The Annals of Statistics
2017, Vol. 45, No. 3, 991–1023
DOI: 10.1214/16-AOS1472

RANDOMIZED SKETCHES FOR KERNELS: FAST AND OPTIMAL NONPARAMETRIC REGRESSION[1]

BY YUN YANG*, MERT PILANCI[†,2] AND MARTIN J. WAINWRIGHT[†]

*Florida State University** *and University of California, Berkeley*[†]

Kernel ridge regression (KRR) is a standard method for performing nonparametric regression over reproducing kernel Hilbert spaces. Given n samples, the time and space complexity of computing the KRR estimate scale as $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$, respectively, and so is prohibitive in many cases. We propose approximations of KRR based on m-dimensional randomized sketches of the kernel matrix, and study how small the projection dimension m can be chosen while still preserving minimax optimality of the approximate KRR estimate. For various classes of randomized sketches, including those based on Gaussian and randomized Hadamard matrices, we prove that it suffices to choose the sketch dimension m proportional to the statistical dimension (modulo logarithmic factors). Thus, we obtain fast and minimax optimal approximations to the KRR estimate for nonparametric regression. In doing so, we prove a novel lower bound on the minimax risk of kernel regression in terms of the localized Rademacher complexity.

1. Introduction. The goal of nonparametric regression is to make predictions of a response variable $Y \in \mathbb{R}$ based on observing a covariate vector $X \in \mathcal{X}$. In practice, we are given a collection of n samples, say $\{(x_i, y_i)\}_{i=1}^n$ of covariate-response pairs and our goal is to estimate the regression function $f^*(x) = \mathbb{E}[Y|X = x]$. In the standard Gaussian model, it is assumed that the covariate-response pairs are related via the model

$$y_i = f^*(x_i) + \sigma w_i \qquad \text{for } i = 1, \ldots, n, \tag{1}$$

where the sequence $\{w_i\}_{i=1}^n$ consists of i.i.d. standard Gaussian variates. It is typical to assume that the regression function f^* has some regularity properties, and one way of enforcing such structure is to require f^* to belong to a reproducing kernel Hilbert space, or RKHS for short [4, 12, 33]. Given such an assumption, it is natural to estimate f^* by minimizing a combination of the least-squares fit to the data and a penalty term involving the squared Hilbert norm, leading to an

Received September 2015; revised April 2016.

[1]Supported in part by Office of Naval Research MURI Grant N00014-11-1-0688, National Science Foundation Grants CIF-31712-23800 and DMS-11-07000, and Air Force Office of Scientific Research Grant AFOSR-FA9550-14-1-0016.

[2]Supported in part by a Microsoft Research Fellowship.

MSC2010 subject classifications. Primary 62G08; secondary 68W20.

Key words and phrases. Nonparametric regression, random projection, kernel method, dimensionality reduction, convex optimization.

estimator known as *kernel ridge regression*, or KRR for short [14, 28]. From a statistical point of view, the behavior of KRR can be characterized using existing results on M-estimation and empirical processes (e.g., [17, 22, 32]). When the regularization parameter is set appropriately, it is known to yield a function estimate with minimax prediction error for various classes of kernels.

Despite these attractive statistical properties, the computational complexity of computing the KRR estimate prevents it from being routinely used in large-scale problems. More precisely, in a standard implementation [27], the time complexity and space complexity of KRR scales as $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$, respectively, where n refers to the number of samples. As a consequence, it becomes important to design methods for computing approximate forms of the KRR estimate, while retaining guarantees of optimality in terms of statistical minimaxity. Various authors have taken different approaches to this problem. Zhang et al. [35] analyze a distributed implementation of KRR, in which a set of t machines each compute a separate estimate based on a random t-way partition of the full data set, and combine it into a global estimate by averaging. This divide-and-conquer approach has time complexity and space complexity $\mathcal{O}(n^3/t^3)$ and $\mathcal{O}(n^2/t^2)$, respectively. Zhang et al. [35] give conditions on the number of splits t, as a function of the kernel, under which minimax optimality of the resulting estimator can be guaranteed.

In this paper, we consider approximations to KRR based on random projections, also known as sketches, of the data. The random projection method is classical way of performing dimensionality reduction; see the papers [13, 20] and references therein for its uses in numerical linear algebra and low-rank approximation. In this spirit, our proposal is to approximate n-dimensional kernel matrix by projecting its row and column subspaces to a randomly chosen m-dimensional subspace with $m \ll n$. By doing so, an approximate form of the KRR estimate can be obtained by solving an m-dimensional quadratic program, which involves time and space complexity $\mathcal{O}(m^3)$ and $\mathcal{O}(m^2)$. Computing the approximate kernel matrix is a pre-processing step that has time complexity $\mathcal{O}(n^2 \log(m))$ for suitably chosen projections; this pre-processing step is trivially parallelizable, meaning it can be reduced to $\mathcal{O}(n^2 \log(m)/t)$ by using $t \leq n$ clusters.

Given such an approximation, we pose the following question: how small can the projection dimension m be chosen while still retaining minimax optimality of the approximate KRR estimate? We answer this question by connecting it to the *statistical dimension* d_n of the n-dimensional kernel matrix, a quantity that measures the effective number of degrees of freedom. (See Section 2.3 for a precise definition.) In our earlier work on sketching constrained squares problems [23, 24], we have studied how the projection dimension required for an accurate approximation varies as a function of the geometry of the constraint set. In the setting of kernel ridge regression, the constraint set is elliptical in nature, with its geometry determining a quantity known as the statistical dimension of the problem. Thus, it is natural to conjecture that it should be possible to project the kernel matrix

down to the statistical dimension while preserving minimax optimality of the resulting estimator. The main contribution of this paper is to confirm this conjecture for several classes of random projection matrices.

It is worth mentioning that our sketching approach is different in some important ways from the classical form of sketching [20] for unconstrained least squares problems. The classical method would apply the sketch to both the data vector and the kernel matrix, whereas our sketch is applied only to the kernel matrix. Moreover, our sketch operates on the right-hand side of the kernel matrix, thereby reducing the effective dimensionality of the parameter space, whereas the classical sketch applies to the left-hand side of the kernel matrix. These differences are important in establishing optimality of the method. As shown in the paper [24], although the classical least-squares sketch yields a good approximation to the value of the quadratic objective function, it is provably suboptimal for approximating the solution in terms of some distance measure between the approximate minimizer and the true minimizer. In contrast, the sketching approach developed here is carefully constructed so as to retain minimax optimality of the approximate KRR estimate with a relatively small projection dimension.

There are a related class of methods that form low-rank approximations $\widetilde{K}$ to the n-dimensional kernel matrix K, such as randomized singular value decompositions (e.g., [8, 13, 26]) and the Nyström methods (e.g., [10, 11]). The time complexity of such low-rank approximations is either $\mathcal{O}(nr^2)$ or $\mathcal{O}(n^2 r)$, depending on the specific approach (excluding the time for factorization), where r is the maintained rank, and the space complexity is $\mathcal{O}(nr)$. For the randomized singular value decomposition, if we use $\widehat{\mu}_k$ to denote kth eigenvalue of the empirical kernel matrix K, the results of Halko et al. [13] guarantee an operator norm bound of the form $|||K - \widetilde{K}|||_{\mathrm{op}} \leq (c_1 n/r)\widehat{\mu}_k$ for $r \geq c_2 k \log(k) + c_2 \log(kn)\log(k)$ using the sub-sampled randomized Hadamard transform (SRHT). However, in the regime $r \ll n$ that is needed to reduce the complexity of KRR, this bound does not seem strong enough to guarantee minimax optimality of the associated low rank approximated kernel method. Subsequent work [8] gives an improved bound for SRHT when the tail eigenvalue sum $\sum_{j=k+1}^{n} \widehat{\mu}_j$ has a rapid decay. In contrast, the theory developed in this paper requires no assumption on the tail eigenvalue sum in order for our approximate KRR estimate to achieve minimax optimality with the projection dimension proportional to the statistical dimension. Also related is a recent line of work [2, 5] that analyzes the tradeoff between the rank r and the resulting statistical performance of the estimator; we discuss it at more length in Section 3.3.

The remainder of this paper is organized as follows. Section 2 is devoted to further background on nonparametric regression, reproducing kernel Hilbert spaces and associated measures of complexity, as well as the notion of statistical dimension of a kernel. In Section 3, we turn to statements of our main results. Theorem 2 provides a general sufficient condition on a random sketch for the associated approximate form of KRR to achieve the minimax risk. In Corollary 1, we derive

some consequences of this general result for particular classes of random sketch matrices, and confirm these theoretical predictions with some simulations. We also compare at more length to methods based on the Nyström approximation in Section 3.3. Section 4 is devoted to the proofs of our main results, with the proofs of more technical results deferred to the Appendices. We conclude with a discussion in Section 5.

2. Problem formulation and background. We begin by introducing some background on nonparametric regression and reproducing kernel Hilbert spaces, before formulating the problem discussed in this paper.

2.1. *Regression in reproducing kernel Hilbert spaces.* Given n samples $\{(x_i, y_i)\}_{i=1}^n$ from the nonparametric regression model (1), our goal is to estimate the unknown regression function f^*. Our results apply to both fixed and random design, where in the latter the results can be viewed as conditioning on the design points $\{x_i\}_{i=1}^n$. The quality of an estimate $\widehat{f}$ can be measured in different ways: in this paper, we focus on the squared $L^2(\mathbb{P}_n)$ error

$$\|\widehat{f} - f^*\|_n^2 := \frac{1}{n}\sum_{i=1}^n \big(\widehat{f}(x_i) - f^*(x_i)\big)^2. \tag{2}$$

Naturally, the difficulty of nonparametric regression is controlled by the structure in the function f^*, and one way of modeling such structure is within the framework of a reproducing kernel Hilbert space (or RKHS for short). Here, we provide a very brief introduction referring the reader to the books [7, 12, 33] for more details and background.

Given a space $\mathcal{X}$ endowed with a probability distribution $\mathbb{P}$, the space $L^2(\mathbb{P})$ consists of all functions that are square-integrable with respect to $\mathbb{P}$. In abstract terms, a space $\mathcal{H} \subset L^2(\mathbb{P})$ is an RKHS if for each $x \in \mathcal{X}$, the evaluation function $f \mapsto f(x)$ is a bounded linear functional. In more concrete terms, any RKHS is generated by a positive semidefinite (PSD) kernel function in the following way. A PSD kernel function is a symmetric function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that, for any positive integer N, collections of points $\{v_1, \ldots, v_N\}$ and weight vector $\omega \in \mathbb{R}^N$, the sum $\sum_{i,j=1}^N \omega_i \omega_j \mathcal{K}(v_i, v_j)$ is nonnegative. Suppose moreover that for each fixed $v \in \mathcal{X}$, the function $u \mapsto \mathcal{K}(u, v)$ belongs to $L^2(\mathbb{P})$. We can then consider the vector space of all functions $g : \mathcal{X} \to \mathbb{R}$ of the form

$$g(\cdot) = \sum_{i=1}^N \omega_i \mathcal{K}(\cdot, v_i)$$

for some integer N, points $\{v_1, \ldots, v_N\} \subset \mathcal{X}$ and weight vector $w \in \mathbb{R}^N$, and define its norm by $\|g\|_{\mathcal{H}}^2 := \sum_{i,j=1}^N \omega_i \omega_j \mathcal{K}(v_i, v_j)$. By taking the closure of all such linear combinations, it can be shown [4] that we generate an RKHS $\mathcal{H}$ equipped

with a norm $\|\cdot\|_{\mathcal{H}}$, and one that is uniquely associated with the kernel $\mathcal{K}$. We provide some examples of various kernels and the associated function classes in Section 2.3 to follow.

2.2. *Kernel ridge regression and its sketched form.* Given the dataset $\{(x_i, y_i)\}_{i=1}^n$, a natural method for estimating unknown function $f^* \in \mathcal{H}$ is known as kernel ridge regression (KRR): it is based on the convex program

$$f^{\diamondsuit} := \arg\min_{f\in\mathcal{H}} \left\{ \frac{1}{2n}\sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2 + \lambda_n \|f\|_{\mathcal{H}}^2 \right\}, \tag{3}$$

where λ_n is a regularization parameter corresponding to the Hilbert space norm $\|\cdot\|_{\mathcal{H}}$.

As stated, this optimization problem can be infinite-dimensional in nature, since it takes place over the Hilbert space. However, as a straightforward consequence of the representer theorem [16], the solution to this optimization problem can be obtained by solving the n-dimensional convex program. In particular, let us define the *empirical kernel matrix*, namely the n-dimensional symmetric matrix K with entries $K_{ij} = n^{-1}\mathcal{K}(x_i, x_j)$. Here, we adopt the n^{-1} scaling for later theoretical convenience. In terms of this matrix, the KRR estimate can be obtained by first solving the quadratic program

$$\omega^{\dagger} = \arg\min_{\omega\in\mathbb{R}^n} \left\{ \frac{1}{2}\omega^T K^2 \omega - \omega^T \frac{Ky}{\sqrt{n}} + \lambda_n \omega^T K \omega \right\}, \tag{4a}$$

and then outputting the function

$$f^{\diamondsuit}(\cdot) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \omega_i^{\dagger}\mathcal{K}(\cdot, x_i). \tag{4b}$$

In principle, the original KRR optimization problem (4a) is simple to solve: it is an n dimensional quadratic program, and can be solved exactly using $\mathcal{O}(n^3)$ via a QR decomposition. However, in many applications, the number of samples may be large, so that this type of cubic scaling is prohibitive. In addition, the n-dimensional kernel matrix K is dense in general, and so requires storage of order n^2 numbers, which can also be problematic in practice.

