Journal of Machine Learning Research 16 (2015) 3299-3340 Submitted 1/15; Revised 7/15; Published 12/15

Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm with Minimax Optimal Rates

Yuchen Zhang YUCZHANG@BERKELEY.EDU
Department of Electrical Engineering and Computer Science
University of California, Berkeley, Berkeley, CA 94720, USA

John Duchi JDUCHI@STANFORD.EDU
Departments of Statistics and Electrical Engineering
Stanford University, Stanford, CA 94305, USA

Martin Wainwright WAINWRIG@BERKELEY.EDU
Departments of Statistics and Electrical Engineering and Computer Science
University of California, Berkeley, Berkeley, CA 94720, USA

Editor: Hui Zou

Abstract

We study a decomposition-based scalable approach to kernel ridge regression, and show that it achieves minimax optimal convergence rates under relatively mild conditions. 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 using a careful choice of the regularization parameter, 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 two main theorems establish that despite the computational speed-up, statistical optimality is retained: as long as m is not too large, the partition-based estimator achieves the statistical minimax rate over all estimators using the set of N samples. As concrete examples, our theory guarantees that the number of subsets m may grow nearly linearly for finite-rank or Gaussian kernels and polynomially in N for Sobolev spaces, which in turn allows for substantial reductions in computational cost. We conclude with experiments on both simulated data and a music-prediction task that complement our theoretical results, exhibiting the computational and statistical benefits of our approach.

Keywords: kernel ridge regression, divide and conquer, computation complexity

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 a researchers have studied a wide range of estimators (see, for example, the books of Gyorfi et al. (2002), Wasserman (2006), or van de Geer (2000)). One class of methods, known as regularized M-estimators (van de Geer, 2000), are based on minimizing the combination 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 (Koltchinskii, 2006; Mendelson, 2002a; 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 $\mathcal{O}(N^3)$ time and $\mathcal{O}(N^2)$ memory. 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. The associated prediction error has only been studied very recently. Concurrent work by Bach (2013) establishes conditions on the maintained rank that still guarantee optimal convergence rates; see the discussion in Section 7 for more detail. 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. (2013) for parametric smooth convex optimization 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 and discussing the assumptions that underlie our analysis. 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 provide both a result when the regression function f^* belongs to the Hilbert space $\mathcal{H}$ associated with the kernel, as well as a more general oracle inequality that holds for a general 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 Sections 4 and 5 to the proofs of our results, deferring more technical aspects of the analysis to appendices. Lastly, we present simulation results in Section 6.1 to further explore our theoretical results, while Section 6.2 contains experiments with a reasonably large music prediction experiment.

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, referring the reader to one of 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 in $L^2(\mathbb{P})$. For each $x \in \mathcal{X}$, the function $z \mapsto K(z, x)$ is contained with 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

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

Consequently, we see that the RKHS can be viewed as an elliptical subset of the sequence space $\ell^2(\mathbb{N})$ as 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,

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

where $\lambda > 0$ is a regularization parameter. When $\mathcal{H}$ is a reproducing kernel Hilbert space, then 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.

By the representer theorem for reproducing kernel Hilbert spaces (Wahba, 1990), any solution to the KRR program (2) must belong to the linear span of the kernel functions $\{K(\cdot, x_i),\, i = 1, \ldots, N\}$. This fact allows the computation of the KRR estimate to be reduced to an N-dimensional quadratic program, involving the N^2 entries of the kernel matrix $\{K(x_i, x_j),\, i, j = 1, \ldots, n\}$. On the statistical side, a line of past work (van de Geer, 2000; Zhang, 2005; Caponnetto and De Vito, 2007; Steinwart et al., 2009; Hsu et al., 2012) has provided bounds on the estimation error of $\widehat{f}$ as a function of N and λ.

3. Main Results and Their Consequences

We now turn to the description of our algorithm, followed by the statements of our main results, namely Theorems 1 and 2. Each theorem provides an upper bound on the mean-squared prediction error for any trace class kernel. The second theorem is of "oracle type," meaning that it applies even when the true regression function f^* does not belong to the Hilbert space $\mathcal{H}$, and hence involves a combination of approximation and estimation error terms. The first theorem requires that $f^* \in \mathcal{H}$, and provides somewhat sharper bounds on the estimation error in this case. Both of these theorems apply to any trace class kernel, but as we illustrate, they provide concrete results when applied to specific classes of kernels. Indeed, as a corollary, we establish that our distributed KRR algorithm achieves minimax-optimal rates for three different kernel classes, namely finite-rank, Gaussian, and Sobolev.

3.1 Algorithm and Assumptions

The divide-and-conquer algorithm Fast-KRR is easy to describe. 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 random into the m disjoint subsets $S_1, \ldots, S_m \subset \mathcal{X} \times \mathbb{R}$, such that every subset contains N/m samples.

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 $j \in \mathbb{N}$.*

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 $j \in \mathbb{N}$.*

Assumption A′ holds, for example, when the input x is drawn from a closed interval and the kernel is translation invariant, i.e. $K(x, x') = \psi(x - x')$ for some even function ψ. Given input space $\mathcal{X}$ and kernel K, the assumption is verifiable without the data.

Recalling that $f^*(x) := \mathbb{E}[Y \mid X = x]$, our second assumption involves the deviations of the zero-mean noise variables $Y - f^*(x)$. In the simplest case, when $f^* \in \mathcal{H}$, we require only a bounded variance condition:

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

When the function $f^* \notin \mathcal{H}$, we require a slightly stronger variant of this assumption. For each $\lambda \geq 0$, define

$$f_\lambda^* = \operatorname*{argmin}_{f \in \mathcal{H}} \left\{ \mathbb{E}\left[(f(X) - Y)^2\right] + \lambda \|f\|_{\mathcal{H}}^2 \right\}. \tag{4}$$

Note that $f^* = f_0^*$ corresponds to the usual regression function. As $f^* \in L^2(\mathbb{P})$, for each $\lambda \geq 0$, the associated mean-squared error $\sigma_\lambda^2(x) := \mathbb{E}[(Y - f_\lambda^*(x))^2 \mid x]$ is finite for almost every x. In this more general setting, the following assumption replaces Assumption B:

Assumption B′ *For any $\lambda \geq 0$, there exists a constant $\tau_\lambda < \infty$ such that $\tau_\lambda^4 = \mathbb{E}[\sigma_\lambda^4(X)]$.*

3.2 Statement of Main Results

With these assumptions in place, we are now ready for the statements of our main results. All of our results give bounds 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. Both theorem statements involve the following three kernel-related quantities:

$$\operatorname{tr}(K) := \sum_{j=1}^{\infty} \mu_j, \quad \gamma(\lambda) := \sum_{j=1}^{\infty} \frac{1}{1 + \lambda/\mu_j}, \quad \text{and} \quad \beta_d = \sum_{j=d+1}^{\infty} \mu_j. \tag{5}$$

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 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 $\beta_0 = \operatorname{tr} K$. Finally, both theorems involve a quantity that depends on the number of moments k in Assumption A:

$$b(n, d, k) := \max \left\{ \sqrt{\max\{k, \log(d)\}}, \frac{\max\{k, \log(d)\}}{n^{1/2 - 1/k}} \right\}. \tag{6}$$

Here the integer $d \in \mathbb{N}$ is a free parameter 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*

$$\mathbb{E}\left[\|\bar{f} - f^*\|_2^2\right] \leq \left(8 + \frac{12}{m}\right) \lambda \|f^*\|_{\mathcal{H}}^2 + \frac{12\sigma^2\gamma(\lambda)}{N} + \inf_{d \in \mathbb{N}} \left\{T_1(d) + T_2(d) + T_3(d)\right\}, \tag{7}$$

where

$$T_1(d) = \frac{8\rho^4 \|f^*\|_{\mathcal{H}}^2 \operatorname{tr}(K)\beta_d}{\lambda}, \quad T_2(d) = \frac{4\|f^*\|_{\mathcal{H}}^2 + 2\sigma^2/\lambda}{m}\left(\mu_{d+1} + \frac{12\rho^4 \operatorname{tr}(K)\beta_d}{\lambda}\right), \quad \textit{and}$$

$$T_3(d) = \left(Cb(n, d, k)\frac{\rho^2\gamma(\lambda)}{\sqrt{n}}\right)^k \mu_0 \|f^*\|_{\mathcal{H}}^2 \left(1 + \frac{2\sigma^2}{m\lambda} + \frac{4\|f^*\|_{\mathcal{H}}^2}{m}\right),$$

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 make a few heuristic arguments in order to provide intuition. In typical settings, the term $T_3(d)$ goes to zero quickly: if the number of moments k is suitably large and number of partitions m is small—say enough to guarantee that $(b(n,d,k)\gamma(\lambda)/\sqrt{n})^k = \mathcal{O}(1/N)$—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 (7) 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 (7). In such settings, Theorem 1 guarantees an upper bound of the form

$$\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right] = \mathcal{O}(1) \cdot \Big[\underbrace{\lambda \left\|f^*\right\|_{\mathcal{H}}^2}_{\text{Squared bias}} + \underbrace{\frac{\sigma^2 \gamma(\lambda)}{N}}_{\text{Variance}} \Big]. \tag{8}$$

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

$$\lambda \left\|f^*\right\|_{\mathcal{H}}^2 \simeq \sigma^2 \frac{\gamma(\lambda)}{N}. \tag{9}$$

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 (9) 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 (9), followed by application of Theorem 1, yields minimax-optimal prediction error (for the Fast-KRR algorithm) for many kernel classes.

We now turn to an error bound that applies without requiring that $f^* \in \mathcal{H}$. In order to do so, we introduce an auxiliary variable $\bar{\lambda} \in [0, \lambda]$ for use in our analysis (the algorithm's execution does not depend on $\bar{\lambda}$, and in our ensuing bounds we may choose any $\bar{\lambda} \in [0, \lambda]$ to give the sharpest possible results). Let the radius $R = \left\|f^*_{\bar{\lambda}}\right\|_{\mathcal{H}}$, where the population (regularized) regression function $f^*_{\bar{\lambda}}$ was previously defined (4). The theorem requires a few additional conditions to those in Theorem 1, involving the quantities $\mathrm{tr}(K)$, $\gamma(\lambda)$ and β_d defined in Eq. (5), as well as the error moment $\tau_{\bar{\lambda}}$ from Assumption B′. We assume that the triplet (m, d, k) of positive integers satisfy the conditions

$$\begin{aligned} &\beta_d \le \frac{\lambda}{(R^2 + \tau_{\bar{\lambda}}^2/\lambda)N}, \qquad \mu_{d+1} \le \frac{1}{(R^2 + \tau_{\bar{\lambda}}^2/\lambda)N}, \\ &m \le \min\left\{ \frac{\sqrt{N}}{\rho^2 \gamma(\lambda)\log(d)}, \frac{N^{1-\frac{2}{k}}}{(R^2 + \tau_{\bar{\lambda}}^2/\lambda)^{2/k}(b(n,d,k)\rho^2\gamma(\lambda))^2} \right\}. \end{aligned} \tag{10}$$

We then have the following result:

Theorem 2 *Under condition* (10)*, Assumption A with $k \geq 4$, and Assumption B$'$, for any $\bar{\lambda} \in [0, \lambda]$ and $q > 0$ we have*

$$\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right] \leq \left(1 + \frac{1}{q}\right) \inf_{\|f\|_{\mathcal{H}} \leq R} \|f - f^*\|_2^2 + (1 + q)\, \mathcal{E}_{N,m}(\lambda, \bar{\lambda}, R, \rho) \tag{11}$$

where the residual term is given by

$$\mathcal{E}_{N,m}(\lambda, \bar{\lambda}, R, \rho) := \left(\left(4 + \frac{C}{m}\right)(\lambda - \bar{\lambda})R^2 + \frac{C\gamma(\lambda)\rho^2\tau_{\bar{\lambda}}^2}{N} + \frac{C}{N}\right), \tag{12}$$

and C denotes a universal (numerical) constant.

Remarks: Theorem 2 is an oracle inequality, as it upper bounds the mean-squared error in terms of the error $\inf_{\|f\|_{\mathcal{H}} \leq R} \|f - f^*\|_2^2$, which may only be obtained by an oracle knowing the sampling distribution $\mathbb{P}$, along with the residual error term (12).

