JMLR: Workshop and Conference Proceedings vol 30 (2013) 1–26

Divide and Conquer Kernel Ridge Regression

Yuchen Zhang YUCZHANG@EECS.BERKELEY.EDU
University of California, Berkeley

John Duchi JDUCHI@EECS.BERKELEY.EDU
University of California, Berkeley

Martin Wainwright WAINWRIG@STAT.BERKELEY.EDU
University of California, Berkeley

Abstract

We study a decomposition-based scalable approach to performing kernel ridge regression. The method is simple to describe: it randomly partitions a dataset of size N into m subsets of equal size, computes an independent kernel ridge regression estimator for each subset, then averages the local solutions into a global predictor. This partitioning leads to a substantial reduction in computation time versus the standard approach of performing kernel ridge regression on all N samples. Our main theorem establishes that despite the computational speed-up, statistical optimality is retained: if m is not too large, the partition-based estimate achieves optimal rates of convergence for the full sample size N. As concrete examples, our theory guarantees that m may grow polynomially in N for Sobolev spaces, and nearly linearly for finite-rank kernels and Gaussian kernels. We conclude with simulations complementing our theoretical results and exhibiting the computational and statistical benefits of our approach.

1. Introduction

In non-parametric regression, the statistician receives N samples of the form $\{(x_i, y_i)\}_{i=1}^N$, where each $x_i \in \mathcal{X}$ is a covariate and $y_i \in \mathbb{R}$ is a real-valued response, and the samples are drawn i.i.d. from some unknown joint distribution $\mathbb{P}$ over $\mathcal{X} \times \mathbb{R}$. The goal is to estimate a function $\widehat{f} : \mathcal{X} \to \mathbb{R}$ that can be used to predict future responses based on observing only the covariates. Frequently, the quality of an estimate $\widehat{f}$ is measured in terms of the mean-squared prediction error $\mathbb{E}[(\widehat{f}(X) - Y)^2]$, in which case the conditional expectation $f^*(x) = \mathbb{E}[Y \mid X = x]$ is optimal. The problem of non-parametric regression is a classical one, and researchers have studied a wide range of estimators (see, e.g., the books by Gyorfi et al. (2002),Wasserman (2006), and van de Geer (2000)). One class of methods, known as regularized M-estimators (van de Geer, 2000), are based on minimizing the sum of a data-dependent loss function with a regularization term. The focus of this paper is a popular M-estimator that combines the least-squares loss with a squared Hilbert norm penalty for regularization. When working in a reproducing kernel Hilbert space (RKHS), the resulting method is known as *kernel ridge regression*, and is widely used in practice (Hastie et al., 2001; Shawe-Taylor and Cristianini, 2004). Past work has established bounds on the estimation error for RKHS-based methods (e.g., Koltchinskii, 2006;

Mendelson, 2002; van de Geer, 2000; Zhang, 2005), which have been refined and extended in more recent work (e.g., Steinwart et al., 2009).

Although the statistical aspects of kernel ridge regression (KRR) are well-understood, the computation of the KRR estimate can be challenging for large datasets. In a standard implementation (Saunders et al., 1998), the kernel matrix must be inverted, which requires costs $\mathcal{O}(N^3)$ and $\mathcal{O}(N^2)$ in time and memory respectively. Such scalings are prohibitive when the sample size N is large. As a consequence, approximations have been designed to avoid the expense of finding an exact minimizer. One family of approaches is based on low-rank approximation of the kernel matrix; examples include kernel PCA (Schölkopf et al., 1998), the incomplete Cholesky decomposition (Fine and Scheinberg, 2002), or Nyström sampling (Williams and Seeger, 2001). These methods reduce the time complexity to $\mathcal{O}(dN^2)$ or $\mathcal{O}(d^2N)$, where $d \ll N$ is the preserved rank. To our knowledge, however, there are no results showing that such low-rank versions of KRR still achieve minimax-optimal rates in estimation error. A second line of research has considered early-stopping of iterative optimization algorithms for KRR, including gradient descent (Yao et al., 2007; Raskutti et al., 2011) and conjugate gradient methods (Blanchard and Krämer, 2010), where early-stopping provides regularization against over-fitting and improves run-time. If the algorithm stops after t iterations, the aggregate time complexity is $\mathcal{O}(tN^2)$.

In this work, we study a different decomposition-based approach. The algorithm is appealing in its simplicity: we partition the dataset of size N randomly into m equal sized subsets, and we compute the kernel ridge regression estimate $\widehat{f}_i$ for each of the $i = 1, \ldots, m$ subsets independently, with a *careful choice* of the regularization parameter. The estimates are then averaged via $\bar{f} = (1/m) \sum_{i=1}^{m} \widehat{f}_i$. Our main theoretical result gives conditions under which the average $\bar{f}$ achieves the minimax rate of convergence over the underlying Hilbert space. Even using naive implementations of KRR, this decomposition gives time and memory complexity scaling as $\mathcal{O}(N^3/m^2)$ and $\mathcal{O}(N^2/m^2)$, respectively. Moreover, our approach dovetails naturally with parallel and distributed computation: we are guaranteed superlinear speedup with m parallel processors (though we must still communicate the function estimates from each processor). Divide-and-conquer approaches have been studied by several authors, including McDonald et al. (2010) for perceptron-based algorithms, Kleiner et al. (2012) in distributed versions of the bootstrap, and Zhang et al. (2012) for parametric smooth convex optimization objectives arising out of statistical estimation problems. This paper demonstrates the potential benefits of divide-and-conquer approaches for nonparametric and infinite-dimensional regression problems.

One difficulty in solving each of the sub-problems independently is how to choose the regularization parameter. Due to the infinite-dimensional nature of non-parametric problems, the choice of regularization parameter must be made with care (e.g., Hastie et al., 2001). An interesting consequence of our theoretical analysis is in demonstrating that, even though each partitioned sub-problem is based only on the fraction N/m of samples, it is nonetheless *essential to regularize the partitioned sub-problems as though they had all N samples*. Consequently, from a local point of view, each sub-problem is under-regularized. This "under-regularization" allows the bias of each local estimate to be very small, but it causes a detrimental blow-up in the variance. However, as we prove, the m-fold averaging underlying the method reduces variance enough that the resulting estimator $\bar{f}$ still attains optimal convergence rate.

The remainder of this paper is organized as follows. We begin in Section 2 by providing background on the kernel ridge regression estimate. In Section 3, we present our main theorems on the mean-squared error between the averaged estimate $\bar{f}$ and the optimal regression function f^*. We then provide several corollaries that exhibit concrete consequences of the results, including convergence rates of r/N for kernels with finite rank r, and convergence rates of $N^{-2\nu/(2\nu+1)}$ for estimation of functionals in a Sobolev space with ν degrees of smoothness. As we discuss, both of these estimation rates are minimax-optimal and hence unimprovable. We devote Section 4 to the proofs of our results, deferring more technical aspects of the analysis to appendices. Lastly, we present simulation results in Section 5 to further explore our theoretical results.

2. Background and problem formulation

We begin with the background and notation required for a precise statement of our problem.

2.1. Reproducing kernels

The method of kernel ridge regression is based on the idea of a reproducing kernel Hilbert space. We provide only a very brief coverage of the basics here; see the many books on the topic (Wahba, 1990; Shawe-Taylor and Cristianini, 2004; Berlinet and Thomas-Agnan, 2004; Gu, 2002) for further details. Any symmetric and positive semidefinite kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defines a reproducing kernel Hilbert space (RKHS for short). For a given distribution $\mathbb{P}$ on $\mathcal{X}$, the Hilbert space is strictly contained within $L^2(\mathbb{P})$. For each $x \in \mathcal{X}$, the function $z \mapsto K(z, x)$ is contained in the Hilbert space $\mathcal{H}$; moreover, the Hilbert space is endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ such that $K(\cdot, x)$ acts as the representer of evaluation, meaning

$$\langle f, K(x, \cdot) \rangle_{\mathcal{H}} = f(x) \quad \text{for } f \in \mathcal{H}. \tag{1}$$

We let $\|g\|_{\mathcal{H}} := \sqrt{\langle g, g \rangle_{\mathcal{H}}}$ denote the norm in $\mathcal{H}$, and similarly $\|g\|_2 := (\int_{\mathcal{X}} g(x)^2 d\mathbb{P}(x))^{1/2}$ denotes the norm in $L^2(\mathbb{P})$. Under suitable regularity conditions, Mercer's theorem guarantees that the kernel has an eigen-expansion of the form

$$K(x, x') = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(x'),$$

where $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ are a non-negative sequence of eigenvalues, and $\{\phi_j\}_{j=1}^{\infty}$ is an orthonormal basis for $L^2(\mathbb{P})$.

From the reproducing relation (1), we have $\langle \phi_j, \phi_j \rangle_{\mathcal{H}} = 1/\mu_j$ for any j and $\langle \phi_j, \phi_{j'} \rangle_{\mathcal{H}} = 0$ for any $j \neq j'$. For any $f \in \mathcal{H}$, by defining the basis coefficients $\theta_j = \langle f, \phi_j \rangle_{L^2(\mathbb{P})}$ for $j = 1, 2, \ldots$, we can expand the function in terms of these coefficients as $f = \sum_{j=1}^{\infty} \theta_j \phi_j$, and simple calculations show that

$$\|f\|_2^2 = \int_{\mathcal{X}} f^2(x) d\mathbb{P}(x) = \sum_{j=1}^{\infty} \theta_j^2, \quad \text{and} \quad \|f\|_{\mathcal{H}}^2 = \langle f, f \rangle = \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j}.$$

Consequently, we see that the RKHS can be viewed as an elliptical subset of the sequence space $\ell^2(\mathbb{N})$ defined by the non-negative eigenvalues $\{\mu_j\}_{j=1}^{\infty}$.

2.2. Kernel ridge regression

Suppose that we are given a data set $\{(x_i, y_i)\}_{i=1}^N$ consisting of N i.i.d. samples drawn from an unknown distribution $\mathbb{P}$ over $\mathcal{X} \times \mathbb{R}$, and our goal is to estimate the function that minimizes the mean-squared error $\mathbb{E}[(f(X) - Y)^2]$, where the expectation is taken jointly over (X, Y) pairs. It is well-known that the optimal function is the conditional mean $f^*(x) := \mathbb{E}[Y \mid X = x]$. In order to estimate the unknown function f^*, we consider an M-estimator that is based on minimizing a combination of the least-squares loss defined over the dataset with a weighted penalty based on the squared Hilbert norm,

$$\widehat{f} := \operatorname*{argmin}_{f \in \mathcal{H}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \tag{2}$$

where $\lambda > 0$ is a regularization parameter. When $\mathcal{H}$ is a reproducing kernel Hilbert space, the estimator (2) is known as the *kernel ridge regression estimate*, or KRR for short. It is a natural generalization of the ordinary ridge regression estimate (Hoerl and Kennard, 1970) to the non-parametric setting.

3. Main results and their consequences

We now turn to the description of our algorithm, which we follow with our main result (Theorem 1), which provides a general upper bound on the resulting prediction error for any trace class kernel. We illustrate the application of this general result to three different classes of kernels, showing that it leads to minimax-optimal rates in all three cases.

3.1. Algorithm and assumptions

The divide-and-conquer algorithm Fast-KRR is easy to describe. We are given N samples drawn i.i.d. according to the distribution $\mathbb{P}$. Rather than solving the kernel ridge regression problem (2) on all N samples, the Fast-KRR method executes the following three steps:

1. Divide the set of samples $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ evenly and uniformly at randomly into the m disjoint subsets $S_1, \ldots, S_m \subset \mathcal{X} \times \mathbb{R}$.