In this paper, we consider an approximation based on limiting the original parameter $\omega \in \mathbb{R}^n$ to an m-dimensional subspace of $\mathbb{R}^n$, where $m \ll n$ is the *projection dimension*. We define this approximation via a sketch matrix $S \in \mathbb{R}^{m\times n}$, such that the m-dimensional subspace is generated by the row span of S. More precisely, the *sketched kernel ridge regression* estimate is given by first solving

$$\widehat{\alpha} = \arg\min_{\theta\in\mathbb{R}^m} \left\{ \frac{1}{2}\alpha^T (SK)(KS^T)\alpha - \alpha^T S\frac{Ky}{\sqrt{n}} + \lambda_n \alpha^T SKS^T\alpha \right\}, \tag{5a}$$

and then outputting the function

$$\widehat{f}(\cdot) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(S^T \widehat{\alpha}\big)_i \mathcal{K}(\cdot, x_i). \tag{5b}$$

Note that the sketched program (5a) is a quadratic program in m dimensions: it takes as input the m-dimensional matrices (SK^2S^T, SKS^T) and the m-dimensional vector SKy. Consequently, it can be solved efficiently via QR decomposition with computational complexity $\mathcal{O}(m^3)$. Moreover, the computation of the sketched kernel matrix $SK = [SK_1, \ldots, SK_n]$ in the input can be parallelized across its columns. In passing, it is worth mentioning that this sketching idea can be extended to other kernel methods based on more general loss functions, such as kernel SVM, using the representer theorem. Characterizing theoretical properties of general sketched kernel methods is an interesting direction for future work

In this paper, we analyze various forms of random sketch matrices S. Let us consider a few of them here.

Sub-Gaussian sketches: We say the row s_i of the sketch matrix is zero-mean 1-sub-Gaussian if for any fixed unit vector $u \in \mathcal{S}^{n-1}$, we have

$$\mathbb{P}\big[\big|\langle u, s_i\rangle \geq t\big|\big] \leq 2e^{-\frac{nt^2}{2}} \qquad \text{for all } t \geq 0.$$

Many standard choices of sketch matrices have i.i.d. 1-sub-Gaussian rows in this sense; examples include matrices with i.i.d. Gaussian entries, i.i.d. Bernoulli entries or random matrices with independent rows drawn uniformly from a rescaled sphere. For convenience, the sub-Gaussian sketch matrices considered in this paper are all rescaled so that their rows have the covariance matrix $\frac{1}{\sqrt{m}} I_{n\times n}$.

Randomized orthogonal system (ROS) sketches: This class of sketches are based on randomly sampling and rescaling the rows of a fixed orthonormal matrix $H \in \mathbb{R}^{n\times n}$. Examples of such matrices include the discrete Fourier transform (DFT) matrix, and the Hadamard matrix. More specifically, a ROS sketch matrix $S \in \mathbb{R}^{m\times n}$ is formed with i.i.d. rows of the form

$$s_i = \sqrt{\frac{n}{m}} R H^T p_i \qquad \text{for } i = 1, \ldots, m,$$

where R is a random diagonal matrix whose entries are i.i.d. Rademacher variables and $\{p_1, \ldots, p_m\}$ is a random subset of m rows sampled uniformly from the $n \times n$ identity matrix without replacement. An advantage of using ROS sketches is that for suitably chosen orthonormal matrices, including the DFT and Hadamard cases among others, a matrix-vector product (say of the form Su for some vector $u \in \mathbb{R}^n$) can be computed in $\mathcal{O}(n \log m)$ time, as opposed to $\mathcal{O}(nm)$ time required for the same operation with generic dense sketches. For instance, see Ailon and Liberty [1] and [23] for further details. Throughout this paper, we focus on ROS sketches based on orthonormal matrices H with uniformly bounded entries,

meaning that $|H_{ij}| \leq \frac{c}{\sqrt{n}}$ for all entries (i, j). This entrywise bound is satisfied by Hadamard and DFT matrices, among others.

Sub-sampling sketches: This class of sketches are even simpler, based on sub-sampling the rows of the identity matrix without replacement. In particular, the sketch matrix $S \in \mathbb{R}^{m \times n}$ has rows of the form $s_i = \sqrt{\frac{n}{m}} p_i$, where the vectors $\{p_1, \ldots, p_m\}$ are drawn uniformly at random without replacement from the n-dimensional identity matrix. It can be understood as related to a ROS sketch, based on the identity matrix as an orthonormal matrix, and not using the Rademacher randomization nor satisfying the entrywise bound. In Appendix A, we show that the sketched KRR estimate (5a) based on a sub-sampling sketch matrix is equivalent to the Nyström approximation.

2.3. *Kernel complexity measures and statistical guarantees*. So as to set the stage for later results, let us characterize an appropriate choice of the regularization parameter λ, and the resulting bound on the prediction error $\|f^{\diamondsuit} - f^*\|_n$. Recall the empirical kernel matrix K defined in the previous section: since it is symmetric and positive definite, it has an eigendecomposition of the form $K = UDU^T$, where $U \in \mathbb{R}^{n \times n}$ is an orthonormal matrix, and $D \in \mathbb{R}^{n \times n}$ is diagonal with elements $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \cdots \geq \widehat{\mu}_n \geq 0$. Using these eigenvalues, consider the *kernel complexity function*

$$\widehat{\mathcal{R}}(\delta) = \sqrt{\frac{1}{n} \sum_{j=1}^{n} \min\{\delta^2, \widehat{\mu}_j\}}, \tag{6}$$

corresponding to a rescaled sum of the eigenvalues, truncated at level δ^2. This function arises via analysis of the local Rademacher complexity of the kernel class (e.g., [6, 17, 22, 25]). For a given kernel matrix and noise variance $\sigma > 0$, the *critical radius* is defined to be the smallest positive solution $\delta_n > 0$ to the inequality

$$\frac{\widehat{\mathcal{R}}(\delta)}{\delta} \leq \frac{\delta}{\sigma}. \tag{7}$$

Note that the existence and uniqueness of this critical radius is guaranteed for any kernel class [6].

Bounds on ordinary KRR: The significance of the critical radius is that it can be used to specify bounds on the prediction error in kernel ridge regression. More precisely suppose that we compute the KRR estimate (3) with any regularization parameter $\lambda \geq 2\delta_n^2$. If f^* is in $\mathcal{H}$, then with probability at least $1 - c_1 e^{-c_2 n \delta_n^2}$ with respect to the noise $\{w_i\}_{i=1}^n$, we are guaranteed that

$$\|f^{\diamondsuit} - f^*\|_n^2 \leq c_u \{\lambda_n + \delta_n^2\}, \tag{8}$$

where $c_u > 0$ is a constant (depends on $\|f^*\|_{\mathcal{H}}$ but independent of n, σ and the kernel). This known result follows from standard techniques in empirical process

theory (e.g., [6, 32]); we also note that it can be obtained as a corollary of our more general theorem on sketched KRR estimates to follow (namely, Theorem 2).

To illustrate, let us consider a few examples of reproducing kernel Hilbert spaces, and compute the critical radius in different cases. In working through these examples, so as to determine explicit rates, we assume that the design points $\{x_i\}_{i=1}^n$ are sampled i.i.d. from some underlying distribution $\mathbb{P}$, and we make use of the useful fact that, up to constant factors, we can always work with the population-level kernel complexity function

$$\mathcal{R}(\delta) = \sqrt{\frac{1}{n}\sum_{j=1}^{\infty} \min\{\delta^2, \mu_j\}}, \tag{9}$$

where $\{\mu_j\}_{j=1}^{\infty}$ are the eigenvalues of the kernel integral operator (assumed to be uniformly bounded). This equivalence follows from standard results on the population and empirical Rademacher complexities [6, 22].

EXAMPLE 1 (Polynomial kernel). For some integer $D \geq 1$, consider the kernel function on $[0, 1] \times [0, 1]$ given by $\mathcal{K}_{\text{poly}}(u, v) = (1 + \langle u, v\rangle)^D$. For $D = 1$, it generates the class of all linear functions of the form $f(x) = a_0 + a_1 x$ for some scalars (a_0, a_1), and corresponds to a linear kernel. More generally, for larger integers D, it generates the class of all polynomial functions of degree at most D, that is, functions of the form $f(x) = \sum_{j=0}^{D} a_j x^j$.

Let us now compute a bound on the critical radius δ_n. It is straightforward to show that the polynomial kernel is of finite rank at most $D + 1$, meaning that the kernel matrix K always has at most $\min\{D + 1, n\}$ nonzero eigenvalues. Consequently, as long $n > D + 1$, there is a universal constant c such that

$$\widehat{\mathcal{R}}(\delta) \leq c\sqrt{\frac{D+1}{n}}\delta,$$

which implies that $\delta_n^2 \precsim \sigma^2 \frac{D+1}{n}$. Here, we use the notation $A(n) \precsim B(n)$ to mean $A(n) \leq cB(n)$ for some universal constant $c > 0$. Consequently, we conclude that the KRR estimate satisfies the bound $\|\widehat{f} - f^*\|_n^2 \precsim \sigma^2 \frac{D+1}{n}$ with high probability. Note that this bound is intuitive, since a polynomial of degree D has $D + 1$ free parameters.

EXAMPLE 2 (Gaussian kernel). The Gaussian kernel with bandwidth $h > 0$ takes the form $\mathcal{K}_{\text{Gau}}(u, v) = e^{-\frac{1}{2h^2}(u-v)^2}$. When defined with respect to Lebesgue measure on the real line, the eigenvalues of the kernel integral operator scale as $\mu_j \asymp \exp(-\pi h^2 j^2)$ as $j \to \infty$. Based on this fact, it can be shown that the critical radius scales as $\delta_n^2 \asymp \frac{\sigma^2}{nh^2}\sqrt{\log(\frac{nh^2}{\sigma^2})}$. Thus, even though the Gaussian kernel is nonparametric (since it cannot be specified by a fixed number of parameters), it is still a relatively small function class.

EXAMPLE 3 (First-order Sobolev space). As a final example, consider the kernel defined on the unit square $[0, 1] \times [0, 1]$ given by $\mathcal{K}_{\mathrm{sob}}(u, v) = \min\{u, v\}$. It generates the function class

$$\mathcal{H}^1[0,1] = \Big\{ f : [0,1] \to \mathbb{R} \,|\, f(0) = 0, \text{ and } f \text{ is abs. cts. with } \int_0^1 \big[f'(x)\big]^2 dx < \infty \Big\}, \tag{10}$$

a class that contains all Lipschitz functions on the unit interval $[0, 1]$. Roughly speaking, we can think of the first-order Sobolev class as functions that are almost everywhere differentiable with derivative in $L^2[0, 1]$. Note that this is a much larger kernel class than the Gaussian kernel class. The first-order Sobolev space can be generalized to higher order Sobolev spaces, in which functions have additional smoothness. See the book [12] for further details on these and other reproducing kernel Hilbert spaces.

If the kernel integral operator is defined with respect to Lebesgue measure on the unit interval, then the population level eigenvalues are given by $\mu_j = (\frac{2}{(2j-1)\pi})^2$ for $j = 1, 2, \ldots$. Given this relation, some calculation shows that the critical radius scales as $\delta_n^2 \asymp (\frac{\sigma^2}{n})^{2/3}$. This is the familiar minimax risk for estimating Lipschitz functions in one dimension [29].

Lower bounds for nonparametric regression: For future reference, it is also convenient to provide a lower bound on the prediction error achievable by *any estimator*. In order to do so, we first define the *statistical dimension* of the kernel as

$$d_n := \min\{ j \in [n] : \widehat{\mu}_j \leq \delta_n^2 \}, \tag{11}$$

and $d_n = n$ if no such index j exists. By definition, we are guaranteed that $\widehat{\mu}_j > \delta_n^2$ for all $j \in \{1, 2, \ldots, d_n\}$. Our definition of the statistical dimension may differ from some others in the literature, such as the one for cones in [3]. In terms of this statistical dimension, we have

$$\widehat{\mathcal{R}}(\delta_n) = \left[\frac{d_n}{n} \delta_n^2 + \frac{1}{n} \sum_{j=d_n+1}^{n} \widehat{\mu}_j \right]^{1/2},$$

showing that the statistical dimension controls a type of bias-variance tradeoff.

It is reasonable to expect that the critical rate δ_n should be related to the statistical dimension as $\delta_n^2 \asymp \frac{\sigma^2 d_n}{n}$. This scaling relation holds whenever the tail sum satisfies a bound of the form $\sum_{j=d_n+1}^{n} \widehat{\mu}_j \precsim d_n \delta_n^2$. Although it is possible to construct pathological examples in which this scaling relation does not hold, it is true for most kernels of interest, including all examples considered in this paper. For any such regular kernel, the critical radius provides a fundamental lower bound on the performance of *any estimator*, as summarized in the following theorem.

THEOREM 1 (Critical radius and minimax risk). *Given n i.i.d. samples $\{(y_i, x_i)\}_{i=1}^n$ from the standard nonparametric regression model over any regular kernel class, any estimator $\widetilde{f}$ has prediction error lower bounded as*

$$\sup_{\|f^*\|_{\mathcal{H}} \leq 1} \mathbb{E}\|\widetilde{f} - f^*\|_n^2 \geq c_\ell \delta_n^2, \tag{12}$$

where $c_\ell > 0$ is a numerical constant, and δ_n is the critical radius (7).

The proof of this claim, provided in Appendix B.1, is based on a standard application of Fano's inequality, combined with a random packing argument. It establishes that the critical radius is a fundamental quantity, corresponding to the appropriate benchmark to which sketched kernel regression estimates should be compared.

