Journal of Machine Learning Research 15 (2014) 283-314 Submitted 8/13; Revised 12/13; Published 1/14

Early Stopping and Non-parametric Regression: An Optimal Data-dependent Stopping Rule

Garvesh Raskutti RASKUTTI@STAT.WISC.EDU
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706-1799, USA

Martin J. Wainwright WAINWRIG@BERKELEY.EDU
Bin Yu BINYU@STAT.BERKELEY.EDU
Department of Statistics[*]
University of California
Berkeley, CA 94720-1776, USA

Editor: Sara van de Geer

Abstract

Early stopping is a form of regularization based on choosing when to stop running an iterative algorithm. Focusing on non-parametric regression in a reproducing kernel Hilbert space, we analyze the early stopping strategy for a form of gradient-descent applied to the least-squares loss function. We propose a data-dependent stopping rule that does not involve hold-out or cross-validation data, and we prove upper bounds on the squared error of the resulting function estimate, measured in either the $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ norm. These upper bounds lead to minimax-optimal rates for various kernel classes, including Sobolev smoothness classes and other forms of reproducing kernel Hilbert spaces. We show through simulation that our stopping rule compares favorably to two other stopping rules, one based on hold-out data and the other based on Stein's unbiased risk estimate. We also establish a tight connection between our early stopping strategy and the solution path of a kernel ridge regression estimator.

Keywords: early stopping, non-parametric regression, kernel ridge regression, stopping rule, reproducing kernel hilbert space, rademacher complexity, empirical processes

1. Introduction

The phenomenon of overfitting is ubiquitous throughout statistics. It is especially problematic in nonparametric problems, where some form of regularization is essential in order to prevent it. In the non-parametric setting, the most classical form of regularization is that of Tikhonov regularization, where a quadratic smoothness penalty is added to the least-squares loss. An alternative and algorithmic approach to regularization is based on early stopping of an iterative algorithm, such as gradient descent applied to the unregularized loss function. The main advantage of early stopping for regularization, as compared to penalized forms, is lower computational complexity.

∗. Also in the Department of Electrical Engineering and Computer Science.

The idea of early stopping has a fairly lengthy history, dating back to the 1970's in the context of the Landweber iteration. For instance, see the paper by Strand (1974) as well as the subsequent papers (Anderssen and Prenter, 1981; Wahba, 1987). Early stopping has also been widely used in neural networks (Morgan and Bourlard, 1990), for which stochastic gradient descent is used to estimate the network parameters. Past work has provided intuitive arguments for the benefits of early stopping. Roughly speaking, it is clear that each step of an iterative algorithm will reduce bias but increase variance, so early stopping ensures the variance of the estimator is not too high. However, prior to the 1990s, there had been little theoretical justification for these claims. A more recent line of work has developed a theory for various forms of early stopping, including boosting algorithms (Bartlett and Traskin, 2007; Buhlmann and Yu, 2003; Freund and Schapire, 1997; Jiang, 2004; Mason et al., 1999; Yao et al., 2007; Zhang and Yu, 2005), greedy methods (Barron et al., 2008), gradient descent over reproducing kernel Hilbert spaces (Caponneto, 2006; Caponetto and Yao, 2006; Vito et al., 2010; Yao et al., 2007), the conjugate gradient algorithm (Blanchard and Kramer, 2010), and the power method for eigenvalue computation (Orecchia and Mahoney, 2011). Most relevant to our work is the paper of Buhlmann and Yu (2003), who derived optimal mean-squared error bounds for L^2-boosting with early stopping in the case of fixed design regression. However, these optimal rates are based on an "oracle" stopping rule, one that cannot be computed based on the data. Thus, their work left open the following natural question: is there a data-dependent and easily computable stopping rule that produces a minimax-optimal estimator?

The main contribution of this paper is to answer this question in the affirmative for a certain class of non-parametric regression problems, in which the underlying regression function belongs to a reproducing kernel Hilbert space (RKHS). In this setting, a standard estimator is the method of kernel ridge regression (Wahba, 1990), which minimizes a weighted sum of the least-squares loss with a squared Hilbert norm penalty as a regularizer. Instead of a penalized form of regression, we analyze early stopping of an iterative update that is equivalent to gradient descent on the least-squares loss in an appropriately chosen coordinate system. By analyzing the mean-squared error of our iterative update, we derive a data-dependent stopping rule that provides the optimal trade-off between the estimated bias and variance at each iteration. In particular, our stopping rule is based on the first time that a running sum of step-sizes after t steps increases above the critical trade-off between bias and variance. For Sobolev spaces and other types of kernel classes, we show that the function estimate obtained by this stopping rule achieves minimax-optimal estimation rates in both the empirical and generalization norms. Importantly, our stopping rule does not require the use of cross-validation or hold-out data.

In more detail, our first main result (Theorem 1) provides bounds on the squared prediction error for all iterates prior to the stopping time, and a lower bound on the squared error for all iterations after the stopping time. These bounds are applicable to the case of fixed design, where as our second main result (Theorem 2) provides similar types of upper bounds for randomly sampled covariates. These bounds are stated in terms of the squared $L^2(\mathbb{P})$ norm or generalization error, as opposed to the in-sample prediction error, or equivalently, the $L^2(\mathbb{P}_n)$ seminorm defined by the data. Both of these theorems apply to any reproducing kernel, and lead to specific predictions for different kernel classes, depending on their eigendecay. For the case of low rank kernel classes and Sobolev spaces, we prove that

our stopping rule yields a function estimate that achieves the minimax optimal rate (up to a constant pre-factor), so that the bounds from our analysis are essentially unimprovable. Our proof is based on a combination of analytic techniques (Buhlmann and Yu, 2003) with techniques from empirical process theory (van de Geer, 2000). We complement these theoretical results with simulation studies that compare its performance to other rules, in particular a method using hold-out data to estimate the risk, as well as a second method based on Stein's Unbiased Risk Estimate (SURE). In our experiments for first-order Sobolev kernels, we find that our stopping rule performs favorably compared to these alternatives, especially as the sample size grows. In Section 3.4, we provide an explicit link between our early stopping strategy and the kernel ridge regression estimator.

2. Background and Problem Formulation

We begin by introducing some background on non-parametric regression and reproducing kernel Hilbert spaces, before turning to a precise formulation of the problem studied in this paper.

2.1 Non-parametric Regression and Kernel Classes

Suppose that our goal is to use a covariate $X \in \mathcal{X}$ to predict a real-valued response $Y \in \mathbb{R}$. We do so by using a function $f : \mathcal{X} \to \mathbb{R}$, where the value $f(x)$ represents our prediction of Y based on the realization $X = x$. In terms of mean-squared error, the optimal choice is the *regression function* defined by $f^*(x) := \mathbb{E}[Y \mid x]$. In the problem of non-parametric regression with random design, we observe n samples of the form $\{(x_i, y_i), i = 1, \ldots, n\}$, each drawn independently from some joint distribution on the Cartesian product $\mathcal{X} \times \mathbb{R}$, and our goal is to estimate the regression function f^*. Equivalently, we observe samples of the form

$$y_i = f^*(x_i) + w_i, \quad \text{for } i = 1, 2, \ldots, n,$$

where $w_i := y_i - f^*(x_i)$ are independent zero-mean noise random variables. Throughout this paper, we assume that the random variables w_i are *sub-Gaussian* with parameter σ, meaning that

$$\mathbb{E}[e^{tw_i}] \leq e^{t^2\sigma^2/2} \quad \text{for all } t \in \mathbb{R}.$$

For instance, this sub-Gaussian condition is satisfied for normal variates $w_i \sim N(0, \sigma^2)$, but it also holds for various non-Gaussian random variables. Parts of our analysis also apply to the fixed design setting, in which we condition on a particular realization $\{x_i\}_{i=1}^n$ of the covariates.

In order to estimate the regression function, we make use of the machinery of reproducing kernel Hilbert spaces (Aronszajn, 1950; Wahba, 1990; Gu and Zhu, 2001). Using $\mathbb{P}$ to denote the marginal distribution of the covariates, we consider a Hilbert space $\mathcal{H} \subset L^2(\mathbb{P})$, meaning a family of functions $g : \mathcal{X} \to \mathbb{R}$, with $\|g\|_{L^2(\mathbb{P})} < \infty$, and an associated inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ under which $\mathcal{H}$ is complete. The space $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) if there exists a symmetric function $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ such that: (a) for each $x \in \mathcal{X}$, the function $\mathbb{K}(\cdot, x)$ belongs to the Hilbert space $\mathcal{H}$, and (b) we have the reproducing relation

$f(x) = \langle f, \mathbb{K}(\cdot, x)\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. Any such kernel function must be positive semidefinite. Moreover, under suitable regularity conditions, Mercer's theorem (1909) guarantees that the kernel has an eigen-expansion of the form

$$\mathbb{K}(x, x') = \sum_{k=1}^{\infty} \lambda_k \phi_k(x)\phi_k(x'),$$

where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq 0$ are a non-negative sequence of eigenvalues, and $\{\phi_k\}_{k=1}^{\infty}$ are the associated eigenfunctions, taken to be orthonormal in $L^2(\mathbb{P})$. The decay rate of the eigenvalues will play a crucial role in our analysis.

Since the eigenfunctions $\{\phi_k\}_{k=1}^{\infty}$ form an orthonormal basis, any function $f \in \mathcal{H}$ has an expansion of the form $f(x) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} a_k \phi_k(x)$, where for all k such that $\lambda_k > 0$, the coefficients

$$a_k := \frac{1}{\sqrt{\lambda_k}} \langle f, \phi_k \rangle_{L^2(\mathbb{P})} = \int_{\mathcal{X}} f(x)\phi_k(x)\, d\mathbb{P}(x)$$

are rescaled versions of the generalized Fourier coefficients.[1] Associated with any two functions in $\mathcal{H}$—where $f = \sum_{k=1}^{\infty} \sqrt{\lambda_k} a_k \phi_k$ and $g = \sum_{k=1}^{\infty} \sqrt{\lambda_k} b_k \phi_k$—are two distinct inner products. The first is the usual inner product in the space $L^2(\mathbb{P})$—namely, $\langle f, g\rangle_{L^2(\mathbb{P})} := \int_{\mathcal{X}} f(x)g(x)\, d\mathbb{P}(x)$. By Parseval's theorem, it has an equivalent representation in terms of the rescaled expansion coefficients and kernel eigenvalues—that is,

$$\langle f, g\rangle_{L^2(\mathbb{P})} = \sum_{k=1}^{\infty} \lambda_k a_k b_k.$$

The second inner product, denoted by $\langle f, g\rangle_{\mathcal{H}}$, is the one that defines the Hilbert space; it can be written in terms of the rescaled expansion coefficients as

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{k=1}^{\infty} a_k b_k.$$

Using this definition, the unit ball for the Hilbert space $\mathcal{H}$ with eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$ and eigenfunctions $\{\phi_k\}_{k=1}^{\infty}$ takes the form

$$\mathbb{B}_{\mathcal{H}}(1) := \big\{ f = \sum_{k=1}^{\infty} \sqrt{\lambda_k} b_k \phi_k \quad \text{for some} \quad \sum_{k=1}^{\infty} b_k^2 \leq 1 \big\}.$$

The class of reproducing kernel Hilbert spaces contains many interesting classes that are widely used in practice, including polynomials of degree d, Sobolev spaces of varying smoothness, and Gaussian kernels. For more background and examples on reproducing kernel Hilbert spaces, we refer the reader to various standard references (Aronszajn, 1950; Saitoh, 1988; Schölkopf and Smola, 2002; Wahba, 1990; Weinert, 1982).