In some situations, it may be difficult to verify Assumption B$'$. In such scenarios, an alternative condition suffices. For instance, if there exists a constant $\kappa < \infty$ such that $\mathbb{E}[Y^4] \leq \kappa^4$, then under condition (10), the bound (11) holds with $\tau_{\bar{\lambda}}^2$ replaced by $\sqrt{8\,\mathrm{tr}(K)^2 R^4 \rho^4 + 8\kappa^4}$—that is, with the alternative residual error

$$\widetilde{\mathcal{E}}_{N,m}(\lambda, \bar{\lambda}, R, \rho) := \left(\left(2 + \frac{C}{m}\right)(\lambda - \bar{\lambda})R^2 + \frac{C\gamma(\lambda)\rho^2\sqrt{8\,\mathrm{tr}(K)^2 R^4 \rho^4 + 8\kappa^4}}{N} + \frac{C}{N}\right). \tag{13}$$

In essence, if the response variable Y has sufficiently many moments, the prediction mean-square error $\tau_{\bar{\lambda}}^2$ in the statement of Theorem 2 can be replaced by constants related to the size of $\left\|f_{\bar{\lambda}}^*\right\|_{\mathcal{H}}$. See Section 5.2 for a proof of inequality (13).

In comparison with Theorem 1, Theorem 2 provides somewhat looser bounds. It is, however, instructive to consider a few special cases. For the first, we may assume that $f^* \in \mathcal{H}$, in which case $\|f^*\|_{\mathcal{H}} < \infty$. In this setting, the choice $\bar{\lambda} = 0$ (essentially) recovers Theorem 1, since there is no approximation error. Taking $q \to 0$, we are thus left with the bound

$$\mathbb{E}\|\bar{f} - f^*\|_2^2] \;\lesssim\; \lambda\, \|f^*\|_{\mathcal{H}}^2 + \frac{\gamma(\lambda)\rho^2\tau_0^2}{N}, \tag{14}$$

where $\lesssim$ denotes an inequality up to constants. By inspection, this bound is roughly equivalent to Theorem 1; see in particular the decomposition (8). On the other hand, when the condition $f^* \in \mathcal{H}$ fails to hold, we can take $\bar{\lambda} = \lambda$, and then choose q to balance between the familiar approximation and estimation errors: we have

$$\mathbb{E}[\|\bar{f} - f^*\|_2^2] \lesssim \left(1 + \frac{1}{q}\right) \underbrace{\inf_{\|f\|_{\mathcal{H}} \leq R} \|f - f^*\|_2^2}_{\text{approximation}} \;+\; (1 + q) \underbrace{\left(\frac{\gamma(\lambda)\rho^2\tau_{\lambda}^2}{N}\right)}_{\text{estimation}}. \tag{15}$$

Relative to Theorem 1, the condition (10) required to apply Theorem 2 involves constraints on the number m of subsampled data sets that are more restrictive. In particular, when ignoring constants and logarithm terms, the quantity m may grow at rate $\sqrt{N/\gamma^2(\lambda)}$. By contrast, Theorem 1 allows m to grow as quickly as $N/\gamma^2(\lambda)$ (recall the remarks on $T_3(d)$ following Theorem 1 or look ahead to condition (28)). Thus—at least in our current analysis—generalizing to the case that $f^* \notin \mathcal{H}$ prevents us from dividing the data into finer subsets.

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. Throughout this section, we assume that $f^* \in \mathcal{H}$.

3.3.1 Finite-rank Kernels

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

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

where c is a universal (numerical) constant. For suitably large N, the mean-squared error is bounded as

$$\mathbb{E}\left[\left\| \bar{f} - f^* \right\|_2^2 \right] = \mathcal{O}(1) \frac{\sigma^2 r}{N}. \tag{16}$$

For finite-rank kernels, the rate (16) is known to be minimax-optimal, meaning that there is a universal constant $c' > 0$ such that

$$\inf_{\widetilde{f}} \sup_{\|f^*\|_{\mathcal{H}} \le 1} \mathbb{E}[\| \widetilde{f} - f^* \|_2^2] \ge c' \, \frac{r}{N}, \tag{17}$$

where the infimum ranges over all estimators $\widetilde{f}$ based on observing all N samples (and with no constraints on memory and/or computation). This lower bound follows from Theorem 2(a) of Raskutti et al. (2012) with $s = d = 1$.

3.3.2 Polynomially Decaying Eigenvalues

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

$$\mu_j \le C\, j^{-2\nu} \quad \text{for all } j = 1, 2, \ldots, \tag{18}$$

where C is a universal constant, and $\nu > 1/2$ parameterizes the decay rate. We note that equation (5) assumes a finite kernel trace $\mathrm{tr}(K) := \sum_{j=1}^{\infty} \mu_j$. Since $\mathrm{tr}(K)$ appears in Theorem 1, it is natural to use $\sum_{j=1}^{\infty} C j^{-2\nu}$ as an upper bound on $\mathrm{tr}(K)$. This upper bound is finite if and only if $\nu > 1/2$.

Kernels with polynomial decaying eigenvalues include those that underlie for the Sobolev spaces with different orders of smoothness (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 eigendecay with larger values of the parameter ν.

Corollary 4 *For any kernel with ν-polynomial eigendecay* (18)*, 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*

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

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

$$\mathbb{E}\left[\|\bar{f} - f^*\|_2^2 \right] = \mathcal{O}\left(\left(\frac{\sigma^2}{N} \right)^{\frac{2\nu}{2\nu+1}} \right). \tag{19}$$

The upper bound (19) 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).

3.3.3 Exponentially Decaying Eigenvalues

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

$$\mu_j \le c_1 \exp(-c_2 j^2) \quad \text{for all } j = 1, 2, \ldots, \tag{20}$$

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 5 *For a kernel with sub-Gaussian eigendecay* (20)*, 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*

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

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

$$\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right] = \mathcal{O}\left(\sigma^2 \frac{\sqrt{\log N}}{N}\right). \tag{21}$$

The upper bound (21) is minimax optimal; see, for example, Theorem 1 and Example 2 of the recent paper by Yang et al. (2015).

3.3.4 Summary

Each corollary gives a critical threshold for the number m of data partitions: as long as m is below this threshold, 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 than four 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, as it requires solving problems only of size N/m.

3.4 The Choice of Regularization Parameter

In practice, the local sample size on each machine may be different and the optimal choice for the regularization λ may not be known *a priori*, so that an adaptive choice of the regularization parameter λ is desirable (e.g. Tsybakov, 2009, Chapters 3.5–3.7). We recommend using cross-validation to choose the regularization parameter, and we now sketch a heuristic argument that an adaptive algorithm using cross-validation may achieve optimal rates of convergence. (We leave fuller analysis to future work.)

Let λ_n be the (oracle) optimal regularization parameter given knowledge of the sampling distribution $\mathbb{P}$ and eigen-structure of the kernel K. We assume (cf. Corollary 4) that there is a constant $\nu > 0$ such that $\lambda_n \asymp n^{-\nu}$ as $n \to \infty$. Let n_i be the local sample size for each machine i and N the global sample size; we assume that $n_i \gg \sqrt{N}$ (clearly, $N \geq n_i$). First, use local cross-validation to choose regularization parameters $\widehat{\lambda}_{n_i}$ and $\widehat{\lambda}_{n_i^2/N}$ corresponding to samples of size n_i and n_i^2/N, respectively. Heuristically, if cross validation is successful, we expect to have $\widehat{\lambda}_{n_i} \simeq n_i^{-\nu}$ and $\widehat{\lambda}_{n_i^2/N} \simeq N^\nu n_i^{-2\nu}$, yielding that $\widehat{\lambda}_{n_i}^2/\widehat{\lambda}_{n_i^2/N} \simeq N^{-\nu}$. With this intuition, we then compute local estimates

$$\widehat{f}_i := \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n_i} \sum_{(x,y) \in S_i} (f(x) - y)^2 + \widehat{\lambda}_{(i)} \left\|f\right\|_{\mathcal{H}}^2 \quad \text{where } \widehat{\lambda}_{(i)} := \frac{\widehat{\lambda}_{n_i}^2}{\widehat{\lambda}_{n_i^2/N}} \tag{22}$$

and global average estimate $\bar{f} = \sum_{i=1}^m \frac{n_i}{N} \widehat{f}_i$ as usual. Notably, we have $\widehat{\lambda}_{(i)} \simeq \lambda_N$ in this heuristic setting. Using formula (22) and the average $\bar{f}$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right] &= \mathbb{E}\left[\left\|\sum_{i=1}^m \frac{n_i}{N}\left(\widehat{f}_i - \mathbb{E}[\widehat{f}_i]\right)\right\|_2^2\right] + \left\|\sum_{i=1}^m \frac{n_i}{N}\left(\mathbb{E}[\widehat{f}_i] - f^*\right)\right\|_2^2 \\ &\leq \sum_{i=1}^m \frac{n_i^2}{N^2} \mathbb{E}\left[\left\|\widehat{f}_i - \mathbb{E}[\widehat{f}_i]\right\|_2^2\right] + \max_{i \in [m]} \left\{\left\|\mathbb{E}[\widehat{f}_i] - f^*\right\|_2^2\right\}. \end{aligned} \tag{23}$$

Using Lemmas 6 and 7 from the proof of Theorem 1 to come and assuming that $\widehat{\lambda}_n$ is concentrated tightly enough around λ_n, we obtain $\|\mathbb{E}[\widehat{f}_i] - f^*\|_2^2 = \mathcal{O}(\lambda_N \|f^*\|_{\mathcal{H}}^2)$ by Lemma 6 and that $\mathbb{E}[\|\widehat{f}_i - \mathbb{E}[\widehat{f}_i]\|_2^2] = \mathcal{O}(\frac{\gamma(\lambda_N)}{n_i})$ by Lemma 7. Substituting these bounds into inequality (23) and noting that $\sum_i n_i = N$, we may upper bound the overall estimation error as

$$\mathbb{E}\Big[\, \|\bar{f} - f^*\|_2^2 \Big] \le \mathcal{O}(1) \cdot \left(\lambda_N \|f^*\|_{\mathcal{H}}^2 + \frac{\gamma(\lambda_N)}{N} \right).$$

While the derivation of this upper bound was non-rigorous, we believe that it is roughly accurate, and in comparison with the previous upper bound (8), it provides optimal rates of convergence.

4. Proofs of Theorem 1 and Related Results

We now turn to the proofs of Theorem 1 and Corollaries 3 through 5. 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

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

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

$$\begin{aligned}
\mathbb{E}\left[\|\bar{f} - f^*\|_2^2 \right] &= \frac{1}{m}\mathbb{E}\left[\|\widehat{f}_1 - \mathbb{E}[\widehat{f}_1]\|_2^2 \right] + \|\mathbb{E}[\widehat{f}_1] - f^*\|_2^2 \\
&\le \frac{1}{m}\mathbb{E}\left[\|\widehat{f}_1 - f^*\|_2^2 \right] + \|\mathbb{E}[\widehat{f}_1] - f^*\|_2^2,
\end{aligned} \tag{24}$$

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 (24) suggests our strategy: we upper 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 (5). In each lemma, C represents a universal (numerical) constant.

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

$$\|\mathbb{E}[\widehat{f}] - f^*\|_2^2 \le 8\lambda \|f^*\|_{\mathcal{H}}^2 + \frac{8\rho^4 \|f^*\|_{\mathcal{H}}^2 \operatorname{tr}(K)\beta_d}{\lambda} + \left(Cb(n,d,k)\frac{\rho^2\gamma(\lambda)}{\sqrt{n}} \right)^k \mu_0 \|f^*\|_{\mathcal{H}}^2. \tag{25}$$

Lemma 7 (Variance bound) *Under Assumptions A and B, for each $d = 1, 2, \ldots$, we have*

$$\mathbb{E}[\|\widehat{f} - f^*\|_2^2] \leq 12\lambda \|f^*\|_{\mathcal{H}}^2 + \frac{12\sigma^2\gamma(\lambda)}{n} + \left(\frac{2\sigma^2}{\lambda} + 4\|f^*\|_{\mathcal{H}}^2\right)\left(\mu_{d+1} + \frac{12\rho^4 \operatorname{tr}(K)\beta_d}{\lambda} + \left(Cb(n,d,k)\frac{\rho^2\gamma(\lambda)}{\sqrt{n}}\right)^k \|f^*\|_2^2\right). \quad (26)$$

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 (24) with Lemmas 6 and 7 yields the claim of Theorem 1.