2. For each $i = 1, 2, \ldots, m$, compute the *local KRR estimate*

$$\widehat{f}_i := \operatorname*{argmin}_{f \in \mathcal{H}} \left\{ \frac{1}{|S_i|} \sum_{(x,y) \in S_i} (f(x) - y)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}. \tag{3}$$

3. Average together the local estimates and output $\bar{f} = \frac{1}{m} \sum_{i=1}^m \widehat{f}_i$.

This description actually provides a family of estimators, one for each choice of the regularization parameter $\lambda > 0$. Our main result applies to any choice of λ, while our corollaries for specific kernel classes optimize λ as a function of the kernel.

We now describe our main assumptions. Our first assumption, for which we have two variants, deals with the tail behavior of the basis functions $\{\phi_j\}_{j=1}^\infty$.

Assumption A *For some $k \geq 2$, there is a constant $\rho < \infty$ such that $\mathbb{E}[\phi_j(X)^{2k}] \leq \rho^{2k}$ for all $j = 1, 2, \ldots$.*

In certain cases, we show that sharper error guarantees can be obtained by enforcing a stronger condition of uniform boundedness:

Assumption A′ *There is a constant $\rho < \infty$ such that $\sup_{x \in \mathcal{X}} |\phi_j(x)| \leq \rho$ for all $j = 1, 2, \ldots$.*

Recalling that $f^*(x) := \mathbb{E}[Y \mid X = x]$, our second assumption involves the regression function f^* and the deviations of the zero-mean noise variables $Y - f^*(x)$:[1]

Assumption B *The function $f^* \in \mathcal{H}$, and for $x \in \mathcal{X}$, we have $\mathbb{E}[(Y - f^*(x))^2 \mid x] \leq \sigma^2$.*

3.2. Statement of main results

With these assumptions in place, we are now ready for the statement of our main result. Our result gives bound on the mean-squared estimation error $\mathbb{E}[\|\bar{f} - f^*\|_2^2]$ associated with the averaged estimate $\bar{f}$ based on an assigning $n = N/m$ samples to each of m machines. The theorem statement involves the following three kernel-related quantities:

$$\operatorname{tr}(K) := \sum_{j=1}^{\infty} \mu_j, \quad \gamma(\lambda) := \sum_{j=1}^{\infty} \frac{1}{1 + \lambda/\mu_j}, \quad \text{and} \quad \beta_d = \sum_{j=d+1}^{\infty} \mu_j. \tag{4}$$

The first quantity is the kernel trace, which serves a crude estimate of the "size" of the kernel operator, and assumed to be finite. The second quantity $\gamma(\lambda)$, familiar from previous work on kernel regression (Zhang, 2005), is known as the "effective dimensionality" of the kernel K with respect to $L^2(\mathbb{P})$. Finally, the quantity β_d is parameterized by a positive integer d that we may choose in applying the bounds, and it describes the tail decay of the eigenvalues of K. For $d = 0$, note that β_0 reduces to the ordinary trace. Finally, Theorem 1 involves one further quantity that depends on the number of moments k in Assumption A, namely

$$b(n, d, k) := \max\left\{\sqrt{\max\{k, \log(2d)\}}, \frac{\max\{k, \log(2d)\}}{n^{1/2 - 1/k}}\right\}. \tag{5}$$

Here the parameter $d \in \mathbb{N}$ is a quantity that may be optimized to obtain the sharpest possible upper bound. (The algorithm's execution is independent of d.)

Theorem 1 *With $f^* \in \mathcal{H}$ and under Assumptions A and B, the mean-squared error of the averaged estimate $\bar{f}$ is upper bounded as*

$$\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right] \leq \left(4 + \frac{6}{m}\right) \lambda \|f^*\|_{\mathcal{H}}^2 + \frac{12\sigma^2\gamma(\lambda)}{N} + \inf_{d \in \mathbb{N}} \left\{T_1(d) + T_2(d) + T_3(d)\right\}, \tag{6}$$

1. We can extend our results to the more general setting in which $f^* \notin \mathcal{H}$. In this paper, we limit ourselves to the case $f^* \in \mathcal{H}$ to facilitate comparisons with minimax rates and for conciseness in presentation. The assumption $f^* \in \mathcal{H}$ is, in any case, fairly standard (Gyorfi et al., 2002; Wasserman, 2006).

where

$$T_1(d) = \frac{2\rho^4 \|f^*\|_{\mathcal{H}}^2 \operatorname{tr}(K)\beta_d}{\lambda}, \quad T_2(d) = \frac{4\|f^*\|_{\mathcal{H}}^2 + 2\sigma^2/\lambda}{m}\left(\mu_{d+1} + \frac{3\rho^4 \operatorname{tr}(K)\beta_d}{\lambda}\right), \quad \text{and}$$

$$T_3(d) = \left(Cb(n,d,k)\frac{\rho^2\gamma(\lambda)+1}{\sqrt{n}}\right)^k \|f^*\|_2^2 \left(1 + \frac{2\sigma^2}{m\lambda} + \frac{4\|f^*\|_{\mathcal{H}}^2}{m}\right),$$

and C denotes a universal (numerical) constant.

Theorem 1 is a general result that applies to any trace-class kernel. Although the statement appears somewhat complicated at first sight, it yields concrete and interpretable guarantees on the error when specialized to particular kernels, as we illustrate in Section 3.3.

Before doing so, let us provide a few heuristic arguments for intuition. In typical settings, the term $T_3(d)$ goes to zero quickly: if the number of moments k is large and number of partitions m is small—say enough to guarantee that $(\gamma(\lambda)^2 N/m)^{-k/2} = \mathcal{O}(1/N)$—it will be of lower order. As for the remaining terms, at a high level, we show that an appropriate choice of the free parameter d leaves the first two terms in the upper bound (6) dominant. Note that the terms μ_{d+1} and β_d are decreasing in d while the term $b(n,d,k)$ increases with d. However, the increasing term $b(n,d,k)$ grows only logarithmically in d, which allows us to choose a fairly large value without a significant penalty. As we show in our corollaries, for many kernels of interest, as long as the number of machines m is not "too large," this tradeoff is such that $T_1(d)$ and $T_2(d)$ are also of lower order compared to the two first terms in the bound (6). In such settings, Theorem 1 guarantees an upper bound of the form

$$\mathbb{E}\left[\|\bar{f} - f^*\|_2^2\right] = \mathcal{O}\Big(\underbrace{\lambda\|f^*\|_{\mathcal{H}}^2}_{\text{Squared bias}} + \underbrace{\frac{\sigma^2\gamma(\lambda)}{N}}_{\text{Variance}}\Big). \tag{7}$$

This inequality reveals the usual bias-variance trade-off in non-parametric regression; choosing a smaller value of $\lambda > 0$ reduces the first squared bias term, but increases the second variance term. Consequently, the setting of λ that minimizes the sum of these two terms is defined by the relationship

$$\lambda\|f^*\|_{\mathcal{H}}^2 \simeq \sigma^2\frac{\gamma(\lambda)}{N}. \tag{8}$$

This type of fixed point equation is familiar from work on oracle inequalities and local complexity measures in empirical process theory (Bartlett et al., 2005; Koltchinskii, 2006; van de Geer, 2000; Zhang, 2005), and when λ is chosen so that the fixed point equation (8) holds this (typically) yields minimax optimal convergence rates (Bartlett et al., 2005; Koltchinskii, 2006; Zhang, 2005; Caponnetto and De Vito, 2007). In Section 3.3, we provide detailed examples in which the choice λ^* specified by equation (8), followed by application of Theorem 1, yields minimax-optimal prediction error (for the Fast-KRR algorithm) for many kernel classes.

3.3. Some consequences

We now turn to deriving some explicit consequences of our main theorems for specific classes of reproducing kernel Hilbert spaces. In each case, our derivation follows the broad outline

given the the remarks following Theorem 1: we first choose the regularization parameter λ to balance the bias and variance terms, and then show, by comparison to known minimax lower bounds, that the resulting upper bound is optimal. Finally, we derive an upper bound on the number of subsampled data sets m for which the minimax optimal convergence rate can still be achieved.

Our first corollary applies to problems for which the kernel has finite rank r, meaning that its eigenvalues satisfy $\mu_j = 0$ for all $j > r$. Examples of such finite rank kernels include the linear kernel $K(x, x') = \langle x, x' \rangle_{\mathbb{R}^d}$, which has rank at most $r = d$; and the kernel $K(x, x') = (1 + x\, x')^m$ generating polynomials of degree m, which has rank at most $r = m + 1$.

Corollary 2 *For a kernel with rank r, consider the output of the Fast-KRR algorithm with $\lambda = r/N$. Suppose that Assumption B and Assumption A (or A′) hold, and that the number of processors m satisfy the bound*

$$m \leq c \frac{N^{\frac{k-4}{k-2}}}{r^2 \rho^{\frac{4k}{k-2}} \log^{\frac{k}{k-2}} r} \quad \text{(Assumption A)} \quad \text{or} \quad m \leq c \frac{N}{r^2 \rho^4 \log N} \quad \text{(Assumption A′)},$$

where c is a universal (numerical) constant. Then the mean-squared error is bounded as

$$\mathbb{E}\left[\|\bar{f} - f^*\|_2^2\right] = \mathcal{O}\left(\frac{\sigma^2 r}{N}\right). \tag{9}$$

For finite-rank kernels, the rate (9) is minimax-optimal: if $\mathbb{B}_{\mathcal{H}}(1)$ denotes the 1-norm ball in $\mathcal{H}$, there is a universal constant $c' > 0$ such that $\inf_f \sup_{f^* \in \mathbb{B}_{\mathcal{H}}(1)} \mathbb{E}[\|f - f^*\|_2^2] \geq c' \frac{r}{N}$. This lower bound follows from Theorem 2(a) of Raskutti et al. (2012) with $s = d = 1$.

Our next corollary applies to kernel operators with eigenvalues that obey a bound of the form

$$\mu_j \leq C\, j^{-2\nu} \quad \text{for all } j = 1, 2, \ldots, \tag{10}$$

where C is a universal constant, and $\nu > 1/2$ parameterizes the decay rate. Kernels with polynomially-decaying eigenvalues include those that underlie for the Sobolev spaces with different smoothness orders (e.g. Birman and Solomjak, 1967; Gu, 2002). As a concrete example, the first-order Sobolev kernel $K(x, x') = 1 + \min\{x, x'\}$ generates an RKHS of Lipschitz functions with smoothness $\nu = 1$. Other higher-order Sobolev kernels also exhibit polynomial eigen-decay with larger values of the parameter ν.

Corollary 3 *For any kernel with ν-polynomial eigen-decay (10), consider the output of the Fast-KRR algorithm with $\lambda = (1/N)^{\frac{2\nu}{2\nu+1}}$. Suppose that Assumption B and Assumption A (or A′) hold, and that the number of processors satisfy the bound*

$$m \leq c \left(\frac{N^{\frac{2(k-4)\nu - k}{(2\nu+1)}}}{\rho^{4k} \log^k N} \right)^{\frac{1}{k-2}} \quad \text{(Assumption A)} \quad \text{or} \quad m \leq c \frac{N^{\frac{2\nu-1}{2\nu+1}}}{\rho^4 \log N} \quad \text{(Assumption A′)},$$

where c is a constant only depending on ν. Then the mean-squared error is bounded as

$$\mathbb{E}\left[\|\bar{f} - f^*\|_2^2\right] = \mathcal{O}\left(\left(\frac{\sigma^2}{N}\right)^{\frac{2\nu}{2\nu+1}}\right). \tag{11}$$

The upper bound (11) is unimprovable up to constant factors, as shown by known minimax bounds on estimation error in Sobolev spaces (Stone, 1982; Tsybakov, 2009); see also Theorem 2(b) of Raskutti et al. (2012).