Throughout this paper, we assume that any function f in the unit ball of the Hilbert space is uniformly bounded, meaning that there is some constant $B < \infty$ such that

$$\|f\|_{\infty} := \sup_{x \in \mathcal{X}} |f(x)| \leq B \qquad \text{for all } f \in \mathbb{B}_{\mathcal{H}}(1). \tag{1}$$

1. We have chosen this particular rescaling for later theoretical convenience.

This boundedness condition (1) is satisfied for any RKHS with a kernel such that $\sup_{x \in \mathcal{X}} \mathbb{K}(x, x) \leq B$. Kernels of this type include the Gaussian and Laplacian kernels, the kernels underlying Sobolev and other spline classes, as well as as well as any trace class kernel with trignometric eigenfunctions. The boundedness condition (1) is quite standard in non-asymptotic analysis of non-parametric regression procedures (e.g., van de Geer, 2000). We study non-parametric regression when the unknown function f^* is viewed as fixed, meaning that no prior is imposed on the function space.

2.2 Gradient Update Equation

We now turn to the form of the gradient update that we study in this paper. Given the samples $\{(x_i, y_i)\}_{i=1}^n$, consider minimizing the least-squares loss function

$$\mathcal{L}(f) := \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2$$

over some subset of the Hilbert space $\mathcal{H}$. By the representer theorem (Kimeldorf and Wahba, 1971), it suffices to restrict attention to functions f belonging to the span of the kernel functions defined on the data points—namely, the span of $\{\mathbb{K}(\cdot, x_i), i = 1, \ldots, n\}$. Accordingly, we adopt the parameterization

$$f(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_i \mathbb{K}(\cdot, x_i), \tag{2}$$

for some coefficient vector $\omega \in \mathbb{R}^n$. Here the rescaling by $1/\sqrt{n}$ is for later theoretical convenience.

Our gradient descent procedure is based on a parameterization of the least-squares loss that involves the *empirical kernel matrix* $K \in \mathbb{R}^{n \times n}$ with entries

$$[K]_{ij} = \frac{1}{n} \mathbb{K}(x_i, x_j) \qquad \text{for } i, j = 1, 2, \ldots, n.$$

For any positive semidefinite kernel function, this matrix must be positive semidefinite, and so has a unique symmetric square root denoted by $\sqrt{K}$. By first introducing the convenient shorthand $y_1^n := \big(y_1 \; y_2 \; \cdots y_n\big) \in \mathbb{R}^n$, we can write the least-squares loss in the form

$$\mathcal{L}(\omega) = \frac{1}{2n} \|y_1^n - \sqrt{n} K \omega\|_2^2.$$

A direct approach would be to perform gradient descent on this form of the least-squares loss. For our purposes, it turns out to be more natural to perform gradient descent in the transformed co-ordinate system $\theta = \sqrt{K}\, \omega$. Some straightforward calculations (see Appendix A for details) yield that the gradient descent algorithm in this new co-ordinate system generates a sequence of vectors $\{\theta_t\}_{t=0}^{\infty}$ via the recursion

$$\theta_{t+1} = \theta_t - \alpha_t \Big(K\, \theta_t - \frac{1}{\sqrt{n}} \sqrt{K}\, y_1^n \Big), \tag{3}$$

where $\{\alpha_t\}_{t=0}^{\infty}$ is a sequence of positive step sizes (to be chosen by the user). We assume throughout that the gradient descent procedure is initialized with $\theta_0 = 0$.

The parameter estimate θ_t at iteration t defines a function estimate f_t in the following way. We first compute[2] the weight vector $\omega^t = \sqrt{K^{-1}}\,\theta_t$, which then defines the function estimate $f_t(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \omega_i^t \mathbb{K}(\cdot, x_i)$ as before. In this paper, our goal is to study how the sequence $\{f_t\}_{t=0}^{\infty}$ evolves as an approximation to the true regression function f^*. We measure the error in two different ways: the $L^2(\mathbb{P}_n)$ norm

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

compares the functions only at the observed design points, whereas the $L^2(\mathbb{P})$-norm

$$\|f_t - f^*\|_2^2 := \mathbb{E}\Big[\big(f_t(X) - f^*(X)\big)^2\Big]$$

corresponds to the usual mean-squared error.

2.3 Overfitting and Early Stopping

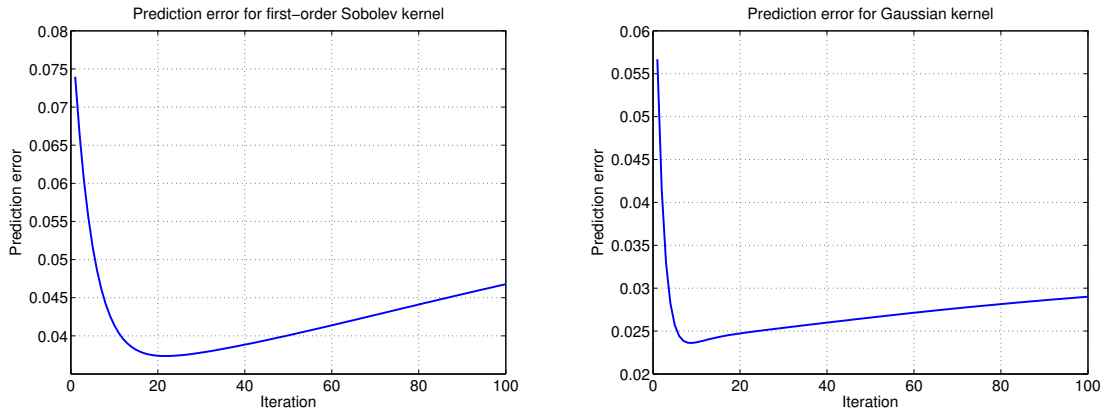
In order to illustrate the phenomenon of interest in this paper, we performed some simulations on a simple problem. In particular, we formed $n = 100$ i.i.d. observations of the form $y = f^*(x_i) + w_i$, where $w_i \sim N(0,1)$, and using the fixed design $x_i = i/n$ for $i = 1, \ldots, n$. We then implemented the gradient descent update (3) with initialization $\theta_0 = 0$ and constant step sizes $\alpha_t = 0.25$. We performed this experiment with the regression function $f^*(x) = |x - 1/2| - 1/2$, and two different choices of kernel functions. The kernel $\mathbb{K}(x, x') = \min\{x, x'\}$ on the unit square $[0,1] \times [0,1]$ generates an RKHS of Lipschitz functions, whereas the Gaussian kernel $\mathbb{K}(x, x') = \exp(-\frac{1}{2}(x - x')^2)$ generates a smoother class of infinitely differentiable functions.

Figure 1 provides plots of the squared prediction error $\|f_t - f^*\|_n^2$ as a function of the iteration number t. For both kernels, the prediction error decreases fairly rapidly, reaching a minimum before or around $T \approx 20$ iterations, before then beginning to increase. As the analysis of this paper will clarify, too many iterations lead to fitting the noise in the data (i.e., the additive perturbations w_i), as opposed to the underlying function f^*. In a nutshell, the goal of this paper is to quantify precisely the meaning of "too many" iterations, and in a data-dependent and easily computable manner.

3. Main Results and Consequences

In more detail, our main contribution is to formulate a data-dependent stopping rule, meaning a mapping from the data $\{(x_i, y_i)\}_{i=1}^n$ to a positive integer $\widehat{T}$, such that the two forms of prediction error $\|f_{\widehat{T}} - f^*\|_n$ and $\|f_{\widehat{T}} - f^*\|_2$ are minimal. In our formulation of such a

2. If the empirical matrix K is not invertible, then we use the pseudoinverse. Note that it may appear as though a matrix inversion is required to estimate ω^t for each t which is computationally intensive. However, the weights ω^t may be computed directly via the iteration $\omega^{t+1} = \omega^t - \alpha_t K(\omega^t - \frac{y_1^n}{\sqrt{n}})$. However, the equivalent update (3) is more convenient for our analysis.



(a) (b)

Figure 1: Behavior of gradient descent update (3) with constant step size $\alpha = 0.25$ applied to least-squares loss with $n = 100$ with equi-distant design points $x_i = i/n$ for $i = 1, \ldots, n$, and regression function $f^*(x) = |x-1/2|-1/2$. Each panel gives plots the $L^2(\mathbb{P}_n)$ error $\|f_t - f^*\|_n^2$ as a function of the iteration number $t = 1, 2, \ldots, 100$. (a) For the first-order Sobolev kernel $\mathbb{K}(x, x') = \min\{x, x'\}$. (b) For the Gaussian kernel $\mathbb{K}(x, x') = \exp(-\frac{1}{2}(x - x')^2)$.

stopping rule, two quantities play an important role: first, the *running sum* of the step sizes

$$\eta_t := \sum_{\tau=0}^{t-1} \alpha_\tau,$$

and secondly, the eigenvalues $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_n \geq 0$ of the empirical kernel matrix K previously defined (2.2). The kernel matrix and hence these eigenvalues are computable from the data. We also note that there is a large body of work on fast computation of kernel eigenvalues (e.g., see Drineas and Mahoney, 2005 and references therein).

3.1 Stopping Rules and General Error Bounds

Our stopping rule involves the use of a model complexity measure, familiar from past work on uniform laws over kernel classes (Bartlett et al., 2005; Koltchinskii, 2006; Mendelson, 2002), known as the local empirical Rademacher complexity. For the kernel classes studied in this paper, it takes the form

$$\widehat{\mathcal{R}}_K(\varepsilon) := \left[\frac{1}{n}\sum_{i=1}^{n} \min\left\{\widehat{\lambda}_i, \varepsilon^2\right\}\right]^{1/2}. \tag{4}$$

For a given noise variance $\sigma > 0$, a closely related quantity—one of central importance to our analysis—is the *critical empirical radius* $\widehat{\varepsilon}_n > 0$, defined to be the smallest positive

solution to the inequality

$$\widehat{\mathcal{R}}_K(\varepsilon) \leq \varepsilon^2/(2e\sigma). \tag{5}$$

The existence and uniqueness of $\widehat{\varepsilon}_n$ is guaranteed for any reproducing kernel Hilbert space; see Appendix D for details. As clarified in our proof, this inequality plays a key role in trading off the bias and variance in a kernel regression estimate.

Our stopping rule is defined in terms of an analogous inequality that involves the running sum $\eta_t = \sum_{\tau=0}^{t-1} \alpha_\tau$ of the step sizes. Throughout this paper, we assume that the step sizes are chosen to satisfy the following properties:

- Boundedness: $0 \leq \alpha_\tau \leq \min\{1, 1/\widehat{\lambda}_1\}$ for all $\tau = 0, 1, 2, \ldots$.
- Non-increasing: $\alpha_{\tau+1} \leq \alpha_\tau$ for all $\tau = 0, 1, 2, \ldots$.
- Infinite travel: the running sum $\eta_t = \sum_{\tau=0}^{t-1} \alpha_\tau$ diverges as $t \to +\infty$.

We refer to any sequence $\{\alpha_\tau\}_{\tau=0}^{\infty}$ that satisfies these conditions as a *valid stepsize sequence*. We then define the *stopping time*

$$\widehat{T} := \arg\min\left\{ t \in \mathbb{N} \mid \widehat{\mathcal{R}}_K\big(1/\sqrt{\eta_t}\big) > (2e\sigma\eta_t)^{-1} \right\} - 1. \tag{6}$$

As discussed in Appendix D, the integer $\widehat{T}$ belongs to the interval $[0, \infty)$ and is unique for any valid stepsize sequence. As will be clarified in our proof, the intuition underlying the stopping rule (6) is that the sum of the step-sizes η_t acts as a tuning parameter that controls the bias-variance tradeoff. The minimizing value is specified by a fixed point of the local Rademacher complexity, in a manner analogous to certain calculations in empirical process theory (van de Geer, 2000; Mendelson, 2002). The stated choice of $\widehat{T}$ optimizes the bias-variance trade-off.

The following result applies to any sequence $\{f_t\}_{t=0}^{\infty}$ of function estimates generated by the gradient iteration (3) with a valid stepsize sequence.