Remarks: The proofs of Lemmas 6 and 7 are somewhat complex, but to the best of our knowledge, existing literature does not yield significantly simpler proofs. We now discuss this claim to better situate our technical contributions. Define the regularized population minimizer $f_\lambda^* := \operatorname{argmin}_{f\in\mathcal{H}}\{\mathbb{E}[(f(X)-Y)^2] + \lambda\|f\|_{\mathcal{H}}^2\}$. Expanding the decomposition (24) of the $L^2(\mathbb{P})$-risk into bias and variance terms, we obtain the further bound

$$\mathbb{E}\Big[\|\bar{f} - f^*\|_2^2\Big] \leq \|\mathbb{E}[\widehat{f}_1] - f^*\|_2^2 + \frac{1}{m}\mathbb{E}\left[\|\widehat{f}_1 - f^*\|_2^2\right]$$
$$= \underbrace{\|\mathbb{E}[\widehat{f}_1] - f^*\|_2^2}_{:=T_1} + \frac{1}{m}\Big(\underbrace{\|f_\lambda^* - f^*\|_2^2}_{:=T_2} + \underbrace{\mathbb{E}\left[\|\widehat{f}_1 - f^*\|_2^2\right] - \|f_\lambda^* - f^*\|_2^2}_{:=T_3}\Big) = T_1 + \frac{1}{m}(T_2 + T_3).$$

In this decomposition, T_1 and T_2 are bias and approximation error terms induced by the regularization parameter λ, while T_3 is an excess risk (variance) term incurred by minimizing the empirical loss.

This upper bound illustrates three trade-offs in our subsampled and averaged kernel regression procedure:

- The trade-off between T_2 and T_3: when the regularization parameter λ grows, the bias term T_2 increases while the variance term T_3 converges to zero.
- The trade-off between T_1 and T_3: when the regularization parameter λ grows, the bias term T_1 increases while the variance term T_3 converges to zero.
- The trade-off between T_1 and the computation time: when the number of machines m grows, the bias term T_1 increases (as the local sample size $n = N/m$ shrinks), while the computation time N^3/m^2 decreases.

Theoretical results in the KRR literature focus on the trade-off between T_2 and T_3, but in the current context, we also need an upper bound on the bias term T_1, which is not relevant for classical (centralized) analyses.

With this setting in mind, Lemma 6 tightly upper bounds the bias T_1 as a function of λ and n. An essential part of the proof is to characterize the properties of $\mathbb{E}[\widehat{f}_1]$, which is the expectation of a nonparametric empirical loss minimizer. We are not aware of existing literature on this problem, and the proof of Lemma 6 introduces novel techniques for this purpose.

On the other hand, Lemma 7 upper bounds $\mathbb{E}[\|\widehat{f}_1 - f^*\|_2^2]$ as a function of λ and n. Past work has focused on bounding a quantity of this form, but for technical reasons, most work (e.g. van de Geer, 2000; Mendelson, 2002b; Bartlett et al., 2002; Zhang, 2005) focuses on analyzing the constrained form

$$\widehat{f}_i := \operatorname*{argmin}_{\|f\|_{\mathcal{H}} \le C} \frac{1}{|S_i|} \sum_{(x,y)\in S_i} (f(x) - y)^2, \tag{27}$$

of kernel ridge regression. While this problem traces out the same set of solutions as that of the regularized kernel ridge regression estimator (3), it is non-trivial to determine a matched setting of λ for a given C. Zhang (2003) provides one of the few analyses of the regularized ridge regression estimator (3) (or (2)), providing an upper bound of the form $\mathbb{E}[\|\widehat{f} - f^*\|_2^2] = \mathcal{O}(\lambda + \frac{1/\lambda}{n})$, which is at best $\mathcal{O}(\frac{1}{\sqrt{n}})$. In contrast, Lemma 7 gives upper bound $\mathcal{O}(\lambda + \frac{\gamma(\lambda)}{n})$; the effective dimension $\gamma(\lambda)$ is often much smaller than $1/\lambda$, yielding a stronger convergence guarantee.

4.2 Proof of Corollary 3

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 such that we have $\log(d) \ge k$ and $d \ge N$. In the rest of the proof, our assignment to d will satisfy these inequalities. In this case, inspection of Theorem 1 shows that if m is small enough that

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

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

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

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

$$m \le \frac{\lambda^{\frac{2}{k-2}} N}{\gamma(\lambda)^{2\frac{k-1}{k-2}} \rho^{\frac{4k}{k-2}} \log^{\frac{k}{k-2}} d}, \tag{28}$$

then the term $T_3(d)$ of the bound (7) is $\mathcal{O}(\gamma(\lambda)/N)$.

Now we apply the bound (28) in the case in the corollary. Let us take $d = \max\{r, N\}$. Notice that $\beta_d = \beta_r = \mu_{r+1} = 0$. We find that $\gamma(\lambda) \le r$ since each of its terms is bounded by 1, and we take $\lambda = r/N$. Evaluating the expression (28) with this value, we arrive at

$$m \le \frac{N^{\frac{k-4}{k-2}}}{r^{2\frac{k-1}{k-2}} \rho^{\frac{4k}{k-2}} \log^{\frac{k}{k-2}} d}.$$

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 k can be chosen arbitrarily large), then we may take $k = \log N$, which implies that $N^{\frac{k-4}{k-2}} = \Omega(N)$, $r^{2\frac{k-1}{k-2}} = \mathcal{O}(r^2)$ and $\rho^{\frac{4k}{k-2}} = \mathcal{O}(\rho^4)$; and we replace $\log d$ with $\log N$. 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 4

We follow the program outlined in our remarks following Theorem 1. We must first choose λ on the order of $\gamma(\lambda)/N$. 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 (7) are thus negligible. Moreover, we have for any finite k that $\log d \gtrsim k$.

Applying the general bound (28) 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. Then taking $k = \log N$, the above bound becomes (to a multiplicative constant factor) $N^{\frac{2\nu-1}{2\nu+1}}/\rho^4 \log N$ as claimed.

4.4 Proof of Corollary 5

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 (7) are negligible.

Recalling our inequality (28), 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 3, we take $k = \log N$ in Theorem 1. By inspection, this yields the second statement of the corollary.

5. Proof of Theorem 2 and Related Results

In this section, we provide the proofs of Theorem 2, as well as the bound (13) based on the alternative form of the residual error. As in the previous section, we present a high-level proof, deferring more technical arguments to the appendices.

5.1 Proof of Theorem 2

We begin by stating and proving two auxiliary claims:

$$\mathbb{E}\left[(Y - f(X))^2\right] = \mathbb{E}\left[(Y - f^*(X))^2\right] + \|f - f^*\|_2^2 \quad \text{for any } f \in L^2(\mathbb{P}), \quad \text{and} \tag{29a}$$

$$f_{\bar{\lambda}}^* = \operatorname*{argmin}_{\|f\|_{\mathcal{H}} \leq R} \|f - f^*\|_2^2 . \tag{29b}$$

Let us begin by proving equality (29a). By adding and subtracting terms, we have

$$\begin{aligned} \mathbb{E}\left[(Y - f^*(X))^2\right] &= \mathbb{E}\left[(Y - f^*(X))^2\right] + \|f - f^*\|_2^2 + 2\mathbb{E}[(f(X) - f^*(X))\mathbb{E}[Y - f^*(X) \mid X]] \\ &\overset{(i)}{=} \mathbb{E}\left[(Y - f^*(X))^2\right] + \|f - f^*\|_2^2 , \end{aligned}$$

where equality (i) follows since the random variable $Y - f^*(X)$ is mean-zero given $X = x$.

For the second equality (29b), consider any function f in the RKHS that satisfies the bound $\|f\|_{\mathcal{H}} \leq R$. The definition of the minimizer $f_{\bar{\lambda}}^*$ guarantees that

$$\mathbb{E}\left[(f_{\bar{\lambda}}^*(X) - Y)^2\right] + \bar{\lambda} R^2 \leq \mathbb{E}[(f(X) - Y)^2] + \bar{\lambda}\, \|f\|_{\mathcal{H}}^2 \leq \mathbb{E}[(f(X) - Y)^2] + \bar{\lambda} R^2.$$

This result combined with equation (29a) establishes the equality (29b).

We now turn to the proof of the theorem. Applying Hölder's inequality yields that

$$\begin{aligned} \left\|\bar{f} - f^*\right\|_2^2 &\leq \left(1 + \frac{1}{q}\right) \left\|f_{\bar{\lambda}}^* - f^*\right\|_2^2 + (1+q) \left\|\bar{f} - f_{\bar{\lambda}}^*\right\|_2^2 \\ &= \left(1 + \frac{1}{q}\right) \inf_{\|f\|_{\mathcal{H}} \leq R} \|f - f^*\|_2^2 + (1+q) \left\|\bar{f} - f_{\bar{\lambda}}^*\right\|_2^2 \qquad \text{for all } q > 0, \end{aligned} \tag{30}$$

where the second step follows from equality (29b). It thus suffices to upper bound $\left\|\bar{f} - f_{\bar{\lambda}}^*\right\|_2^2$, and following the deduction of inequality (24), we immediately obtain the decomposition formula

$$\mathbb{E}\left[\left\|\bar{f} - f_{\bar{\lambda}}^*\right\|_2^2\right] \leq \frac{1}{m}\mathbb{E}[\|\widehat{f}_1 - f_{\bar{\lambda}}^*\|_2^2] + \|\mathbb{E}[\widehat{f}_1] - f_{\bar{\lambda}}^*\|_2^2, \tag{31}$$

where $\widehat{f}_1$ denotes the empirical minimizer for *one* of the subsampled datasets (i.e. the standard KRR solution on a sample of size $n = N/m$ with regularization λ). This suggests our strategy, which parallels our proof of Theorem 1: we upper bound $\mathbb{E}[\|\widehat{f}_1 - f^*_{\bar{\lambda}}\|_2^2]$ and $\|\mathbb{E}[\widehat{f}_1] - f^*_{\bar{\lambda}}\|_2^2$, respectively. In the rest of the proof, we let $\widehat{f} = \widehat{f}_1$ denote this solution.

Let the estimation error for a subsample be given by $\Delta = \widehat{f} - f^*_{\bar{\lambda}}$. Under Assumptions A and B$'$, we have the following two lemmas bounding expression (31), which parallel Lemmas 6 and 7 in the case when $f^* \in \mathcal{H}$. In each lemma, C denotes a universal constant.

Lemma 8 *For all $d = 1, 2, \ldots$, we have*

$$\mathbb{E}\left[\|\Delta\|_2^2\right] \le \frac{16(\bar{\lambda}-\lambda)^2 R^2}{\lambda} + \frac{8\gamma(\lambda)\rho^2\tau_{\bar{\lambda}}^2}{n} + \sqrt{32R^4 + 8\tau_{\bar{\lambda}}^4/\lambda^2}\left(\mu_{d+1} + \frac{16\rho^4 \operatorname{tr}(K)\beta_d}{\lambda} + \left(Cb(n,d,k)\frac{\rho^2\gamma(\lambda)}{\sqrt{n}}\right)^k\right). \quad (32)$$

Denoting the right hand side of inequality (32) by D^2, we have

Lemma 9 *For all $d = 1, 2, \ldots$, we have*

$$\|\mathbb{E}[\Delta]\|_2^2 \le \frac{4(\bar{\lambda}-\lambda)^2 R^2}{\lambda} + \frac{C\log^2(d)(\rho^2\gamma(\lambda))^2}{n} D^2 + \sqrt{32R^4 + 8\tau_{\bar{\lambda}}^4/\lambda^2}\left(\mu_{d+1} + \frac{4\rho^4 \operatorname{tr}(K)\beta_d}{\lambda}\right). \quad (33)$$

See Appendices C and D for the proofs of these two lemmas.