Our final corollary applies to kernel operators with eigenvalues that obey a bound of the form

$$\mu_j \leq c_1 \exp(-c_2 j^2) \quad \text{for all } j = 1, 2, \ldots, \tag{12}$$

for strictly positive constants (c_1, c_2). Such classes include the RKHS generated by the Gaussian kernel $K(x, x') = \exp(-\|x - x'\|_2^2)$.

Corollary 4 *For a kernel with exponential eigen-decay (12), consider the output of the Fast-KRR algorithm with $\lambda = 1/N$. Suppose that Assumption B and Assumption A (or A′) hold, and that the number of processors satisfy the bound*

$$m \leq c \frac{N^{\frac{k-4}{k-2}}}{\rho^{\frac{4k}{k-2}} \log^{\frac{2k-1}{k-2}} N} \quad \textit{(Assumption A)} \qquad \text{or} \qquad m \leq c \frac{N}{\rho^4 \log^2 N} \quad \textit{(Assumption A′)},$$

where c is a constant only depending on c_2. Then the mean-squared error is bounded as

$$\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right] = \mathcal{O}\left(\sigma^2 \frac{\sqrt{\log N}}{N}\right). \tag{13}$$

The upper bound (13) is also minimax optimal for the exponential kernel classes, which behave like a finite-rank kernel with effective rank $\sqrt{\log N}$.

Summary: Each corollary gives a critical threshold for the number m of data partitions: as long as m is below this threshold, we see that the decomposition-based Fast-KRR algorithm gives the optimal rate of convergence. It is interesting to note that the number of splits may be quite large: each grows asymptotically with N whenever the basis functions have more sufficiently many moments (viz. Assumption A). Moreover, the Fast-KRR method can attain these optimal convergence rates while using substantially less computation than standard kernel ridge regression methods.

4. Proofs of Theorem 1 and related results

We now turn to the proof of Theorem 1 and Corollaries 2 through 4. This section contains only a high-level view of proof of Theorem 1; we defer more technical aspects to the appendices.

4.1. Proof of Theorem 1

Using the definition of the averaged estimate $\bar{f} = \frac{1}{m}\sum_{i=1}^m \widehat{f}_i$, a bit of algebra yields

$$\begin{aligned}
\mathbb{E}[\left\|\bar{f} - f^*\right\|_2^2] &= \mathbb{E}[\left\|(\bar{f} - \mathbb{E}[\bar{f}]) + (\mathbb{E}[\bar{f}] - f^*)\right\|_2^2] \\
&= \mathbb{E}[\left\|\bar{f} - \mathbb{E}[\bar{f}]\right\|_2^2] + \left\|\mathbb{E}[\bar{f}] - f^*\right\|_2^2 + 2\mathbb{E}[\langle \bar{f} - \mathbb{E}[\bar{f}], \mathbb{E}[\bar{f}] - f^* \rangle_{L^2(\mathbb{P})}] \\
&= \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m (\widehat{f}_i - \mathbb{E}[\widehat{f}_i])\right\|_2^2\right] + \left\|\mathbb{E}[\bar{f}] - f^*\right\|_2^2,
\end{aligned}$$

where we used the fact that $\mathbb{E}[\widehat{f}_i] = \mathbb{E}[\bar{f}]$ for each $i \in [m]$. Using this unbiasedness once more, we bound the variance of the terms $\widehat{f}_i - \mathbb{E}[\bar{f}]$ to see that

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right] &\leq \frac{1}{m}\mathbb{E}\left[\|\widehat{f}_1 - \mathbb{E}[\widehat{f}_1]\|_2^2\right] + \|\mathbb{E}[\widehat{f}_1] - f^*\|_2^2 \\ &\leq \frac{1}{m}\mathbb{E}\left[\|\widehat{f}_1 - f^*\|_2^2\right] + \|\mathbb{E}[\widehat{f}_1] - f^*\|_2^2, \end{aligned} \tag{14}$$

where we have used the fact that $\mathbb{E}[\widehat{f}_i]$ minimizes $\mathbb{E}[\|\widehat{f}_i - f\|_2^2]$ over $f \in \mathcal{H}$.

The error bound (14) suggests our strategy: we bound $\mathbb{E}[\|\widehat{f}_1 - f^*\|_2^2]$ and $\|\mathbb{E}[\widehat{f}_1] - f^*\|_2^2$ respectively. Based on equation (3), the estimate $\widehat{f}_1$ is obtained from a standard kernel ridge regression with sample size $n = N/m$ and ridge parameter λ. Accordingly, the following two auxiliary results provide bounds on these two terms, where the reader should recall the definitions of $b(n, d, k)$ and β_d from equation (4). In each lemma, C represents a universal (numerical) constant.

Lemma 5 (Bias bound) *Under Assumptions A and B, for $d = 1, 2, \ldots$, we have*

$$\|\mathbb{E}[\widehat{f}] - f^*\|_2^2 \leq 4\lambda \|f^*\|_{\mathcal{H}}^2 + \frac{2\rho^4 \|f^*\|_{\mathcal{H}}^2 \operatorname{tr}(K)\beta_d}{\lambda} + \left(Cb(n, d, k)\frac{\rho^2\gamma(\lambda) + 1}{\sqrt{n}}\right)^k \|f^*\|_2^2. \tag{15}$$

Lemma 6 (Variance bound) *Under Assumptions A and B, for $d = 1, 2, \ldots$, we have*

$$\begin{aligned}\mathbb{E}[\|\widehat{f} - f^*\|_2^2] &\leq 6\lambda \|f^*\|_{\mathcal{H}}^2 + \frac{12\sigma^2\gamma(\lambda)}{n} \\ &+ \left(\frac{2\sigma^2}{\lambda} + 4\|f^*\|_{\mathcal{H}}^2\right)\left(\mu_{d+1} + \frac{3\rho^4 \operatorname{tr}(K)\beta_d}{\lambda} + \left(Cb(n, d, k)\frac{\rho^2\gamma(\lambda) + 1}{\sqrt{n}}\right)^k \|f^*\|_2^2\right). \end{aligned} \tag{16}$$

The proofs of these lemmas, contained in Appendices A and B respectively, constitute one main technical contribution of this paper.

Given these two lemmas, the remainder of the theorem proof is straightforward. Combining the inequality (14) with Lemmas 5 and 6 yields the claim of Theorem 1.

4.2. Proof of Corollary 2

We first present a general inequality bounding the size of m for which optimal convergence rates are possible. We assume that d is chosen large enough that for some constant c, we have $c\log(2d) \geq k$ in Theorem 1, and that the regularization λ has been chosen. In this case, inspection of Theorem 1 shows that if m is small enough that

$$\left(\sqrt{\frac{\log d}{N/m}}\rho^2(\gamma(\lambda) + 1)\right)^k \frac{1}{m\lambda} \leq \frac{\gamma(\lambda) + 1}{N},$$

then the term $T_3(d)$ provides a convergence rate given by $(\gamma(\lambda) + 1)/N$. Thus, solving the expression above for m, we find

$$\frac{m\log d}{N}\rho^4(\gamma(\lambda) + 1)^2 = \frac{\lambda^{2/k} m^{2/k}(\gamma(\lambda) + 1)^{2/k}}{N^{2/k}} \quad \text{or} \quad m^{\frac{k-2}{k}} = \frac{\lambda^{\frac{2}{k}} N^{\frac{k-2}{k}}}{(\gamma(\lambda) + 1)^{2\frac{k-1}{k}}\rho^4 \log d}.$$

Taking $(k-2)/k$-th roots of both sides, we obtain that if

$$m \leq \frac{\lambda^{\frac{2}{k-2}} N}{(\gamma(\lambda)+1)^{2\frac{k-1}{k-2}} \rho^{\frac{4k}{k-2}} \log^{\frac{k}{k-2}} d}, \tag{17}$$

then the term $T_3(d)$ of the bound (6) is $\mathcal{O}(\gamma(\lambda)/N + 1/N)$.

Now we apply the bound (17) in the case in the corollary. Let us take $d = r$; then $\beta_d = \mu_{d+1} = 0$. We find that $\gamma(\lambda) \leq r$ since each of its terms is bounded by 1, and we take $\lambda = r/N$. Evaluating the expression (17) with this value, we arrive at

$$m \leq \frac{N^{\frac{k-4}{k-2}}}{4r^2 \rho^{\frac{4k}{k-2}} \log^{\frac{k}{k-2}} r} \quad \Rightarrow \quad m \leq \frac{N^{\frac{k-4}{k-2}}}{(r+1)^2 \rho^{\frac{4k}{k-2}} \log^{\frac{k}{k-2}} r}.$$

If we have sufficiently many moments that $k \geq \log N$, and $N \geq r$ (for example, if the basis functions ϕ_j have a uniform bound ρ), then we may take $k = \log N$, which implies that $N^{\frac{k-4}{k-2}} = \Omega(N)$, and we replace $\log d = \log r$ with $\log N$ (we assume $N \geq r$), by recalling Theorem 1. Then so long as

$$m \leq c \frac{N}{r^2 \rho^4 \log N}$$

for some constant $c > 0$, we obtain an identical result.

4.3. Proof of Corollary 3

We follow the program outlined in our remarks following Theorem 1. We must first choose λ so that $\lambda = \gamma(\lambda)/N$. To that end, we note that setting $\lambda = N^{-\frac{2\nu}{2\nu+1}}$ gives

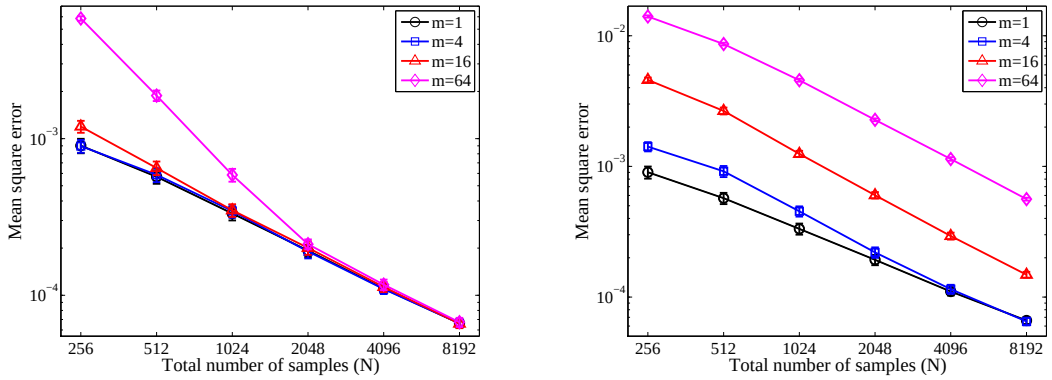
$$\begin{aligned}\gamma(\lambda) = \sum_{j=1}^{\infty} \frac{1}{1 + j^{2\nu} N^{-\frac{2\nu}{2\nu+1}}} &\leq N^{\frac{1}{2\nu+1}} + \sum_{j > N^{\frac{1}{2\nu+1}}} \frac{1}{1 + j^{2\nu} N^{-\frac{2\nu}{2\nu+1}}} \\ &\leq N^{\frac{1}{2\nu+1}} + N^{\frac{2\nu}{2\nu+1}} \int_{N^{\frac{1}{2\nu+1}}} \frac{1}{u^{2\nu}} du = N^{\frac{1}{2\nu+1}} + \frac{1}{2\nu - 1} N^{\frac{1}{2\nu+1}}.\end{aligned}$$

Dividing by N, we find that $\lambda \approx \gamma(\lambda)/N$, as desired. Now we choose the truncation parameter d. By choosing $d = N^t$ for some $t \in \mathbb{R}_+$, then we find that $\mu_{d+1} \lesssim N^{-2\nu t}$ and an integration yields $\beta_d \lesssim N^{-(2\nu-1)t}$. Setting $t = 3/(2\nu - 1)$ guarantees that $\mu_{d+1} \lesssim N^{-3}$ and $\beta_d \lesssim N^{-3}$; the corresponding terms in the bound (6) are thus negligible. Moreover, we have for any finite k that $\log d \gtrsim k$.