Theorem 1 *Given the stopping time $\widehat{T}$ defined by the rule* (6) *and critical radius $\widehat{\varepsilon}_n$ defined in Equation* (5)*, there are universal positive constants (c_1, c_2) such that the following events hold with probability at least $1 - c_1 \exp(-c_2 n \widehat{\varepsilon}_n^2)$:*

(a) For all iterations $t = 1, 2, ..., \widehat{T}$:

$$\|f_t - f^*\|_n^2 \leq \frac{4}{e\,\eta_t}.$$

(b) At the iteration $\widehat{T}$ chosen according to the stopping rule (6)*, we have*

$$\|f_{\widehat{T}} - f^*\|_n^2 \leq 12\,\widehat{\varepsilon}_n^2.$$

(c) Moreover, for all $t > \widehat{T}$,

$$\mathbb{E}[\|f_t - f^*\|_n^2] \geq \frac{\sigma^2}{4} \eta_t \widehat{\mathcal{R}}_K^2(\eta_t^{-1/2}).$$

3.1.1 Remarks

Although the bounds (a) and (b) are stated as high probability claims, a simple integration argument can be used to show that the expected mean-squared error (over the noise variables, with the design fixed) satisfies a bound of the form

$$\mathbb{E}\big[\|f_t - f^*\|_n^2\big] \leq \frac{4}{e\,\eta_t} \quad \text{for all } t \leq \widehat{T}.$$

To be clear, note that the critical radius $\widehat{\varepsilon}_n$ cannot be made arbitrarily small, since it must satisfy the defining inequality (5). But as will be clarified in corollaries to follow, this critical radius is essentially optimal: we show how the bounds in Theorem 1 lead to minimax-optimal rates for various function classes. The interpretation of Theorem 1 is as follows: if the sum of the step-sizes η_t remains below the threshold defined by (6), applying the gradient update (3) reduces the prediction error. Moreover, note that for Hilbert spaces with a larger kernel complexity, the stopping time $\widehat{T}$ is smaller, since fitting functions in a larger class incurs a greater risk of overfitting.

Finally, the lower bound (c) shows that for large t, running the iterative algorithm beyond the optimal stopping point leads to inconsistent estimators for infinite rank kernels. More concretely, let us suppose that $\widehat{\lambda}_i > 0$ for all $1 \leq i \leq n$. In this case, we have

$$\eta_t \widehat{\mathcal{R}}_K^2(\eta_t^{-1/2}) = \frac{1}{n}\sum_{i=1}^{n} \min(\hat{\lambda}_i \eta_t, 1),$$

which converges to 1 as $t \to$. Consequently, part (c) implies that $\liminf_{t\to\infty} \mathbb{E}[\|f_t - f^*\|_n^2] \geq \frac{\sigma^2}{4}$ as $t \to \infty$, thereby showing the inconsistency of the method.

The statement of Theorem 1 is for the case of fixed design points $\{x_i\}_{i=1}^n$, so that the probability is taken only over the sub-Gaussian noise variables $\{w_i\}_{i=1}^n$ In the case of random design point $x_i \sim \mathbb{P}$ i.i.d. , we can also provide bounds on generalization error in the form the $L^2(\mathbb{P})$-norm $\|f_t - f^*\|_2$. In this setting, for the purposes of comparing to minimax lower bounds, it is also useful to state some results in terms of the population analog of the local empirical Rademacher complexity (4), namely the quantity

$$\mathcal{R}_{\mathbb{K}}(\varepsilon) := \left[\frac{1}{n}\sum_{j=1}^{\infty} \min\left\{\lambda_j, \varepsilon^2\right\}\right]^{1/2}, \tag{7}$$

where λ_j correspond to the eigenvalues of the population kernel $\mathbb{K}$ defined in (3). Using this complexity measure, we define the *critical population rate* ε_n to be the smallest positive solution to the inequality

$$40\,\mathcal{R}_{\mathbb{K}}(\varepsilon) \leq \frac{\varepsilon^2}{\sigma}. \tag{8}$$

(Our choice of the pre-factor 40 is for later theoretical convenience.) In contrast to the critical empirical rate $\widehat{\varepsilon}_n$, this quantity is not data-dependent, since it is specified by the population eigenvalues of kernel operator underlying the RKHS.

Theorem 2 (Random design) *Suppose that in addition to the conditions of Theorem 1, the design variables $\{x_i\}_{i=1}^n$ are sampled i.i.d. according to $\mathbb{P}$ and that the population critical radius ε_n satisfies inequality* (8)*. Then there are universal constants $c_j, j = 1, 2, 3$ such that*

$$\|f_{\widehat{T}} - f^*\|_2^2 \leq c_3 \varepsilon_n^2$$

with probability at least $1 - c_1 \exp(-c_2 n \varepsilon_n^2)$.

Theorems 1 and 2 are general results that apply to any reproducing kernel Hilbert space. Their proofs involve combination of direct analysis of our iterative update (3), combined with techniques from empirical process theory and concentration of measure (van de Geer, 2000; Ledoux, 2001); see Section 4 for the details.

It is worthwhile to compare with the past work of Buhlmann and Yu (2003) (hereafter BY), who also provide some theory for gradient descent, referred to as L^2-boosting in their paper, but focusing exclusively on the fixed design case. Our theory applies to random as well as fixed design, and a broader set of stepsize choices. The most significant difference between Theorem 1 in our paper and Theorem 3 of BY is that we provide a data-dependent stopping rule, whereas their analysis does not lead to a stopping rule that can be computed from the data.

3.2 Consequences for Specific Kernel Classes

Let us now illustrate some consequences of our general theory for special choices of kernels that are of interest in practice.

3.2.1 Kernels with Polynomial Eigendecay

We begin with the class of RKHSs whose eigenvalues satisfy a *polynomial decay condition*, meaning that

$$\lambda_k \leq C\big(\frac{1}{k}\big)^{2\beta} \qquad \text{for some } \beta > 1/2 \text{ and constant } C. \tag{9}$$

Among other examples, this type of scaling covers various types of Sobolev spaces, consisting of functions with β derivatives (Birman and Solomjak, 1967; Gu, 2002). As a very special case, the first-order Sobolev kernel $\mathbb{K}(x, x') = \min\{x, x'\}$ on the unit square $[0, 1] \times [0, 1]$ generates an RKHS of functions that are differentiable almost everywhere, given by

$$\mathcal{H} := \big\{f : [0, 1] \to \mathbb{R} \mid \; f(0) = 0, \quad \int_0^1 (f'(x))^2 dx < \infty\big\}, \tag{10}$$

For the uniform measure on $[0, 1]$, this class exhibits polynomial eigendecay (9) with $\beta = 1$. For any class that satisfies the polynomial decay condition, we have the following corollary:

Corollary 3 *Suppose that in addition to the assumptions of Theorem 2, the kernel class $\mathcal{H}$ satisfies the polynomial eigenvalue decay* (9) *for some parameter $\beta > 1/2$. Then there is a universal constant c_5 such that*

$$\mathbb{E}\big[\|f_{\widehat{T}} - f^*\|_2^2\big] \leq c_5 \big(\frac{\sigma^2}{n}\big)^{\frac{2\beta}{2\beta+1}}. \tag{11}$$

Moreover, if $\lambda_k \geq c\,(1/k)^{2\beta}$ *for all* $k = 1, 2, \ldots$, *then*

$$\mathbb{E}\big[\|f_t - f^*\|_2^2\big] \geq \frac{1}{4}\min\big\{1,\ \sigma^2\frac{(\eta_t)^{\frac{1}{2\beta}}}{n}\big\} \quad \textit{for all iterations } t = 1, 2, \ldots.$$

The proof, provided in Section 4.3, involves showing that the population critical rate (7) is of the order $\mathcal{O}(n^{-\frac{2\beta}{2\beta+1}})$. By known results on non-parametric regression (Stone, 1985; Yang and Barron, 1999), the error bound (11) is minimax-optimal.

In the special case of the first-order spline family (10), Corollary 3 guarantees that

$$\mathbb{E}[\|f_{\widehat{T}} - f^*\|_2^2] \precsim \big(\frac{\sigma^2}{n}\big)^{2/3}. \tag{12}$$

In order to test the accuracy of this prediction, we performed the following set of simulations. First, we generated samples from the observation model

$$y_i = f^*(x_i) + w_i, \qquad \text{for } i = 1, 2, \ldots, n, \tag{13}$$

where $x_i = i/n$, and $w_i \sim N(0, \sigma^2)$ are i.i.d. noise terms. We present results for the function $f^*(x) = |x - 1/2| - 1/2$, a piecewise linear function belonging to the first-order Sobelev class. For all our experiments, the noise variance σ^2 was set to one, but so as to have a data-dependent method, this knowledge was not provided to the estimator. There is a large body of work on estimating the noise variance σ^2 in non-parametric regression. For our simulations, we use a simple method due to Hall and Marron (1990). They proved that their estimator is ratio consistent, which is sufficient for our purposes.

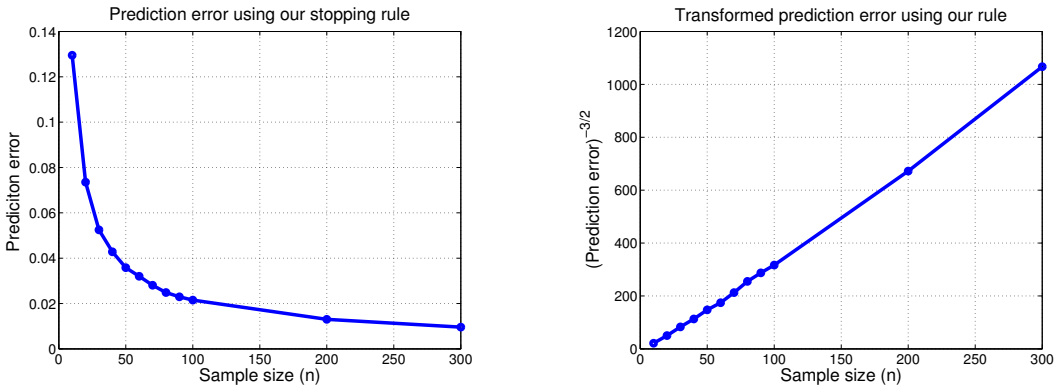
For a range of sample sizes n between 10 and 300, we performed the updates (3) with constant stepsize $\alpha = 0.25$, stopping at the specified time $\widehat{T}$. For each sample size, we performed 10,000 independent trials, and averaged the resulting prediction errors. In panel (a) of Figure 2, we plot the mean-squared error versus the sample size, which shows consistency of the method. The bound (12) makes a more specific prediction: the mean-squared error raised to the power $-3/2$ should scale linearly with the sample size. As shown in panel (b) of Figure 2, the simulation results do indeed reveal this predicted linear relationship. We also performed the same experiments for the case of randomly drawn designs $x_i \sim \text{Unif}(0, 1)$. In this case, we observed similar results, but with more trials required to average out the additional randomness in the design.

3.2.2 Finite Rank Kernels

We now turn to the class of RKHSs based on finite-rank kernels, meaning that there is some finite integer $m < \infty$ such that $\lambda_j = 0$ for all $j \geq m + 1$. For instance, the kernel function $\mathbb{K}(x, x') = (1 + xx')^2$ is a finite rank kernel with $m = 2$, and it generates the RKHS of all quadratic functions. More generally, for any integer $d \geq 2$, the kernel $\mathbb{K}(x, x') = (1 + xx')^d$ generates the RKHS of all polynomials with degree at most d. For any such kernel, we have the following corollary:

Corollary 4 *If, in addition to the conditions of Theorem 2, the kernel has finite rank* m, *then*

$$\mathbb{E}\big[\|\widehat{f}_{\widehat{T}} - f^*\|_2^2\big] \leq c_5\,\sigma^2\frac{m}{n}.$$



(a) (b)

Figure 2: Prediction error obtained from the stopping rule (6) applied to a regression model with n samples of the form $f^*(x_i) + w_i$ at equidistant design points $x_i = i/n$ for $i = 0, 1, \ldots 99$, and i.i.d. Gaussian noise $w_i \sim N(0, 1)$. For these simulations, the true regression function is given by $f^*(x) = |x - \frac{1}{2}| - \frac{1}{2}$. (a) Mean-squared error (MSE) using the stopping rule (6) versus the sample size n. Each point is based on $10,000$ independent realizations of the noise variables $\{w_i\}_{i=1}^n$. (b) Plots of the quantity $MSE^{-3/2}$ versus sample size n. As predicted by the theory, this form of plotting yields a straight line.