Given these two lemmas, we can now complete the proof of the theorem. If the conditions (10) hold, we have

$$\beta_d \le \frac{\lambda}{(R^2 + \tau_{\bar{\lambda}}^2/\lambda)N}, \qquad \mu_{d+1} \le \frac{1}{(R^2 + \tau_{\bar{\lambda}}^2/\lambda)N},$$
$$\frac{\log^2(d)(\rho^2\gamma(\lambda))^2}{n} \le \frac{1}{m} \quad \text{and} \quad \left(b(n,d,k)\frac{\rho^2\gamma(\lambda)}{\sqrt{n}}\right)^k \le \frac{1}{(R^2 + \tau_{\bar{\lambda}}^2/\lambda)N},$$

so there is a universal constant C' satisfying

$$\sqrt{32R^4 + 8\tau_{\bar{\lambda}}^4/\lambda^2}\left(\mu_{d+1} + \frac{16\rho^4 \operatorname{tr}(K)\beta_d}{\lambda} + \left(Cb(n,d,k)\frac{\rho^2\gamma(\lambda)}{\sqrt{n}}\right)^k\right) \le \frac{C'}{N}.$$

Consequently, Lemma 8 yields the upper bound

$$\mathbb{E}[\|\Delta\|_2^2] \le \frac{8(\bar{\lambda}-\lambda)^2 R^2}{\lambda} + \frac{8\gamma(\lambda)\rho^2\tau_{\bar{\lambda}}^2}{n} + \frac{C'}{N}.$$

Since $\log^2(d)(\rho^2\gamma(\lambda))^2/n \le 1/m$ by assumption, we obtain

$$\mathbb{E}\left[\|\bar{f} - f^*_{\bar{\lambda}}\|_2^2\right] \le \frac{C(\bar{\lambda}-\lambda)^2 R^2}{\lambda m} + \frac{C\gamma(\lambda)\rho^2\tau^2_{\bar{\lambda}}}{N} + \frac{C}{Nm} + \frac{4(\bar{\lambda}-\lambda)^2R^2}{\lambda} + \frac{C(\bar{\lambda}-\lambda)^2R^2}{\lambda m} + \frac{C\gamma(\lambda)\rho^2\tau^2_{\bar{\lambda}}}{N} + \frac{C}{Nm} + \frac{C}{N},$$

where C is a universal constant (whose value is allowed to change from line to line). Summing these bounds and using the condition that $\lambda \ge \bar{\lambda}$, we conclude that

$$\mathbb{E}\left[\|\bar{f} - f^*_{\bar{\lambda}}\|_2^2\right] \le \left(4 + \frac{C}{m}\right)(\lambda - \bar{\lambda})R^2 + \frac{C\gamma(\lambda)\rho^2\tau^2_{\bar{\lambda}}}{N} + \frac{C}{N}.$$

Combining this error bound with inequality (30) completes the proof.

5.2 Proof of Bound (13)

Using Theorem 2, it suffices to show that

$$\tau^4_{\bar{\lambda}} \le 8\,\mathrm{tr}(K)^2\|f^*_{\bar{\lambda}}\|^4_{\mathcal{H}}\rho^4 + 8\kappa^4. \tag{34}$$

By the tower property of expectations and Jensen's inequality, we have

$$\tau^4_{\bar{\lambda}} = \mathbb{E}[(\mathbb{E}[(f^*_{\bar{\lambda}}(x) - Y)^2 \mid X = x])^2] \le \mathbb{E}[(f^*_{\bar{\lambda}}(X) - Y)^4] \;\le\; 8\mathbb{E}[(f^*_{\bar{\lambda}}(X))^4] + 8\mathbb{E}[Y^4].$$

Since we have assumed that $\mathbb{E}[Y^4] \le \kappa^4$, the only remaining step is to upper bound $\mathbb{E}[(f^*_{\bar{\lambda}}(X))^4]$. Let $f^*_{\bar{\lambda}}$ have expansion $(\theta_1, \theta_2, \ldots)$ in the basis $\{\phi_j\}$. For any $x \in \mathcal{X}$, Hölder's inequality applied with the conjugates $4/3$ and 4 implies the upper bound

$$f^*_{\bar{\lambda}}(x) = \sum_{j=1}^{\infty}(\mu_j^{1/4}\theta_j^{1/2})\frac{\theta_j^{1/2}\phi_j(x)}{\mu_j^{1/4}} \le \left(\sum_{j=1}^{\infty}\mu_j^{1/3}\theta_j^{2/3}\right)^{3/4}\left(\sum_{j=1}^{\infty}\frac{\theta_j^2}{\mu_j}\phi_j^4(x)\right)^{1/4}. \tag{35}$$

Again applying Hölder's inequality—this time with conjugates $3/2$ and 3—to upper bound the first term in the product in inequality (35), we obtain

$$\sum_{j=1}^{\infty}\mu_j^{1/3}\theta_j^{2/3} = \sum_{j=1}^{\infty}\mu_j^{2/3}\left(\frac{\theta_j^2}{\mu_j}\right)^{1/3} \le \left(\sum_{j=1}^{\infty}\mu_j\right)^{2/3}\left(\sum_{j=1}^{\infty}\frac{\theta_j^2}{\mu_j}\right)^{1/3} = \mathrm{tr}(K)^{2/3}\|f^*_{\bar{\lambda}}\|^{2/3}_{\mathcal{H}}. \tag{36}$$

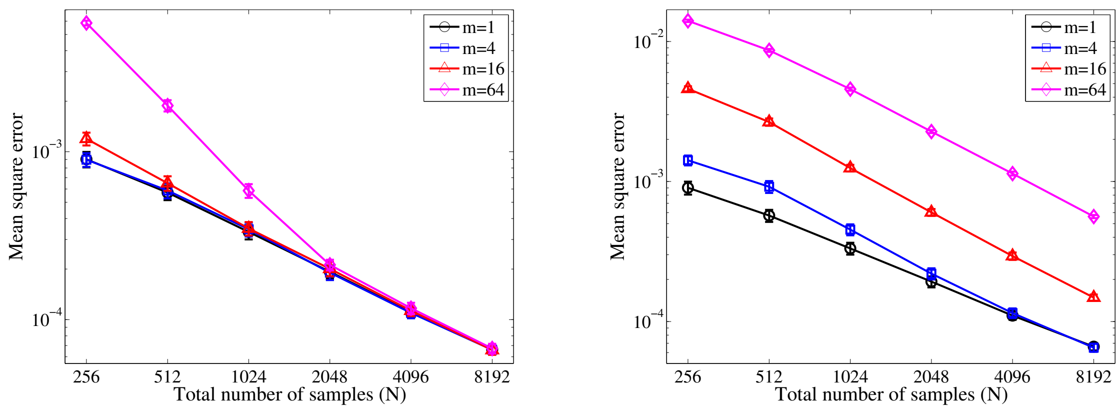
Combining inequalities (35) and (36), we find that

$$\mathbb{E}[(f^*_{\bar{\lambda}}(X))^4] \le \mathrm{tr}(K)^2\|f^*_{\bar{\lambda}}\|^2_{\mathcal{H}}\sum_{j=1}^{\infty}\frac{\theta_j^2}{\mu_j}\mathbb{E}[\phi_j^4(X)] \le \mathrm{tr}(K)^2\|f^*_{\bar{\lambda}}\|^4_{\mathcal{H}}\rho^4,$$

where we have used Assumption A. This completes the proof of inequality (34).

6. Experimental Results

In this section, we report the results of experiments on both simulated and real-world data designed to test the sharpness of our theoretical predictions.



(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 $\lambda \simeq N^{-2/3}$. (b) Analogous plots when each sub-problem is *not* under-regularized—that is, with $\lambda = n^{-2/3} = (N/m)^{-2/3}$ chosen as if there were only a single dataset of size n.

6.1 Simulation Studies

We begin by exploring 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 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

$$\widehat{f} = \sum_{i=1}^{N} \alpha_i K(x_i, \cdot), \quad \text{where} \quad \alpha = (K + \lambda N I)^{-1} y, \tag{37}$$

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 $N^{-2/3}$ (see e.g. Tsybakov (2009); Stone (1982); Caponnetto and De Vito (2007)).

By Corollary 4 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 (37), where the Gram matrix is constructed using the ith batch of samples (and n replaces N). We then compute $\bar{f} = (1/m) \sum_{i=1}^{m} \widehat{f}_i$.

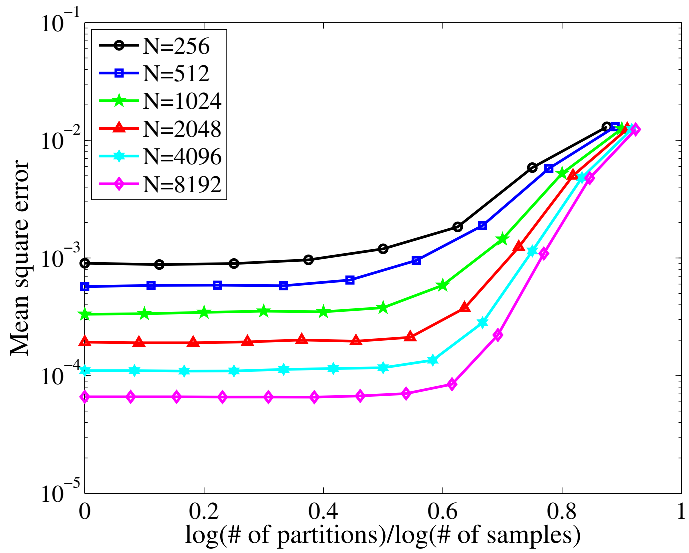


Figure 2: The mean-square error curves for fixed sample size but varied number of partitions. We are interested in the threshold of partitioning number m under which the optimal rate of convergence is achieved.

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 6.1(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 6.1(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 6.1(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 4 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-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 4, but it does confirm that the number of partitions m can scale polynomially with N while retaining minimax optimality.

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

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 (37) 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.

6.2 Real Data Experiments

We now turn to the results of experiments studying the performance of Fast-KRR on the task of predicting the year in which a song was released based on audio features associated with the song. We use the Million Song Dataset (Bertin-Mahieux et al., 2011), which consists of 463,715 training examples and a second set of 51,630 testing examples. Each example is a song (track) released between 1922 and 2011, and the song is represented as a vector of timbre information computed about the song. Each sample consists of the pair $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, where $x_i \in \mathbb{R}^d$ is a $d = 90$-dimensional vector and $y_i \in [1922, 2011]$ is the year in which the song was released. (For further details, see Bertin-Mahieux et al. (2011)).

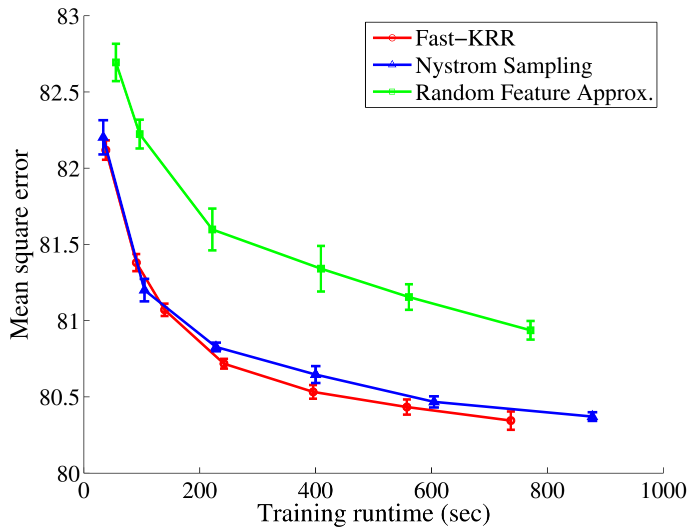


Figure 3: Results on year prediction on held-out test songs for Fast-KRR, Nyström sampling, and random feature approximation. Error bars indicate standard deviations over ten experiments.

Our experiments with this dataset use the Gaussian radial basis kernel

$$K(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{2\sigma^2}\right). \tag{38}$$

We normalize the feature vectors x so that the timbre signals have standard deviation 1, and select the bandwidth parameter $\sigma = 6$ via cross-validation. For regularization, we set $\lambda = N^{-1}$; since the Gaussian kernel has exponentially decaying eigenvalues (for typical distributions on X), Corollary 5 shows that this regularization achieves the optimal rate of convergence for the Hilbert space.

In Figure 3, we compare the time-accuracy curve of Fast-KRR with two approximation-based methods, plotting the mean-squared error between the predicted release year and the actual year on test songs. The first baseline is Nyström subsampling (Williams and Seeger, 2001), where the kernel matrix is approximated by a low-rank matrix of rank $r \in \{1, \ldots, 6\} \times 10^3$. The second baseline approach is an approximate form of kernel ridge regression using random features (Rahimi and Recht, 2007). The algorithm approximates the Gaussian kernel (38) by the inner product of two random feature vectors of dimensions $D \in \{2, 3, 5, 7, 8.5, 10\} \times 10^3$, and then solves the resulting linear regression problem. For the Fast-KRR algorithm, we use seven partitions $m \in \{32, 38, 48, 64, 96, 128, 256\}$ to test the algorithm. Each algorithm is executed 10 times to obtain standard deviations (plotted as error-bars in Figure 3).