Applying the general bound (17) on m, we arrive at the inequality

$$m \leq c \frac{N^{-\frac{4\nu}{(2\nu+1)(k-2)}} N}{N^{\frac{2(k-1)}{(2\nu+1)(k-2)}} \rho^{\frac{4k}{k-2}} \log^{\frac{k}{k-2}} N} = c \frac{N^{\frac{2(k-4)\nu - k}{(2\nu+1)(k-2)}}}{\rho^{\frac{4k}{k-2}} \log^{\frac{k}{k-2}} N}.$$

Whenever this holds, we have convergence rate $\lambda = N^{-\frac{2\nu}{2\nu+1}}$. Now, let Assumption A′ hold, and take $k = \log N$. Then the above bound becomes (to a multiplicative constant factor) $N^{\frac{2\nu-1}{2\nu+1}}/\rho^4 \log N$, as claimed.



(a) With under-regularization (b) Without under-regularization

Figure 1. The squared $L^2(\mathbb{P})$-norm between between the averaged estimate $\bar{f}$ and the optimal solution f^*. (a) These plots correspond to the output of the Fast-KRR algorithm: each sub-problem is under-regularized by using $\lambda \sim N^{-2/3}$. (b) Analogous plots when each sub-problem is *not* under-regularized—that is, with $\lambda \sim n^{-2/3}$ is chosen as usual.

4.4. Proof of Corollary 4

First, we set $\lambda = 1/N$. Considering the sum $\gamma(\lambda) = \sum_{j=1}^{\infty} \mu_j/(\mu_j + \lambda)$, we see that for $j \leq \sqrt{(\log N)/c_2}$, the elements of the sum are bounded by 1. For $j > \sqrt{(\log N)/c_2}$, we make the approximation

$$\sum_{j \geq \sqrt{(\log N)/c_2}} \frac{\mu_j}{\mu_j + \lambda} \leq \frac{1}{\lambda} \sum_{j \geq \sqrt{(\log N)/c_2}} \mu_j \lesssim N \int_{\sqrt{(\log N)/c_2}}^{\infty} \exp(-c_2 t^2) dt = \mathcal{O}(1).$$

Thus we find that $\gamma(\lambda) + 1 \leq c\sqrt{\log N}$ for some constant c. By choosing $d = N^2$, we have that the tail sum and $(d+1)$-th eigenvalue both satisfy $\mu_{d+1} \leq \beta_d \lesssim c_2^{-1} N^{-4}$. As a consequence, all the terms involving β_d or μ_{d+1} in the bound (6) are negligible.

Recalling our inequality (17), we thus find that (under Assumption A), as long as the number of partitions m satisfies

$$m \leq c \frac{N^{\frac{k-4}{k-2}}}{\rho^{\frac{4k}{k-2}} \log^{\frac{2k-1}{k-2}} N},$$

the convergence rate of $\bar{f}$ to f^* is given by $\gamma(\lambda)/N \simeq \sqrt{\log N}/N$. Under the boundedness assumption A′, as we did in the proof of Corollary 2, we take $k = \log N$ in Theorem 1. By inspection, this yields the second statement of the corollary.

5. Simulation studies

In this section, we explore the empirical performance of our subsample-and-average methods for a non-parametric regression problem on simulated datasets. For all experiments in this section, we simulate data from the regression model $y = f^*(x) + \varepsilon$ for $x \in [0,1]$, where $f^*(x) := \min\{x, 1-x\}$ is 1-Lipschitz, the noise variables $\varepsilon \sim \mathsf{N}(0, \sigma^2)$ are normally distributed with variance $\sigma^2 = 1/5$, and the samples $x_i \sim \mathsf{Uni}[0,1]$. The Sobolev space

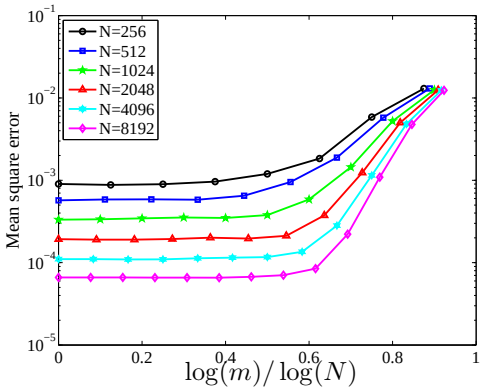


Figure 2. The mean-square error curves for fixed sample size but varied number of partitions. We are interested in the threshold of partitioning number m under which the optimal rate of convergence is achieved.

of Lipschitz functions on $[0,1]$ has reproducing kernel $K(x,x') = 1 + \min\{x,x'\}$ and norm $\|f\|_{\mathcal{H}}^2 = f^2(0) + \int_0^1 (f'(z))^2 dz$. By construction, the function $f^*(x) = \min\{x, 1-x\}$ satisfies $\|f^*\|_{\mathcal{H}} = 1$. The kernel ridge regression estimator $\widehat{f}$ takes the form

$$\widehat{f} = \sum_{i=1}^{N} \alpha_i K(x_i, \cdot), \quad \text{where} \quad \alpha = (K + \lambda N I)^{-1} y, \tag{18}$$

and K is the $N \times N$ Gram matrix and I is the $N \times N$ identity matrix. Since the first-order Sobolev kernel has eigenvalues (Gu, 2002) that scale as $\mu_j \simeq (1/j)^2$, the minimax convergence rate in terms of squared $L^2(\mathbb{P})$-error is $N^{-2/3}$ (e.g. Tsybakov, 2009; Stone, 1982; Caponnetto and De Vito, 2007).

By Corollary 3 with $\nu = 1$, this optimal rate of convergence can be achieved by Fast-KRR with regularization parameter $\lambda \approx N^{-2/3}$ as long as the number of partitions m satisfies $m \lesssim N^{1/3}$. In each of our experiments, we begin with a dataset of size $N = mn$, which we partition uniformly at random into m disjoint subsets. We compute the local estimator $\widehat{f}_i$ for each of the m subsets using n samples via (18), where the Gram matrix is constructed using the ith batch of samples (and n replaces N). We then compute $\bar{f} = (1/m)\sum_{i=1}^m \widehat{f}_i$. Our experiments compare the error of $\bar{f}$ as a function of sample size N, the number of partitions m, and the regularization λ.

In Figure 5(a), we plot the error $\|\bar{f} - f^*\|_2^2$ versus the total number of samples N, where $N \in \{2^8, 2^9, \ldots, 2^{13}\}$, using four different data partitions $m \in \{1, 4, 16, 64\}$. We execute each simulation 20 times to obtain standard errors for the plot. The black circled curve ($m = 1$) gives the baseline KRR error; if the number of partitions $m \leq 16$, Fast-KRR has accuracy comparable to the baseline algorithm. Even with $m = 64$, Fast-KRR's performance closely matches the full estimator for larger sample sizes ($N \geq 2^{11}$). In the right plot Figure 5(b), we perform an identical experiment, but we over-regularize by choosing $\lambda = n^{-2/3}$ rather than $\lambda = N^{-2/3}$ in each of the m sub-problems, combining the local estimates by averaging as usual. In contrast to Figure 5(a), there is an obvious gap between the performance of the algorithms when $m = 1$ and $m > 1$, as our theory predicts.

It is also interesting to understand the number of partitions m into which a dataset of size N may be divided while maintaining good statistical performance. According to Corollary 3 with $\nu = 1$, for the first-order Sobolev kernel, performance degradation should be limited as long as $m \lesssim N^{1/3}$. In order to test this prediction, Figure 2 plots the mean-

N		$m = 1$	$m = 16$	$m = 64$	$m = 256$	$m = 1024$
2^{12}	Error	$1.26 \cdot 10^{-4}$	$1.33 \cdot 10^{-4}$	$1.38 \cdot 10^{-4}$	N/A	N/A
	Time	1.12 (0.03)	0.03 (0.01)	0.02 (0.00)		
2^{13}	Error	$6.40 \cdot 10^{-5}$	$6.29 \cdot 10^{-5}$	$6.72 \cdot 10^{-5}$	N/A	N/A
	Time	5.47 (0.22)	0.12 (0.03)	0.04 (0.00)		
2^{14}	Error	$3.95 \cdot 10^{-5}$	$4.06 \cdot 10^{-5}$	$4.03 \cdot 10^{-5}$	$3.89 \cdot 10^{-5}$	N/A
	Time	30.16 (0.87)	0.59 (0.11)	0.11 (0.00)	0.06 (0.00)	
2^{15}	Error	Fail	$2.90 \cdot 10^{-5}$	$2.84 \cdot 10^{-5}$	$2.78 \cdot 10^{-5}$	N/A
	Time		2.65 (0.04)	0.43 (0.02)	0.15 (0.01)	
2^{16}	Error	Fail	$1.75 \cdot 10^{-5}$	$1.73 \cdot 10^{-5}$	$1.71 \cdot 10^{-5}$	$1.67 \cdot 10^{-5}$
	Time		16.65 (0.30)	2.21 (0.06)	0.41 (0.01)	0.23 (0.01)
2^{17}	Error	Fail	$1.19 \cdot 10^{-5}$	$1.21 \cdot 10^{-5}$	$1.25 \cdot 10^{-5}$	$1.24 \cdot 10^{-5}$
	Time		90.80 (3.71)	10.87 (0.19)	1.88 (0.08)	0.60 (0.02)

Table 1. Timing experiment giving $\|\bar{f} - f^*\|_2^2$ as a function of number of partitions m and data size N, providing mean run-time (measured in second) for each number m of partitions and data size N.

square error $\|\bar{f} - f^*\|_2^2$ versus the ratio $\log(m)/\log(N)$. Our theory predicts that even as the number of partitions m may grow polynomially in N, the error should grow only above some constant value of $\log(m)/\log(N)$. As Figure 2 shows, the point that $\|\bar{f} - f^*\|_2$ begins to increase appears to be around $\log(m) \approx 0.45 \log(N)$ for reasonably large N. This empirical performance is somewhat better than the (1/3) thresholded predicted by Corollary 3, but it does confirm that the number of partitions m can scale polynomially with N while retaining minimax optimality.

Our final experiment gives evidence for the improved time complexity partitioning provides. Here we compare the amount of time required to solve the KRR problem using the naive matrix inversion (18) for different partition sizes m and provide the resulting squared errors $\|\bar{f} - f^*\|_2^2$. Although there are more sophisticated solution strategies, we believe this is a reasonable proxy to exhibit Fast-KRR's potential. In Table 1, we present the results of this simulation, which we performed in Matlab using a Windows machine with 16GB of memory and a single-threaded 3.4Ghz processor. In each entry of the table, we give the mean error of Fast-KRR and the mean amount of time it took to run (with standard deviation over 10 simulations in parentheses; the error rate standard deviations are an order of magnitude smaller than the errors, so we do not report them). The entries "Fail" correspond to out-of-memory failures because of the large matrix inversion, while entries "N/A" indicate that $\|\bar{f} - f^*\|_2$ was significantly larger than the optimal value (rendering time improvements meaningless). The table shows that without sacrificing accuracy, decomposition via Fast-KRR can yield substantial computational improvements.

References

P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, 2004.