For any rank m-kernel, the rate $\frac{m}{n}$ is minimax optimal in terms of squared $L^2(\mathbb{P})$ error; this fact follows as a consequence of more general lower bounds due to Raskutti et al. (2012).

3.3 Comparison with Other Stopping Rules

In this section, we provide a comparison of our stopping rule to two other stopping rules, as well as a oracle method that involves knowledge of f^*, and so cannot be computed in practice.

3.3.1 Hold-out Method

We begin by comparing to a simple hold-out method that performs gradient descent using 50% of the data, and uses the other 50% of the data to estimate the risk. In more detail, assuming that the sample size is even for simplicity, we split the full data set $\{x_i\}_{i=1}^n$ into two equally sized subsets S_{tr} and S_{te}. The data indexed by the training set S_{tr} is used to estimate the function $f_{\text{tr,t}}$ using the gradient descent update (3). At each iteration $t = 0, 1, 2, \ldots$, the data indexed by S_{te} is used to estimate the risk via $R_{\text{HO}}(f_t) = \frac{1}{n} \sum_{i \in S_{te}} \big(y_i - f_{\text{tr,t}}(x_i)\big)^2$, which defines the stopping rule

$$\widehat{T}_{\text{HO}} := \arg\min \Big\{ t \in \mathbb{N} \mid R_{\text{HO}}(f_{\text{tr},\, t+1}) > R_{\text{HO}}(f_{\text{tr,t}}) \Big\} - 1. \tag{14}$$

A line of past work (Yao et al., 2007; Bauer et al., 2007; Caponneto, 2006; Caponetto and Yao, 2006, 2010; Vito et al., 2010) has analyzed stopping rules based on this type of hold-out rule. For instance, Caponneto (2006) analyzes a hold-out method, and shows that it yields rates that are optimal for Sobolev spaces with $\beta \leq 1$ but not in general. A major drawback of using a hold-out rule is that it "wastes" a constant fraction of the data, thereby leading to inflated mean-squared error.

3.3.2 SURE Method

Alternatively, we can use Stein's Unbiased Risk estimate (SURE) to define another stopping rule. Gradient descent is based on the shrinkage matrix $\tilde{S}_t = \prod_{\tau=0}^{t-1}(I - \alpha_\tau K)$. Based on this fact, it can be shown that the SURE estimator (Stein, 1981) takes the form

$$R_{\text{SU}}(f_t) = \frac{1}{n}\{n\sigma^2 + (y_1^n)^T(\tilde{S}_t)^2 y_1^n - 2\sigma^2 \operatorname{trace}(\tilde{S}_t)\}.$$

This risk estimate can be used to define the associated stopping rule

$$\widehat{T}_{\text{SU}} := \arg\min\left\{t \in \mathbb{N} \mid R_{\text{SU}}(f_{t+1}) > R_{\text{SU}}(f_t)\right\} - 1. \tag{15}$$

In contrast with hold-out, the SURE stopping rule (15) makes use of all the data. However, we are not aware of any theoretical guarantees for early stopping based on the SURE rule.

For any valid sequence of stepsizes, it can be shown that both stopping rules (14) and (15) define a unique stopping time. Note that our stopping rule $\widehat{T}$ based on (6) requires estimation of both the empirical eigenvalues, and the noise variance σ^2. In contrast, the SURE-based rule requires estimation of σ^2 but not the empirical eigenvalues, whereas the hold-out rule requires no parameters to be estimated, but a percentage of the data is used to estimate the risk.

3.3.3 Oracle Method

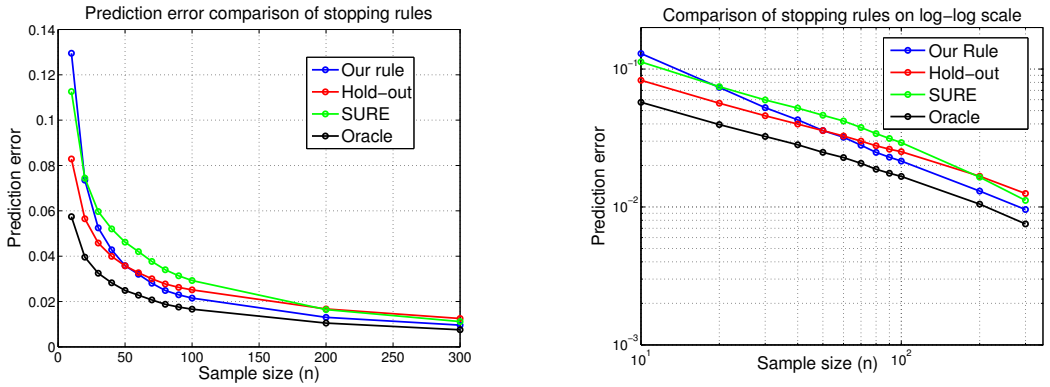
As a third point of reference, we also plot the mean-squared error for an "oracle" method. It is allowed to base its stopping time on the exact in-sample prediction error $R_{\text{OR}}(f_t) = \|f_t - f^*\|_n^2$, which defines the oracle stopping rule

$$\widehat{T}_{\text{OR}} := \arg\min\left\{t \in \mathbb{N} \mid R_{\text{OR}}(f_{t+1}) > R_{\text{OR}}(f_t)\right\} - 1. \tag{16}$$

Note that this stopping rule is not computable from the data, since it assumes exact knowledge of the function f^* that we are trying to estimate.

In order to compare our stopping rule (6) with these alternatives, we generated i.i.d. samples from the previously described model (see Equation (13) and the following discussion). We varied the sample size n from 10 to 300, and for each sample size, we performed $M = 10,000$ independent trials (randomizations of the noise variables $\{w_i\}_{i=1}^n$), and computed the average of squared prediction error over these M trials.

Figure 3 compares the resulting mean-squared errors of our stopping rule (6), the hold-out stopping rule (14), the SURE-based stopping rule (15), and the oracle stopping rule (16). Panel (a) shows the mean-squared error versus sample size, whereas panel (b) shows the



(a) (b)

Figure 3: Illustration of the performance of different stopping rules for kernel gradient descent with the kernel $\mathbb{K}(x,x) = \min\{|x|, |x'|\}$ and noisy samples of the function $f^*(x) = |x - \frac{1}{2}| - \frac{1}{2}$. In each case, we applied the gradient update (3) with constant stepsizes $\alpha_t = 1$ for all t. Each curve corresponds to the mean-squared error, estimated by averaging over $M = 10,000$ independent trials, versus the sample size for $n \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300\}$. Each panel shows MSE curves for four different stopping rules: (i) the stopping rule (6); (ii) holding out 50% of the data and using (14); (iii) the SURE stopping rule (15); and (iv) the oracle stopping rule (14). (a) MSE versus sample size on a standard scale. (b) MSE versus sample size on a log-log scale.

same curves in terms of logarithm of mean-squared error. Our proposed rule exhibits better performance than the hold-out and SURE-based rules for sample sizes n larger than 50. On the flip side, since the construction of our stopping rule is based on the assumption that f^* belongs to a known RKHS, it is unclear how robust it would be to model mis-specification. In contrast, the hold-out and SURE-based stopping rules are generic methods, not based directly on the RKHS structure, so might be more robust to model mis-specification. Thus, one interesting direction is to explore the robustness of our stopping rule. On the theoretical front, it would be interesting to determine whether the hold-out and/or SURE-based stopping rules can be proven to achieve minimax optimal rates for general kernels, as we have established for our stopping rule.

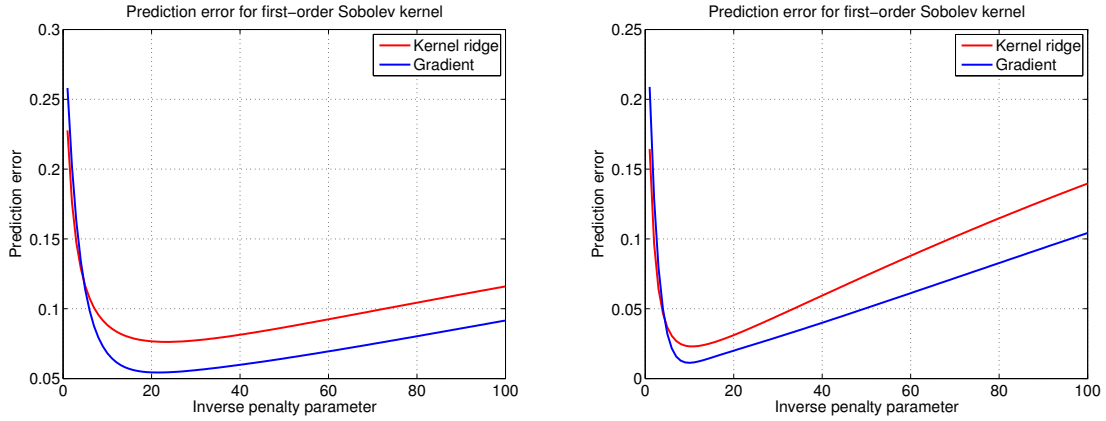
3.4 Connections to Kernel Ridge Regression

We conclude by presenting an interesting link between our early stopping procedure and kernel ridge regression. The kernel ridge regression (KRR) estimate is defined as

$$\widehat{f}_\nu := \arg\min_{f \in \mathcal{H}} \Big\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \frac{1}{2\nu} \|f\|_{\mathcal{H}}^2 \Big\}, \tag{17}$$

where ν is the (inverse) regularization parameter. For any $\nu < \infty$, the objective is strongly convex, so that the KRR solution is unique.

Friedman and Popescu (2004) observed through simulations that the regularization paths for early stopping of gradient descent and ridge regression are similar, but did not provide any theoretical explanation of this fact. As an illustration of this empirical phenomenon, Figure 4 compares the prediction error $\|\widehat{f}_\nu - f^*\|_n^2$ of the kernel ridge regression estimate over the interval $\nu \in [1, 100]$ versus that of the gradient update (3) over the first 100 iterations. Note that the curves, while not identical, are qualitatively very similar.



(a) (b)

Figure 4: Comparison of the prediction error of the path of kernel ridge regression estimates (17) obtained by varying $\nu \in [1, 100]$ to those of the gradient updates (3) over 100 iterations with constant step size. All simulations were performed with the kernel $\mathbb{K}(x, x') = \min\{|x|, |x'|\}$ based on $n = 100$ samples at the design points $x_i = i/n$ with $f^*(x) = |x - \frac{1}{2}| - \frac{1}{2}$. (a) Noise variance $\sigma^2 = 1$. (b) Noise variance $\sigma^2 = 2$.

From past theoretical work (van de Geer, 2000; Mendelson, 2002), kernel ridge regression with the appropriate setting of the penalty parameter ν is known to achieve minimax-optimal error for various kernel classes. These classes include the Sobolev and finite-rank kernels for which we have previously established that our stopping rule (6) yields optimal rates. In this section, we provide a theoretical basis for these connections. More precisely, we prove that if the inverse penalty parameter ν is chosen using the same criterion as our stopping rule, then the prediction error satisfies the same type of bounds, with ν now playing the role of the running sum η_t.

Define $\widehat{\nu} > 0$ to be the smallest positive solution to the inequality

$$\big(4\sigma\nu\big)^{-1} < \widehat{\mathcal{R}}_K(1/\sqrt{\nu}). \tag{18}$$

Note that this criterion is identical to the one underlying our stopping rule, except that the continuous parameter ν replaces the discrete parameter $\eta_t = \sum_{\tau=0}^{t-1} \alpha_\tau$.