As we see in Figure 3, for a fixed time budget, Fast-KRR enjoys the best performance, though the margin between Fast-KRR and Nyström sampling is not substantial. In spite of this close performance between Nyström sampling and the divide-and-conquer Fast-KRR

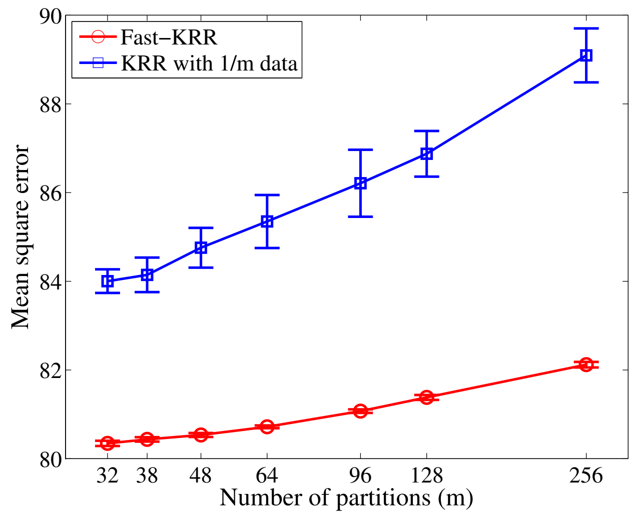


Figure 4: Comparison of the performance of Fast-KRR to a standard KRR estimator using a fraction $1/m$ of the data.

algorithm, it is worth noting that with parallel computation, it is trivial to accelerate Fast-KRR m times; parallelizing approximation-based methods appears to be a non-trivial task. Moreover, as our results in Section 3 indicate, Fast-KRR is minimax optimal in many regimes. At the same time the conference version of this paper was submitted, Bach (2013) published the first results we know of establishing convergence results in ℓ_2-error for Nyström sampling; see the discussion for more detail. We note in passing that standard linear regression with the original 90 features, while quite fast with runtime on the order of 1 second (ignoring data loading), has mean-squared-error 90.44, which is significantly worse than the kernel-based methods.

Our final experiment provides a sanity check: is the final averaging step in Fast-KRR even necessary? To this end, we compare Fast-KRR with standard KRR using a fraction $1/m$ of the data. For the latter approach, we employ the standard regularization $\lambda \approx (N/m)^{-1}$. As Figure 4 shows, Fast-KRR achieves much lower error rates than KRR using only a fraction of the data. Moreover, averaging stabilizes the estimators: the standard deviations of the performance of Fast-KRR are negligible compared to those for standard KRR.

7. Discussion

In this paper, we present results establishing that our decomposition-based algorithm for kernel ridge regression achieves minimax optimal convergence rates whenever the number of splits m of the data is not too large. The error guarantees of our method depend on the *effective dimensionality* $\gamma(\lambda) = \sum_{j=1}^{\infty} \mu_j/(\mu_j + \lambda)$ of the kernel. For any number of splits

$m \lesssim N/\gamma(\lambda)^2$, our method achieves estimation error decreasing as

$$\mathbb{E}\left[\|\bar{f} - f^*\|_2^2\right] \lesssim \lambda \|f^*\|_{\mathcal{H}}^2 + \frac{\sigma^2 \gamma(\lambda)}{N}.$$

(In particular, recall the bound (8) following Theorem 1). Notably, this convergence rate is minimax optimal, and we achieve substantial computational benefits from the subsampling schemes, in that computational cost scales (nearly) linearly in N.

It is also interesting to consider the number of kernel evaluations required to implement our method. Our estimator requires m sub-matrices of the full kernel (Gram) matrix, each of size $N/m \times N/m$. Since the method may use $m \asymp N/\gamma^2(\lambda)$ machines, in the best case, it requires at most $N\gamma^2(\lambda)$ kernel evaluations. By contrast, Bach (2013) shows that Nyström-based subsampling can be used to form an estimator within a constant factor of optimal as long as the number of N-dimensional subsampled columns of the kernel matrix scales roughly as the *marginal dimension* $\widetilde{\gamma}(\lambda) = N \left\|\mathrm{diag}(K(K + \lambda N I)^{-1})\right\|_\infty$. Consequently, using roughly $N\widetilde{\gamma}(\lambda)$ kernel evaluations, Nyström subsampling can achieve optimal convergence rates. These two scalings–namely, $N\gamma^2(\lambda)$ versus $N\widetilde{\gamma}(\lambda)$—are currently not comparable: in some situations, such as when the data is not compactly supported, $\widetilde{\gamma}(\lambda)$ can scale linearly with N, while in others it appears to scale roughly as the true effective dimensionality $\gamma(\lambda)$. A natural question arising from these lines of work is to understand the true optimal scaling for these different estimators: is one fundamentally better than the other? Are there natural computational tradeoffs that can be leveraged at large scale? As datasets grow substantially larger and more complex, these questions should become even more important, and we hope to continue to study them.

Acknowledgments

We thank Francis Bach for interesting and enlightening conversations on the connections between this work and his paper (Bach, 2013) and Yining Wang for pointing out a mistake in an earlier version of this manuscript. We also thank two reviewers for useful feedback and comments. JCD was partially supported by a National Defense Science and Engineering Graduate Fellowship (NDSEG) and a Facebook PhD fellowship. This work was partially supported by ONR MURI grant N00014-11-1-0688 to MJW.

Appendix A. Proof of Lemma 6

This appendix is devoted to the bias bound stated in Lemma 6. 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 $\|\mathbb{E}[\Delta]\|_2 \le \mathbb{E}[\|\mathbb{E}[\Delta \mid X]\|_2]$, 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

$$\frac{1}{n}\sum_{i=1}^n \left(\langle \xi_{x_i}, f\rangle_{\mathcal{H}} - y_i\right)^2 + \lambda \|f\|_{\mathcal{H}}^2 . \tag{39}$$

$\widehat{f}$	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 $M = \mathrm{diag}(\mu_1, \ldots, \mu_d)$
Q	Diagonal matrix with $Q = (I_{d\times d} + \lambda M^{-1})^{\frac{1}{2}}$
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 $\max\{\sqrt{\max\{k, \log(d)\}}, \max\{k, \log(d)\}/n^{1/2-1/k}\}$

Table 2: Notation used in proofs

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

$$\frac{1}{n}\sum_{i=1}^n \xi_{x_i}(\langle \xi_{x_i}, \widehat{f} - f^* \rangle_{\mathcal{H}} - \varepsilon_i) + \lambda \widehat{f} = \frac{1}{n}\sum_{i=1}^n \xi_{x_i}(\langle \xi_{x_i}, \widehat{f} \rangle_{\mathcal{H}} - y_i) + \lambda \widehat{f} = 0, \tag{40}$$

where the last equation uses the fact that $y_i = \langle \xi_{x_i}, f^* \rangle_{\mathcal{H}} + \varepsilon_i$. 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

$$(\widehat{\Sigma} + \lambda I)\mathbb{E}[\Delta \mid X] = -\lambda f^*. \tag{41}$$

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 $\{\phi_j\}_{j=1}^{\infty}$:

$$\mathbb{E}[\Delta \mid X] = \sum_{j=1}^{\infty} \delta_j \phi_j, \quad \text{with } \delta_j = \langle \mathbb{E}[\Delta \mid X], \phi_j \rangle_{L^2(\mathbb{P})}. \tag{42}$$

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

$$\|\mathbb{E}[\Delta \mid X]\|_2^2 = \|\delta\|_2^2 = \|\delta^{\downarrow}\|_2^2 + \|\delta^{\uparrow}\|_2^2. \tag{43}$$

We control each of the elements of the sum (43) in turn.

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

$$\|\delta^{\uparrow}\|_2^2 = \frac{\mu_{d+1}}{\mu_{d+1}} \sum_{j=d+1}^{\infty} \delta_j^2 \le \mu_{d+1} \sum_{j=d+1}^{\infty} \frac{\delta_j^2}{\mu_j} \overset{(i)}{\le} \mu_{d+1} \|\mathbb{E}[\Delta \mid X]\|_{\mathcal{H}}^2 \overset{(ii)}{\le} \mu_{d+1} \|f^*\|_{\mathcal{H}}^2, \tag{44}$$

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 $\|\mathbb{E}[\Delta \mid X]\|_{\mathcal{H}} \le \|f^*\|_{\mathcal{H}}$, which is a consequence of equality (41).

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

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

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 (41) 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

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

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

$$\left(\frac{1}{n} \Phi^T \Phi + \lambda M^{-1} \right) \delta^{\downarrow} = -\lambda M^{-1} \theta^{\downarrow} - \frac{1}{n} \Phi^T v. \tag{45}$$

We now show how the expression (45) gives us the desired bound in the lemma. By defining the shorthand matrix $Q = (I + \lambda M^{-1})^{1/2}$, we have

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

As a consequence, we can rewrite expression (45) to

$$\left(I + Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)Q^{-1}\right)Q\delta^{\downarrow} = -\lambda Q^{-1}M^{-1}\theta^{\downarrow} - \frac{1}{n}Q^{-1}\Phi^T v. \tag{46}$$

We now present a lemma bounding the terms in equality (46) to control $\delta^{\downarrow}$.

Lemma 10 *The following bounds hold:*

$$\left\|\lambda Q^{-1}M^{-1}\theta^{\downarrow}\right\|_2^2 \le \lambda \|f^*\|_{\mathcal{H}}^2, \quad \text{and} \tag{47a}$$

$$\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^T v\right\|_2^2\right] \le \frac{\rho^4 \|f^*\|_{\mathcal{H}}^2 \operatorname{tr}(K)\beta_d}{\lambda}. \tag{47b}$$

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

$$\mathbb{P}(\mathcal{E}^c) \le \left(\max\left\{\sqrt{k \vee \log(d)}, \frac{k \vee \log(d)}{n^{1/2-1/k}}\right\}\frac{C\rho^2\gamma(\lambda)}{\sqrt{n}}\right)^k. \tag{48}$$

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 $I + Q^{-1}((1/n)\Phi^T\Phi - I)Q^{-1} \succeq (1/2)I$. In particular, we have

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

on $\mathcal{E}$, by Eq. (46). Since $\|Q\delta^{\downarrow}\|_2^2 \ge \|\delta^{\downarrow}\|_2^2$, the above inequality implies that

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

Since $\mathcal{E}$ is X-measurable, we thus obtain

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

Applying the bounds (47a) and (47b), along with the elementary inequality $(a+b)^2 \le 2a^2 + 2b^2$, we have

$$\mathbb{E}\left[\|\delta^{\downarrow}\|_2^2\right] \le 8\lambda\|f^*\|_{\mathcal{H}}^2 + \frac{8\rho^4\|f^*\|_{\mathcal{H}}^2 \operatorname{tr}(K)\beta_d}{\lambda} + \mathbb{E}\left[1(\mathcal{E}^c)\,\|\delta^{\downarrow}\|_2^2\right]. \tag{49}$$

Now we use the fact that by the gradient optimality condition (41),

$$\|\mathbb{E}[\Delta \mid X]\|_2^2 \le \mu_0 \|\mathbb{E}[\Delta \mid X]\|_{\mathcal{H}}^2 \le \mu_0 \|f^*\|_{\mathcal{H}}^2$$

Recalling the shorthand (6) for $b(n,d,k)$, we apply the bound (48) to see

$$\mathbb{E}\left[1(\mathcal{E}^c) \|\delta^{\downarrow}\|_2^2\right] \le \mathbb{P}(\mathcal{E}^c)\mu_0 \|f^*\|_{\mathcal{H}}^2 \le \left(\frac{Cb(n,d,k)\rho^2\gamma(\lambda)}{\sqrt{n}}\right)^k \mu_0 \|f^*\|_{\mathcal{H}}^2$$

Combining this with the inequality (49), we obtain the desired statement of Lemma 6.

A.1 Proof of Lemma 10

Proof of bound (47a)*:* Beginning with the proof of the bound (47a), we have

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

Multiplying both sides by λ^2 gives the result.

Proof of bound (47b)*:* Next we turn to the proof of the bound (47b). We begin by re-writing $Q^{-1}\Phi^T v$ as the product of two components:

$$\frac{1}{n}Q^{-1}\Phi^T v = (M+\lambda I)^{-1/2}\left(\frac{1}{n}M^{1/2}\Phi^T v\right). \tag{50}$$

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

$$\left|\!\left|\!\left|(M+\lambda I)^{-1/2}\right|\!\right|\!\right| = \max_{j\in[d]} \frac{1}{\sqrt{\mu_j+\lambda}} \le \frac{1}{\sqrt{\lambda}}. \tag{51}$$

For the second factor in the product (50), 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,

$$\left\|M^{1/2}\Phi^T v\right\|_2^2 = \sum_{\ell=1}^d \mu_\ell(\Phi_\ell^T v)^2 \le \sum_{\ell=1}^d \mu_\ell \|\Phi_\ell\|_2^2 \|v\|_2^2, \tag{52}$$

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

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

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

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

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

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

by Cauchy-Schwarz. As a consequence, we find

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

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

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

Combining our bounds on $\|\Phi_\ell\|_2$ and $\|v\|_2$ with our initial bound (52), we obtain the inequality

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

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

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

Combining this inequality with our expansion (50) and the bound (51) yields the claim (47b).