3. Main results and their consequences. We now turn to statements of our main theorems on kernel sketching, as well as a discussion of some of their consequences. We first introduce the notion of a K-satisfiable sketch matrix, and then show (in Theorem 2) that any sketched KRR estimate based on a K-satisfiable sketch also achieves the minimax risk. We illustrate this achievable result with several corollaries for different types of randomized sketches. For Gaussian and ROS sketches, we show that choosing the sketch dimension proportional to the statistical dimension of the kernel (with additional log factors in the ROS case) is sufficient to guarantee that the resulting sketch will be K-satisfiable with high probability. In addition, we illustrate the sharpness of our theoretical predictions via some experimental simulations.

3.1. *General conditions for sketched kernel optimality.* Recall the definition (11) of the statistical dimension d_n, and consider the eigendecomposition $K = UDU^T$ of the kernel matrix, where $U \in \mathbb{R}^{n \times n}$ is an orthonormal matrix of eigenvectors, and $D = \operatorname{diag}\{\widehat{\mu}_1, \ldots, \widehat{\mu}_n\}$ is a diagonal matrix of eigenvalues. Let $U_1 \in \mathbb{R}^{n \times d_n}$ denote the left block of U, and similarly, $U_2 \in \mathbb{R}^{n \times (n-d_n)}$ denote the right block. Note that the columns of the left block U_1 correspond to the eigenvectors of K associated with the leading d_n eigenvalues, whereas the columns of the right block U_2 correspond to the eigenvectors associated with the remaining $n - d_n$ smallest eigenvalues. Intuitively, a sketch matrix $S \in \mathbb{R}^{m \times n}$ is "good" if the sub-matrix $SU_1 \in \mathbb{R}^{m \times d_n}$ is relatively close to an isometry, whereas the sub-matrix $SU_2 \in \mathbb{R}^{m \times (n-d_n)}$ has a relatively small operator norm.

This intuition can be formalized in the following way. For a given kernel matrix K, a sketch matrix S is said to be *K-satisfiable* if there is a universal constant c such that

$$\left|\left|\left|(SU_1)^T SU_1 - I_{d_n}\right|\right|\right|_{\mathrm{op}} \leq 1/2, \quad \text{and} \quad \left|\left|\left|SU_2 D_2^{1/2}\right|\right|\right|_{\mathrm{op}} \leq c\delta_n, \tag{13}$$

where $D_2 = \operatorname{diag}\{\widehat{\mu}_{d_n+1}, \ldots, \widehat{\mu}_n\}$.

Given this definition, the following theorem shows that any sketched KRR estimate based on a K-satisfiable matrix achieves the minimax risk (with high probability over the noise in the observation model).

THEOREM 2 (Upper bound). *Given n i.i.d. samples $\{(y_i, x_i)\}_{i=1}^n$ from the standard nonparametric regression model, consider the sketched KRR problem (5a) based on a K-satisfiable sketch matrix S. If $f^* \in \mathcal{H}$, then any for $\lambda_n \geq 2\delta_n^2$, the sketched regression estimate $\widehat{f}$ from equation (5b) satisfies the bound*

$$\|\widehat{f} - f^*\|_n^2 \leq c_u\{\lambda_n + \delta_n^2\}$$

with probability greater than $1 - c_1 e^{-c_2 n \delta_n^2}$. Here, constant c_u only depends on $\|f^*\|_{\mathcal{H}}$.

We emphasize that in the case of fixed design regression and for a fixed sketch matrix, the K-satisfiable condition on the sketch matrix S is a deterministic statement: apart from the sketch matrix, it only depends on the properties of the kernel function $\mathcal{K}$ and design variables $\{x_i\}_{i=1}^n$. Thus, when using randomized sketches, the algorithmic randomness can be completely decoupled from the randomness in the noisy observation model (1). In fact, since our work was first posted, some other researchers [9] have used the conditions underlying our Theorem 2 to show that a rather different class of sketch matrices can be used to perform optimal KRR regression. This illustrates that the decoupling approach in Theorem 2 is a fruitful one.

Proof intuition: The proof of Theorem 2 is given in Section 4.1. At a high-level, it is based on an upper bound on the prediction error $\|\widehat{f} - f^*\|_n^2$ that involves two sources of error: the *approximation error* associated with solving a zero-noise version of the KRR problem in the projected m-dimensional space, and the *estimation error* between the noiseless and noisy versions of the projected problem. In more detail, letting $z^* := (f^*(x_1), \ldots, f^*(x_n))$ denote the vector of function evaluations defined by f^*, consider the quadratic program

$$\alpha^\dagger := \arg\min_{\alpha \in \mathbb{R}^m} \left\{ \frac{1}{2n} \|z^* - \sqrt{n} K S^T \alpha\|_2^2 + \lambda_n \|K^{1/2} S^T \alpha\|_2^2 \right\}, \tag{14}$$

as well as the associated fitted function $f^\dagger = \frac{1}{\sqrt{n}} \sum_{i=1}^n (S\alpha^\dagger)_i \mathcal{K}(\cdot, x_i)$. The vector $\alpha^\dagger \in \mathbb{R}^m$ is the solution of the sketched problem in the case of zero noise, whereas the fitted function $f^\dagger$ corresponds to the best penalized approximation of f^* within the range space of S^T.

Given this definition, we then have the elementary inequality

$$\frac{1}{2}\|\widehat{f} - f^*\|_n^2 \leq \underbrace{\|f^\dagger - f^*\|_n^2}_{\text{Approximation error}} + \underbrace{\|f^\dagger - \widehat{f}\|_n^2}_{\text{Estimation error}}. \tag{15}$$

For a fixed sketch matrix, the approximation error term is deterministic: it corresponds to the error induced by approximating f^* over the range space of S^T. On the other hand, the estimation error depends both on the sketch matrix and the observation noise. In Section 4.1, we state and prove two lemmas that control the approximation and error terms, respectively.

As a corollary, Theorem 2 implies the stated upper bound (8) on the prediction error of the original (unsketched) KRR estimate (3). Indeed, this estimator can be obtained using the "sketch matrix" $S = I_{n\times n}$, which is easily seen to be K-satisfiable. In practice, however, we are interested in $m \times n$ sketch matrices with $m \ll n$, so as to achieve computational savings. In particular, a natural conjecture is that it should be possible to efficiently generate K-satisfiable sketch matrices with the projection dimension m proportional to the statistical dimension d_n of the kernel. Of course, one such K-satisfiable matrix is given by $S = U_1^T \in \mathbb{R}^{d_n\times n}$, but it is not easy to generate, since it requires computing the eigendecomposition of K. Nonetheless, as we now show, there are various randomized constructions that lead to K-satisfiable sketch matrices with high probability.

3.2. *Corollaries for randomized sketches.* When combined with additional probabilistic analysis, Theorem 2 implies that various forms of randomized sketches achieve the minimax risk using a sketch dimension proportional to the statistical dimension d_n. Here, we analyze the Gaussian and ROS families of random sketches, as previously defined in Section 2.2. Throughout our analysis, we require that the sketch dimension satisfies a lower bound of the form

$$m \geq \begin{cases} cd_n, & \text{for Gaussian sketches, and} \\ cd_n \log^4(n), & \text{for ROS sketches,} \end{cases} \tag{16a}$$

where d_n is the *statistical dimension* as previously defined in equation (11). Here, it should be understood that the constant c can be chosen sufficiently large (but finite). In addition, for the purposes of stating high probability results, we define the function

$$\begin{aligned} &\phi(m, d_n, n) \\ &:= \begin{cases} c_1 e^{-c_2 m}, & \text{for Gaussian sketches, and} \\ c_1\big[e^{-c_2 \frac{m}{d_n \log^2(n)}} + e^{-c_2 d_n \log^2(n)}\big], & \text{for ROS sketches,} \end{cases} \end{aligned} \tag{16b}$$

where c_1, c_2 are universal constants. With this notation, the following result provides a high probability guarantee for both Gaussian and ROS sketches.

COROLLARY 1 (Guarantees for Gaussian and ROS sketches). *Given n i.i.d. samples $\{(y_i, x_i)\}_{i=1}^n$ from the standard nonparametric regression model (1), consider the sketched KRR problem (5a) based on a sketch dimension m satisfying the lower bound (16a). If $f^* \in \mathcal{H}$, then there is a constant c'_u only depending on*

$\|f^*\|_{\mathcal{H}}$ such that for any $\lambda_n \geq 2\delta_n^2$, the sketched regression estimate (5b) satisfies the bound

$$\|\widehat{f} - f^*\|_n^2 \leq c_u'\{\lambda_n + \delta_n^2\}$$

with probability greater than $1 - \phi(m, d_n, n) - c_3 e^{-c_4 n \delta_n^2}$.

As a remark, in Corollary 1, ROS sketches require an additional $\log^4(n)$ factor over d_n in the sketch dimension m. For some kernels (such as the Gaussian, which has a logarithmic statistical dimension), these logarithmic factors are significant. Based on our simulation results below, we suspect that some of these additional poly-log terms may be an artifact of our proof technique. We leave as an important open problem whether or not the scaling for ROS sketches can be sharpened.

In order to illustrate Corollary 1, let us return to the three examples previously discussed in Section 2.3. To be concrete, we derive the consequences for Gaussian sketches, noting that ROS sketches incur only an additional $\log^4(n)$ overhead:

- For the Dth-order polynomial kernel from Example 1, the statistical dimension d_n for any sample size n is at most $D + 1$, so that a sketch size of order $D + 1$ is sufficient. This is a very special case, since the kernel is finite rank and so the required sketch dimension has no dependence on the sample size.
- For the Gaussian kernel from Example 2, the statistical dimension satisfies the scaling $d_n \asymp h^{-2}\sqrt{\log(nh^2)}$, so that it suffices to take a sketch dimension scaling logarithmically with the sample size.
- For the first-order Sobolev kernel from Example 3, the statistical dimension scales as $d_n \asymp n^{1/3}$, so that a sketch dimension scaling as the cube root of the sample size is required.

In order to illustrate these theoretical predictions, we performed some simulations. Beginning with the Sobolev kernel $\mathcal{K}_{\text{sob}}(u, v) = \min\{u, v\}$ on the unit square, as introduced in Example 3, we generated n i.i.d. samples from the model (1) with noise standard deviation $\sigma = 0.5$, the unknown regression function

$$f^*(x) = 1.6\big|(x - 0.4)(x - 0.6)\big| - 0.3, \tag{17}$$

and uniformly spaced design points $x_i = \frac{i}{n}$ for $i = 1, \ldots, n$. By construction, the function f^* belongs to the first-order Sobolev space with $\|f^*\|_{\mathcal{H}} \approx 1.3$. As suggested by our theory for the Sobolev kernel, we set the projection dimension $m = \lceil n^{1/3} \rceil$, and then solved the sketched version of kernel ridge regression, for both Gaussian sketches and ROS sketches based on the fast Hadamard transform. We performed simulations for n in the set $\{32, 64, 128, \ldots, 16{,}384\}$ so as to study scaling with the sample size. As noted above, our theory predicts that the squared prediction loss $\|\widehat{f} - f^*\|_n^2$ should tend to zero at the same rate $n^{-2/3}$ as that of the unsketched estimator $f^{\diamondsuit}$. Figure 1 confirms this theoretical prediction. In panel

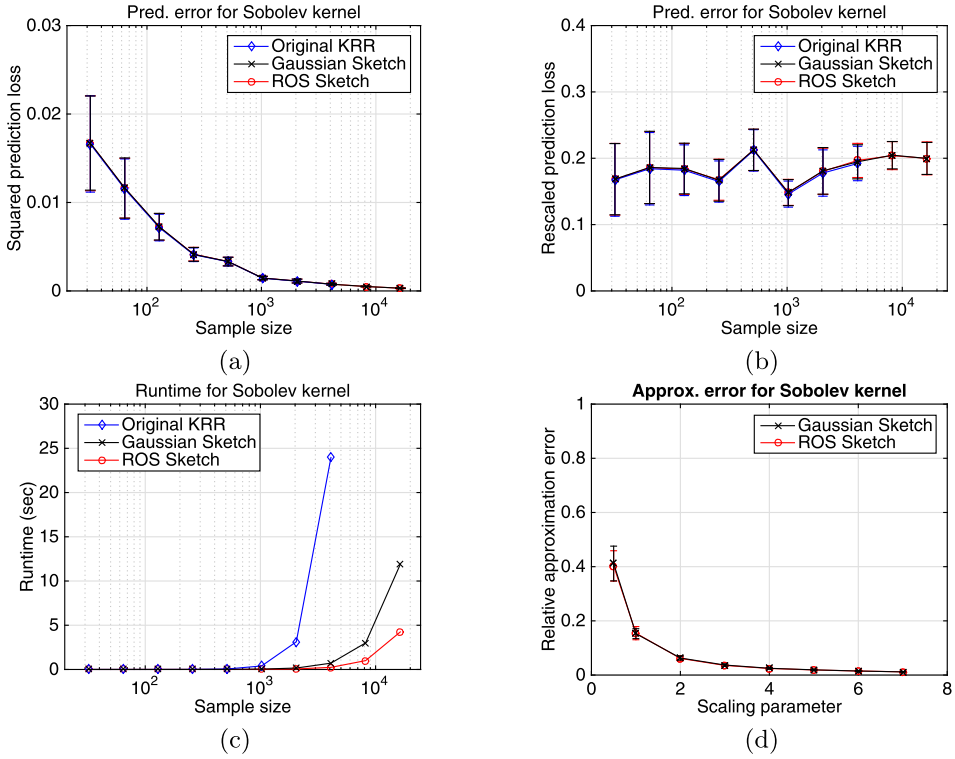

FIG. 1. *Prediction error versus sample size for original KRR, Gaussian sketch and ROS sketches for the Sobolev one kernel for the function* $f^*(x) = 1.6|(x-0.4)(x-0.6)| - 0.3$ *with* $\lambda = 2\delta_n^2 = 0.5n^{-2/3}$. *In all cases, each point corresponds to the average of* 100 *trials, with standard errors also shown.* (a) *Squared prediction error* $\|\widehat{f} - f^*\|_n^2$ *versus the sample size* $n \in \{32, 64, 128, \ldots, 16{,}384\}$ *for projection dimension* $m = \lceil n^{1/3} \rceil$. (b) *Rescaled prediction error* $n^{2/3}\|\widehat{f} - f^*\|_n^2$ *versus the sample size.* (c) *Runtime versus the sample size. The original KRR for* $n = 8192$ *and* $n = 16{,}384$ *samples are not computed due to out-of-memory failures.* (d) *For a problem of size* $n = 1024$, *plots of the ratios* $\|\widehat{f}_c - f^\diamondsuit\|_n^2 / \|f^\diamondsuit - f^*\|_n^2$ *versus the scaling parameter* c, *where the sketched estimate is computed using a projection dimension* $m = \lceil cn^{1/3} \rceil$. *The constant* c *ranges over* $\{0.5, 1, 2, \ldots, 7\}$.