R. Bhatia. *Matrix Analysis*. Springer, 1997.

M. Birman and M. Solomjak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Sbornik: Mathematics*, 2(3):295–317, 1967.

G. Blanchard and N. Krämer. Optimal learning rates for kernel conjugate gradient regression. In *Advances in Neural Information Processing Systems 24*, 2010.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

R. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference*, to appear, 2012.

S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2002.

C. Gu. *Smoothing spline ANOVA models*. Springer, 2002.

L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, 2002.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

A. Kleiner, A. Talwalkar, P. Sarkar, and M. Jordan. Bootstrapping big data. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

D. Luenberger. *Optimization by Vector Space Methods*. Wiley, 1969.

R. McDonald, K. Hall, and G. Mann. Distributed training strategies for the structured perceptron. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2010.

S. Mendelson. Geometric parameters of kernel machines. In *Proceedings of COLT*, pages 29–43, 2002.

G. Raskutti, M. Wainwright, and B. Yu. Early stopping for non-parametric regression: An optimal data-dependent stopping rule. In *49th Annual Allerton Conference on Communication, Control, and Computing*, pages 1318–1325, 2011.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 12: 389–427, March 2012.

C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning*, pages 515–521. Morgan Kaufmann, 1998.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *IEEE Transactions on Information Theory*, 10(5):1299–1319, 1998.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93, 2009.

C. J. Stone. Optimal global rates of convergence for non-parametric regression. *Annals of Statistics*, 10(4):1040–1053, 1982.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

G. Wahba. *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PN, 1990.

L. Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems 14*, pages 682–688, 2001.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

Y. Zhang, J. C. Duchi, and M. J. Wainwright. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems 26*, 2012.

$\widehat{f}$	Empirical KRR minimizer based on n samples
f^*	Optimal function generating data, where $y_i = f^*(x_i) + \varepsilon_i$
Δ	Error $\widehat{f} - f^*$
ξ_x	RKHS evaluator $\xi_x := K(x, \cdot)$, so $\langle f, \xi_x \rangle = \langle \xi_x, f \rangle = f(x)$
$\widehat{\Sigma}$	Operator mapping $\mathcal{H} \to \mathcal{H}$ defined as the outer product $\widehat{\Sigma} := \frac{1}{n} \sum_{i=1}^n \xi_{x_i} \otimes \xi_{x_i}$, so that $\widehat{\Sigma} f = \frac{1}{n} \sum_{i=1}^n \langle \xi_{x_i}, f \rangle \, \xi_{x_i}$
ϕ_j	jth orthonormal basis vector for $L^2(\mathbb{P})$
δ_j	Basis coefficients of Δ or $\mathbb{E}[\Delta \mid X]$ (depending on context), i.e. $\Delta = \sum_{j=1}^\infty \delta_j \phi_j$
θ_j	Basis coefficients of f^*, i.e. $f^* = \sum_{j=1}^\infty \theta_j \phi_j$
d	Integer-valued truncation point
M	Diagonal matrix with $M = \operatorname{diag}(\mu_1, \ldots, \mu_d)$
Q	Diagonal matrix with $Q = I_{d \times d} + \lambda M^{-1}$
Φ	$n \times d$ matrix with coordinates $\Phi_{ij} = \phi_j(x_i)$
$v^\downarrow$	Truncation of vector v. For $v = \sum_j \nu_j \phi_j \in \mathcal{H}$, defined as $v^\downarrow = \sum_{j=1}^d \nu_j \phi_j$; for $v \in \ell_2(\mathbb{N})$ defined as $v^\downarrow = (v_1, \ldots, v_d)$
$v^\uparrow$	Untruncated part of vector v, defined as $v^\uparrow = (v_{d+1}, v_{d+1}, \ldots)$
β_d	The tail sum $\sum_{j>d} \mu_j$
$\gamma(\lambda)$	The sum $\sum_{j=1}^\infty 1/(1 + \lambda/\mu_j)$
$b(n, d, k)$	The maximum $\max\{\sqrt{\max\{k, \log(2d)\}}, \max\{k, \log(2d)\}/n^{1/2-1/k}\}$

Table 2: Notation used in proofs

Appendix A. Proof of Lemma 5

This appendix is devoted to the bias bound stated in Lemma 5. Let $X = \{x_i\}_{i=1}^n$ be shorthand for the design matrix, and define the error vector $\Delta = \widehat{f} - f^*$. By Jensen's inequality, we have $\|\mathbb{E}[\Delta]\|_2 \leq \mathbb{E}[\|\mathbb{E}[\Delta \mid X]\|_2]$, so it suffices to provide a bound on $\|\mathbb{E}[\Delta \mid X]\|_2$. Throughout this proof and the remainder of the paper, we represent the kernel evaluator by the function ξ_x, where $\xi_x := K(x, \cdot)$ and $f(x) = \langle \xi_x, f \rangle$ for any $f \in \mathcal{H}$. Using this notation, the estimate $\widehat{f}$ minimizes the empirical objective

$$\frac{1}{n} \sum_{i=1}^n \left(\langle \xi_{x_i}, f \rangle_{\mathcal{H}} - y_i \right)^2 + \lambda \, \|f\|_{\mathcal{H}}^2 . \tag{19}$$

This objective is Fréchet differentiable, and as a consequence, the necessary and sufficient conditions for optimality (Luenberger, 1969) of $\widehat{f}$ are that

$$\frac{1}{n} \sum_{i=1}^n (\langle \xi_{x_i}, \widehat{f} - f^* \rangle_{\mathcal{H}} - \varepsilon_i) + \lambda \widehat{f} = \frac{1}{n} \sum_{i=1}^n (\langle \xi_{x_i}, \widehat{f} \rangle_{\mathcal{H}} - y_i) + \lambda \widehat{f} = 0. \tag{20}$$

Taking conditional expectations over the noise variables $\{\varepsilon_i\}_{i=1}^n$ with the design $X = \{x_i\}_{i=1}^n$ fixed, we find that

$$\frac{1}{n} \sum_{i=1}^n \xi_{x_i} \langle \xi_{x_i}, \mathbb{E}[\Delta \mid X] \rangle + \lambda \mathbb{E}[\widehat{f} \mid X] = 0.$$

Define the sample covariance operator $\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n} \xi_{x_i} \otimes \xi_{x_i}$. Adding and subtracting λf^* from the above equation yields

$$(\widehat{\Sigma} + \lambda I)\mathbb{E}[\Delta \mid X] = -\lambda f^*. \tag{21}$$

Consequently, we see we have $\|\mathbb{E}[\Delta \mid X]\|_{\mathcal{H}} \leq \|f^*\|_{\mathcal{H}}$, since $\widehat{\Sigma} \succeq 0$.

We now use a truncation argument to reduce the problem to a finite dimensional problem. To do so, we let $\delta \in \ell_2(\mathbb{N})$ denote the coefficients of $\mathbb{E}[\Delta \mid X]$ when expanded in the basis $\{\phi_j\}_{j=1}^{\infty}$:

$$\mathbb{E}[\Delta \mid X] = \sum_{j=1}^{\infty} \delta_j \phi_j, \quad \text{with } \delta_j = \langle \mathbb{E}[\Delta \mid X], \phi_j \rangle_{L^2(\mathbb{P})}. \tag{22}$$

For a fixed $d \in \mathbb{N}$, define the vectors $\delta^{\downarrow} := (\delta_1, \ldots, \delta_d)$ and $\delta^{\uparrow} := (\delta_{d+1}, \delta_{d+2}, \ldots)$ (we suppress dependence on d for convenience). By the orthonormality of the collection $\{\phi_j\}$, we have

$$\|\mathbb{E}[\Delta \mid X]\|_2^2 = \|\delta\|_2^2 = \|\delta^{\downarrow}\|_2^2 + \|\delta^{\uparrow}\|_2^2. \tag{23}$$

We control each of the elements of the sum (23) in turn.

Control of the term $\|\delta^{\uparrow}\|_2^2$: By definition, we have

$$\|\delta^{\uparrow}\|_2^2 = \frac{\mu_{d+1}}{\mu_{d+1}} \sum_{j=d+1}^{\infty} \delta_j^2 \leq \mu_{d+1} \sum_{j=d+1}^{\infty} \frac{\delta_j^2}{\mu_j} \overset{(i)}{\leq} \mu_{d+1} \|\mathbb{E}[\Delta \mid X]\|_{\mathcal{H}}^2 \overset{(ii)}{\leq} \mu_{d+1} \|f^*\|_{\mathcal{H}}^2, \tag{24}$$

where inequality (i) follows since $\|\mathbb{E}[\Delta \mid X]\|_{\mathcal{H}}^2 = \sum_{j=1}^{\infty} \frac{\delta_j^2}{\mu_j}$; and inequality (ii) follows from the bound $\|\mathbb{E}[\Delta \mid X]\|_{\mathcal{H}} \leq \|f^*\|_{\mathcal{H}}$, which is a consequence of equality (21).

Control of the term $\|\delta^{\downarrow}\|_2^2$: Let $(\theta_1, \theta_2, \ldots)$ be the coefficients of f^* in the basis $\{\phi_j\}$. In addition, define the matrices $\Phi \in \mathbb{R}^{n \times d}$ by

$$\Phi_{ij} = \phi_j(x_i) \quad \text{for } i \in \{1, \ldots, n\}, \text{ and } j \in \{1, \ldots, d\}$$

and $M = \operatorname{diag}(\mu_1, \ldots, \mu_d) \succ 0 \in \mathbb{R}^{d \times d}$. Lastly, define the tail error vector $v \in \mathbb{R}^n$ by

$$v_i := \sum_{j > d} \delta_j \phi_j(x_i) = \mathbb{E}[\Delta^{\uparrow}(x_i) \mid X].$$

Let $l \in \mathbb{N}$ be arbitrary. Computing the (Hilbert) inner product of the terms in equation (21) with ϕ_l, we obtain

$$\begin{aligned} -\lambda \frac{\theta_l}{\mu_l} &= \langle \phi_l, -\lambda f^* \rangle = \left\langle \phi_l, (\widehat{\Sigma} + \lambda)\mathbb{E}[\Delta \mid X] \right\rangle \\ &= \frac{1}{n} \sum_{i=1}^{n} \langle \phi_l, \xi_{x_i} \rangle \langle \xi_{x_i}, \mathbb{E}[\Delta \mid X] \rangle + \lambda \langle \phi_l, \mathbb{E}[\Delta \mid X] \rangle = \frac{1}{n} \sum_{i=1}^{n} \phi_l(x_i) \mathbb{E}[\Delta(x_i) \mid X] + \lambda \frac{\delta_l}{\mu_l}. \end{aligned}$$

We can rewrite the final sum above using the fact that $\Delta = \Delta^{\downarrow} + \Delta^{\uparrow}$, which implies

$$\frac{1}{n}\sum_{i=1}^{n}\phi_l(x_i)\mathbb{E}[\Delta(x_i) \mid X] = \frac{1}{n}\sum_{i=1}^{n}\phi_l(x_i)\Big(\sum_{j=1}^{d}\phi_j(x_i)\delta_j + \sum_{j>d}\phi_j(x_i)\delta_j\Big)$$

Applying this equality for $l = 1, 2, \ldots, d$ yields

$$\left(\frac{1}{n}\Phi^T\Phi + \lambda M^{-1}\right)\delta^{\downarrow} = -\lambda M^{-1}\theta^{\downarrow} - \frac{1}{n}\Phi^T v. \tag{25}$$

We now show how the expression (25) gives us the desired bound in the lemma. By defininingthe shorthand matrix $Q = (I + \lambda M^{-1})$, we have

$$\frac{1}{n}\Phi^T\Phi + \lambda M^{-1} = I + \lambda M^{-1} + \frac{1}{n}\Phi^T\Phi - I = Q\left(I + Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)\right).$$

As a consequence, we can rewrite expression (25) to

$$\left(I + Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)\right)\delta^{\downarrow} = -\lambda Q^{-1}M^{-1}\theta^{\downarrow} - \frac{1}{n}Q^{-1}\Phi^T v. \tag{26}$$

We now present a lemma bounding the terms in equality (26) to control $\delta^{\downarrow}$.