Proof of bound (48)*:* We consider the expectation of the norm of $Q^{-1}(\frac{1}{n}\Phi^T\Phi - I)Q^{-1}$. For each $i \in [n]$, $\pi_i := (\phi_1(x_i), \ldots, \phi_d(x_i))^T \in \mathbb{R}^d$, then π_i^T is the i-th row of the matrix $\Phi \in \mathbb{R}^{n\times d}$. Then we know that

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

Define the sequence of matrices

$$A_i := Q^{-1}(\pi_i\pi_i^T - I)Q^{-1}$$

Then the matrices $A_i = A_i^T \in \mathbb{R}^{d\times d}$. 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 $k \ge 1$, we have

$$\mathbb{E}\left[\left|\left|\left|Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)Q^{-1}\right|\right|\right|^k\right] = \mathbb{E}\left[\left|\left|\left|\frac{1}{n}\sum_{i=1}^n A_i\right|\right|\right|^k\right] \le 2^k\mathbb{E}\left[\left|\left|\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i A_i\right|\right|\right|^k\right]. \tag{53}$$

Lemma 11 *The quantity* $\mathbb{E}\left[\left\|\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i A_i\right\|\right\|^k\right]^{1/k}$ *is upper bounded by*

$$\sqrt{e(k \vee 2\log(d))}\frac{\rho^2 \sum_{j=1}^d \frac{1}{1+\lambda/\mu_j}}{\sqrt{n}} + \frac{4e(k \vee 2\log(d))}{n^{1-1/k}}\left(\sum_{j=1}^d \frac{\rho^2}{1+\lambda/\mu_j}\right). \tag{54}$$

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 (53), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left\|Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)Q^{-1}\right\|\right\|^k\right] \\
&\leq 2^k\left(\sqrt{e(k \vee \log(d))}\frac{\rho^2\gamma(\lambda)}{\sqrt{n}} + \frac{4e(k \vee 2\log(d))}{n^{1-1/k}}\rho^2\gamma(\lambda)\right)^k \\
&\leq \max\left\{\sqrt{k \vee \log(d)}, \frac{k \vee \log(d)}{n^{1/2-1/k}}\right\}^k \left(\frac{C\rho^2\gamma(\lambda)}{\sqrt{n}}\right)^k, \qquad (55)
\end{aligned}$$

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

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

This completes the proof of the bound (48).

It remains to prove Lemma 11, for which we make use of the following result, due to Chen et al. (2012, Theorem A.1(2)).

Lemma 12 *Let $X_i \in \mathbb{R}^{d\times d}$ be independent symmetrically distributed Hermitian matrices. Then*

$$\mathbb{E}\left[\left\|\left\|\sum_{i=1}^n X_i\right\|\right\|^k\right]^{1/k} \leq \sqrt{e(k \vee 2\log d)}\left\|\left\|\sum_{i=1}^n \mathbb{E}[X_i^2]\right\|\right\|^{1/2} + 2e(k \vee 2\log d)\left(\mathbb{E}[\max_i \|\!|X_i|\!\|^k]\right)^{1/k}. \tag{56}$$

The proof of Lemma 11 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 (56).

We begin with the first term. Note that for any symmetric matrix Z, we have the matrix inequalities $0 \preceq \mathbb{E}[(Z - \mathbb{E}[Z])^2] = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 \preceq \mathbb{E}[Z^2]$, so

$$\mathbb{E}[A_i^2] = \mathbb{E}[Q^{-1}(\pi_i\pi_i^T - I)Q^{-2}(\pi_i\pi_i^T - I)Q^{-1}] \preceq \mathbb{E}[Q^{-1}\pi_i\pi_i^T Q^{-2}\pi_i\pi_i^T Q^{-1}].$$

Instead of computing these moments directly, we provide bounds on their norms. Since $\pi_i\pi_i^T$ is rank one and Q is diagonal, we have

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

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

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

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

$$\mathbb{E}\left[\left(\sum_{j=1}^{d} \frac{\phi_j(x_i)^2}{1+\lambda/\mu_j}\right)^k\right] \leq \left(\sum_{j=1}^{d} \frac{1}{1+\lambda/\mu_j}\right)^k \rho^{2k}. \tag{57}$$

The sub-multiplicativity of matrix norms implies $|||(Q^{-1}\pi_i\pi_i^T Q^{-1})^2||| \leq |||Q^{-1}\pi_i\pi_i^T Q^{-1}|||^2$, and consequently we have

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

where the final step follows from inequality (57). Combined with first term on the right-hand side of Lemma 12, we have thus obtained the first term on the right-hand side of expression (54).

We now turn to the second term in expression (54). 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

$$\|A_i\|^k \leq 2^{k-1}\left(\sum_{j=1}^{d} \frac{\phi_j(x_i)^2}{1+\lambda/\mu_j}\right)^k + 2^{k-1}|||Q^{-2}|||^k = 2^{k-1}\left(\sum_{j=1}^{d} \frac{\phi_j(x_i)^2}{1+\lambda/\mu_j}\right)^k + 2^{k-1}\left(\frac{1}{1+\lambda/\mu_1}\right)^k.$$

Since $\rho \geq 1$ (recall that the ϕ_j are an orthonormal basis), we apply inequality (57), to find that

$$\mathbb{E}[\max_i \|\varepsilon_i A_i/n\|^k] \leq \frac{1}{n^{k-1}}\left[2^{k-1}\left(\sum_{j=1}^{d} \frac{1}{1+\lambda/\mu_j}\right)^k \rho^{2k} + 2^{k-1}\left(\frac{1}{1+\lambda/\mu_1}\right)^k \rho^{2k}\right].$$

Taking kth roots yields the second term in the expression (54).

Appendix B. Proof of Lemma 7

This proof follows an outline similar to Lemma 6. We begin with a simple bound on $\|\Delta\|_{\mathcal{H}}$:

Lemma 13 *Under Assumption B, we have* $\mathbb{E}[\|\Delta\|_{\mathcal{H}}^2 \mid X] \le 2\sigma^2/\lambda + 4\|f^*\|_{\mathcal{H}}^2$.

Proof We have

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

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

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

which completes the proof. ■

With Lemma 13 in place, we now proceed to the proof of the theorem proper. Recall from Lemma 6 the optimality condition

$$\frac{1}{n}\sum_{i=1}^n \xi_{x_i}(\langle \xi_{x_i}, \widehat{f} - f^* \rangle - \varepsilon_i) + \lambda \widehat{f} = 0. \tag{58}$$

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 6), 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

$$\mathbb{E}[\|\Delta\|_2^2] = \mathbb{E}[\|\Delta^{\downarrow}\|_2^2] + \mathbb{E}[\|\Delta^{\uparrow}\|_2^2] = \mathbb{E}[\|\delta^{\downarrow}\|_2^2] + \mathbb{E}[\|\delta^{\uparrow}\|_2^2]. \tag{59}$$

We bound each of the terms (59) in turn.

By Lemma 13, the second term is upper bounded as

$$\mathbb{E}[\|\Delta^{\uparrow}\|_2^2] = \sum_{j>d} \mathbb{E}[\delta_j^2] \le \sum_{j>d} \frac{\mu_{d+1}}{\mu_j} \mathbb{E}[\delta_j^2] = \mu_{d+1}\mathbb{E}[\|\Delta^{\uparrow}\|_{\mathcal{H}}^2] \le \mu_{d+1}\left(\frac{2\sigma^2}{\lambda} + 4\|f^*\|_{\mathcal{H}}^2\right). \tag{60}$$

The remainder of the proof is devoted the bounding the term $\mathbb{E}[\|\Delta^{\downarrow}\|_2^2]$ in the decomposition (59). By taking the Hilbert inner product of ϕ_k with the optimality condition (58), we find as in our derivation of the matrix equation (45) 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

$$\left(\frac{1}{n}\Phi^T\Phi + \lambda M^{-1}\right)\delta^{\downarrow} = -\lambda M^{-1}\theta^{\downarrow} - \frac{1}{n}\Phi^T v + \frac{1}{n}\Phi^T\varepsilon. \tag{61}$$

As in the proof of Lemma 6, we find that

$$\left(I + Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)Q^{-1}\right)Q\delta^{\downarrow} = -\lambda Q^{-1}M^{-1}\theta^{\downarrow} - \frac{1}{n}Q^{-1}\Phi^T v + \frac{1}{n}Q^{-1}\Phi^T\varepsilon, \tag{62}$$

where we recall that $Q = (I + \lambda M^{-1})^{1/2}$.

We now recall the bounds (47a) and (48) from Lemma 10, as well as the previously defined event $\mathcal{E} := \{|\!|\!| Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)Q^{-1} |\!|\!| \le 1/2\}$. When $\mathcal{E}$ occurs, the expression (62) implies the inequality

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

When $\mathcal{E}$ fails to hold, Lemma 13 may still be applied since $\mathcal{E}$ is measurable with respect to $\{x_i\}_{i=1}^n$. Doing so yields

$$\begin{aligned}
&\mathbb{E}[\|\Delta^{\downarrow}\|_2^2] = \mathbb{E}[1(\mathcal{E})\,\|\Delta^{\downarrow}\|_2^2] + \mathbb{E}[1(\mathcal{E}^c)\,\|\Delta^{\downarrow}\|_2^2] \\
&\le 4\mathbb{E}\left[\left\|-\lambda Q^{-1}M^{-1}\theta^{\downarrow} - (1/n)Q^{-1}\Phi^T v + (1/n)Q^{-1}\Phi^T\varepsilon\right\|_2^2\right] + \mathbb{E}\left[1(\mathcal{E}^c)\,\mathbb{E}[\|\Delta^{\downarrow}\|_2^2 \mid \{x_i\}_{i=1}^n]\right] \\
&\le 4\mathbb{E}\left[\left\|\lambda Q^{-1}M^{-1}\theta^{\downarrow} + \frac{1}{n}Q^{-1}\Phi^T v - \frac{1}{n}Q^{-1}\Phi^T\varepsilon\right\|_2^2\right] + \mathbb{P}(\mathcal{E}^c)\left(\frac{2\sigma^2}{\lambda} + 4\,\|f^*\|_{\mathcal{H}}^2\right).
\end{aligned} \tag{63}$$

Since the bound (48) still holds, it remains to provide a bound on the first term in the expression (63).

As in the proof of Lemma 6, we have $\|\lambda Q^{-1}M^{-1}\theta^{\downarrow}\|_2^2 \le \lambda\,\|f^*\|_{\mathcal{H}}^2$ via the bound (47a). Turning to the second term inside the norm, we claim that, under the conditions of Lemma 7, the following bound holds:

$$\mathbb{E}\left[\left\|(1/n)Q^{-1}\Phi^T v\right\|_2^2\right] \le \frac{\rho^4 \operatorname{tr}(K)\beta_d(2\sigma^2/\lambda + 4\,\|f^*\|_{\mathcal{H}}^2)}{\lambda}. \tag{64}$$

This claim is an analogue of our earlier bound (47b), 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 $1/\sqrt{1 + \lambda/\mu_j}$, we have

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

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

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

The last sum is bounded by $(\sigma^2/n)\gamma(\lambda)$. Applying the inequality $(a+b+c)^2 \le 3a^2+3b^2+3c^2$ to inequality (63), we obtain

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

Applying the bound (48) to control $\mathbb{P}(\mathcal{E}^c)$ and bounding $\mathbb{E}[\|\Delta^{\uparrow}\|_2^2]$ using inequality (60) completes the proof of the lemma.

It remains to prove bound (64). Recalling the inequality (51), we see that

$$\left\|(1/n)Q^{-1}\Phi^T v\right\|_2^2 \le \left\|Q^{-1}M^{-1/2}\right\|^2 \left\|(1/n)M^{1/2}\Phi^T v\right\|_2^2 \le \frac{1}{\lambda}\left\|(1/n)M^{1/2}\Phi^T v\right\|_2^2. \tag{65}$$

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

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

Now consider the inner expectation. Applying the Cauchy-Schwarz inequality as in the proof of the bound (47b), we have

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

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

$$\mathbb{E}\left[\left\|M^{1/2}\Phi^T v\right\|_2^2\right] \le \sum_{i=1}^n \sum_{l=1}^d \mu_\ell \mathbb{E}\left[\|\Phi_\ell\|_2^2 \left(\sum_{j>d} \mu_j \phi_j^2(X_i)\right) \mathbb{E}[\|\Delta\|_{\mathcal{H}}^2 \mid X]\right]. \tag{66}$$

Lemma 13 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 (47b). 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

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

Dividing by λn^2 completes the proof.