(a), we plot the squared prediction error versus the sample size, showing that all three curves (original, Gaussian sketch and ROS sketch) tend to zero. Panel (b) plots the *rescaled* prediction error $n^{2/3}\|\widehat{f} - f^*\|_n^2$ versus the sample size, with the relative flatness of these curves confirming the $n^{-2/3}$ decay predicted by our theory. Panel (c) plots the running time versus the sample size and the squared prediction error, showing that kernel sketching considerably speeds up KRR.

In our second experiment, we repeated the same set of simulations this time for the 3-d Gaussian kernel $\mathcal{K}_{\mathrm{Gau}}(u, v) = e^{-\frac{1}{2h^2}\|u-v\|_2^2}$ with bandwidth $h = 1$, and the function $f^*(x) = 0.5e^{-x_1+x_2} - x_2x_3$. In this case, as suggested by our theory,

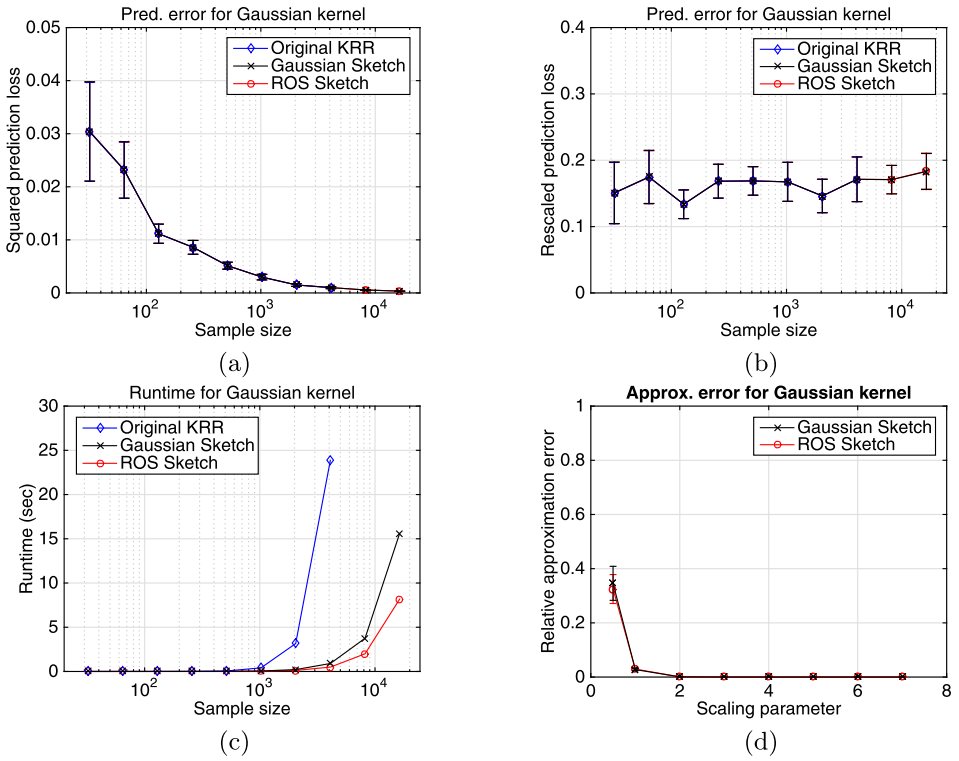

FIG. 2. *Prediction error versus sample size for original KRR, Gaussian sketch and ROS sketches for the Gaussian kernel with the function* $f^*(x) = 0.5e^{-x_1+x_2} - x_2x_3$ *with* $\lambda = 2\delta_n^2 = 0.5\log^{1.5}(n)/n$. *In all cases, each point corresponds to the average of* 100 *trials, with standard errors also shown.* (a) *Squared prediction error* $\|\widehat{f} - f^*\|_n^2$ *versus the sample size* $n \in \{32, 64, 128, \ldots, 16{,}384\}$ *for projection dimension* $m = \lceil 1.25(\log n)^{3/2} \rceil$. (b) *Rescaled prediction error* $\frac{n}{(\log n)^{3/2}}\|\widehat{f} - f^*\|_n^2$ *versus the sample size.* (c) *Runtime versus the sample size. The original KRR under* $n = 8192$ *and* 16,384 *are not computed due to out-of-memory failures.* (d) *For a problem of size* $n = 1024$, *plots of the ratios* $\|\widehat{f}_c - f^{\diamondsuit}\|_n^2 / \|f^{\diamondsuit} - f^*\|_n^2$ *versus the scaling parameter* c, *where the sketched estimate is computed using a projection dimension* $m = \lceil cn^{1/3} \rceil$. *The constant* c *ranges over* $\{0.5, 1, 2, \ldots, 7\}$.

we choose the sketch dimension $m = \lceil 1.25(\log n)^{3/2} \rceil$. Figure 2 shows the same types of plots with the prediction error. In this case, we expect that the squared prediction error will decay at the rate $\frac{(\log n)^{3/2}}{n}$. This prediction is confirmed by the plot in panel (b), showing that the rescaled error $\frac{n}{(\log n)^{3/2}}\|\widehat{f} - f^*\|_n^2$, when plotted versus the sample size, remains relatively constant over a wide range.

REMARK. In practice, the target sketch dimension m is only known up to a multiplicative constant. To determine this multiplicative constant, one can implement the randomized algorithm in an adaptive fashion where the multiplica-

tive constant is increased until the squared prediction norm of the change in the sketched function estimate $\widehat{f}$ falls below a desired tolerance. More precisely, letting $\{m_1, m_2, \ldots\}$ denote a sequence of projection dimensions, we compute the sketched estimates $\widehat{f}_t$ based on a projection dimension of size m_t. This adaptive procedure only increases the computational complexity by a constant multiple: when increasing the sketch dimension from m_t to m_{t+1}, we only need to sample additional $m_{t+1} - m_t$ rows to form the new sketch matrix S'. Correspondingly, forming the new sketched kernel matrix $S'K$ only requires computing the product of the new rows of S' and the kernel matrix K, and computing the new sketched estimate $\widehat{f}_{t+1}$ is a quadratic program in the low-dimensional space. The plots in panels (d) of Figure 1 and Figure 2 show that the approximation error $\|\widehat{f}_t - f^{\diamondsuit}\|_n^2$ rapidly approaches zero relative to the squared error $\|f^{\diamondsuit} - f^*\|_n^2$ in the original KRR estimate $f^{\diamondsuit}$ as the projection dimension m_t grows, which justifies the validity of the adaptive procedure. In practice, we cannot compute the differences $\|\widehat{f}_t - f^{\diamondsuit}\|_n^2$, but the differences $\|\widehat{f}_t - \widehat{f}_{t+1}\|_n^2$ exhibit a rapid decay that is qualitatively very similar, and can be used as a stopping criterion.

3.3. *Comparison with Nyström-based approaches.* It is interesting to compare the convergence rate and computational complexity of our methods with guarantees based on the Nyström approximation. As shown in Appendix A, this Nyström approximation approach can be understood as a particular form of our sketched estimate, one in which the sketch corresponds to a random row-sampling matrix.

Bach [5] analyzed the prediction error of the Nyström approximation to KRR based on uniformly sampling a subset of p-columns of the kernel matrix K, leading to an overall computational complexity of $\mathcal{O}(np^2)$. In order for the approximation to match the performance of KRR, the number of sampled columns must be lower bounded as

$$p \gtrsim n \big\| \operatorname{diag}\big(K(K + \lambda_n I)^{-1}\big) \big\|_\infty \log n,$$

a quantity which can be substantially larger than the statistical dimension required by our methods. Moreover, as shown in the following example, there are many classes of kernel matrices for which the performance of the Nyström approximation will be poor.

EXAMPLE 4 (Failure of Nyström approximation). Given a sketch dimension $m \leq n \log 2$, consider an empirical kernel matrix K that has a block diagonal form $\operatorname{diag}(K_1, K_2)$, where $K_1 \in \mathbb{R}^{(n-k)\times(n-k)}$ and $K_2 \in \mathbb{R}^{k\times k}$ for any integer $k \leq \frac{n}{m} \log 2$. Then the probability of not sampling any of the last k columns/rows is at least $1 - (1 - k/n)^m \geq 1 - e^{-km/n} \geq 1/2$. This means that with probability at least $1/2$, the sub-sampling sketch matrix can be expressed as $S = (S_1, 0)$, where $S_1 \in \mathbb{R}^{m\times(n-k)}$. Under such an event, the sketched KRR (5a) takes on a degenerate form, namely

$$\widehat{\alpha} = \arg\min_{\theta\in\mathbb{R}^m} \left\{ \frac{1}{2}\alpha^T S_1 K_1^2 S_1^T \alpha - \alpha^T S_1 \frac{K_1 y_1}{\sqrt{n}} + \lambda_n \alpha^T S_1 K_1 S_1^T \alpha \right\},$$

and objective that depends only on the first $n-k$ observations. Since the values of the last k observations can be arbitrary, this degeneracy has the potential to lead to substantial approximation error.

The previous example suggests that the Nyström approximation is likely to be very sensitive to non-inhomogeneity in the sampling of covariates. In order to explore this conjecture, we performed some additional simulations, this time comparing both Gaussian and ROS sketches with the uniform Nyström approximation sketch. Returning again to the Gaussian kernel $\mathcal{K}_{\mathrm{Gau}}(u,v) = e^{-\frac{1}{2h^2}(u-v)^2}$ with bandwidth $h = 0.25$, and the function $f^*(x) = -1 + 2x^2$, we first generated n i.i.d. samples that were uniform on the unit interval $[0,1]$. We then implemented sketches of various types (Gaussian, ROS or Nyström) using a sketch dimension $m = \lceil 4\sqrt{\log n} \rceil$ and regularization parameter $\lambda = 0.5\sqrt{\log(n)}/n$. As shown in the top row [panels (a) and (b)] of Figure 3, all three sketch types perform very well for this regular design, with prediction error that is essentially indistinguishable from the original KRR estimate. Keeping the same kernel and function, we then considered an irregular form of design, namely with $k = \lceil \sqrt{n} \rceil$ samples perturbed as follows:

$$x_i \sim \begin{cases} \mathrm{Unif}[0, 1/2], & \text{if } i = 1, \ldots, n-k, \\ 1 + z_i, & \text{for } i = k+1, \ldots, n, \end{cases}$$

where each $z_i \sim N(0, 1/n)$. The performance of the sketched estimators in this case are shown in the bottom row [panels (c) and (d)] of Figure 3. As before, both the Gaussian and ROS sketches track the performance of the original KRR estimate very closely; in contrast, the Nyström approximation behaves very poorly for this regression problem, consistent with the intuition suggested by the preceding example.