Lemma 7 *The following bounds hold:*

$$\left\|\lambda Q^{-1}M^{-1}\theta^{\downarrow}\right\|_2^2 \leq \lambda\left\|f^*\right\|_{\mathcal{H}}^2/2, \quad \text{and} \tag{27a}$$

$$\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^T v\right\|_2^2\right] \leq \frac{\rho^4\left\|f^*\right\|_{\mathcal{H}}^2\operatorname{tr}(K)\beta_d}{4\lambda}. \tag{27b}$$

Define the event $\mathcal{E} := \left\{\left\|\left\|Q^{-1}((1/n)\Phi^T\Phi - I)\right\|\right\| \leq 1/2\right\}$. *Under Assumption A with moment bound* $\mathbb{E}[\phi_j(X)^{2k}] \leq \rho^{2k}$, *there exists a universal constant* C *such that*

$$\mathbb{P}(\mathcal{E}^c) \leq \left(\max\left\{\sqrt{k \vee \log(2d)}, \frac{k \vee \log(2d)}{n^{1/2-1/k}}\right\}\frac{C(\rho^2\gamma(\lambda)+1)}{\sqrt{n}}\right)^k. \tag{28}$$

We defer the proof of this lemma to Appendix A.1.

Based on this lemma, we can now complete the proof. Whenever the event $\mathcal{E}$ holds, we know that $(I + Q^{-1}((1/n)\Phi^T\Phi - I)) \succeq (1/2)I$. In particular, we have

$$\|\delta^{\downarrow}\|_2^2 \leq 4\left\|\lambda Q^{-1}M^{-1}\theta^{\downarrow} + (1/n)Q^{-1}\Phi^T v\right\|_2^2$$

on $\mathcal{E}$, by Eq. (26). Since $\mathcal{E}$ is X-measureable, we thus obtain

$$\begin{aligned}\mathbb{E}\left[\|\delta^{\downarrow}\|_2^2\right] &= \mathbb{E}\left[1(\mathcal{E})\|\delta^{\downarrow}\|_2^2\right] + \mathbb{E}\left[1(\mathcal{E}^c)\|\delta^{\downarrow}\|_2^2\right] \\ &\leq 4\mathbb{E}\left[1(\mathcal{E})\left\|\lambda Q^{-1}M^{-1}\theta^{\downarrow} + (1/n)Q^{-1}\Phi^T v\right\|_2^2\right] + \mathbb{E}\left[1(\mathcal{E}^c)\|\delta^{\downarrow}\|_2^2\right].\end{aligned}$$

Applying the bounds (27a) and (27b), along with the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\mathbb{E}\left[\|\delta^{\downarrow}\|_2^2\right] \leq 4\lambda \|f^*\|_{\mathcal{H}}^2 + \frac{2\rho^4 \|f^*\|_{\mathcal{H}}^2 \operatorname{tr}(K)\beta_d}{\lambda} + \mathbb{E}\left[1(\mathcal{E}^c)\|\delta^{\downarrow}\|_2^2\right]. \tag{29}$$

Now we use the fact that by the gradient optimality condition (21), $\|\mathbb{E}[\Delta \mid X]\|_2^2 \leq \|f^*\|_2^2$. Recalling the shorthand (5) for $b(n,d,k)$, we apply the bound (28) to see

$$\mathbb{E}\left[1(\mathcal{E}^c)\|\delta^{\downarrow}\|_2^2\right] \leq \mathbb{P}(\mathcal{E}^c)\|f^*\|_2^2 \leq \left(\frac{Cb(n,d,k)(\rho^2\gamma(\lambda)+1)}{\sqrt{n}}\right)^k \|f^*\|_2^2.$$

Combining this with the inequality (29), we obtain the desired statement of Lemma 5.

A.1. Proof of Lemma 7

Proof of bound (27a): Beginning with the proof of the bound (27a), we have

$$\begin{aligned}\left\|Q^{-1}M^{-1}\theta^{\downarrow}\right\|_2^2 &= (\theta^{\downarrow})^T(M+\lambda I)^{-2}\theta^{\downarrow} = (\theta^{\downarrow})^T(M^2+\lambda^2 I + 2\lambda M)^{-1}\theta^{\downarrow} \\ &\leq (\theta^{\downarrow})^T(2\lambda M)^{-1}\theta^{\downarrow} \leq \frac{1}{2\lambda}(\theta^{\downarrow})^T M^{-1}\theta^{\downarrow} \leq \frac{1}{2\lambda}\|f^*\|_{\mathcal{H}}^2.\end{aligned}$$

Multiplying both sides by λ^2 gives the result.

Proof of bound (27b): Next we turn to the proof of the bound (27b). We begin by re-writing $Q^{-1}\Phi^T v$ as the product of two components:

$$\frac{1}{n}Q^{-1}\Phi^T v = (M^{1/2} + \lambda M^{-1/2})^{-1}\left(\frac{1}{n}M^{1/2}\Phi^T v\right). \tag{30}$$

The first matrix is a diagonal matrix whose operator norm is bounded:

$$\left|\left|\left|(M^{\frac{1}{2}} + \lambda M^{-\frac{1}{2}})^{-1}\right|\right|\right| = \max_{j\in[d]} \frac{1}{\sqrt{\mu_j} + \lambda/\sqrt{\mu_j}} = \max_{j\in[d]} \frac{\sqrt{\mu_j}}{\mu_j + \lambda} \leq \frac{1}{2\sqrt{\lambda}}, \tag{31}$$

the final inequality coming because $\sqrt{\mu_j}/(\mu_j+\lambda)$ is maximized at $\mu_j = \lambda$.

For the second factor in the product (30), the analysis is a little more complicated. Let $\Phi_\ell = (\phi_l(x_1), \ldots, \phi_l(x_n))$ be the ℓth column of Φ. In this case,

$$\left\|M^{1/2}\Phi^T v\right\|_2^2 = \sum_{\ell=1}^d \mu_\ell(\Phi_\ell^T v)^2 \leq \sum_{\ell=1}^d \mu_\ell \|\Phi_\ell\|_2^2 \|v\|_2^2, \tag{32}$$

using the Cauchy-Schwarz inequality. Taking expectations with respect to the design $\{x_i\}_{i=1}^n$ and applying Hölder's inequality yields

$$\mathbb{E}[\|\Phi_\ell\|_2^2 \|v\|_2^2] \leq \sqrt{\mathbb{E}[\|\Phi_\ell\|_2^4]}\sqrt{\mathbb{E}[\|v\|_2^4]}.$$

We bound each of the terms in this product in turn. For the first, we have

$$\mathbb{E}[\|\Phi_\ell\|_2^4] = \mathbb{E}\Big[\Big(\sum_{i=1}^n \phi_\ell^2(X_i)\Big)^2\Big] = \mathbb{E}\Big[\sum_{i,j=1}^n \phi_\ell^2(X_i)\phi_\ell^2(X_j)\Big] \le n^2\mathbb{E}[\phi_\ell^4(X_1)] \le n^2\rho^4$$

since the X_i are i.i.d., $\mathbb{E}[\phi_\ell^2(X_1)] \le \sqrt{\mathbb{E}[\phi_\ell^4(X_1)]}$, and $\mathbb{E}[\phi_\ell^4(X_1)] \le \rho^4$ by assumption. Turning to the term involving v, we have

$$v_i^2 = \Big(\sum_{j>d} \delta_j \phi_j(x_i)\Big)^2 \le \Big(\sum_{j>d} \frac{\delta_j^2}{\mu_j}\Big)\Big(\sum_{j>d} \mu_j \phi_j^2(x_i)\Big)$$

by Cauchy-Schwarz. As a consequence, we find

$$\begin{aligned}\mathbb{E}[\|v\|_2^4] = \mathbb{E}\Big[\Big(n\frac{1}{n}\sum_{i=1}^n v_i^2\Big)^2\Big] &\le n^2\frac{1}{n}\sum_{i=1}^n \mathbb{E}[v_i^4] \le n\sum_{i=1}^n \mathbb{E}\Big[\Big(\sum_{j>d}\frac{\delta_j^2}{\mu_j}\Big)^2\Big(\sum_{j>d}\mu_j\phi_j^2(X_i)\Big)^2\Big] \\ &\le n^2\mathbb{E}\Big[\|\mathbb{E}[\Delta \mid X]\|_{\mathcal{H}}^4\Big(\sum_{j>d}\mu_j\phi_j^2(X_1)\Big)^2\Big],\end{aligned}$$

since the X_i are i.i.d. Using the fact that $\|\mathbb{E}[\Delta \mid X]\|_{\mathcal{H}} \le \|f^*\|_{\mathcal{H}}$, we expand the second square to find

$$\frac{1}{n^2}\mathbb{E}[\|v\|_2^4] \le \|f^*\|_{\mathcal{H}}^4 \sum_{j,k>d} \mathbb{E}\left[\mu_j\mu_k\phi_j^2(X_1)\phi_k^2(X_1)\right] \le \|f^*\|_{\mathcal{H}}^4\,\rho^4 \sum_{j,k>d}\mu_j\mu_k = \|f^*\|_{\mathcal{H}}^4\,\rho^4\Big(\sum_{j>d}\mu_j\Big)^2.$$

Combining our bounds on $\|\Phi_\ell\|_2$ and $\|v\|_2$ with our initial bound (32), we obtain the inequality

$$\mathbb{E}\left[\left\|M^{1/2}\Phi^T v\right\|_2^2\right] \le \sum_{l=1}^d \mu_\ell\sqrt{n^2\rho^4}\sqrt{n^2\,\|f^*\|_{\mathcal{H}}^4\,\rho^4\Big(\sum_{j>d}\mu_j\Big)^2} = n^2\rho^4\,\|f^*\|_{\mathcal{H}}^2\Big(\sum_{j>d}\mu_j\Big)\sum_{l=1}^d\mu_\ell.$$

Dividing by n^2, recalling the definition of $\beta_d = \sum_{j>d}\mu_j$, and noting that $\operatorname{tr}(K) \ge \sum_{l=1}^d \mu_\ell$ shows that

$$\mathbb{E}\left[\left\|\frac{1}{n}M^{1/2}\Phi^T v\right\|_2^2\right] \le \rho^4\,\|f^*\|_{\mathcal{H}}^2\,\beta_d\operatorname{tr}(K).$$

Combining this inequality with our expansion (30) and the bound (31) yields the claim (27b).

