions of large dimensional random matrices. *J. Indian Statist. Assoc.* **41** 221–259. MR2101995

[8] CAPONNETTO, A. and DE VITO, E. (2007). Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **7** 331–368. MR2335249 https://doi.org/10.1007/s10208-006-0196-8

[9] CRESSIE, N. (1990). The origins of kriging. *Math. Geol.* **22** 239–252. MR1047810 https://doi.org/10.1007/BF00889887

[10] CUCKER, F. and SMALE, S. (2002). Best choices for regularization parameters in learning theory: On the bias-variance problem. *Found. Comput. Math.* **2** 413–428. MR1930945 https://doi.org/10.1007/s102080010030

[11] DE VITO, E., CAPONNETTO, A. and ROSASCO, L. (2005). Model selection for regularized least-squares algorithm in learning theory. *Found. Comput. Math.* **5** 59–85. MR2125691 https://doi.org/10.1007/s10208-004-0134-1

[12] DOU, X. and LIANG, T. (2019). Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. ArXiv preprint. Available at arXiv:1901.07114.

[13] EL KAROUI, N. (2010). The spectrum of kernel random matrices. *Ann. Statist.* **38** 1–50. MR2589315 https://doi.org/10.1214/08-AOS648

[14] EVGENIOU, T., PONTIL, M. and POGGIO, T. (2000). Regularization networks and support vector machines. *Adv. Comput. Math.* **13** 1–50. MR1759187 https://doi.org/10.1023/A:1018946025316

[15] GOLUB, G. H., HEATH, M. and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223. MR0533250 https://doi.org/10.2307/1268518

[16] GUNASEKAR, S., WOODWORTH, B. E., BHOJANAPALLI, S., NEYSHABUR, B. and SREBRO, N. (2017). Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems* 6151–6159.

[17] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. *Springer Series in Statistics*. Springer, New York. MR1920390 https://doi.org/10.1007/b97848

[18] KOLTCHINSKII, V. and BEZNOSOVA, O. (2005). Exponential convergence rates in classification. In *Learning Theory*. *Lecture Notes in Computer Science* **3559** 295–307. Springer, Berlin. MR2203269 https://doi.org/10.1007/11503415_20

[19] LeCun, Y., Cortes, C. and Burges, C. J. (2010). Mnist handwritten digit database. AT&T Labs [Online]. Available at http://yann.Lecun.Com/exdb/mnist.
[20] Li, Y., Ma, T. and Zhang, H. (2017). Algorithmic regularization in over-parameterized matrix recovery. ArXiv preprint. Available at arXiv:1712.09203.
[21] Liang, T. and Rakhlin, A. (2019). Supplement to "Just interpolate: Kernel "Ridgeless" regression can generalize". https://doi.org/10.1214/19-AOS1849SUPP.
[22] Neyshabur, B., Tomioka, R. and Srebro, N. (2014). In search of the real inductive bias: On the role of implicit regularization in deep learning. ArXiv preprint. Available at arXiv:1412.6614.
[23] Pedregosa, F., Varoquaux, G., Gramfort, A. et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12** 2825–2830. MR2854348
[24] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.
[25] Smola, A. J. and Schölkopf, B. (1998). *Learning with Kernels* **4**. Citeseer.
[26] Vapnik, V. N. (1998). *Statistical Learning Theory*. *Adaptive and Learning Systems for Signal Processing, Communications, and Control*. Wiley, New York. MR1641250
[27] Vovk, V. (2013). Kernel ridge regression. In *Empirical Inference* 105–116. Springer, Heidelberg. MR3236860 https://doi.org/10.1007/978-3-642-41136-6_11
[28] Wahba, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442 https://doi.org/10.1137/1.9781611970128
[29] Yao, Y., Rosasco, L. and Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constr. Approx.* **26** 289–315. MR2327601 https://doi.org/10.1007/s00365-006-0663-2
[30] Yin, Y. Q., Bai, Z. D. and Krishnaiah, P. R. (1988). On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probab. Theory Related Fields* **78** 509–521. MR0950344 https://doi.org/10.1007/BF00353874
[31] Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. ArXiv preprint. Available at arXiv:1611.03530.