As is known from general theory on the Nyström approximation, its performance can be improved by knowledge of the so-called leverage scores of the underlying matrix. In this vein, recent work by Alaoui and Mahoney [2] suggests a Nyström approximation nonuniform sampling of the columns of kernel matrix involving the leverage scores. Assuming that the leverage scores are known, they show that their method matches the performance of original KRR using a nonuniform sub-sample of the order $\mathrm{trace}(K(K + \lambda_n I)^{-1}) \log n$ columns. When the regularization parameter λ_n is set optimally—that is, proportional to δ_n^2—then apart from the extra logarithmic factor; this sketch size scales with the statistical dimension, as defined here. However, the leverage scores are *not known*, and their method for obtaining a sufficiently approximation requires sampling $\tilde{p}$ columns of the kernel matrix K, where

$$\tilde{p} \gtrsim \lambda_n^{-1} \mathrm{trace}(K) \log n.$$

For a typical (normalized) kernel matrix K, we have $\mathrm{trace}(K) \gtrsim 1$; moreover, in order to achieve the minimax rate, the regularization parameter λ_n should scale

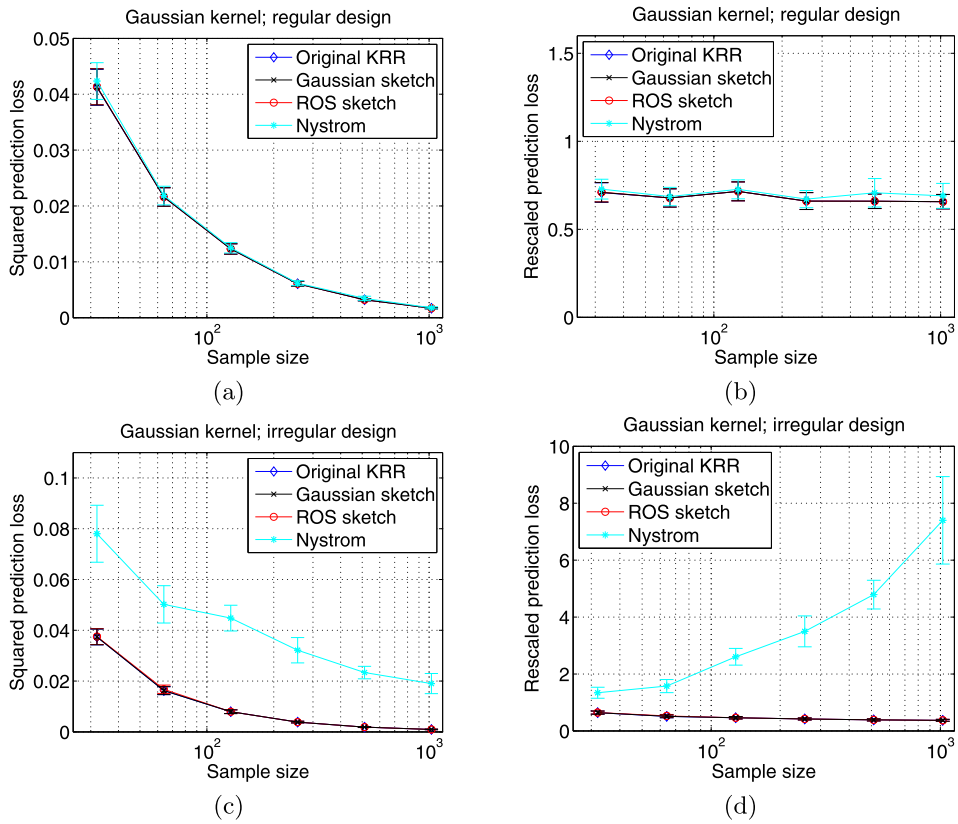

FIG. 3. *Prediction error versus sample size for original KRR, Gaussian sketch, ROS sketch and Nyström approximation. Left panels* (a) *and* (c) *show* $\|\widehat{f} - f^*\|_n^2$ *versus the sample size* $n \in \{32, 64, 128, 256, 512, 1024\}$ *for projection dimension* $m = \lceil 4\sqrt{\log n} \rceil$. *In all cases, each point corresponds to the average of* 100 *trials, with standard errors also shown. Right panels* (b) *and* (d) *show the rescaled prediction error* $\frac{n}{\sqrt{\log n}} \|\widehat{f} - f^*\|_n^2$ *versus the sample size. Top row correspond to covariates arranged uniformly on the unit interval, whereas bottom row corresponds to an irregular design* (*see text for details*).

with δ_n^2. Putting together the pieces, we see that the sampling parameter $\tilde{p}$ must satisfy the lower bound $\tilde{p} \gtrsim \delta_n^{-2} \log n$. This requirement is much larger than the statistical dimension, and prohibitive in many cases:

- For the Gaussian kernel, we have $\delta_n^2 \asymp \frac{\sqrt{\log(n)}}{n}$, and so $\tilde{p} \gtrsim n \log^{1/2}(n)$, meaning that all rows of the kernel matrix are sampled. In contrast, the statistical dimension scales as $\sqrt{\log n}$.
- For the first-order Sobolev kernel, we have $\delta_n^2 \asymp n^{-2/3}$, so that $\tilde{p} \gtrsim n^{2/3} \log n$. In contrast, the statistical dimension for this kernel scales as $n^{1/3}$.

It remains an open question as to whether a more efficient procedure for approximating the leverage scores might be devised, which would allow a method of this type to be statistically optimal in terms of the sampling dimension.

4. Proofs. In this section, we provide the proofs of our main theorems. Some technical proofs of the intermediate results are provided in the Appendices.

4.1. *Proof of Theorem 2.* Recall the definition (14) of the estimate $f^\dagger$, as well as the upper bound (15) in terms of approximation and estimation error terms. The remainder of our proof consists of two technical lemmas used to control these two terms.

LEMMA 1 (Control of estimation error). *Under the conditions of Theorem 2, we have*

$$\|f^\dagger - \widehat{f}\|_n^2 \leq c\delta_n^2 \tag{18}$$

with probability at least $1 - c_1 e^{-c_2 n\delta_n^2}$.

LEMMA 2 (Control of approximation error). *For any K-satisfiable sketch matrix S, we have*

$$\|f^\dagger - f^*\|_n^2 \leq c\{\lambda_n + \delta_n^2\} \quad \text{and} \quad \|f^\dagger\|_{\mathcal{H}} \leq c\left\{1 + \frac{\delta_n^2}{\lambda_n}\right\}. \tag{19}$$

These two lemmas, in conjunction with the upper bound (15), yield the claim in the theorem statement. Accordingly, it remains to prove the two lemmas.

4.1.1. *Proof of Lemma 1.* So as to simplify notation, we assume throughout the proof that $\sigma = 1$. (A simple rescaling argument can be used to recover the general statement.) Since $\alpha^\dagger$ is optimal for the quadratic program (14), it must satisfy the zero gradient condition

$$-SK\left(\frac{1}{\sqrt{n}} f^* - KS^T\alpha^\dagger\right) + 2\lambda_n SKS^T\alpha^\dagger = 0. \tag{20}$$

By the optimality of $\widehat{\alpha}$ and feasibility of $\alpha^\dagger$ for the sketched problem (5a), we have

$$\begin{aligned}\frac{1}{2}\|KS^T\widehat{\alpha}\|_2^2 - \frac{1}{\sqrt{n}} y^T KS^T\widehat{\alpha} + \lambda_n \|K^{1/2}S^T\widehat{\alpha}\|_2^2 \\ \leq \frac{1}{2}\|KS^T\alpha^\dagger\|_2^2 - \frac{1}{\sqrt{n}} y^T KS^T\alpha^\dagger + \lambda_n \|K^{1/2}S^T\alpha^\dagger\|_2^2.\end{aligned}$$

Defining the error vector $\widehat{\Delta} := S^T(\widehat{\alpha} - \alpha^\dagger)$, some algebra leads to the following inequality:

$$\begin{aligned}\frac{1}{2}\|K\widehat{\Delta}\|_2^2 \le -\langle K\widehat{\Delta}, KS^T\alpha^\dagger\rangle + \frac{1}{\sqrt{n}} y^T K\widehat{\Delta} \\ + \lambda_n \|K^{1/2}S^T\alpha^\dagger\|_2^2 - \lambda_n \|K^{1/2}S^T\widehat{\alpha}\|_2^2.\end{aligned} \tag{21}$$

Consequently, by plugging in $y = z^* + w$ and applying the optimality condition (20), we obtain the basic inequality

$$\frac{1}{2}\|K\widehat{\Delta}\|_2^2 \le \left|\frac{1}{\sqrt{n}} w^T K\widehat{\Delta}\right| - \lambda_n \|K^{1/2}\widehat{\Delta}\|_2^2. \tag{22}$$

The following lemma provides control on the right-hand side.

LEMMA 3. *With probability at least* $1 - c_1 e^{-c_2 n\delta_n^2}$, *we have that for all* $\Delta \in \mathbb{R}^n$,

$$\begin{aligned}&\left|\frac{1}{\sqrt{n}} w^T K\Delta\right| \\ &\le \begin{cases} 6\delta_n \|K\Delta\|_2 + 2\delta_n^2, & \text{if } \|K^{1/2}\Delta\|_2 \le 1, \\ 2\delta_n \|K\Delta\|_2 + 2\delta_n^2 \|K^{1/2}\Delta\|_2 + \frac{1}{16}\delta_n^2, & \text{if } \|K^{1/2}\Delta\|_2 \ge 1. \end{cases}\end{aligned} \tag{23}$$

See Appendix B.2 for the proof of this lemma.

Based on this auxiliary result, we divide the remainder of our analysis into two cases.

Case 1: If $\|K^{1/2}\widehat{\Delta}\|_2 \le 1$, then the basic inequality (22) and the top inequality in Lemma 3 imply

$$\frac{1}{2}\|K\widehat{\Delta}\|_2^2 \le \left|\frac{1}{\sqrt{n}} w^T K\widehat{\Delta}\right| \le 6\delta_n \|K\widehat{\Delta}\|_2 + 2\delta_n^2 \tag{24}$$

with probability at least $1 - c_1 e^{-c_2 n\delta_n^2}$. Note that we have used that fact that the randomness in the sketch matrix S is independent of the randomness in the noise vector w. The quadratic inequality (24) implies that $\|K\widehat{\Delta}\|_2 \le c\delta_n$ for some universal constant c.

Case 2: If $\|K^{1/2}\widehat{\Delta}\|_2 > 1$, then the basic inequality (22) and the bottom inequality in Lemma 3 imply

$$\frac{1}{2}\|K\widehat{\Delta}\|_2^2 \le 2\delta_n \|K\widehat{\Delta}\|_2 + 2\delta_n^2 \|K^{1/2}\widehat{\Delta}\|_2 + \frac{1}{16}\delta_n^2 - \lambda_n \|K^{1/2}\widehat{\Delta}\|_2^2$$

with probability at least $1 - c_1 e^{-c_2 n\delta_n^2}$. If $\lambda_n \ge 2\delta_n^2$, then under the assumed condition $\|K^{1/2}\widehat{\Delta}\|_2 > 1$, the above inequality gives

$$\frac{1}{2}\|K\widehat{\Delta}\|_2^2 \le 2\delta_n \|K\widehat{\Delta}\|_2 + \frac{1}{16}\delta_n^2 \le \frac{1}{4}\|K\widehat{\Delta}\|_2^2 + 4\delta_n^2 + \frac{1}{16}\delta_n^2.$$

By rearranging terms in the above, we obtain $\|K\widehat{\Delta}\|_2^2 \leq c\delta_n^2$ for a universal constant, which completes the proof.

4.1.2. *Proof of Lemma 2.* Without loss of generality, we may assume $\|f^*\|_{\mathcal{H}} \leq 1$. Our goal is to show that the bound

$$\frac{1}{2n}\|z^* - \sqrt{n}KS^T\alpha^\dagger\|_2^2 + \lambda_n\|K^{1/2}S^T\alpha^\dagger\|_2^2 \leq c\{\lambda_n + \delta_n^2\}.$$

In fact, since $\alpha^\dagger$ is a minimizer, it suffices to exhibit some $\alpha \in \mathbb{R}^m$ for which this inequality holds. Recalling the eigendecomposition $K = UDU^T$, it is equivalent to exhibit some $\alpha \in \mathbb{R}^m$ such that

$$\frac{1}{2}\|\theta^* - D\widetilde{S}^T\alpha\|_2^2 + \lambda_n\alpha^T\widetilde{S}D\widetilde{S}^T\alpha \leq c\{\lambda_n + \delta_n^2\}, \tag{25}$$

where $\widetilde{S} = SU$ is the transformed sketch matrix, and the vector $\theta^* = n^{-1/2}Uz^* \in \mathbb{R}^n$ satisfies the ellipse constraint $\|D^{-1/2}\theta^*\|_2 \leq 1$.

We do so via a constructive procedure. First, we partition the vector $\theta^* \in \mathbb{R}^n$ into two sub-vectors, namely $\theta_1^* \in \mathbb{R}^{d_n}$ and $\theta_2^* \in \mathbb{R}^{n-d_n}$. Similarly, we partition the diagonal matrix D into two blocks, D_1 and D_2, with dimensions d_n and $n - d_n$, respectively. Under the condition $m > d_n$, we may let $\widetilde{S}_1 \in \mathbb{R}^{m\times d_n}$ denote the left block of the transformed sketch matrix, and similarly, let $\widetilde{S}_2 \in \mathbb{R}^{m\times(n-d_n)}$ denote the right block. In terms of this notation, the assumption that S is K-satisfiable corresponds to the inequalities

$$\|\widetilde{S}_1^T\widetilde{S}_1 - I_{d_n}\|_{\mathrm{op}} \leq \frac{1}{2} \quad \text{and} \quad \|\widetilde{S}_2\sqrt{D_2}\|_{\mathrm{op}} \leq c\delta_n. \tag{26}$$

As a consequence, we are guarantee that the matrix $\widetilde{S}_1^T\widetilde{S}_1$ is invertible, so that we may define the m-dimensional vector

$$\widehat{\alpha} = \widetilde{S}_1(\widetilde{S}_1^T\widetilde{S}_1)^{-1}(D_1)^{-1}\theta_1^* \in \mathbb{R}^m.$$

Recalling the disjoint partition of our vectors and matrices, we have

$$\begin{aligned}\|\theta^* - D\widetilde{S}^T\widehat{\alpha}\|_2^2 \\ = \underbrace{\|\theta_1^* - D_1\widetilde{S}_1^T\widehat{\alpha}\|_2^2}_{=0} + \underbrace{\|\theta_2^* - D_2\widetilde{S}_2^T\widetilde{S}_1(\widetilde{S}_1^T\widetilde{S}_1)^{-1}D_1^{-1}\theta_1^*\|_2^2}_{T_1^2}.\end{aligned} \tag{27a}$$

By the triangle inequality, we have

$$\begin{aligned}T_1 &\leq \|\theta_2^*\|_2 + \|D_2\widetilde{S}_2^T\widetilde{S}_1(\widetilde{S}_1^T\widetilde{S}_1)^{-1}D_1^{-1}\theta_1^*\|_2 \\ &\leq \|\theta_2^*\|_2 + \|D_2\widetilde{S}_2^T\|_{\mathrm{op}}\|\widetilde{S}_1\|_{\mathrm{op}}\|(\widetilde{S}_1^T\widetilde{S}_1)^{-1}\|_{\mathrm{op}}\|D_1^{-1/2}\|_{\mathrm{op}}\|D_1^{-1/2}\theta_1^*\|_2 \\ &\leq \|\theta_2^*\|_2 + \|\sqrt{D_2}\|_{\mathrm{op}}\|\widetilde{S}_2\sqrt{D_2}\|_{\mathrm{op}}\|\widetilde{S}_1\|_{\mathrm{op}} \\ &\quad \times \|(\widetilde{S}_1^T\widetilde{S}_1)^{-1}\|_{\mathrm{op}}\|D_1^{-1/2}\|_{\mathrm{op}}\|D_1^{-1/2}\theta_1^*\|_2.\end{aligned}$$