Proof of bound (28): Let us consider the expectation of the norm of the matrix $Q^{-1}((1/n)\Phi^T\Phi - I)$. For each $i \in [n]$, let $\pi_i = (\phi_1(x_i), \ldots, \phi_d(x_i)) \in \mathbb{R}^d$ denote the ith row of the matrix $\Phi \in \mathbb{R}^{n\times d}$. Then we know that

$$Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right) = \frac{1}{n}Q^{-1}\sum_{i=1}^n(\pi_i\pi_i^T - I).$$

Define the sequence of matrices

$$A_i := \begin{bmatrix} 0 & Q^{-1}\pi_i\pi_i^T - I \\ I - \pi_i\pi_i^T Q^{-1} & 0 \end{bmatrix}$$

Then the matrices $A_i = A_i^T \in \mathbb{R}^{2d\times 2d}$, and moreover $\lambda_{\max}(A_i) = |||A_i||| = |||Q^{-1}\pi_i\pi_i^T - I|||$, and similarly for their averages Bhatia (1997). Note that $\mathbb{E}[A_i] = 0$ and let ε_i be i.i.d. $\{-1, 1\}$-valued Rademacher random variables. Applying a standard symmetrization argument (Ledoux and Talagrand, 1991), we find that for any $k \geq 1$, we have

$$\mathbb{E}\left[\left|\left|\left|Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)\right|\right|\right|^k\right] = \mathbb{E}\left[\left|\left|\left|\frac{1}{n}\sum_{i=1}^n A_i\right|\right|\right|^k\right] \leq 2^k\mathbb{E}\left[\left|\left|\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i A_i\right|\right|\right|^k\right]. \tag{33}$$

Lemma 8 *The quantity* $\mathbb{E}\left[\left|\left|\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i A_i\right|\right|\right|^k\right]^{1/k}$ *is upper bounded by*

$$\sqrt{e(k \vee 2\log(2d))}\frac{\rho^2\sum_{j=1}^d \frac{1}{1+\lambda/\mu_j} + 1}{\sqrt{n}} + \frac{4e(k \vee 2\log(2d))}{n^{1-1/k}}\left(\sum_{j=1}^d \frac{\rho^2}{1+\lambda/\mu_j} + 1\right). \tag{34}$$

We take this lemma as given for the moment, returning to prove it shortly. Recall the definition of the constant $\gamma(\lambda) = \sum_{j=1}^\infty 1/(1+\lambda/\mu_j) \geq \sum_{j=1}^d 1/(1+\lambda/\mu_j)$. Then using our symmetrization inequality (33), we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\left|\left|Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)\right|\right|\right|^k\right] \\
&\leq 2^k\left(\sqrt{e(k \vee \log(2d))}\frac{\rho^2\gamma(\lambda)+1}{\sqrt{n}} + \frac{4e(k \vee 2\log(2d))}{n^{1-1/k}}(\rho^2\gamma(\lambda)+1)\right)^k \\
&\leq \max\left\{\sqrt{k \vee \log(2d)}, \frac{k \vee \log(2d)}{n^{1/2-1/k}}\right\}^k\left(\frac{C(\rho^2\gamma(\lambda)+1)}{\sqrt{n}}\right)^k,
\end{aligned} \tag{35}$$

where C is a numerical constant. By definition of the event $\mathcal{E}$, we see by Markov's inequality that for any $k \in \mathbb{R}, k \geq 1$,

$$\mathbb{P}(\mathcal{E}^c) \leq \frac{\mathbb{E}\left[\left|\left|\left|Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)\right|\right|\right|^k\right]}{2^{-k}} \leq \max\left\{\sqrt{k \vee \log(2d)}, \frac{k \vee \log(2d)}{n^{1/2-1/k}}\right\}^k\left(\frac{2C(\rho^2\gamma(\lambda)+1)}{\sqrt{n}}\right)^k.$$

This completes the proof of the bound (28).

It remains to prove Lemma 8, for which we make use of the following result, due to Chen et al. (2012, Theorem A.1(2)).

Lemma 9 *Let* $X_i \in \mathbb{R}^{d\times d}$ *be independent symmetrically distributed Hermitian matrices. Then*

$$\mathbb{E}\left[\left|\left|\left|\sum_{i=1}^n X_i\right|\right|\right|^k\right]^{1/k} \leq \sqrt{e(k \vee 2\log d)}\left|\left|\left|\sum_{i=1}^n \mathbb{E}[X_i^2]\right|\right|\right|^{1/2} + 2e(k \vee 2\log d)\left(\mathbb{E}[\max_i |||X_i|||^k]\right)^{1/k}. \tag{36}$$

The proof of Lemma 8 is based on applying this inequality with $X_i = \varepsilon_i A_i/n$, and then bounding the two terms on the right-hand side of inequality (36).

We begin with the first term. Note that

$$\mathbb{E}[A_i^2] = \mathbb{E}\,\mathrm{diag}\left(\begin{bmatrix} Q^{-1}\pi_i\pi_i^T + \pi_i\pi_i^T Q^{-1} - \pi_i\pi_i^T Q^{-2}\pi_i\pi_i^T - I \\ Q^{-1}\pi_i\pi_i^T + \pi_i\pi_i^T Q^{-1} - Q^{-1}\pi_i\pi_i^T\pi_i\pi_i^T Q^{-1} - I \end{bmatrix}\right).$$

Moreover, we have $\mathbb{E}[\pi_i\pi_i^T] = I_{d\times d}$, which leaves us needing to compute $\mathbb{E}[\pi_i\pi_i^T Q^{-2}\pi_i\pi_i^T]$ and $\mathbb{E}[Q^{-1}\pi_i\pi_i^T\pi_i\pi_i^T Q^{-1}]$. Instead of computing them directly, we provide bounds on their norms. Since $\pi_i\pi_i^T$ is rank one and Q is diagonal, we have

$$\left|\!\left|\!\left| Q^{-1}\pi_i\pi_i^T \right|\!\right|\!\right| = \left|\!\left|\!\left| \pi_i\pi_i^T Q^{-1} \right|\!\right|\!\right| = \pi_i^T(I + \lambda M^{-1})^{-1}\pi_i = \sum_{j=1}^d \frac{\phi_j(x_i)^2}{1 + \lambda/\mu_j}.$$

We also note that, for any $k \in \mathbb{R}, k \geq 1$, convexity implies that

$$\begin{aligned}\left(\sum_{j=1}^d \frac{\phi_j(x_i)^2}{1+\lambda/\mu_j}\right)^k &= \left(\frac{\sum_{l=1}^d 1/(1+\lambda/\mu_\ell)}{\sum_{l=1}^d 1/(1+\lambda/\mu_\ell)} \sum_{j=1}^d \frac{\phi_j(x_i)^2}{1+\lambda/\mu_j}\right)^k \\ &\leq \left(\sum_{l=1}^d \frac{1}{1+\lambda/\mu_\ell}\right)^k \frac{1}{\sum_{l=1}^d 1/(1+\lambda/\mu_\ell)} \sum_{j=1}^d \frac{\phi_j(x_i)^{2k}}{1+\lambda/\mu_j},\end{aligned}$$

so if $\mathbb{E}[\phi_j(X_i)^{2k}] \leq \rho^{2k}$, we obtain

$$\mathbb{E}\left[\left(\sum_{j=1}^d \frac{\phi_j(x_i)^2}{1+\lambda/\mu_j}\right)^k\right] \leq \left(\sum_{j=1}^d \frac{1}{1+\lambda/\mu_j}\right)^k \rho^{2k}. \tag{37}$$

The sub-multiplicativity of the matrix norm implies $\left|\!\left|\!\left| Q^{-1}\pi_i\pi_i^T\pi_i\pi_i^T Q^{-1} \right|\!\right|\!\right| \leq \left|\!\left|\!\left| Q^{-1}\pi_i\pi_i^T \right|\!\right|\!\right|^2$, and consequently we have

$$\mathbb{E}[\left|\!\left|\!\left| Q^{-1}\pi_i\pi_i^T\pi_i\pi_i^T Q^{-1} \right|\!\right|\!\right|] \leq \mathbb{E}\left[\left(\pi_i^T(I+\lambda M^{-1})^{-1}\pi_i\right)^2\right] \leq \rho^4\left(\sum_{j=1}^d \frac{1}{1+\lambda/\mu_j}\right)^2,$$

where the final step follows from inequality (37).

Returning to our expectation of A_i^2, we note that $0 \preceq Q^{-1} \preceq I$ and $\mathbb{E}[\pi_i\pi_i^T] = I$, and hence

$$\left|\!\left|\!\left| \mathbb{E}[Q^{-1}\pi_i\pi_i^T + \pi_i\pi_i^T Q^{-1} - I] \right|\!\right|\!\right| = \left|\!\left|\!\left| 2Q^{-1} - I \right|\!\right|\!\right| \leq 1.$$

Consequently,

$$\left|\!\left|\!\left| \left(\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[A_i^2]\right)^{1/2} \right|\!\right|\!\right| = \left|\!\left|\!\left| \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[A_i^2] \right|\!\right|\!\right|^{1/2} \leq \frac{1}{\sqrt{n}}\sqrt{\rho^4\left(\sum_{j=1}^d \frac{1}{1+\lambda/\mu_j}\right)^2 + 1}.$$

We have thus obtained the first term on the right-hand side of expression (34).

We now turn to the second term in expression (34). For real $k \geq 1$, we have

$$\mathbb{E}[\max_i |||\varepsilon_i A_i/n|||^k] = \frac{1}{n^k}\mathbb{E}[\max_i |||A_i|||^k] \leq \frac{1}{n^k}\sum_{i=1}^n \mathbb{E}[|||A_i|||^k]$$

Since norms are sub-additive, we find that

$$|||A_i|||^k \leq 2^{k-1}\left(\sum_{j=1}^d \frac{\phi_j(x_i)^2}{1+\lambda/\mu_j}\right)^k + 2^{k-1}.$$

Thus, applying inequality (37), we find that

$$\mathbb{E}[\max_i |||\varepsilon_i A_i/n|||^k] \leq \frac{1}{n^{k-1}}\left[2^{k-1}\left(\sum_{j=1}^d \frac{1}{1+\lambda/\mu_j}\right)^k \rho^{2k} + 2^{k-1}\right].$$

Taking kth roots yields the second term in the expression (34).

Appendix B. Proof of Lemma 6

This proof follows an outline similar to Lemma 5. We begin with a simple bound on $\|\Delta\|_{\mathcal{H}}$:

Lemma 10 *Under Assumption B, we have* $\mathbb{E}[\|\Delta\|_{\mathcal{H}}^2 \mid X] \leq 2\sigma^2/\lambda + 4\|f^*\|_{\mathcal{H}}^2$.

Proof We have

$$\begin{aligned}\lambda\,\mathbb{E}[\,\|\widehat{f}\|_{\mathcal{H}}^2 \mid \{x_i\}_{i=1}^n] &\leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\widehat{f}(x_i) - f^*(x_i) - \varepsilon_i)^2 + \lambda\|\widehat{f}\|_{\mathcal{H}}^2 \mid \{x_i\}_{i=1}^n\right] \\ &\overset{(i)}{\leq} \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\varepsilon_i^2 \mid x_i] + \lambda\,\|f^*\|_{\mathcal{H}}^2 \\ &\overset{(ii)}{\leq} \sigma^2 + \lambda\,\|f^*\|_{\mathcal{H}}^2\,,\end{aligned}$$

where inequality (i) follows since $\widehat{f}$ minimizes the objective function (2); and inequality (ii) uses the fact that $\mathbb{E}[\varepsilon_i^2 \mid x_i] \leq \sigma^2$. Applying the triangle inequality to $\|\Delta\|_{\mathcal{H}}$ along with the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we find that

$$\mathbb{E}[\|\Delta\|_{\mathcal{H}}^2 \mid \{x_i\}_{i=1}^n] \leq 2\,\|f^*\|_{\mathcal{H}}^2 + 2\mathbb{E}[\|\widehat{f}\|_{\mathcal{H}}^2 \mid \{x_i\}_{i=1}^n] \;\leq\; \frac{2\sigma^2}{\lambda} + 4\,\|f^*\|_{\mathcal{H}}^2\,,$$

which completes the proof. ■

With Lemma 10 in place, we now proceed to the proof of the theorem proper. Recall from Lemma 5 the optimality condition

$$\frac{1}{n}\sum_{i=1}^n \xi_{x_i}(\langle \xi_{x_i}, \widehat{f} - f^*\rangle - \varepsilon_i) + \lambda\widehat{f} = 0. \tag{38}$$

Now, let $\delta \in \ell_2(\mathbb{N})$ be the expansion of the error Δ in the basis $\{\phi_j\}$, so that $\Delta = \sum_{j=1}^{\infty} \delta_j \phi_j$, and (again, as in Lemma 5), we choose $d \in \mathbb{N}$ and truncate Δ via

$$\Delta^{\downarrow} := \sum_{j=1}^{d} \delta_j \phi_j \quad \text{and} \quad \Delta^{\uparrow} := \Delta - \Delta^{\downarrow} = \sum_{j>d} \delta_j \phi_j.$$

Let $\delta^{\downarrow} \in \mathbb{R}^d$ and $\delta^{\uparrow}$ denote the corresponding vectors for the above. As a consequence of the orthonormality of the basis functions, we have

$$\mathbb{E}[\|\Delta\|_2^2] = \mathbb{E}[\|\Delta^{\downarrow}\|_2^2] + \mathbb{E}[\|\Delta^{\uparrow}\|_2^2] = \mathbb{E}[\|\delta^{\downarrow}\|_2^2] + \mathbb{E}[\|\delta^{\uparrow}\|_2^2]. \tag{39}$$

We bound each of the terms (39) in turn.