Proposition 5 *Consider the kernel ridge regression estimator* (17) *applied to n i.i.d. samples $\{(x_i, y_i)\}$ with σ-sub Gaussian noise. Then there are universal constants (c_1, c_2, c_3) such that with probability at least $1 - c_1 \exp(-c_2\, n\, \widehat{\varepsilon}_n^2)$:*

(a) For all $0 < \nu \leq \widehat{\nu}$, we have

$$\|\widehat{f}_\nu - f^*\|_n^2 \leq \frac{2}{\nu}$$

(b) With $\widehat{\nu}$ chosen according to the rule (18)*, we have*

$$\|\widehat{f}_{\widehat{\nu}} - f^*\|_n^2 \leq c_3\, \widehat{\varepsilon}_n^2.$$

(c) Moreover, for all $\nu > \widehat{\nu}$, we have

$$\mathbb{E}[\|\widehat{f}_\nu - f^*\|_n^2] \geq \frac{\sigma^2}{4} \nu \widehat{\mathcal{R}}_K^2(\nu^{-1/2}).$$

Note that apart from a slightly different leading constant, the upper bound (a) is *identical* to the upper bound in Theorem 1 part (a). The only difference is that the inverse regularization parameter ν replaces the running sum $\eta_t = \sum_{\tau=0}^{t-1} \alpha_\tau$. Similarly, part (b) of Proposition 5 guarantees that the kernel ridge regression (17) has prediction error that is upper bounded by the empirical critical rate $\widehat{\varepsilon}_n^2$, as in part (b) of Theorem 1. Let us emphasize that bounds of this type on kernel ridge regression have been derived in past work (Mendelson, 2002; Zhang, 2005; van de Geer, 2000). The novelty here is that the structure of our result reveals the intimate connection to early stopping, and in fact, the proofs follow a parallel thread.

In conjunction, Proposition 5 and Theorem 1 provide a theoretical explanation for why, as shown in Figure 4, the paths of the gradient descent update (3) and kernel ridge regression estimate (17) are so similar. However, it is important to emphasize that from a computational point of view, early stopping has certain advantages over kernel ridge regression. In general, solving a quadratic program of the form (17) requires on the order of $\mathcal{O}(n^3)$ basic operations, and this must be done repeatedly for each new choice of ν. On the other hand, by its very construction, the iterates of the gradient algorithm correspond to the desired path of solutions, and each gradient update involves multiplication by the kernel matrix, incurring $\mathcal{O}(n^2)$ operations.

4. Proofs

We now turn to the proofs of our main results. The main steps in each proof are provided in the main text, with some of the more technical results deferred to the appendix.

4.1 Proof of Theorem 1

In order to derive upper bounds on the $L^2(\mathbb{P}_n)$-error in Theorem 1, we first rewrite the gradient update (3) in an alternative form. For each iteration $t = 0, 1, 2, \ldots$, let us introduce the shorthand

$$f_t(x_1^n) := \begin{bmatrix} f_t(x_1) & f_t(x_2) & \cdots & f_t(x_n) \end{bmatrix} \in \mathbb{R}^n,$$

corresponding to the n-vector obtained by evaluating the function f^t at all design points, and the short-hand

$$w := \big[w_1, w_2, ..., w_n\big] \in \mathbb{R}^n,$$

corresponding to the vector of zero mean sub-Gaussian noise random variables. From Equation (2), we have the relation

$$f^t(x_1^n) = \frac{1}{\sqrt{n}} K\,\omega^t \;=\; \frac{1}{\sqrt{n}}\sqrt{K}\,\theta_t.$$

Consequently, by multiplying both sides of the gradient update (3) by $\sqrt{K}$, we find that the sequence $\{f_t(x_1^n)\}_{t=0}^{\infty}$ evolves according to the recursion

$$f_{t+1}(x_1^n) = f_t(x_1^n) - \alpha_t K\,(f_t(x_1^n) - y_1^n) \;=\; \Big(I_{n\times n} - \alpha_t K\Big) f_t(x_1^n) + \alpha_t K\, y_1^n. \tag{19}$$

Since $\theta_0 = 0$, the sequence is initialized with $f_0(x_1^n) = 0$. The recursion (19) lies at the heart of our analysis.

Letting $r = \operatorname{rank}(K)$, the empirical kernel matrix has the eigendecomposition $K = U\Lambda U^T$, where $U \in \mathbb{R}^{n\times n}$ is an orthonormal matrix (satisfying $UU^T = U^TU = I_{n\times n}$) and

$$\Lambda := \operatorname{diag}(\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_r, 0, 0, \cdots, 0)$$

is the diagonal matrix of eigenvalues, augmented with $n-r$ zero eigenvalues as needed. We then define a sequence of diagonal *shrinkage matrices* S^t as follows:

$$S^t := \prod_{\tau=0}^{t-1} (I_{n\times n} - \alpha_\tau \Lambda) \in \mathbb{R}^{n\times n}.$$

The matrix S^t indicates the extent of shrinkage towards the origin; since $0 \;\leq\; \alpha_t \;\leq \min\{1, 1/\widehat{\lambda}_1\}$ for all iterations t, in the positive semodefinite ordering, we have the sandwich relation

$$0 \preceq S^{t+1} \preceq S^t \preceq I_{n\times n}.$$

Moreover, the following lemma shows that the $L^2(\mathbb{P}_n)$-error at each iteration can be bounded in terms of the eigendecomposition and these shrinkage matrices:

Lemma 6 (Bias/variance decomposition) *At each iteration $t = 0, 1, 2, \ldots$,*

$$\|f_t - f^*\|_n^2 \leq \underbrace{\frac{2}{n}\sum_{j=1}^{r} (S^t)_{jj}^2 [U^T f^*(x_1^n)]_j^2 + \frac{2}{n}\sum_{j=r+1}^{n} [U^T f^*(x_1^n)]_j^2}_{\textit{Squared Bias } B_t^2} + \underbrace{\frac{2}{n}\sum_{j=1}^{r} (1 - S_{jj}^t)^2 [U^T w]_j^2}_{\textit{Variance } V_t}. \tag{20}$$

Moreover, we have the lower bound $\mathbb{E}[\|f_t - f^*\|_n^2] \geq \mathbb{E}[V_t]$.

See Appendix B.1 for the proof of this intermediate claim.

In order to complete the proof of the upper bound in Theorem 1, our next step is to obtain high probability upper bounds on these two terms. We summarize our conclusions in an additional lemma, and use it to complete the proof of Theorem 1(a) before returning to prove it.

Lemma 7 (Bounds on the bias and variance) *For all iterations $t = 1, 2, \ldots$, the squared bias is upper bounded as*

$$B_t^2 \leq \frac{1}{e\,\eta_t}, \tag{21}$$

Moreover, there is a universal constant $c_1 > 0$ such that, for any iteration $t = 1, 2, \ldots, \widehat{T}$,

$$V_t \;\leq\; 5\sigma^2\,\eta_t \mathcal{R}_K^2\big(1/\sqrt{\eta_t}\big) \tag{22}$$

with probability at least $1 - \exp\big(-c_1\, n\widehat{\varepsilon}_n^2\big)$. Moreover we have $\mathbb{E}[V_t] \geq \frac{\sigma^2}{4}\,\eta_t \mathcal{R}_K^2\big(1/\sqrt{\eta_t}\big)$.

We can now complete the proof of Theorem 1(a). Conditioned on the event $V_t \leq 5\sigma^2 \eta_t \mathcal{R}_K^2\big(1/\sqrt{\eta_t}\big)$, we have

$$\|f_t - f^*\|_n^2 \overset{(i)}{\leq} B_t^2 + V_t \overset{(ii)}{\leq} \frac{1}{e\,\eta_t} + 5\sigma^2\,\eta_t \mathcal{R}_K^2\big(1/\sqrt{\eta_t}\big) \overset{(iii)}{\leq} \frac{4}{e\,\eta_t},$$

where inequality (i) follows from (20) in Lemma 6, and inequality (ii) follows from the bounds in Lemma 7 and (iii) follows since $t \leq \widehat{T}$. The lower bound (c) follows from (22).

Turning to the proof of part (b), using the upper bound from (a)

$$\|f_{\widehat{T}} - f^*\|_n^2 \leq \frac{1}{e\,\eta_{\widehat{T}}} + \frac{5}{\eta_{\widehat{T}}} \;\leq\; \frac{4}{e\eta_{\widehat{T}}}.$$

Based on the definition of $\widehat{T}$ and $\widehat{\varepsilon}_n$, we are guaranteed that $\frac{1}{\eta_{\widehat{T}+1}} \leq \widehat{\varepsilon}_n^2$, Moreover, by the non-decreasing nature of our step sizes, we have $\alpha_{\widehat{T}+1} \leq \alpha_{\widehat{T}}$, which implies that $\eta_{\widehat{T}+1} \leq 2\eta_{\widehat{T}}$, and hence

$$\frac{1}{\eta_{\widehat{T}}} \leq \frac{2}{\eta_{\widehat{T}+1}} \;\leq\; 2\widehat{\varepsilon}_n^2.$$

Putting together the pieces establishes the bound claimed in part (b).

It remains to establish the bias and variance bounds stated in Lemma 7, and we do so in the following subsections. The following auxiliary lemma plays a role in both proofs:

Lemma 8 (Properties of shrinkage matrices) *For all indices $j \in \{1, 2, \ldots, r\}$, the shrinkage matrices S^t satisfy the bounds*

$$0 \leq\; (S^t)_{jj}^2 \leq \frac{1}{2e\eta_t\widehat{\lambda}_j}, \quad \textit{and} \tag{23}$$

$$\frac{1}{2}\min\{1, \eta_t\widehat{\lambda}_j\} \leq 1 - S_{jj}^t \;\leq \min\{1, \eta_t\widehat{\lambda}_j\}. \tag{24}$$

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

4.1.1 Bounding the Squared Bias

Let us now prove the upper bound (21) on the squared bias. We bound each of the two terms in the definition (20) of B_t^2 in term. Applying the upper bound (23) from Lemma 8, we see that

$$\frac{2}{n}\sum_{j=1}^{r}(S^t)_{jj}^2[U^T f^*(x_1^n)]_j^2 \leq \frac{1}{e\ n\ \eta_t}\ \sum_{j=1}^{r}\frac{[U^T f^*(x_1^n)]_j^2}{\widehat{\lambda}_j}.$$

Now consider the linear operator $\Phi_X : \ell^2(\mathbb{N}) \to \mathbb{R}^n$ defined element-wise via $[\Phi_X]_{jk} = \phi_j(x_k)$. Similarly, we define a (diagonal) linear operator $D : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ with entries $[D]_{jj} = \lambda_j$ and $[D]_{jk} = 0$ for $j \neq k$. With these definitions, the vector $f(x_1^n) \in \mathbb{R}^n$ can be expressed in terms of some sequence $a \in \ell^2(\mathbb{N})$ in the form

$$f(x_1^n) = \Phi_X D^{1/2} a.$$

In terms of these quantities, we can write $K = \frac{1}{n}\Phi_X D \Phi_X^T$. Moreover, as previously noted, we also have $K = U\Lambda U^T$ where $\Lambda = \operatorname{diag}\{\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_n\}$, and $U \in \mathbb{R}^{n\times n}$ is orthonormal. Combining the two representations, we conclude that

$$\frac{\Phi_X D^{1/2}}{\sqrt{n}} = U\Lambda^{1/2}\Psi^*,$$

for some linear operator $\Psi : \mathbb{R}^n \to \ell^2(\mathbb{N})$ (with adjoint Ψ^*) such that $\Psi^*\Psi = I_{n\times n}$. Using this equality, we have

$$\begin{aligned}
\frac{1}{e\,\eta_t\,n}\sum_{j=1}^{r}\frac{[U^T f^*(X)]_j^2}{\widehat{\lambda}_j} &= \frac{1}{e\,\eta_t\,n}\ \sum_{j=1}^{r}\frac{[U^T\Phi_X D^{1/2}a]_j^2}{\widehat{\lambda}_j} \\
&= \frac{1}{e\,\eta_t}\ \sum_{j=1}^{r}\frac{[U^T U\Lambda^{1/2}V^* a]_j^2}{\widehat{\lambda}_j} \\
&= \frac{1}{e\,\eta_t}\ \sum_{j=1}^{r}\frac{\widehat{\lambda}_j\,[\Psi^* a]_j^2}{\widehat{\lambda}_j} \\
&\leq \frac{1}{e\,\eta_t}\ \|\Psi^* a\|_2^2 \\
&\leq \frac{1}{e\,\eta_t}, \qquad (25)
\end{aligned}$$

Here the final step follows from the fact that Ψ is a unitary operator, so that $\|\Psi^* a\|_2^2 \leq \|a\|_2^2 = \|f^*\|_{\mathcal{H}}^2 \leq 1$.