Since $\|D^{-1/2}\theta^*\|_2 \le 1$, we have $\|D_1^{-1/2}\theta_1^*\|_2 \le 1$ and, moreover,

$$\|\theta_2^*\|_2^2 = \sum_{j=d_n+1}^{n} (\theta_{2,j}^*)^2 \le \delta_n^2 \sum_{j=d_n+1}^{n} \frac{(\theta_{2,j}^*)^2}{\widehat{\mu}_j} \le \delta_n^2,$$

since $\widehat{\mu}_j \le \delta_n^2$ for all $j \ge d_n + 1$. Similarly, we have $\|\sqrt{D_2}\|_{\mathrm{op}} \le \sqrt{\widehat{\mu}_{d_n+1}} \le \delta_n$, and $\|D_1^{-1/2}\|_{\mathrm{op}} \le \delta_n^{-1}$. Putting together the pieces, we have

$$T_1 \le \delta_n + \|\widetilde{S}_2\sqrt{D_2}\|_{\mathrm{op}}\|\widetilde{S}_1\|_{\mathrm{op}}\|(\widetilde{S}_1^T\widetilde{S}_1)^{-1}\|_{\mathrm{op}} \le (c\delta_n)\sqrt{\frac{3}{2}}2 = c'\delta_n, \tag{27b}$$

where we have invoked the K-satisfiability of the sketch matrix to guarantee the bounds $\|\widetilde{S}_1\|_{\mathrm{op}} \le \sqrt{3/2}$, $\|(\widetilde{S}_1^T\widetilde{S})\|_{\mathrm{op}} \ge 1/2$ and $\|\widetilde{S}_2\sqrt{D_2}\|_{\mathrm{op}} \le c\delta_n$. Bounds (27a) and (27b) in conjunction guarantee that

$$\|\theta^* - D\widetilde{S}^T\widehat{\alpha}\|_2^2 \le c\delta_n^2, \tag{28a}$$

where the value of the universal constant c may change from line to line.

Turning to the remaining term on the left-hand side of inequality (25), applying the triangle inequality and the previously stated bounds leads to

$$\begin{aligned}\widehat{\alpha}^T\widetilde{S}D\widetilde{S}^T\widehat{\alpha} &\le \|D_1^{-1/2}\theta_1^*\|_2^2 + \|D_2^{1/2}\widetilde{S}_2^T\|_{\mathrm{op}}\|\widetilde{S}_1\|_{\mathrm{op}} \\ &\quad \times \|(\widetilde{S}_1^T\widetilde{S}_1)^{-1}\|_{\mathrm{op}}\|D_1^{-1/2}\|_{\mathrm{op}}\|D_1^{-1/2}\theta_1^*\|_2 \\ &\le 1 + (c\delta_n)\sqrt{3/2}\frac{1}{2}\delta_n^{-1}(1) \le c'.\end{aligned} \tag{28b}$$

Combining the two bounds (28a) and (28b) yields the claim (25).

5. Discussion. In this paper, we have analyzed randomized sketching methods for kernel ridge regression. Our main theorem gives sufficient conditions on any sketch matrix for the sketched estimate to achieve the minimax risk for nonparametric regression over the underlying kernel class. We specialized this general result to two broad classes of sketches, namely those based on Gaussian random matrices and randomized orthogonal systems (ROS), for which we proved that a sketch size proportional to the statistical dimension is sufficient to achieve the minimax risk. More broadly, we suspect that sketching methods of the type analyzed here have the potential to save time and space in other forms of statistical computation, and we hope that the results given here are useful for such explorations.

APPENDIX A: SUB-SAMPLING SKETCHES YIELD NYSTRÖM APPROXIMATION

In this Appendix, we show that the sub-sampling sketch matrix described at the end of Section 2.2 coincides with applying Nyström approximation [34] to the kernel matrix.

We begin by observing that the original KRR quadratic program (4a) can be written in the equivalent form $\min_{\omega\in\mathbb{R}^n, u\in\mathbb{R}^n}\{\frac{1}{2n}\|u\|^2 + \lambda_n \omega^T K\omega\}$ such that $y - \sqrt{n}K\omega = u$. The dual of this constrained quadratic program (QP) is given by

$$\xi^\dagger = \arg\max_{\xi\in\mathbb{R}^n}\left\{-\frac{n}{4\lambda_n}\xi^T K\xi + \xi^T y - \frac{1}{2}\xi^T\xi\right\}. \tag{29}$$

The KRR estimate $f^\dagger$ and the original solution $\omega^\dagger$ can be recovered from the dual solution $\xi^\dagger$ via the relation $f^\dagger(\cdot) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \omega^\dagger{}_i \mathcal{K}(\cdot, x_i)$ and $\omega^\dagger = \frac{\sqrt{n}}{2\lambda_n}\xi^\dagger$.

Now turning to the sketched KRR program (5a), note that it can be written in the equivalent form $\min_{\alpha\in\mathbb{R}^n, u\in\mathbb{R}^n}\{\frac{1}{2n}\|u\|^2 + \lambda_n \alpha^T SKS^T\alpha\}$ subject to the constraint $y - \sqrt{n}KS^T\alpha = u$. The dual of this constrained QP is given by

$$\xi^\ddagger = \arg\max_{\xi\in\mathbb{R}^n}\left\{-\frac{n}{4\lambda_n}\xi^T \widetilde{K}\xi + \xi^T y - \frac{1}{2}\xi^T\xi\right\}, \tag{30}$$

where $\widetilde{K} = KS^T(SKS^T)^{-1}SK$ is a rank-m matrix in $\mathbb{R}^{n\times n}$. In addition, the sketched KRR estimate $\widehat{f}$, the original solution $\widehat{\alpha}$ and the dual solution $\xi^\ddagger$ are related by $\widehat{f}(\cdot) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (S^T\widehat{\alpha})_i \mathcal{K}(\cdot, x_i)$ and $\widehat{\alpha} = \frac{\sqrt{n}}{2\lambda_n}(SKS^T)^{-1}SK\xi^\ddagger$.

When S is the sub-sampling sketch matrix, the matrix $\widetilde{K} = KS^T(SKS^T)^{-1}SK$ is known as the Nyström approximation [34]. Consequently, the dual formulation of sketched KRR based on a sub-sampling matrix can be viewed as the Nyström approximation as applied to the dual formulation of the original KRR problem.

APPENDIX B: TECHNICAL PROOFS

Without loss of generality, we assume that $\{x_i\}_{i=1}^n$ are fixed and otherwise we can view everything as conditioning on $\{x_i\}_{i=1}^n$.

B.1. Proof of Theorem 1. We begin by converting the problem to an instance of the normal sequence model [15]. Recall that the kernel matrix can be decomposed as $K = U^T DU$, where $U \in \mathbb{R}^{n\times n}$ is orthonormal, and $D = \operatorname{diag}\{\widehat{\mu}_1, \ldots, \widehat{\mu}_n\}$. Any function $f^* \in \mathcal{H}$ can be decomposed as

$$f^* = \frac{1}{\sqrt{n}}\sum_{j=1}^n \mathcal{K}(\cdot, x_j)\big(U^T\beta^*\big)_j + g, \tag{31}$$

for some vector $\beta^* \in \mathbb{R}^n$, and some function $g \in \mathcal{H}$ is orthogonal to $\operatorname{span}\{\mathcal{K}(\cdot, x_j), j = 1, \ldots, n\}$. Consequently, the inequality $\|f^*\|_{\mathcal{H}} \le 1$ implies that

$$\left\|\frac{1}{\sqrt{n}}\sum_{j=1}^n \mathcal{K}(\cdot, x_j)\big(U^T\beta^*\big)_j\right\|_{\mathcal{H}}^2 = \big(U^T\beta^*\big)^T U^T DU\big(U^T\beta^*\big) = \|\sqrt{D}\beta^*\|_2^2 \le 1.$$

Moreover, we have $f^*(x_1^n) = \sqrt{n}U^T D\beta^*$, and so the original observation model (1) has the equivalent form $y = \sqrt{n}U^T\theta^* + w$, where $\theta^* = D\beta^*$. In fact,

due to the rotation invariance of the Gaussian, it is equivalent to consider the normal sequence model

$$\widetilde{y} = \theta^* + \frac{w}{\sqrt{n}}. \tag{32}$$

Any estimate $\widetilde{\theta}$ of θ^* defines the function estimate $\widetilde{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathcal{K}(\cdot, x_i) \times (U^T D^{-1} \widetilde{\theta})_i$, and by construction, we have $\|\widetilde{f} - f^*\|_n^2 = \|\widetilde{\theta} - \theta^*\|_2^2$. Finally, the original constraint $\|\sqrt{D}\beta^*\|_2^2 \leq 1$ is equivalent to $\|D^{-1/2}\theta^*\|_2 \leq 1$. Thus, we have a version of the normal sequence model subject to an ellipse constraint.

After this reduction, we can assume that we are given n i.i.d. observations $\widetilde{y}_1^n = \{\widetilde{y}_1, \ldots, \widetilde{y}_n\}$, and our goal is to lower bound the Euclidean error $\|\widetilde{\theta} - \theta^*\|_2^2$ of any estimate of θ^*. In order to do so, we first construct a $\delta/2$-packing of the set $\mathcal{B} = \{\theta \in \mathbb{R}^n \mid \|D^{-1/2}\theta\|_2 \leq 1\}$, say $\{\theta^1, \ldots, \ldots, \theta^M\}$. Now consider the random ensemble of regression problems in which we first draw an index A uniformly at random from the index set $[M]$, and then conditioned on $A = a$, we observe n i.i.d. samples from the nonparametric regression model with $f^* = f^a$. Given this setup, a standard argument using Fano's inequality implies that

$$\mathbb{P}\Big[\|\widetilde{f} - f^*\|_n^2 \geq \frac{\delta^2}{4}\Big] \geq 1 - \frac{I(\widetilde{y}_1^n; A) + \log 2}{\log M},$$

where $I(\widetilde{y}_1^n; A)$ is the mutual information between the samples $\widetilde{y}_1^n$ and the random index A. It remains to construct the desired packing and to upper bound the mutual information.

For a given $\delta > 0$, define the ellipse

$$\mathcal{E}(\delta) := \Bigg\{\theta \in \mathbb{R}^n \Big| \underbrace{\sum_{j=1}^n \frac{\theta_j^2}{\min\{\delta^2, \widehat{\mu}_j\}}}_{\|\theta\|_{\mathcal{E}}^2} \leq 1\Bigg\}. \tag{33}$$

By construction, observe that $\mathcal{E}(\delta)$ is contained within Hilbert ball of unit radius. Consequently, it suffices to construct a $\delta/2$-packing of this ellipse in the Euclidean norm.

LEMMA 4. *For any $\delta \in (0, \delta_n]$, there is a $\delta/2$-packing of the ellipse $\mathcal{E}(\delta)$ with cardinality*

$$\log M = \frac{1}{64} d_n. \tag{34}$$

Taking this packing as given, note that by construction we have

$$\|\theta^a\|_2^2 = \delta^2 \sum_{j=1}^n \frac{(\theta^a)_j^2}{\delta^2} \leq \delta^2, \quad \text{and hence} \quad \|\theta^a - \theta^b\|_2^2 \leq 4\delta^2.$$

In conjunction with concavity of the KL diveregence, we have

$$I(y_1^n; J) \leq \frac{1}{M^2} \sum_{a,b=1}^{M} D(\mathbb{P}^a \| \mathbb{P}^b) = \frac{1}{M^2} \frac{n}{2\sigma^2} \sum_{a,b=1}^{M} \|\theta^a - \theta^b\|_2^2 \leq \frac{2n}{\sigma^2} \delta^2.$$

For any δ such that $\log 2 \leq \frac{2n}{\sigma^2}\delta^2$ and $\delta \leq \delta_n$ we have

$$\mathbb{P}\left[\|\widetilde{f} - f^*\|_n^2 \geq \frac{\delta^2}{4} \right] \geq 1 - \frac{4n\delta^2/\sigma^2}{d_n/64}.$$

Moreover, since the kernel is regular, we have $\sigma^2 d_n \geq cn\delta_n^2$ for some positive constant c. Thus, setting $\delta^2 = \frac{c\delta_n^2}{512}$ yields the claim.

PROOF OF LEMMA 4. It remains to prove the lemma, and we do so via the probabilistic method. Consider a random vector $\theta \in \mathbb{R}^n$ of the form

$$\theta = \left[\frac{\delta}{\sqrt{2d_n}} w_1 \quad \frac{\delta}{\sqrt{2d_n}} w_2 \quad \cdots \quad \frac{\delta}{\sqrt{2d_n}} w_{d_n} \quad 0 \quad \cdots \quad 0 \right], \tag{35}$$

where $w = (w_1, \ldots, w_{d_n})^T \sim N(0, I_{d_n})$ is a standard Gaussian vector. We claim that a collection of M such random vectors $\{\theta^1, \ldots, \theta^M\}$, generated in an i.i.d. manner, defines the required packing with high probability.