Appendix C. Proof of Lemma 8

As before, we let $\{x_i\}_{i=1}^n := \{x_1, \ldots, x_n\}$ denote the collection of design points. We begin with some useful bounds on $\|f_{\bar{\lambda}}^*\|_{\mathcal{H}}$ and $\|\Delta\|_{\mathcal{H}}$.

Lemma 14 *Under Assumptions A and B′, we have*

$$\mathbb{E}\left[(\mathbb{E}[\|\Delta\|_{\mathcal{H}}^2 \mid \{x_i\}_{i=1}^n])^2\right] \le B_{\lambda,\bar{\lambda}}^4 \quad \text{and} \quad \mathbb{E}[\|\Delta\|_{\mathcal{H}}^2] \le B_{\lambda,\bar{\lambda}}^2, \tag{67}$$

where

$$B_{\lambda,\bar{\lambda}} := \sqrt[4]{32\|f_{\bar{\lambda}}^*\|_{\mathcal{H}}^4 + 8\tau_{\bar{\lambda}}^4/\lambda^2}. \tag{68}$$

See Section C.1 for the proof of this claim.

This proof follows an outline similar to that of Lemma 7. As usual, we 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 we choose $d \in \mathbb{N}$ and define the truncated vectors $\Delta^\downarrow := \sum_{j=1}^d \delta_j \phi_j$ and $\Delta^\uparrow := \Delta - \Delta^\downarrow = \sum_{j>d} \delta_j \phi_j$. As usual, we have the decomposition $\mathbb{E}[\|\Delta\|_2^2] = \mathbb{E}[\|\delta^\downarrow\|_2^2] + \mathbb{E}[\|\delta^\uparrow\|_2^2]$. Recall the definition (68) of the constant $B_{\lambda,\bar{\lambda}} = \sqrt[4]{32\|f_{\bar{\lambda}}^*\|_{\mathcal{H}}^4 + 8\tau_{\bar{\lambda}}^4/\lambda^2}$. As in our deduction of inequalities (60), Lemma 14 implies that $\mathbb{E}[\|\Delta^\uparrow\|_2^2] \le \mu_{d+1}\mathbb{E}[\|\Delta^\uparrow\|_{\mathcal{H}}^2] \le \mu_{d+1} B_{\lambda,\bar{\lambda}}^2$.

The remainder of the proof is devoted to bounding $\mathbb{E}[\|\delta^\downarrow\|_2^2]$. We use identical notation to that in our proof of Lemma 7, which we recap for reference (see also Table 2). We define the tail error vector $v \in \mathbb{R}^n$ by $v_i = \sum_{j>d} \delta_j \phi_j(x_i)$, $i \in [n]$, and recall the definitions of the eigenvalue matrix $M = \operatorname{diag}(\mu_1, \ldots, \mu_d) \in \mathbb{R}^{d\times d}$ and basis matrix Φ with $\Phi_{ij} = \phi_j(x_i)$. We use $Q = (I + \lambda M^{-1})^{1/2}$ for shorthand, and we let $\mathcal{E}$ be the event that

$$|||Q^{-1}((1/n)\Phi^T\Phi - I)Q^{-1}||| \le 1/2.$$

Writing $f_{\bar{\lambda}}^* = \sum_{j=1}^\infty \theta_j \phi_j$, we define the alternate noise vector $\varepsilon_i' = Y_i - f_{\bar{\lambda}}^*(x_i)$. Using this notation, mirroring the proof of Lemma 7 yields

$$\mathbb{E}[\|\Delta^\downarrow\|_2^2] \le \mathbb{E}[\|Q\delta^\downarrow\|_2^2] \le 4\mathbb{E}\left[\left\|\lambda Q^{-1}M^{-1}\theta^\downarrow + \frac{1}{n}Q^{-1}\Phi^T v - \frac{1}{n}Q^{-1}\Phi^T\varepsilon'\right\|_2^2\right] + \mathbb{P}(\mathcal{E}^c)B_{\lambda,\bar{\lambda}}^2, \tag{69}$$

which is an analogue of equation (63). The bound bound (48) controls the probability $\mathbb{P}(\mathcal{E}^c)$, so it remains to control the first term in the expression (69). We first rewrite the expression within the norm as

$$(\lambda - \bar{\lambda})Q^{-1}M^{-1}\theta^\downarrow + \frac{1}{n}Q^{-1}\Phi^T v - \left(\frac{1}{n}Q^{-1}\Phi^T\varepsilon' - \bar{\lambda}Q^{-1}M^{-1}\theta^\downarrow\right)$$

The following lemma provides bounds on the first two terms:

Lemma 15 *The following bounds hold:*

$$\left\|(\bar{\lambda} - \lambda)Q^{-1}M^{-1}\theta^\downarrow\right\|_2^2 \le \frac{(\bar{\lambda} - \lambda)^2 \|f_{\bar{\lambda}}^*\|_{\mathcal{H}}^2}{\lambda}, \tag{70a}$$

$$\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^T v\right\|_2^2\right] \le \frac{\rho^4 B_{\lambda,\bar{\lambda}}^2 \operatorname{tr}(K)\beta_d}{\lambda}, \tag{70b}$$

For the third term, we make the following claim.

Lemma 16 *Under Assumptions A and B′, we have*

$$\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^T\varepsilon' - \bar{\lambda}Q^{-1}M^{-1}\theta^{\downarrow}\right\|_2^2\right] \le \frac{\gamma(\lambda)\rho^2\tau_{\bar{\lambda}}^2}{n}. \tag{71}$$

Deferring the proof of the two lemmas to Sections C.2 and C.3, we apply the inequality $(a+b+c)^2 \le 4a^2 + 4b^2 + 2c^2$ to inequality (69), and we have

$$\begin{aligned}
&\mathbb{E}[\|\Delta^{\downarrow}\|_2^2] - \mathbb{P}(\mathcal{E}^c)B_{\lambda,\bar{\lambda}}^2 \le \mathbb{E}[\|Q\delta^{\downarrow}\|_2^2] - \mathbb{P}(\mathcal{E}^c)B_{\lambda,\bar{\lambda}}^2 \\
&\le 16\mathbb{E}\left[\left\|(\lambda-\bar{\lambda})Q^{-1}M^{-1}\theta^{\downarrow}\right\|_2^2\right] + \frac{16}{n^2}\mathbb{E}\left[\|Q^{-1}\Phi^T v\|_2^2\right] + \frac{8}{n^2}\mathbb{E}\left[\left\|Q^{-1}\Phi^T\varepsilon' - \bar{\lambda}Q^{-1}M^{-1}\theta^{\downarrow}\right\|_2^2\right] \\
&\le \frac{16(\bar{\lambda}-\lambda)^2\left\|f_{\bar{\lambda}}^*\right\|_{\mathcal{H}}^2}{\lambda} + \frac{16\rho^4 B_{\lambda,\bar{\lambda}}^2\operatorname{tr}(K)\beta_d}{\lambda} + \frac{8\gamma(\lambda)\rho^2\tau_{\bar{\lambda}}^2}{n},
\end{aligned} \tag{72}$$

where we have applied the bounds (70a) and (70b) from Lemma 17 and the bound (71) from Lemma 16. Applying the bound (48) to control $\mathbb{P}(\mathcal{E}^c)$ and recalling that $\mathbb{E}[\|\Delta^{\uparrow}\|_2^2] \le \mu_{d+1}B_{\lambda,\bar{\lambda}}^2$ completes the proof.

C.1 Proof of Lemma 14

Recall that $\widehat{f}$ minimizes the empirical objective. Consequently,

$$\begin{aligned}
\lambda\mathbb{E}[\|\widehat{f}\|_{\mathcal{H}}^2 \mid \{x_i\}_{i=1}^n] &\le \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(\widehat{f}(x_i) - Y_i)^2 + \lambda\|\widehat{f}\|_{\mathcal{H}}^2 \mid \{x_i\}_{i=1}^n\right] \\
&\le \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(f_{\bar{\lambda}}^*(x_i) - Y_i)^2 \mid x_i] + \lambda\|f_{\bar{\lambda}}^*\|_{\mathcal{H}}^2 = \frac{1}{n}\sum_{i=1}^n\sigma_{\bar{\lambda}}^2(x_i) + \lambda\|f_{\bar{\lambda}}^*\|_{\mathcal{H}}^2
\end{aligned}$$

The triangle inequality immediately gives us the upper bound

$$\mathbb{E}[\|\Delta\|_{\mathcal{H}}^2 \mid \{x_i\}_{i=1}^n] \le 2\|f_{\bar{\lambda}}^*\|_{\mathcal{H}}^2 + \mathbb{E}[2\|\widehat{f}\|_{\mathcal{H}}^2 \mid \{x_i\}_{i=1}^n] \le \frac{2}{\lambda n}\sum_{i=1}^n\sigma_{\bar{\lambda}}^2(x_i) + 4\|f_{\bar{\lambda}}^*\|_{\mathcal{H}}^2.$$

Since $(a+b)^2 \le 2a^2 + 2b^2$, convexity yields

$$\begin{aligned}
\mathbb{E}[(\mathbb{E}[\|\Delta\|_{\mathcal{H}}^2 \mid \{x_i\}_{i=1}^n])^2] &\le \mathbb{E}\left[\left(\frac{2}{\lambda n}\sum_{i=1}^n\sigma_{\bar{\lambda}}^2(X_i) + 4\|f_{\bar{\lambda}}^*\|_{\mathcal{H}}^2\right)^2\right] \\
&\le \frac{8}{\lambda^2 n}\sum_{i=1}^n\mathbb{E}[\sigma_{\bar{\lambda}}^4(X_i)] + 32\|f_{\bar{\lambda}}^*\|_{\mathcal{H}}^4 = 32\|f_{\bar{\lambda}}^*\|_{\mathcal{H}}^4 + \frac{8\tau_{\bar{\lambda}}^4}{\lambda^2}.
\end{aligned}$$

This completes the proof of the first of the inequalities (67). The second of the inequalities (67) follows from the first by Jensen's inequality.

C.2 Proof of Lemma 15

Our previous bound (47a) immediately implies inequality (70a). To prove the second upper bound, we follow the proof of the bound (64). From inequalities (65) and (66), we obtain that

$$\left\|(1/n)Q^{-1}\Phi^T v\right\|_2^2 \le \frac{1}{\lambda n^2}\sum_{i=1}^n\sum_{l=1}^d\sum_{j>d}\mu_\ell\mu_j\mathbb{E}\left[\|\Phi_\ell\|_2^2\,\phi_j^2(X_i)\mathbb{E}[\|\Delta\|_{\mathcal{H}}^2 \mid \{X_i\}_{i=1}^n]\right]. \tag{73}$$

Applying Hölder's inequality yields

$$\mathbb{E}\left[\|\Phi_\ell\|_2^2\,\phi_j^2(X_i)\mathbb{E}[\|\Delta\|_{\mathcal{H}}^2 \mid \{X_i\}_{i=1}^n]\right] \le \sqrt{\mathbb{E}[\|\Phi_\ell\|_2^4\,\phi_j^4(X_i)]}\sqrt{\mathbb{E}[(\mathbb{E}[\|\Delta\|_{\mathcal{H}}^2 \mid \{X_i\}_{i=1}^n])^2]}.$$

Note that Lemma 14 provides the bound $B_{\lambda,\bar{\lambda}}^4$ on the final expectation. By definition of Φ_ℓ, we find that

$$\mathbb{E}[\|\Phi_\ell\|_2^4\,\phi_j^4(x_i)] = \mathbb{E}\left[\left(\sum_{k=1}^n\phi_\ell^2(x_k)\right)^2\phi_j^4(x_i)\right] \le n^2\mathbb{E}\left[\frac{1}{2}\left(\phi_\ell^8(x_1)+\phi_j^8(x_1)\right)\right] \le n^2\rho^8,$$

where we have used Assumption A with moment $2k \ge 8$, or equivalently $k \ge 4$. Thus

$$\mathbb{E}\left[\|\Phi_\ell\|_2^2\,\phi_j^2(X_i)\mathbb{E}[\|\Delta\|_{\mathcal{H}}^2 \mid \{X_i\}_{i=1}^n]\right] \le n\rho^4 B_{\lambda,\bar{\lambda}}^2. \tag{74}$$

Combining inequalities (73) and (74) yields the bound (70b).