Turning to the second term in the definition (20), we have

$$\begin{aligned}\sum_{j=r+1}^{n} [U^T f^*(x_1^n)]_j^2 &= \frac{2}{n} \sum_{j=r+1}^{n} [U^T \Phi_X D^{1/2} a]_j^2 \\ &= \sum_{j=r+1}^{n} [U^T U \Lambda^{1/2} \Psi^* a]_j^2 \\ &= \sum_{j=r+1}^{n} [\Lambda^{1/2} \Psi^* a]_j^2 \\ &= 0, \end{aligned} \tag{26}$$

where the final step uses the fact that $\Lambda_{jj}^{1/2} = 0$ for all $j \in \{r+1, \ldots, n\}$ by construction. Combining the upper bounds (25) and (26) with the definition (20) of B_t^2 yields the claim (21).

4.1.2 Controlling the Variance

Let us now prove the bounds (22) on the variance term V_t. (To simplify the proof, we assume throughout that $\sigma = 1$; the general case can be recovered by a simple rescaling argument). By the definition of V_t, we have

$$V_t = \frac{2}{n} \sum_{j=1}^{r} (1 - S_{jj}^t)^2 [U^T w]_j^2 \;=\; \frac{2}{n} \operatorname{trace}(UQU^T ww^T),$$

where $Q = \operatorname{diag}\{(1 - S_{jj}^t)^2,\ j = 1, \ldots, n\}$ is a diagonal matrix. Since $\mathbb{E}[ww^T] \leq I_{n \times n}$ by assumption, we have $\mathbb{E}[V_t] = \frac{2}{n} \operatorname{trace}(Q)$. Using the upper bound in Equation (24) from Lemma 8, we have

$$\frac{1}{n} \operatorname{trace}(Q) \leq \frac{1}{n} \sum_{j=1}^{r} \min\{1, (\eta_t \widehat{\lambda}_j)^2\} \;=\; \eta_t \left(\mathcal{R}_K(1/\sqrt{\eta_t}) \right)^2,$$

where the final equality uses the definition of $\mathcal{R}_K$. Putting together the pieces, we see that

$$\mathbb{E}[V_t] \leq 2\, \eta_t \left(\mathcal{R}_K(1/\sqrt{\eta_t}) \right)^2.$$

Similarly, using the lower bound in Equation (24), we can show that

$$\mathbb{E}[V_t] \geq \frac{\sigma^2}{4} \eta_t \left(\mathcal{R}_K(1/\sqrt{\eta_t}) \right)^2.$$

Our next step is to obtain a bound on the two-sided tail probability $\mathbb{P}[|V_t - \mathbb{E}[V_t]| \geq \delta]$, for which we make use of a result on two-sided deviations for quadratic forms in sub-Gaussian variables. In particular, consider a random variable of the form $Q_n = \sum_{i,j=1}^{n} a_{ij}(Z_i Z_j -$

$\mathbb{E}[Z_i Z_j]$) where $\{Z_i\}_{i=1}^n$ are i.i.d. zero-mean and sub-Gaussian variables (with parameter 1). Wright (1973) proves that there is a constant c such that

$$\mathbb{P}\big[|Q - \mathbb{E}[Q]| \geq \delta\big] \leq \exp\Big(- c \ \min\big\{ \frac{\delta}{|||A|||_{\rm op}}, \frac{\delta^2}{|||A|||_{\rm F}^2} \big\}\Big) \quad \text{for all } u > 0, \tag{27}$$

where $(|||A|||_{\rm op}, |||A|||_{\rm F})$ are (respectively) the operator and Frobenius norms of the matrix $A = \{a_{ij}\}_{i,j=1}^n$.

If we apply this result with $A = \frac{2}{n} U Q U^T$ and $Z_i = w_i$, then we have $Q = V_t$, and moreover

$$|||A|||_{\rm op} \leq \frac{2}{n}, \quad \text{and}$$

$$|||A|||_{\rm F}^2 \ = \ \frac{4}{n^2} \operatorname{trace}(U^T Q U^T U Q U^T) = \frac{4}{n^2} \operatorname{trace}(Q^2) \leq \frac{4}{n^2} \operatorname{trace}(Q) \leq \frac{4}{n} \eta_t \left(\mathcal{R}_K(1/\sqrt{\eta_t}) \right).$$

Consequently, the bound (27) implies that

$$\mathbb{P}\big[|V_t - \mathbb{E}[V_t]| \geq \delta\big] \leq \exp\big(- 4c\, n\, \delta \min\{1, \delta \Big(\eta_t \mathcal{R}_K(1/\sqrt{\eta_t}) \Big)^{-1} \}\big).$$

Since $t \leq \widehat{T}$ setting $\delta = 3\sigma^2 \eta_t \left(\mathcal{R}_K(1/\sqrt{\eta_t}) \right)$, the claim (22) follows.

4.2 Proof of Theorem 2

This proof is based on the following two steps:

- first, proving that the error $\|f_{\widehat{T}} - f^*\|_2$ in the $L^2(\mathbb{P})$ norm is, with high probability, close to the error in the $L^2(\mathbb{P}_n)$ norm, and
- second, showing the empirical critical radius $\widehat{\varepsilon}_n$ defined in Equation (5) is upper bounded by the population critical radius ε_n defined in Equation (8).

Our proof is based on a number of more technical auxiliary lemmas, proved in the appendices. The first lemma provides a high probability bound on the Hilbert norm of the estimate $f_{\widehat{T}}$.

Lemma 9 *There exist universal constants c_1 and $c_2 > 0$ such that $\|f_t\|_{\mathcal{H}} \leq 2$ for all $t \leq \widehat{T}$ with probability greater than or equal to $1 - c_1 \exp(-c_2 n \widehat{\varepsilon}_n^2)$.*

See Appendix E.1 for the proof of this claim. Our second lemma shows in any bounded RKHS, the $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ norms are uniformly close up to the population critical radius ε_n over a Hilbert ball of constant radius:

Lemma 10 *Consider a Hilbert space such that $\|g\|_\infty \leq B$ for all $g \in \mathbb{B}_{\mathcal{H}}(3)$. Then there exist universal constants (c_1, c_2, c_3) such that for any $t \geq \varepsilon_n$, we have*

$$|\|g\|_n^2 - \|g\|_2^2| \ \leq \ c_1 t^2,$$

with probability at least $1 - c_2 \exp(-c_3 n t^2)$.

This claim follows from known results on reproducing kernel Hilbert spaces (e.g., Lemma 5.16 in the paper van de Geer, 2000 and Theorem 2.1 in the paper Bartlett et al., 2005). Our final lemma, proved in Appendix E.2, relates the critical empirical radius $\widehat{\varepsilon}_n$ to the population radius ε_n:

Lemma 11 *There exist constants c_1 and c_2 such that $\widehat{\varepsilon}_n \leq \varepsilon_n$ holds with probability at least $1 - c_1 \exp(-c_2 n \varepsilon_n^2)$.*

With these lemmas in hand, the proof of the theorem is straightforward. First, from Lemma 9, we have $\|f_{\widehat{T}}\|_{\mathcal{H}} \leq 2$ and hence by triangle inequality, $\|f_{\widehat{T}} - f^*\|_{\mathcal{H}} \leq 3$ with high probability as well. Next, applying Lemma 10 with $t = \varepsilon_n$, we find that

$$\|f_{\widehat{T}} - f^*\|_2^2 \leq \|f_{\widehat{T}} - f^*\|_n^2 + c_1 \varepsilon_n^2 \leq c_4(\widehat{\varepsilon}_n^2 + \varepsilon_n^2),$$

with probability greater than $1 - c_2 \exp(-c_3 n \varepsilon_n^2)$. Finally, applying Lemma 11 yields that the bound $\|f_{\widehat{T}} - f^*\|_2^2 \leq c \varepsilon_n^2$ holds with the claimed probability.

4.3 Proof of Corollaries

In each case, it suffices to upper bound the generalization rate ε_n^2 previously defined.

4.3.1 Proof of Corollary 4

In this case, we have

$$\mathcal{R}_{\mathbb{K}}(\epsilon) = \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^{m} \min\{\lambda_j, \epsilon^2\}} \;\leq\; \sqrt{\frac{m}{n}}\, \epsilon$$

so that $\varepsilon_n^2 = c' \sigma^2 \frac{m}{n}$.

4.3.2 Proof of Corollary 3

For any $M \geq 1$, we have

$$\begin{aligned} \mathcal{R}_{\mathbb{K}}(\epsilon) \;=\; \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^{\infty} \min\{C\, j^{-2\beta}, \epsilon^2\}} &\leq \sqrt{\frac{M}{n}}\epsilon + \sqrt{\frac{C}{n}} \sqrt{\sum_{j=\lceil M \rceil}^{\infty} j^{-2\beta}} \\ &\leq \sqrt{\frac{M}{n}}\epsilon + \sqrt{\frac{C'}{n}} \sqrt{\int_{M}^{\infty} t^{-2\beta} dt} \\ &\leq \sqrt{\frac{M}{n}}\epsilon + C'' \frac{1}{\sqrt{n}} (1/M)^{\beta - \frac{1}{2}}. \end{aligned}$$

Setting $M = \epsilon^{-1/\beta}$ yields $\mathcal{R}_{\mathbb{K}}(\epsilon) \leq C^* \epsilon^{1 - \frac{1}{2\beta}}$. Consequently, the critical inequality $\mathcal{R}_{\mathbb{K}}(\epsilon) \leq 40\epsilon^2/\sigma$ is satisfied for $\varepsilon_n \asymp (\sigma^2/n)^{\frac{2\beta}{2\beta+1}}$, as claimed.

4.4 Proof of Proposition 5

We now turn to the proof of our results on the kernel ridge regression estimate (17). The proof follows a very similar structure to that of Theorem 1. Recall the eigendecomposition $K = U\Lambda U^T$ of the empirical kernel matrix, and that we use r to denote its rank. For each $\nu > 0$, we define the *ridge shrinkage matrix*

$$R^\nu := \big(I_{n\times n} + \nu\Lambda\big)^{-1}. \tag{28}$$

We then have the following analog of Lemma 7 from the proof of Theorem 1:

Lemma 12 (Bias/variance decomposition for kernel ridge regression) *For any $\nu > 0$, the prediction error for the estimate $\widehat{f}_\nu$ is bounded as*

$$\|\widehat{f}_\nu - f^*\|_n^2 \leq \frac{2}{n}\sum_{j=1}^{r}[R^\nu]_{jj}^2[U^T f^*(x_1^n)]_j^2 + \frac{2}{n}\sum_{j=r+1}^{n}[U^T f^*(x_1^n)]_j^2 + \frac{2}{n}\sum_{j=1}^{r}\big(1 - R^\nu_{jj}\big)^2[U^T w]_j^2.$$

Note that Lemma 12 is identical to Lemma 7 with the shrinkage matrices S^t replaced by their analogues R^ν. See Appendix C.1 for the proof of this claim.