On one hand, for each index $a \in [M]$, since $\delta^2 \leq \delta_n^2 \leq \widehat{\mu}_j$ for each $j \leq d_n$, we have $\|\theta^a\|_{\mathcal{E}}^2 = \frac{\|w^a\|_2^2}{2d_n}$, corresponding to a normalized χ^2-variate. Consequently, by a combination of standard tail bounds and the union bound, we have

$$\mathbb{P}\big[\|\theta^a\|_{\mathcal{E}}^2 \leq 1 \text{ for all } a \in [M]\big] \geq 1 - Me^{-\frac{d_n}{16}}.$$

Now consider the difference vector $\theta^a - \theta^b$. Since the underlying Gaussian noise vectors w^a and w^b are independent, the difference vector $w^a - w^b$ follows a $N(0, 2I_m)$ distribution. Consequently, the event $\|\theta^a - \theta^b\|_2 \geq \frac{\delta}{2}$ is equivalent to the event $\sqrt{2}\|\theta\|_2 \geq \frac{\delta}{2}$, where θ is a random vector drawn from the original ensemble. Note that $\|\theta\|_2^2 = \delta^2 \frac{\|w\|_2^2}{2d_n}$. Then a combination of standard tail bounds for χ^2-distributions and the union bound argument yields

$$\mathbb{P}\left[\|\theta^a - \theta^b\|_2^2 \geq \frac{\delta^2}{4} \text{ for all } a, b \in [M] \right] \geq 1 - M^2 e^{-\frac{d_n}{16}}.$$

Combining the last two display together, we obtain

$$\begin{aligned} &\mathbb{P}\left[\|\theta^a\|_{\mathcal{E}}^2 \leq 1 \text{ and } \|\theta^a - \theta^b\|_2^2 \geq \frac{\delta^2}{4} \text{ for all } a, b \in [M] \right] \\ &\qquad \geq 1 - Me^{-\frac{d_n}{16}} - M^2 e^{-\frac{d_n}{16}}. \end{aligned}$$

This probability is positive for $\log M = d_n/64$. $\square$

B.2. Proof of Lemma 3. For use in the proof, for each $\delta > 0$, let us define the random variable

$$Z_n(\delta) = \sup_{\substack{\|K^{1/2}\Delta\|_2 \le 1 \\ \|K\Delta\|_2 \le \delta}} \left| \frac{1}{\sqrt{n}} w^T K \Delta \right|. \tag{36}$$

Top inequality in the bound (23): If the top inequality is violated, then we claim that we must have $Z_n(\delta_n) > 2\delta_n^2$. On one hand, if the bound (23) is violated by some vector $\Delta \in \mathbb{R}^n$ with $\|K\Delta\|_2 \le \delta_n$, then we have

$$2\delta_n^2 \le \left| \frac{1}{\sqrt{n}} w^T K \Delta \right| \le Z_n(\delta_n).$$

On the other hand, if the bound is violated by some function with $\|K\Delta\|_2 > \delta_n$, then we can define the rescaled vector $\widetilde{\Delta} = \frac{\delta_n}{\|K\Delta\|_2}\Delta$, for which we have

$$\|K\widetilde{\Delta}\|_2 = \delta_n, \quad \text{and} \quad \|K^{1/2}\widetilde{\Delta}\|_2 = \frac{\delta_n}{\|K\Delta\|_2}\|K^{1/2}\Delta\|_2 \le 1$$

showing that $Z_n(\delta_n) \ge 2\delta_n^2$ as well.

When viewed as a function of the standard Gaussian vector $w \in \mathbb{R}^n$, it is easy to see that $Z_n(\delta_n)$ is Lipschitz with parameter $\delta_n/\sqrt{n}$. Consequently, by concentration of measure for Lipschitz functions of Gaussians [19], we have

$$\mathbb{P}\big[Z_n(\delta_n) \ge \mathbb{E}\big[Z_n(\delta_n)\big] + t\big] \le e^{-\frac{nt^2}{2\delta_n^2}}. \tag{37}$$

Moreover, we claim that

$$\mathbb{E}\big[Z_n(\delta_n)\big] \overset{\text{(i)}}{\le} \underbrace{\sqrt{\frac{1}{n}\sum_{i=1}^n \min\{\delta_n^2, \widehat{\mu}_j\}}}_{\widehat{\mathcal{R}}(\delta_n)} \overset{\text{(ii)}}{\le} \delta_n^2, \tag{38}$$

where inequality (ii) follows by definition of the critical radius (recalling that we have set $\sigma = 1$ by a rescaling argument). Setting $t = \delta_n^2$ in the tail bound (37), we see that $\mathbb{P}[Z_n(\delta_n) \ge 2\delta_n^2] \le e^{n\delta_n^2/2}$, which completes the proof of the top bound.

It only remains to prove inequality (i) in equation (38). The kernel matrix K can be decomposed as $K = U^T D U$, where $D = \operatorname{diag}\{\widehat{\mu}_1, \ldots, \widehat{\mu}_n\}$, and U is a unitary matrix. Defining the vector $\beta = DU\Delta$, the two constraints on Δ can be expressed as $\|D^{-1/2}\beta\|_2 \le 1$ and $\|\beta\|_2 \le \delta$. Note that any vector satisfying these two constraints must belong to the ellipse

$$\mathcal{E} := \left\{ \beta \in \mathbb{R}^n \,\Big|\, \sum_{j=1}^n \frac{\beta_j^2}{v_j} \le 2 \text{ where } v_j = \max\{\delta_n^2, \widehat{\mu}_j\} \right\}.$$

Consequently, we have

$$\mathbb{E}\big[Z_n(\delta_n)\big] \le \mathbb{E}\Big[\sup_{\beta\in\mathcal{E}} \frac{1}{\sqrt{n}}\big|\langle U^T w, \beta\rangle\big|\Big] = \mathbb{E}\Big[\sup_{\beta\in\mathcal{E}} \frac{1}{\sqrt{n}}\big|\langle w, \beta\rangle\big|\Big],$$

since $U^T w$ also follows a standard normal distribution. By the Cauchy–Schwarz inequality, we have

$$\mathbb{E}\Big[\sup_{\beta\in\mathcal{E}} \frac{1}{\sqrt{n}}\big|\langle w, \beta\rangle\big|\Big] \le \frac{1}{\sqrt{n}}\mathbb{E}\sqrt{\sum_{j=1}^n \nu_j w_j^2} \le \underbrace{\frac{1}{\sqrt{n}}\sqrt{\sum_{j=1}^n \nu_j}}_{\widehat{\mathcal{R}}(\delta_n)},$$

where the final step follows from Jensen's inequality.

Bottom inequality in the bound (23): We now turn to the proof of the bottom inequality. We claim that it suffices to show that

$$\Big|\frac{1}{\sqrt{n}} w^T K\widetilde{\Delta}\Big| \le 2\delta_n \|K\widetilde{\Delta}\|_2 + 2\delta_n^2 + \frac{1}{16}\|K\widetilde{\Delta}\|_2^2 \tag{39}$$

for all $\widetilde{\Delta} \in \mathbb{R}^n$ such that $\|K^{1/2}\widetilde{\Delta}\|_2 = 1$. Indeed, for any vector $\Delta \in \mathbb{R}^n$ with $\|K^{1/2}\Delta\|_2 > 1$, we can define the rescaled vector $\widetilde{\Delta} = \Delta/\|K^{1/2}\Delta\|_2$, for which we have $\|K^{1/2}\widetilde{\Delta}\|_2 = 1$. Applying the bound (39) to this choice and then multiplying both sides by $\|K^{1/2}\Delta\|_2$, we obtain

$$\begin{aligned}\Big|\frac{1}{\sqrt{n}} w^T K\Delta\Big| &\le 2\delta_n \|K\Delta\|_2 + 2\delta_n^2 \big\|K^{1/2}\Delta\big\|_2 + \frac{1}{16}\frac{\|K\Delta\|_2^2}{\|K^{1/2}\Delta\|_2} \\ &\le 2\delta_n \|K\Delta\|_2 + 2\delta_n^2 \big\|K^{1/2}\Delta\big\|_2 + \frac{1}{16}\|K\Delta\|_2^2,\end{aligned}$$

as required.

Recall the family of random variables Z_n previously defined (36). For any $u \ge \delta_n$, we have

$$\mathbb{E}\big[Z_n(u)\big] = \widehat{\mathcal{R}}(u) = u\frac{\widehat{\mathcal{R}}(u)}{u} \overset{\text{(i)}}{\le} u\frac{\widehat{\mathcal{R}}(\delta_n)}{\delta_n} \overset{\text{(ii)}}{\le} u\delta_n,$$

where inequality (i) follows since the function $u \mapsto \frac{\widehat{\mathcal{R}}(u)}{u}$ is nonincreasing, and step (ii) follows by our choice of δ_n. Setting $t = \frac{u^2}{32}$ in the concentration bound (37), we conclude that

$$\mathbb{P}\Big[Z_n(u) \ge u\delta_n + \frac{u^2}{64}\Big] \le e^{-cnu^2} \qquad \text{for each } u \ge \delta_n. \tag{40}$$

We are now equipped to prove the bound (39) via a "peeling" argument. Let $\mathcal{E}$ denote the event that the bound (39) is violated for some vector $\widetilde{\Delta}$ with $\|K^{1/2}\widetilde{\Delta}\|_2 = 1$. For real numbers $0 \le a < b$, let $\mathcal{E}(a, b)$ denote the event that it

is violated for some vector with $\|K^{1/2}\Delta\|_2 = 1$ and $\|K\widetilde{\Delta}\|_2 \in [a, b]$. For $m = 0, 1, 2, \ldots,$ define $u_m = 2^m \delta_n$. We then have the decomposition $\mathcal{E} = \mathcal{E}(0, u_0) \cup (\bigcup_{m=0}^{\infty} \mathcal{E}(u_m, u_{m+1}))$, and hence by union bound,

$$\mathbb{P}[\mathcal{E}] \leq \mathbb{P}\big[\mathcal{E}(0, u_0)\big] + \sum_{m=0}^{\infty} \mathbb{P}\big[\mathcal{E}(u_m, u_{m+1})\big]. \tag{41}$$

The final step is to bound each of the terms in this summation, Since $u_0 = \delta_n$, we have

$$\mathbb{P}\big[\mathcal{E}(0, u_0)\big] \leq \mathbb{P}\big[Z_n(\delta_n) \geq 2\delta_n^2\big] \leq e^{-cn\delta_n^2}. \tag{42}$$

On the other hand, suppose that $\mathcal{E}(u_m, u_{m+1})$ holds, meaning that there exists some vector $\widetilde{\Delta}$ with $\|K^{1/2}\widetilde{\Delta}\|_2 = 1$ and $\|K\widetilde{\Delta}\|_2 \in [u_m, u_{m+1}]$ such that

$$\begin{aligned}
\left|\frac{1}{\sqrt{n}} w^T K\widetilde{\Delta}\right| &\geq 2\delta_n \|K\widetilde{\Delta}\|_2 + 2\delta_n^2 + \frac{1}{16}\|K\widetilde{\Delta}\|_2^2 \\
&\geq 2\delta_n u_m + 2\delta_n^2 + \frac{1}{16} u_m^2 \\
&\geq \delta_n u_{m+1} + \frac{1}{64} u_{m+1}^2,
\end{aligned}$$

where the second inequality follows since $\|K\widetilde{\Delta}\|_2 \geq u_m$; and the third inequality follows since $u_{m+1} = 2u_m$. This lower bound implies that $Z_n(u_{m+1}) \geq \delta_n u_{m+1} + \frac{u_{m+1}^2}{64}$, whence the bound (40) implies that

$$\mathbb{P}\big[\mathcal{E}(u_m, u_{m+1})\big] \leq e^{-cnu_{m+1}^2} \leq e^{-cn2^{2m}\delta_n^2}.$$

Combining this tail bound with our earlier bound (42) and substituting into the union bound (41) yields

$$\mathbb{P}[\mathcal{E}] \leq e^{-cn\delta_n^2} + \sum_{m=0}^{\infty} \exp\big(-cn2^{2m}\delta_n^2\big) \leq c_1 e^{-c_2 n\delta_n^2},$$

as claimed.

B.3. Proof of Corollary 1. Based on Theorem 2, we need to verify that the stated lower bound (16a) on the projection dimension is sufficient to guarantee that a random sketch matrix is K-satisfiable is high probability. In particular, let us state this guarantee as a formal claim.

LEMMA 5. *Under the lower bound* (16a) *on the sketch dimension, a* {*Gaussian*, *ROS*} *random sketch is* K*-satisfiable with probability at least* $\phi(m, d_n, n)$.

We split our proof into two parts, one for each inequality in the definition (13) of K-satisfiability.

B.3.1. *Proof of inequality* (i). We need to bound the operator norm of the matrix $Q = U_1^T S^T S U_1 - I_{d_n}$, where the matrix $U_1 \in \mathbb{R}^{n \times d_n}$ has orthonormal columns. Let $\{v^1, \ldots, v^N\}$ be a $1/2$-cover of the Euclidean sphere $\mathcal{S}^{d_n - 1}$; by standard arguments [21], we can find such a set with $N \leq e^{2d_n}$ elements. Using this cover, a straightforward discretization argument yields

$$|||Q|||_{\mathrm{op}} \leq 4 \max_{j,k=1,\ldots,N} \langle v^j, Q v^k \rangle = 4 \max_{j,k=1,\ldots,N} (\widetilde{v})^j \{S^T S - I_n\} \widetilde{v}^k,$$

where $\widetilde{v}^j := U_1 v^j \in \mathcal{S}^{n-1}$, and $\widetilde{Q} = S^T S - I_n$. In the Gaussian case, standard subexponential bounds imply that $\mathbb{P}[(\widetilde{v})^j \widetilde{Q} \widetilde{v}^k \geq 1/8] \leq c_1 e^{-c_2 m}$, and consequently, by the union bound, we have

$$\mathbb{P}\big[|||Q|||_{\mathrm{op}} \geq 1/2\big] \leq c_1 e^{-c_2 m + 4 d_n} \leq c_1 e^{-c_2' m},$$

where the second and third steps uses the assumed lower bound on m. In the ROS case, results of Krahmer and Ward [18] imply that

$$\mathbb{P}\big[|||Q|||_{\mathrm{op}} \geq 1/2\big] \leq c_1 e^{-c_2 \frac{m}{\log^4(n)}},$$

where the final step uses the assumed lower bound on m.