By Lemma 10, the second term is upper bounded as

$$\mathbb{E}[\|\Delta^{\uparrow}\|_2^2] = \sum_{j>d} \mathbb{E}[\delta_j^2] \leq \sum_{j>d} \frac{\mu_{d+1}}{\mu_j} \mathbb{E}[\delta_j^2] = \mu_{d+1} \mathbb{E}[\|\Delta^{\uparrow}\|_{\mathcal{H}}^2] \leq \mu_{d+1} \left(\frac{2\sigma^2}{\lambda} + 4 \|f^*\|_{\mathcal{H}}^2 \right). \tag{40}$$

The remainder of the proof is devoted the bounding the term $\mathbb{E}[\|\Delta^{\downarrow}\|_2^2]$ in the decomposition (39). By taking the Hilbert inner product of ϕ_k with the optimality condition (38), we find as in our derivation of the matrix equation (25) that for each $k \in \{1, \ldots, d\}$

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \phi_k(x_i) \phi_j(x_i) \delta_j + \frac{1}{n} \sum_{i=1}^{n} \phi_k(x_i)(\Delta^{\uparrow}(x_i) - \varepsilon_i) + \lambda \frac{\delta_k}{\mu_k} = 0.$$

Given the expansion $f^* = \sum_{j=1}^{\infty} \theta_j \phi_j$, define the tail error vector $v \in \mathbb{R}^n$ by $v_i = \sum_{j>d} \delta_j \phi_j(x_i)$, and recall the definition of the eigenvalue matrix $M = \operatorname{diag}(\mu_1, \ldots, \mu_d) \in \mathbb{R}^{d \times d}$. Given the matrix Φ defined by its coordinates $\Phi_{ij} = \phi_j(x_i)$, we have

$$\left(\frac{1}{n} \Phi^T \Phi + \lambda M^{-1} \right) \delta^{\downarrow} = -\lambda M^{-1} \theta^{\downarrow} - \frac{1}{n} \Phi^T v + \frac{1}{n} \Phi^T \varepsilon. \tag{41}$$

As in the proof of Lemma 5, we find that

$$\left(I + Q^{-1} \left(\frac{1}{n} \Phi^T \Phi - I \right) \right) \delta^{\downarrow} = -\lambda Q^{-1} M^{-1} \theta^{\downarrow} - \frac{1}{n} Q^{-1} \Phi^T v + \frac{1}{n} Q^{-1} \Phi^T \varepsilon, \tag{42}$$

here $Q := (I + \lambda M^{-1})$.

We now recall the bounds (27a) and (28) from Lemma 7, as well as the previously defined event $\mathcal{E} := \{ \left\| Q^{-1}((1/n)\Phi^T \Phi - I) \right\| \leq 1/2 \}$. When $\mathcal{E}$ occurs, the expression (42) implies the inequality

$$\|\Delta^{\downarrow}\|_2^2 \leq 4 \left\| -\lambda Q^{-1} M^{-1} \theta^{\downarrow} - (1/n) Q^{-1} \Phi^T v + (1/n) Q^{-1} \Phi^T \varepsilon \right\|_2^2.$$

When $\mathcal{E}$ fails to hold, Lemma 10 may still be applied since $\mathcal{E}$ is measureable with respect to $\{x_i\}_{i=1}^n$. Doing so yields

$$\begin{aligned}
&\mathbb{E}[\|\Delta^{\downarrow}\|_2^2] = \mathbb{E}[1(\mathcal{E}) \|\Delta^{\downarrow}\|_2^2] + \mathbb{E}[1(\mathcal{E}^c) \|\Delta^{\downarrow}\|_2^2] \\
&\leq 4\mathbb{E} \left[\left\| -\lambda Q^{-1} M^{-1} \theta^{\downarrow} - (1/n) Q^{-1} \Phi^T v + (1/n) Q^{-1} \Phi^T \varepsilon \right\|_2^2 \right] + \mathbb{E} \left[1(\mathcal{E}^c) \, \mathbb{E}[\|\Delta^{\downarrow}\|_2^2 \mid \{x_i\}_{i=1}^n] \right] \\
&\leq 4\mathbb{E} \left[\left\| \lambda Q^{-1} M^{-1} \theta^{\downarrow} + \frac{1}{n} Q^{-1} \Phi^T v - \frac{1}{n} Q^{-1} \Phi^T \varepsilon \right\|_2^2 \right] + \mathbb{P}(\mathcal{E}^c) \left(\frac{2\sigma^2}{\lambda} + 4 \|f^*\|_{\mathcal{H}}^2 \right).
\end{aligned} \tag{43}$$

Since the bound (28) still holds, it remains to provide a bound on the first term in the expression (43).

As in the proof of Lemma 5, we have $\|\lambda Q^{-1}M^{-1}\theta^{\downarrow}\|_2^2 \leq \lambda/2$ via the bound (27a). Turning to the second term inside the norm, we claim that, under the conditions of Lemma 6, the following bound holds:

$$\mathbb{E}\left[\left\|(1/n)Q^{-1}\Phi^T v\right\|_2^2\right] \leq \frac{\rho^4 \operatorname{tr}(K)\beta_d(2\sigma^2/\lambda + 4\|f^*\|_{\mathcal{H}}^2)}{4\lambda}. \tag{44}$$

This claim is an analogue of our earlier bound (27b), and we prove it shortly. Lastly, we bound the norm of $Q^{-1}\Phi^T\varepsilon/n$. Noting that the diagonal entries of Q^{-1} are $1/(1+\lambda/\mu_j)$, we have

$$\mathbb{E}\left[\left\|Q^{-1}\Phi^T\varepsilon\right\|_2^2\right] = \sum_{j=1}^{d}\sum_{i=1}^{n}\frac{1}{(1+\lambda/\mu_j)^2}\mathbb{E}[\phi_j^2(X_i)\varepsilon_i^2]$$

Since $\mathbb{E}[\phi_j^2(X_i)\varepsilon_i^2] = \mathbb{E}[\phi_j^2(X_i)\mathbb{E}[\varepsilon_i^2 \mid X_i]] \leq \sigma^2$ by assumption, we have the inequality

$$\mathbb{E}\left[\left\|(1/n)Q^{-1}\Phi^T\varepsilon\right\|_2^2\right] \leq \frac{\sigma^2}{n}\sum_{j=1}^{d}\frac{1}{(1+\lambda/\mu_j)^2}.$$

Noting that $1/(1+\lambda/\mu_j)^2 \leq 1/(1+\lambda/\mu_j)$, the last sum is bounded by $(\sigma^2/n)\gamma(\lambda)$. Applying the inequality $(a+b+c)^2 \leq 3a^2+3b^2+3c^2$ to inequality (43), we obtain

$$\mathbb{E}\left[\|\Delta^{\downarrow}\|_2^2\right] \leq 6\lambda\|f^*\|_{\mathcal{H}}^2 + \frac{12\sigma^2\gamma(\lambda)}{n} + \left(\frac{2\sigma^2}{\lambda} + 4\|f^*\|_{\mathcal{H}}^2\right)\left(\frac{3\rho^4\operatorname{tr}(K)\beta_d}{\lambda} + \mathbb{P}(\mathcal{E}^c)\right).$$

Applying the bound (28) to control $\mathbb{P}(\mathcal{E}^c)$ and bounding $\mathbb{E}[\|\Delta^{\uparrow}\|_2^2]$ using inequality (40) completes the proof of the lemma.

It remains to prove bound (44). Recalling the inequality (31), we see that

$$\left\|(1/n)Q^{-1}\Phi^T v\right\|_2^2 \leq \left\|\left\|Q^{-1}M^{-1/2}\right\|\right\|^2\left\|(1/n)M^{1/2}\Phi^T v\right\|_2^2 \leq \frac{1}{4\lambda}\left\|(1/n)M^{1/2}\Phi^T v\right\|_2^2. \tag{45}$$

Let Φ_ℓ denote the ℓth column of the matrix Φ. Taking expectations yields

$$\mathbb{E}\left[\left\|M^{1/2}\Phi^T v\right\|_2^2\right] = \sum_{l=1}^{d}\mu_\ell\mathbb{E}[\langle\Phi_\ell, v\rangle^2] \leq \sum_{l=1}^{d}\mu_\ell\mathbb{E}\left[\|\Phi_\ell\|_2^2\|v\|_2^2\right] = \sum_{l=1}^{d}\mu_\ell\mathbb{E}\left[\|\Phi_\ell\|_2^2\,\mathbb{E}\left[\|v\|_2^2 \mid X\right]\right].$$

Now consider the inner expectation. Applying the Cauchy-Schwarz inequality as in the proof of the bound (27b), we have

$$\|v\|_2^2 = \sum_{i=1}^{n}v_i^2 \leq \sum_{i=1}^{n}\left(\sum_{j>d}\frac{\delta_j^2}{\mu_j}\right)\left(\sum_{j>d}\mu_j\phi_j^2(X_i)\right).$$

Notably, the second term is X-measureable, and the first is bounded by $\|\Delta^{\uparrow}\|_{\mathcal{H}}^2 \le \|\Delta\|_{\mathcal{H}}^2$. We thus obtain

$$\mathbb{E}\left[\left\|M^{1/2}\Phi^T v\right\|_2^2\right] \le \sum_{i=1}^{n}\sum_{l=1}^{d}\mu_\ell \mathbb{E}\left[\|\Phi_\ell\|_2^2\left(\sum_{j>d}\mu_j\phi_j^2(X_i)\right)\mathbb{E}[\|\Delta\|_{\mathcal{H}}^2 \mid X]\right]. \tag{46}$$

Lemma 10 provides the bound $2\sigma^2/\lambda + 4\|f^*\|_{\mathcal{H}}^2$ on the final (inner) expectation.

The remainder of the argument proceeds precisely as in the bound (27b). We have

$$\mathbb{E}[\|\Phi_\ell\|_2^2\,\phi_j(X_i)^2] \le n\rho^4$$

by the moment assumptions on ϕ_j, and thus

$$\mathbb{E}\left[\left\|M^{1/2}\Phi^T v\right\|_2^2\right] \le \sum_{l=1}^{d}\sum_{j>d}\mu_\ell\mu_j n^2\rho^4\left(\frac{2\sigma^2}{\lambda} + 4\|f^*\|_{\mathcal{H}}^2\right) \le n^2\rho^4\beta_d\,\mathrm{tr}(K)\left(\frac{2\sigma^2}{\lambda} + 4\|f^*\|_{\mathcal{H}}^2\right).$$

Dividing by $4\lambda n^2$ completes the proof.