Our next step is to show that the diagonal elements of the shrinkage matrices R^ν are bounded:

Lemma 13 (Properties of kernel ridge shrinkage) *For all indices $j \in \{1, 2, \ldots, r\}$, the diagonal entries R^ν satisfy the bounds*

$$\begin{aligned} 0 \;\leq\; (R^\nu_{jj})^2 \leq \frac{1}{4\nu\widehat{\lambda}_j}, \quad \text{and} \\ \frac{1}{2}\min\big\{1, \nu\widehat{\lambda}_j\big\} \;\leq\; 1 - R^\nu_{jj} \;\leq\; \min\big\{1, \nu\widehat{\lambda}_j\big\}. \end{aligned} \tag{29}$$

Note that this is the analog of Lemma 8 from Theorem 1, albeit with the constant $\frac{1}{4}$ in the bound (29) instead of $\frac{1}{2e}$. See Appendix C.2 for the proof of this claim. With these lemmas in place, the remainder of the proof follows as in the proof of Theorem 1.

5. Discussion

In this paper, we have analyzed the early stopping strategy as applied to gradient descent on the non-parametric least squares loss. Our main contribution was to propose an easily computable and data-dependent stopping rule, and to provide upper bounds on the empirical $L^2(\mathbb{P}_n)$ error (Theorem 1) and generlization $L^2(\mathbb{P})$ error (Theorem 2). We demonstrate in Corollaries 3 and 4 that our stopping rule yields minimax optimal rates for both low rank kernel classes and Sobolev spaces. Our simulation results confirm that our stopping rule yields theoretically optimal rates of convergence for Lipschitz kernels, and performs favorably in comparison to stopping rules based on hold-out data and Stein's Unbiased Risk Estimate. We also showed that early stopping with sum of step-sizes $\eta_t = \sum_{k=0}^{t-1}\alpha_k$ has a regularization path that satisfies almost identical mean-squared error bounds as kernel ridge regression indexed by penalty parameter ν.

Our analysis and stopping rule may be improved and extended in a number of ways. First, it would interesting to see how our stopping rule can be adapted to mis-specified models. As specified, our method relies on computation of the eigenvalues of the kernel matrix. A stopping rule based on approximate eigenvalue computations, for instance via some form of sub-sampling (Drineas and Mahoney, 2005), would be interesting to study as well.

Acknowledgments

This work was partially supported by NSF grant DMS-1107000 to MJW and BY. In addition, BY was partially supported by the NSF grant SES-0835531 (CDI), ARO-W911NF-11-1-0114 and the Center for Science of Information (CSoI), an US NSF Science and Technology Center, under grant agreement CCF-0939370, and MJW was also partially supported ONR MURI grant N00014-11-1-086. During this work, GR received partial support from a Berkeley Graduate Fellowship.

Appendix A. Derivation of Gradient Descent Updates

In this appendix, we provide the details of how the gradient descent updates (3) are obtained. In terms of the transformed vector $\theta = \sqrt{K}\,\omega$, the least-squares objective takes the form

$$\widetilde{\mathcal{L}}(\theta) := \frac{1}{2n}\|y_1^n - \sqrt{n}\sqrt{K}\,\theta\|_2^2 \;=\; \frac{1}{2n}\|y_1^n\|_2^2 - \frac{1}{\sqrt{n}}\langle y_1^n,\, \sqrt{K}\,\theta\rangle + \frac{1}{2}(\theta)^T K\theta.$$

Given a sequence $\{\alpha_t\}_{t=0}^{\infty}$, the gradient descent algorithm operates via the recursion $\theta_{t+1} = \theta_t - \alpha_t \nabla\widetilde{\mathcal{L}}(\theta^t)$. Taking the gradient of $\widetilde{\mathcal{L}}$ yields

$$\nabla\widetilde{\mathcal{L}}(\theta) = K\,\theta - \frac{1}{\sqrt{n}}\sqrt{K}\,y_1^n.$$

Substituting into the gradient descent update yields the claim (3).

Appendix B. Auxiliary Lemmas for Theorem 1

In this appendix, we collect together the proofs of the lemmas for Theorem 1.

B.1 Proof of Lemma 6

We prove this lemma by analyzing the gradient descent iteration in an alternative coordinate system. In particular, given a vector $f^t(x_1^n) \in \mathbb{R}^n$ and the SVD $K = U\Lambda U^T$ of the empirical kernel matrix, we define the vector $\gamma^t = \frac{1}{\sqrt{n}}U^T f^t(x_1^n)$. In this new-coordinate system, our goal is to estimate the vector $\gamma^* = \frac{1}{\sqrt{n}}U^T f^*(x_1^n)$. Recalling the alternative form (19) of the gradient recursion, some simple algebra yields that the sequence $\{\gamma^t\}_{t=0}^{\infty}$ evolves as

$$\gamma^{t+1} = \gamma^t + \alpha_t\Lambda\frac{\tilde{w}}{\sqrt{n}} - \alpha_t\Lambda(\gamma^t - \gamma^*),$$

where $\tilde{w} := U^T w$ is a rotated noise vector. Since $\gamma^0 = 0$, unwrapping this recursion then yields $\gamma^t - \gamma^* = \big(I - S^t\big)\frac{\tilde{w}}{\sqrt{n}} - S^t\gamma^*$, where we have made use of the previously defined shrinkage matrices S^t. Using the inequality $\|a+b\|_2^2 \leq 2(\|a\|_2^2 + \|b\|_2^2)$, we find that

$$\begin{aligned}\|\gamma^t - \gamma^*\|_2^2 &\leq \frac{2}{n}\|(I-S^t)\tilde{w}\|_2^2 + 2\|S^t\gamma^*\|_2^2 \\ &= \frac{2}{n}\|(I-S^t)\tilde{w}\|_2^2 + 2\sum_{j=1}^{r}[S^t]_{jj}^2(\gamma_{jj}^*)^2 + 2\sum_{j=r+1}^{n}(\gamma_{jj}^*)^2.\end{aligned}$$

where the equality uses the fact that $\widehat{\lambda}_j = 0$ for all $j \in \{r+1, \ldots, n\}$. Finally, the orthogonality of U implies that $\|\gamma^t - \gamma^*\|_2^2 = \frac{1}{n}\|f^t(x_1^n) - f^*(x_1^n)\|_2^2$, from which the upper bound (20) follows.

B.2 Proof of Lemma 8

Using the definition of S^t and the elementary inequality $1 - u \leq \exp(-u)$, we have

$$[S^t]_{jj}^2 = \Big(\prod_{\tau=0}^{t-1}(1-\alpha_\tau\widehat{\lambda}_j)\Big)^2 \leq \exp(-2\eta_t\widehat{\lambda}_j) \overset{(i)}{\leq} \frac{1}{2e\eta_t\widehat{\lambda}_j},$$

where inequality (i) follows from the fact that $\sup_{u\in\mathbb{R}}\big\{u\exp(-u)\big\} = 1/e$.

Turning to the second set of inequalities, we have $1 - [S^t]_{jj} = 1 - \prod_{\tau=0}^{t-1}(1-\alpha_\tau\widehat{\lambda}_j)$. By induction, it can be shown that

$$1 - [S^t]_{jj} \leq 1 - \max\{0, 1-\eta_t\widehat{\lambda}_j\} = \min\{1, \eta_t\widehat{\lambda}_j\}.$$

As for the remaining claim, we have

$$\begin{aligned}1 - \prod_{\tau=0}^{t-1}(1-\alpha_\tau\widehat{\lambda}_j) &\overset{(i)}{\geq} 1 - \exp(-\eta_t\widehat{\lambda}_i) \\ &\overset{(ii)}{\geq} 1 - (1+\eta_t\widehat{\lambda}_i)^{-1} \\ &= \frac{\eta_t\widehat{\lambda}_i}{1+\eta_t\widehat{\lambda}_i} \\ &\geq \frac{1}{2}\min\{1, \eta_t\widehat{\lambda}_i\},\end{aligned}$$

where step (i) follows from the inequality $1 - u \leq \exp(-u)$; and step (ii) follows from the inequality $\exp(-u) \leq (1+u)^{-1}$, valid for $u > 0$.

Appendix C. Auxiliary Results for Proposition 5

In this appendix, we prove the auxiliary lemmas used in the proof of Proposition 5 on kernel ridge regression.

C.1 Proof of Lemma 12

By definition of the KRR estimate, we have $\left(K + \frac{1}{\nu} I\right) f_\nu(x_1^n) = K y_1^n$. Consequently, some straightforward algebra yields the relation

$$U^T f_\nu(x_1^n) = (I - R^\nu) U^T y_1^n,$$

where the shrinkage matrix R^ν was previously defined (28). The remainder of the proof follows using identical steps to the proof of Lemma 6 with S^t replaced by R^ν.

C.2 Proof of Lemma 13

By definition (28) of the shrinkage matrix, we have $[R^\nu]_{jj}^2 = (1 + \nu\widehat{\lambda}_j)^{-2} \leq \frac{1}{4\nu\widehat{\lambda}_j}$. Moreover, we also have

$$1 - [R^\nu]_{jj} = 1 - (1 + \nu\widehat{\lambda}_j)^{-1} = \frac{\nu\widehat{\lambda}_j}{1 + \nu\widehat{\lambda}_j} \leq \min\{1, \nu\widehat{\lambda}_j\}, \quad \text{and}$$

$$1 - [R^\nu]_{jj} = \frac{\nu\widehat{\lambda}_j}{1 + \nu\widehat{\lambda}_j} \geq \frac{1}{2}\min\{1, \nu\widehat{\lambda}_j\}.$$

Appendix D. Properties of the Empirical Rademacher Complexity

In this section, we prove that the $\widehat{\varepsilon}_n$ lies in the interval $(0, \infty)$, and is unique. Recall that the stopping point $\widehat{T}$ is defined as $\widehat{\varepsilon}_n := \arg\min\left\{\epsilon > 0 \mid \widehat{\mathcal{R}}_K(\epsilon) > \epsilon^2/(2e\sigma)\right\}$. Re-arranging and substituting for $\widehat{\mathcal{R}}_K(\epsilon)$ yields the equivalent expression

$$\widehat{\varepsilon}_n := \arg\min\left\{\epsilon > 0 \mid \sum_{i=1}^n \min\left\{\epsilon^{-2}\widehat{\lambda}_i, 1\right\} > n\epsilon^2/(4e^2\sigma^2)\right\}.$$

Note that $\sum_{i=1}^n \min\left\{\epsilon^{-2}\widehat{\lambda}_i, 1\right\}$ is non-increasing in ϵ while $n\epsilon^2$ is increasing in ϵ. Furthermore when $\epsilon = 0$, $0 = n\epsilon^2 < \sum_{i=1}^n \min\left\{\epsilon^{-2}\widehat{\lambda}_i, 1\right\} > 0$ while for $\epsilon = \infty$, $\sum_{i=1}^n \min\left\{\eta_t\widehat{\lambda}_i, 1\right\} < n\epsilon^2$, recalling that $\eta_t = \sum_{\tau=0}^{t-1}\alpha_\tau$. Hence $\widehat{\varepsilon}_n$ exists. Further, $\widehat{\mathcal{R}}_K(\epsilon)$ is a continuous function of ϵ since it is the sum of n continuous functions, Therefore, the critical radius $\widehat{\varepsilon}_n$ exists, is unique and satisfies the fixed point equation

$$\widehat{\mathcal{R}}_K(\widehat{\varepsilon}_n) = \widehat{\varepsilon}_n^2/(2e\sigma).$$

Finally, we show that the integer $\widehat{T}$ belongs to the interval $[0, \infty)$ and is unique for any valid sequence of step-sizes. Be the definition of $\widehat{T}$ given by the stopping rule (6) and $\widehat{\varepsilon}_n$, we have $\frac{1}{\eta_{\widehat{T}+1}} \leq \widehat{\varepsilon}_n^2 \leq \frac{1}{\eta_{\widehat{T}}}$. Since $\eta_0 = 0$ and $\eta_t \to \infty$ as $t \to \infty$ and $\widehat{\varepsilon}_n \in (0, \infty)$, there exists a unique stopping point $\widehat{T}$ in the interval $[0, \infty)$.