B.3.2. *Proof of inequality* (ii). We split this claim into two sub-parts: one for Gaussian sketches, and the other for ROS sketches. Throughout the proof, we make use of the $n \times n$ diagonal matrix $\overline{D} = \mathrm{diag}(0_{d_n}, D_2)$, with which we have $S U_2 D_2^{1/2} = S U \overline{D}^{1/2}$.

Gaussian case:

By the definition of the matrix spectral norm, we know

$$|||S U \overline{D}^{1/2}|||_{\mathrm{op}} := \sup_{\substack{u \in \mathcal{S}^{m-1} \\ v \in \mathcal{E}}} \langle u, S v \rangle, \tag{43}$$

where $\mathcal{E} = \{v \in \mathbb{R}^n \mid \|U \overline{D} v\|_2 \leq 1\}$, and $\mathcal{S}^{m-1} = \{u \in \mathbb{R}^m \mid \|u\|_2 = 1\}$.

We may choose a $1/2$-cover $\{u^1, \ldots, u^M\}$ of the set $\mathcal{S}^{m-1}$ of the set with $\log M \leq 2m$ elements. We then have

$$\begin{aligned} |||S U \overline{D}^{1/2}|||_{\mathrm{op}} &\leq \max_{j \in [M]} \sup_{v \in \mathcal{E}} \langle u^j, S v \rangle + \frac{1}{2} \sup_{\substack{u \in \mathcal{S}^{d_n - 1} \\ v \in \mathcal{E}}} \langle u, S v \rangle \\ &= \max_{j \in [M]} \sup_{v \in \mathcal{E}} \langle u^j, S v \rangle + \frac{1}{2} |||S U \overline{D}^{1/2}|||_{\mathrm{op}}, \end{aligned}$$

and rearranging implies that

$$|||S U \overline{D}^{1/2}|||_{\mathrm{op}} \leq 2 \underbrace{\max_{j \in [M]} \sup_{v \in \mathcal{E}} \langle u^j, \widetilde{S} v \rangle}_{\widetilde{Z}}.$$

For each fixed $u^j \in \mathcal{S}^{d_n - 1}$, consider the random variable $Z^j := \sup_{v \in \mathcal{E}} \langle u^j, Sv \rangle$. It is equal in distribution to the random variable $V(g) = \frac{1}{\sqrt{m}} \sup_{v \in \mathcal{E}} \langle g, v \rangle$, where $g \in \mathbb{R}^n$ is a standard Gaussian vector. For $g, g' \in \mathbb{R}^n$, we have

$$\begin{aligned} |V(g) - V(g')| &\leq \frac{2}{\sqrt{m}} \sup_{v \in \mathcal{E}} |\langle g - g', v \rangle| \\ &\leq \frac{2 |\!|\!| D_2^{1/2} |\!|\!|_{\mathrm{op}}}{\sqrt{m}} \|g - g'\|_2 \leq \frac{2\delta_n}{\sqrt{m}} \|g - g'\|_2, \end{aligned}$$

where we have used the fact that $\widehat{\mu}_j \leq {\delta_n}^2$ for all $j \geq d_n + 1$. Consequently, by concentration of measure for Lipschitz functions of Gaussian random variables [19], we have

$$\mathbb{P}\big[V(g) \geq \mathbb{E}\big[V(g)\big] + t\big] \leq e^{-\frac{mt^2}{8\delta_n^2}}. \tag{44}$$

Turning to the expectation we have

$$\begin{aligned} \mathbb{E}\big[V(g)\big] &= \frac{2}{\sqrt{m}} \mathbb{E} \| D_2^{1/2} g \|_2 \leq 2 \sqrt{\frac{\sum_{j=d_n+1}^n \mu_j}{m}} \\ &= 2 \sqrt{\frac{n}{m}} \sqrt{\frac{\sum_{j=d_n+1}^n \mu_j}{n}} \leq 2\delta_n, \end{aligned} \tag{45}$$

where the last inequality follows since $m \geq n\delta_n^2$ and $\sqrt{\frac{\sum_{j=d_n+1}^n \mu_j}{n}} \leq \delta_n^2$. Combining the pieces, we have shown have shown that $\mathbb{P}[Z^j \geq c_0(1+\varepsilon)\delta_n] \leq e^{-c_2 m}$ for each $j = 1, \ldots, M$. Finally, setting $t = c\delta_n$ in the tail bound (44) for a constant $c \geq 1$ large enough to ensure that $\frac{c_2 m}{8} \geq 2 \log M$. Taking the union bound over all $j \in [M]$ yields

$$\mathbb{P}\big[|\!|\!| SU \overline{D}^{1/2} |\!|\!|_{\mathrm{op}} \geq 8c\delta_n\big] \leq c_1 e^{-\frac{c_2 m}{8} + \log M} \leq c_1 e^{-c_2' m}$$

which completes the proof.

ROS case: Here we pursue a matrix Chernoff argument analogous to that in the paper [30]. Letting $r \in \{-1, 1\}^n$ denote an i.i.d. sequence of Rademacher variables, the ROS sketch can be written in the form $S = PH\mathrm{diag}(r)$, where P is a partial identity matrix scaled by n/m, and the matrix H is orthonormal with elements bounded as $|H_{ij}| \leq c/\sqrt{n}$ for some constant c. With this notation, we can write

$$|\!|\!| PH\mathrm{diag}(r) \bar{D}^{1/2} |\!|\!|_{\mathrm{op}}^2 = \left| \left| \left| \frac{1}{m} \sum_{i=1}^m v_i v_i^T \right| \right| \right|_{\mathrm{op}},$$

where $v_i \in \mathbb{R}^n$ are random vectors of the form $\sqrt{n}\overline{D}^{1/2}\mathrm{diag}(r)He$, where $e \in \mathbb{R}^n$ is chosen uniformly at random from the standard Euclidean basis.

We first show that the vectors $\{v_i\}_{i=1}^m$ are uniformly bounded with high probability. Note that we certainly have $\max_{i\in[m]} \|v_i\|_2 \le \max_{j\in[n]} F_j(r)$, where

$$F_j(r) := \sqrt{n}\big\|\overline{D}^{1/2}\mathrm{diag}(r)He_j\big\|_2 = \sqrt{n}\big\|\overline{D}^{1/2}\mathrm{diag}(He_j)r\big\|_2.$$

Beginning with the expectation, define the vector $\widetilde{r} = \mathrm{diag}(He_j)r$, and note that it has entries bounded in absolute value by $c/\sqrt{n}$. Thus, we have

$$\mathbb{E}\big[F_j(r)\big] \le \big[n\mathbb{E}\big[\widetilde{r}^T\overline{D}\widetilde{r}\big]\big]^{1/2} \le c\sqrt{\sum_{j=d_n+1}^{n}\widehat{\mu}_j} \le c\sqrt{n}\delta_n^2.$$

For any two vectors $r, r' \in \mathbb{R}^n$, we have

$$\big|F(r) - F\big(r'\big)\big| \le \sqrt{n}\|r - r'\|_2\big\|\overline{D}^{1/2}\mathrm{diag}(He_j)\big\|_2 \le \delta_n.$$

Consequently, by concentration results for convex Lipschitz functions of Rademacher variables [19], we have

$$\mathbb{P}\big[F_j(r) \ge c_0\sqrt{n}\delta_n^2\log n\big] \le c_1 e^{-c_2 n\delta_n^2\log^2 n}.$$

Taking the union bound over all n rows, we see that

$$\max_{i\in[n]}\|v_i\|_2 \le \max_{j\in[n]} F_j(r) \le 4\sqrt{n}\delta_n^2\log(n)$$

with probability at least $1 - c_1 e^{-c_2 n\delta_n^2\log^2(n)}$. Finally, a simple calculation shows that $|||\mathbb{E}[v_1 v_1^T]|||_{\mathrm{op}} \le \delta_n^2$. Consequently, by standard matrix Chernoff bounds [30, 31], we have

$$\mathbb{P}\left[\left|\left|\left|\frac{1}{m}\sum_{i=1}^m v_i v_i^T\right|\right|\right|_{\mathrm{op}} \ge 2\delta_n^2\right] \le c_1 e^{-c_2\frac{m\delta_n{}^2}{n\delta_n^4\log^2(n)}} + c_1 e^{-c_2 n\delta_n^2\log^2(n)}, \tag{46}$$

from which the claim follows.

REFERENCES

[1] AILON, N. and LIBERTY, E. (2009). Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete Comput. Geom.* **42** 615–630. MR2556458

[2] ALAOUI, A. E. and MAHONEY, M. W. (2014). Fast randomized kernel methods with statistical guarantees. Technical Report, UC Berkeley. Available at arXiv:1411.0306.

[3] AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference* **3** 224–294. MR3311453

[4] ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404. MR0051437

[5] BACH, F. (2012). Sharp analysis of low-rank kernel matrix approximations. In *International Conference on Learning Theory* (*COLT*). Edinburgh.

[6] BARTLETT, P. L., BOUSQUET, O. and MENDELSON, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33** 1497–1537. MR2166554

[7] BERLINET, A. and THOMAS-AGNAN, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston, MA. MR2239907

[8] BOUTSIDIS, C. and GITTENS, A. (2013). Improved matrix algorithms via the subsampled randomized Hadamard transform. *SIAM J. Matrix Anal. Appl.* **34** 1301–1340. MR3101094

[9] COHEN, M. B., NELSON, J. and WOODRUFF, D. P. (2015). Optimal approximate matrix product in terms of stable rank. Technical report.

[10] DRINEAS, P. and MAHONEY, M. W. (2005). On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.* **6** 2153–2175. MR2249884

[11] GITTENS, A. and MAHONEY, M. W. (2013). Revisiting the nystrom method for improved large-scale machine learning. Preprint. Available at arXiv:1303.1849.

[12] GU, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York. MR1876599

[13] HALKO, N., MARTINSSON, P. G. and TROPP, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **53** 217–288. MR2806637

[14] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. Springer, New York. MR1851606

[15] JOHNSTONE, I. M. (2016). *Gaussian Estimation*: *Sequence and Wavelet Models*. Springer, New York.

[16] KIMELDORF, G. and WAHBA, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33** 82–95. MR0290013

[17] KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442

[18] KRAHMER, F. and WARD, R. (2011). New and improved Johnson–Lindenstrauss embeddings via the restricted isometry property. *SIAM J. Math. Anal.* **43** 1269–1281. MR2821584

[19] LEDOUX, M. (2001). *The Concentration of Measure Phenomenon*. *Mathematical Surveys and Monographs* **89**. Amer. Math. Soc., Providence, RI. MR1849347

[20] MAHONEY, M. W. (2011). Randomized algorithms for matrices and data. *Faund. Trends Mach. Learn.* **3** 123–224.

[21] MATOUŠEK, J. (2002). *Lectures on Discrete Geometry*. *Graduate Texts in Mathematics* **212**. Springer, New York. MR1899299

[22] MENDELSON, S. (2002). Geometric parameters of kernel machines. In *Computational Learning Theory* (*Sydney*, 2002). *Lecture Notes in Computer Science* **2375** 29–43. Springer, Berlin. MR2040403

[23] PILANCI, M. and WAINWRIGHT, M. J. (2015). Randomized sketches of convex programs with sharp guarantees. *IEEE Trans. Inform. Theory* **61** 5096–5115. MR3386504

[24] PILANCI, M. and WAINWRIGHT, M. J. (2016). Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *J. Mach. Learn. Res.* **17**. Paper No. 53, 38. MR3504613

[25] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.* **13** 389–427. MR2913704

[26] SARLÓS, T. (2006). Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS '06. 47th Annual IEEE Symposium*. Berkeley, CA.

[27] SAUNDERS, C., GAMMERMAN, A. and VOVK, V. (1998). Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, *ICML '*98 515–521. Morgan Kaufmann, San Francisco, CA.

[28] SHAWE-TAYLOR, J. and CRISTIANINI, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, Cambridge.

[29] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

[30] TROPP, J. A. (2011). Improved analysis of the subsampled randomized Hadamard transform. *Adv. Adapt. Data Anal.* **3** 115–126. MR2835584

[31] TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459

[32] VAN DE GEER, S. A. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

[33] WAHBA, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442

[34] WILLIAMS, C. and SEEGER, M. (2001). Using the Nyström method to speed up kernel machines. In *Proceedings of the* 14*th Annual Conference on Neural Information Processing Systems* 682–688. Vancouver, BC, Canada.

[35] ZHANG, Y., DUCHI, J. C. and WAINWRIGHT, M. J. (2013). Divide and conquer kernel ridge regression. In *Computational Learning Theory* (*COLT*) *Conference*. Princeton, NJ.

Y. YANG
DEPARTMENT OF STATISTICS
FLORIDA STATE UNIVERSITY
TALLAHASSEE, FLORIDA, 32306-4330
USA
E-MAIL: yyang@stat.fsu.edu

M. PILANCI
DEPARTMENT OF EECS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA, 94720
USA
E-MAIL: mert@eecs.berkeley.edu

M. J. WAINWRIGHT
DEPARTMENT OF EECS AND STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA, 94720
USA
E-MAIL: wainwrig@eecs.berkeley.edu