C.3 Proof of Lemma 16

Using the fact that Q and M are diagonal, we have

$$\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^T\varepsilon' - \bar{\lambda}Q^{-1}M^{-1}\theta^{\downarrow}\right\|_2^2\right] = \sum_{j=1}^d Q_{jj}^{-2}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n\phi_j(X_i)\varepsilon_i' - \frac{\bar{\lambda}\theta_j}{\mu_j}\right)^2\right]. \tag{75}$$

Fréchet differentiability and the fact that $f_{\bar{\lambda}}^*$ is the global minimizer of the regularized regression problem imply that

$$\mathbb{E}[\xi_{X_i}\varepsilon_i'] + \bar{\lambda}f_{\bar{\lambda}}^* = \mathbb{E}\left[\xi_X\left(\langle\xi_X, f_{\bar{\lambda}}^*\rangle - y\right)\right] + \bar{\lambda}f_{\bar{\lambda}}^* = 0.$$

Taking the (Hilbert) inner product of the preceding display with the basis function ϕ_j, we get

$$\mathbb{E}\left[\phi_j(X_i)\varepsilon_i' - \frac{\bar{\lambda}\theta_j}{\mu_j}\right] = 0. \tag{76}$$

Combining the equalities (75) and (76) and using the i.i.d. nature of $\{x_i\}_{i=1}^n$ leads to

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^T\varepsilon' - \bar{\lambda}Q^{-1}M^{-1}\theta^{\downarrow}\right\|_2^2\right] &= \sum_{j=1}^d Q_{jj}^{-2}\operatorname{var}\left(\frac{1}{n}\sum_{i=1}^n\phi_j(X_i)\varepsilon_i' - \frac{\bar{\lambda}\theta_j}{\mu_j}\right)\\ &= \frac{1}{n}\sum_{j=1}^d Q_{jj}^{-2}\operatorname{var}\left(\phi_j(X_1)\varepsilon_1'\right).\end{aligned} \tag{77}$$

Using the elementary inequality $\mathrm{var}(Z) \leq \mathbb{E}[Z^2]$ for any random variable Z, we have from Hölder's inequality that

$$\mathrm{var}(\phi_j(X_1)\varepsilon_1') \leq \mathbb{E}[\phi_j(X_1)^2(\varepsilon_1')^2] \leq \sqrt{\mathbb{E}[\phi_j(X_1)^4]\mathbb{E}[\sigma_{\bar{\lambda}}^4(X_1)]} \leq \sqrt{\rho^4}\sqrt{\tau_{\bar{\lambda}}^4},$$

where we used Assumption B′ to upper bound the fourth moment $\mathbb{E}[\sigma_{\bar{\lambda}}^4(X_1)]$. Using the fact that $Q_{jj}^{-1} \leq 1$, we obtain the following upper bound on the quantity (77):

$$\frac{1}{n}\sum_{j=1}^{d} Q_{jj}^{-2}\,\mathrm{var}(\phi_j(X_1)\varepsilon_1') = \frac{1}{n}\sum_{j=1}^{d}\frac{\mathrm{var}(\phi_j(X_1)\varepsilon_1')}{1+\lambda/\mu_j} \leq \frac{\gamma(\lambda)\rho^2\tau_{\bar{\lambda}}^2}{n},$$

which establishes the claim.

Appendix D. Proof of Lemma 9

At a high-level, the proof is similar to that of Lemma 6, but we take care since the errors $f_{\bar{\lambda}}^*(x) - y$ are not conditionally mean-zero (or of conditionally bounded variance). Recalling our notation of ξ_x as the RKHS evaluator for x, we have by assumption that $\widehat{f}$ minimizes the empirical objective (39). As in our derivation of equality (40), the Fréchet differentiability of this objective implies the first-order optimality condition

$$\frac{1}{n}\sum_{i=1}^{n}\xi_{x_i}\langle\xi_{x_i},\Delta\rangle + \frac{1}{n}\sum_{i=1}^{n}(\xi_{x_i}\langle\xi_{x_i},f_{\bar{\lambda}}^*\rangle - y_i) + \lambda\Delta + \lambda f_{\bar{\lambda}}^* = 0, \tag{78}$$

where $\Delta := \widehat{f} - f_{\bar{\lambda}}^*$. In addition, the optimality of $f_{\bar{\lambda}}^*$ implies that $\mathbb{E}[\xi_{x_i}(\langle\xi_{x_i},f_{\bar{\lambda}}^*\rangle - y_i)] + \bar{\lambda}f_{\bar{\lambda}}^* = 0$. Using this in equality (78), we take expectations with respect to $\{x_i, y_i\}$ to obtain

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_{X_i}\langle\xi_{X_i},\Delta\rangle + \lambda\Delta\right] + (\lambda - \bar{\lambda})f_{\bar{\lambda}}^* = 0.$$

Recalling the definition of the sample covariance operator $\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n}\xi_{x_i}\otimes\xi_{x_i}$, we arrive at

$$\mathbb{E}[(\widehat{\Sigma} + \lambda I)\Delta] = (\bar{\lambda} - \lambda)f_{\bar{\lambda}}^*, \tag{79}$$

which is the analogue of our earlier equality (41).

We now proceed via a truncation argument similar to that used in our proofs of Lemmas 6 and 7. Let $\delta \in \ell_2(\mathbb{N})$ be the expansion of the error Δ in the basis $\{\phi_j\}$, so that $\Delta = \sum_{j=1}^{\infty}\delta_j\phi_j$. For a fixed (arbitrary) $d \in \mathbb{N}$, define

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

and note that $\|\mathbb{E}[\Delta]\|_2^2 = \|\mathbb{E}[\Delta^{\downarrow}]\|_2^2 + \|\mathbb{E}[\Delta^{\uparrow}]\|_2^2$. By Lemma 14, the second term is controlled by

$$\|E[\Delta^{\uparrow}]\|_2^2 \leq \mathbb{E}[\|\Delta^{\uparrow}\|_2^2] = \sum_{j>d}\mathbb{E}[\delta_j^2] \leq \sum_{j>d}\frac{\mu_{d+1}}{\mu_j}\mathbb{E}[\delta_j^2] = \mu_{d+1}\mathbb{E}[\|\Delta^{\uparrow}\|_{\mathcal{H}}^2] \leq \mu_{d+1}B_{\lambda,\bar{\lambda}}^2. \tag{80}$$

The remainder of the proof is devoted to bounding $\|\mathbb{E}[\Delta^{\downarrow}]\|_2^2$. Let $f_{\bar{\lambda}}^*$ have the expansion $(\theta_1, \theta_2, \ldots)$ in the basis $\{\phi_j\}$. Recall (as in Lemmas 6 and 7) the definition of the matrix $\Phi \in \mathbb{R}^{n \times d}$ by its coordinates $\Phi_{ij} = \phi_j(x_i)$, the diagonal matrix $M = \operatorname{diag}(\mu_1, \ldots, \mu_d) \succ 0 \in \mathbb{R}^{d \times d}$, and the tail error vector $v \in \mathbb{R}^n$ by $v_i = \sum_{j>d} \delta_j \phi_j(x_i) = \Delta^{\uparrow}(x_i)$. Proceeding precisely as in the derivations of equalities (45) and (61), we have the following equality:

$$\mathbb{E}\left[\left(\frac{1}{n}\Phi^T\Phi + \lambda M^{-1}\right)\delta^{\downarrow}\right] = (\bar{\lambda} - \lambda)M^{-1}\theta^{\downarrow} - \mathbb{E}\left[\frac{1}{n}\Phi^T v\right]. \tag{81}$$

Recalling the definition of the shorthand matrix $Q = (I + \lambda M^{-1})^{1/2}$, with some algebra we have

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

so we can expand expression (81) as

$$\begin{aligned}\mathbb{E}\left[Q\delta^{\downarrow} + Q^{-1}\left(\frac{1}{n}\Phi\Phi^T - I\right)\delta^{\downarrow}\right] &= \mathbb{E}\left[Q^{-1}\left(\frac{1}{n}\Phi^T\Phi + \lambda M^{-1}\right)\delta^{\downarrow}\right] \\ &= (\bar{\lambda} - \lambda)Q^{-1}M^{-1}\theta^{\downarrow} - \mathbb{E}\left[\frac{1}{n}Q^{-1}\Phi^T v\right],\end{aligned}$$

or, rewriting,

$$\mathbb{E}[Q\delta^{\downarrow}] = (\bar{\lambda} - \lambda)Q^{-1}M^{-1}\theta^{\downarrow} - \mathbb{E}\left[\frac{1}{n}Q^{-1}\Phi^T v\right] - \mathbb{E}\left[Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)\delta^{\downarrow}\right]. \tag{82}$$

Lemma 15 provides bounds on the first two terms on the right-hand-side of equation (82). The following lemma provides upper bounds on the third term:

Lemma 17 *There exists a universal constant C such that*

$$\left\|\mathbb{E}\left[Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)\delta^{\downarrow}\right]\right\|_2^2 \le \frac{C(\rho^2\gamma(\lambda)\log d)^2}{n}\mathbb{E}\left[\|Q\delta^{\downarrow}\|_2^2\right], \tag{83}$$

We defer the proof to Section D.1.

Applying Lemma 15 and Lemma 17 to equality (82) and using the standard inequality $(a+b+c)^2 \le 4a^2 + 4b^2 + 2c^2$, we obtain the upper bound

$$\left\|\mathbb{E}[\Delta^{\downarrow}]\right\|_2^2 \le \frac{4(\bar{\lambda} - \lambda)^2 \left\|f_{\bar{\lambda}}^*\right\|_{\mathcal{H}}^2}{\lambda} + \frac{4\rho^4 B_{\lambda,\bar{\lambda}}^2 \operatorname{tr}(K)\beta_d}{\lambda} + \frac{C(\rho^2\gamma(\lambda)\log d)^2}{n}\mathbb{E}\left[\|Q\delta^{\downarrow}\|_2^2\right]$$

for a universal constant C. Note that inequality (72) provides a sufficiently tight bound on the term $\mathbb{E}\left[\|Q\delta^{\downarrow}\|_2^2\right]$. Combined with inequality (80), this completes the proof of Lemma 9.

D.1 Proof of Lemma 17

By using Jensen's inequality and then applying Cauchy-Schwarz, we find

$$\left\| \mathbb{E}\left[Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)\delta^{\downarrow}\right]\right\|_2^2 \le \left(\mathbb{E}\left[\left\| Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)\delta^{\downarrow}\right\|_2\right]\right)^2$$
$$\le \mathbb{E}\left[\left|\!\left|\!\left| Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)Q^{-1}\right|\!\right|\!\right|^2\right]\mathbb{E}\left[\|Q\delta^{\downarrow}\|_2^2\right].$$

The first component of the final product can be controlled by the matrix moment bound established in the proof of inequality (48). In particular, applying (55) with $k = 2$ yields a universal constant C such that

$$\mathbb{E}\left[\left|\!\left|\!\left| Q^{-1}\left(\frac{1}{n}\Phi^T\Phi - I\right)Q^{-1}\right|\!\right|\!\right|^2\right] \le \frac{C(\rho^2\gamma(\lambda)\log d)^2}{n},$$

which establishes the claim (83).

References

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *Proceedings of the Twenty Sixth Annual Conference on Computational Learning Theory*, 2013.

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

P. L. Bartlett, O. Bousquet, and S. Mendelson. Localized rademacher complexities. In *Computational Learning Theory*, pages 44–58. Springer, 2002.

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

T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

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.

D. Hsu, S. Kakade, and T. Zhang. Random design analysis of ridge regression. In *Proceedings of the 25nd Annual Conference on Learning Theory*, 2012.

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 the Fifteenth Annual Conference on Computational Learning Theory*, pages 29–43, 2002a.

S. Mendelson. Improving the sample complexity using global data. *Information Theory, IEEE Transactions on*, 48(7):1977–1991, 2002b.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, 2007.

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. Yang, M. Pilanci, and M. J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *arXiv:1501.06195 [stat.ml]*, 2015.

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

T. Zhang. Leave-one-out bounds for kernel methods. *Neural Computation*, 15(6):1397–1437, 2003.

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. *Journal of Machine Learning Research*, 14:3321–3363, 2013.