Appendix E. Auxiliary Results for Theorem 2

This appendix is devoted to the proofs of auxiliary lemmas used in the proof for Theorem 2.

E.1 Proof of Lemma 9

Let us write $f_t = \sum_{k=0}^{\infty} \sqrt{\lambda_k} a_k \phi_k$, so that $\|f_t\|_{\mathcal{H}}^2 = \sum_{k=0}^{\infty} a_k^2$. Recall the linear operator $\Phi_X : \ell^2(\mathbb{N}) \to \mathbb{R}^n$ defined element-wise via $[\Phi_X]_{jk} = \phi_j(x_k)$ and the diagonal operator $D : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ with entries $[D]_{jj} = \lambda_j$ and $[D]_{jk} = 0$ for $j \neq k$. By the definition of the gradient update (3), we have the relation $a = \frac{1}{n} D^{1/2} \Phi_X^T K^{-1} f_t(x_1^n)$. Since $\frac{1}{n} \Phi_X D \Phi_X^T = K$,

$$\|f_t\|_{\mathcal{H}}^2 = \|a\|_2^2 = \frac{1}{n} f_t(x_1^n)^T K^{-1} f_t(x_1^n). \tag{30}$$

Recall the eigendecomposition $K = U \Lambda U^T$ with $\Lambda = \operatorname{diag}(\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots \widehat{\lambda}_r)$, and the relation $U^T f^t(x_1^n) = (I - S^t) U^T y_1^n$. Substituting into Equation (30) yields

$$\begin{aligned}
\|f_t\|_{\mathcal{H}}^2 &= \frac{1}{n} (y_1^n)^T U (I - S^t)^2 \Lambda^{-1} U^T y_1^n \\
&\overset{(i)}{=} \frac{1}{n} (f^*(x_1^n) + w)^T U (I - S_t)^2 \Lambda^{-1} U^T (f^*(x_1^n) + w) \\
&= \underbrace{\frac{2}{n} w^T U (I - S_t)^2 \Lambda^{-1} U^T f^*(x_1^n)}_{A_t} + \underbrace{\frac{1}{n} w^T U (I - S_t)^2 \Lambda^{-1} U^T w}_{B_t} \\
&\quad + \underbrace{\frac{1}{n} f^*(x_1^n)^T U (I - S_t)^2 \Lambda^{-1} U^T f^*(x_1^n)}_{C_t}
\end{aligned}$$

where equality (i) follows from the observation equation $y_1^n = f^*(x_1^n) + w$. From Lemma 8, we have $1 - S_{jj}^t \leq 1$, and hence $C_t \leq \frac{1}{n} f^*(x_1^n)^T U \Lambda^{-1} U^T f^*(x_1^n) \overset{(i)}{\leq} 1$, where the last step follows from the analysis in Section 4.1.1.

It remains to derive upper bounds on the random variables A_t and B_t.

E.1.1 Bounding A_t

Since the elements of w are i.i.d, zero-mean and sub-Gaussian with parameter σ, we have $\mathbb{P}[|A_t| \geq 1] \leq 2 \exp(-\frac{n}{2\sigma^2 \nu^2})$, where $\nu^2 := \frac{4}{n} [f^*(x_1^n)]^T U (I - S_t)^4 \Lambda^{-2} U^T f^*(x_1^n)$. Since $(1 - (S_t)_{jj}) \leq 1$, we have

$$\begin{aligned}
\nu^2 \leq \frac{4}{n} f^*(x_1^n)^T U (I - S_t) \Lambda^{-2} U^T f^*(x_1^n) &\leq \frac{4}{n} \sum_{j=1}^{r} \frac{[U^T f^*(x_1^n)]_j^2}{\widehat{\lambda}_j^2} \min(1, \eta_t \widehat{\lambda}_j) \\
&\leq 4 \frac{\eta_t}{n} \sum_{j=1}^{r} \frac{[U^T f^*(x_1^n)]_j^2}{\widehat{\lambda}_j} \\
&\leq 4 \eta_t,
\end{aligned}$$

where the final inequality follows from the analysis in Section 4.1.1.

E.1.2 Bounding B_t

We begin by noting that

$$B_t = \frac{1}{n} \sum_{j=1}^{r} \frac{(1 - S_{jj}^t)^2}{\widehat{\lambda}_j} [U^T w]_j^2 = \frac{1}{n} \operatorname{trace}(U Q U^T, w w^T),$$

where $Q = \mathrm{diag}\{\frac{(1-S_{jj}^t)^2}{\widehat{\lambda}_j},\ j = 1, 2, \ldots r\}$. Consequently, B_t is a quadratic form in zero-mean sub-Gaussian variables, and using the tail bound (27), we have

$$\mathbb{P}\big[|B_t - \mathbb{E}[B_t]| \geq 1\big[\big] \leq \exp(-c\min\{n|||UQU^T|||_{\mathrm{op}}^{-1}, n^2|||UQU^T|||_{\mathrm{F}}^{-2}\})$$

for a universal constant c. It remains to bound $\mathbb{E}[B_t]$, $|||UQU^T|||_{\mathrm{op}}$ and $|||UQU^T|||_{\mathrm{F}}$.

We first bound the mean. Since $\mathbb{E}[ww^T] \preceq \sigma^2 I_{n\times n}$ by assumption, we have

$$\mathbb{E}[B_t] \ \leq \frac{\sigma^2}{n}\,\mathrm{trace}(Q)\frac{1}{n}\sum_{j=1}^r = \ (\frac{(1-S_{jj}^t)^2}{\widehat{\lambda}_j}) \ \leq \frac{\eta_t}{n}\sum_{j=1}^r \min((\eta_t\widehat{\lambda}_j)^{-1}, \eta_t\widehat{\lambda}_j)$$

But by the definition (6) of the stopping rule and the fact that $t \leq \widehat{T}$, we have

$$\frac{\eta_t}{n}\sum_{j=1}^r \min((\eta_t\widehat{\lambda}_j)^{-1}, \eta_t\widehat{\lambda}_j) \leq \eta_t^2\mathcal{R}_K^2(1/\sqrt{\eta_t}) \ \leq \ \frac{1}{\sigma^2},$$

showing that $\mathbb{E}[B_t] \leq 1$.

Turning to the operator norm, we have

$$|||UQU^T|||_{\mathrm{op}} = \max_{j=1,\ldots,r}(\frac{(1-S_{jj}^t)^2}{\widehat{\lambda}_j}) \leq \max_{j=1,\ldots,r}\min(\widehat{\lambda}_j^{-1}, \eta_t^2\widehat{\lambda}_j) \ \leq \eta_t.$$

As for the Frobenius norm, we have

$$\frac{1}{n}|||UQU^T|||_{\mathrm{F}}^2 = \sum_{j=1}^r(\frac{(1-S_{jj}^t)^4}{\widehat{\lambda}_j^2}) \leq \frac{1}{n}\sum_{j=1}^r \min(\widehat{\lambda}_j^{-2}, \eta_t^4\widehat{\lambda}_j^2) \ \leq \frac{\eta_t^3}{n}\sum_{j=1}^r \min(\eta_t^{-3}\widehat{\lambda}_j^{-2}, \eta_t\widehat{\lambda}_j^2)$$

Using the definition of the empirical kernel complexity, we have

$$\frac{1}{n}|||UQU^T|||_{\mathrm{F}}^2 \leq \eta_t^3\mathcal{R}_K^2(1/\sqrt{\eta_t}) \ \leq \frac{\eta_t}{\sigma^2},$$

where the final inequality holds for $t \leq \widehat{T}$, using the definition of the stopping rule.

Putting together the pieces, we have shown that

$$\mathbb{P}[|B_t| \geq 2 \quad \text{or} \quad |A_t| \geq 1] \leq \exp(-cn/\eta_t)$$

for all $t \leq \widehat{T}$. Since $\frac{1}{\eta_t} \geq \widehat{\varepsilon}_n^2$ for any $t \leq \widehat{T}$, the claim follows.

E.2 Proof of Lemma 11

In this section, we need to show that $\widehat{\varepsilon}_n \leq \varepsilon_n$. Recall that $\widehat{\varepsilon}_n$ and ε_n satisfy

$$\widehat{\mathcal{R}}_K(\widehat{\varepsilon}_n) = \frac{\widehat{\varepsilon}_n^2}{2e\sigma} \ \text{ and } \ \mathcal{R}_{\mathbb{K}}(\varepsilon_n) = \frac{\varepsilon_n^2}{40\sigma}.$$

It suffices to prove that $\widehat{\mathcal{R}}_K(\varepsilon_n) \leq \frac{\varepsilon_n^2}{2e\sigma}$ using the definition of $\widehat{\varepsilon}_n$.

In order to prove the claim, we define the random variables

$$\widehat{Z}_n(w,t) := \sup_{\substack{\|g\|_{\mathcal{H}} \leq 1 \\ \|g\|_n \leq t}} \big|\frac{1}{n}\sum_{i=1}^n w_i g(x_i)\big|, \quad \text{and} \quad Z_n(w,t) := \mathbb{E}_x\Big[\sup_{\substack{\|g\|_{\mathcal{H}} \leq 1 \\ \|g\|_2 \leq t}} \big|\frac{1}{n}\sum_{i=1}^n w_i g(x_i)\big|\Big],$$

where $w_i \sim N(0,1)$ are i.i.d. standard normal, as well as the associated (deterministic) functions

$$\widehat{\mathcal{Q}}_n(t) \ := \mathbb{E}_w\big[\widehat{Z}_n(w;t)\big] \quad \text{and} \quad \mathcal{Q}_n(t) \ := \mathbb{E}_w\big[Z_n(w;t)\big].$$

By results of Mendelson (2002), there are universal constants $0 < c_\ell \leq c_u$ such that for all $t^2 \geq 1/n$, we have

$$c_\ell \mathcal{R}_{\mathbb{K}}(t) \leq \mathcal{Q}_n(t) \leq c_u \mathcal{R}_{\mathbb{K}}(t), \quad \text{and} \quad c_\ell \widehat{\mathcal{R}}_K(t) \leq \widehat{\mathcal{Q}}_n(t) \leq c_u \widehat{\mathcal{R}}_K(t).$$

We first appeal to the concentration of Lipschitz functions for Gaussian random variables to show that $\widehat{Z}_n(w,t)$ and $Z_n(w,t)$ are concentrated around their respective means. For any $t > 0$ and vectors $w, w' \in \mathbb{R}^n$, we have

$$|\widehat{Z}_n(w,t) - \widehat{Z}_n(w',t)| \leq \sup_{\substack{\|g\|_n \leq t \\ \|g\|_{\mathcal{H}} \leq 1}} \frac{1}{n}|\sum_{i=1}^n (w_i - w_i')g(x_i)| \leq \frac{t}{\sqrt{n}}\|w - w'\|_2,$$

showing that $w \mapsto \widehat{Z}_n(w,t)$ is $\frac{t}{\sqrt{n}}$-Lipschitz with respect to the ℓ_2 norm. A similar calculation for $w \mapsto Z_n(w,t)$ shows that

$$|\mathbb{E}_x[\widehat{Z}_n(w,t)] - \mathbb{E}_x[\widehat{Z}_n(w',t)]| \leq \mathbb{E}_x[\sup_{\substack{\|g\|_2 \leq t \\ \|g\|_{\mathcal{H}} \leq 1}} \frac{1}{n}|\sum_{i=1}^n (w_i - w_i')g(x_i)|] \leq \frac{t}{\sqrt{n}}\|w - w'\|_2,$$

so that it is also Lipschitz $\frac{t}{\sqrt{n}}$. Consequently, standard concentration results (Ledoux, 2001) imply that

$$\begin{aligned} \mathbb{P}\big[|\widehat{Z}_n(w,t) - \widehat{\mathcal{Q}}_n(t)| \geq t_0\big] &\leq 2\exp\Big(-\frac{nt_0^2}{2t^2}\Big), \quad \text{and} \\ \mathbb{P}\big[|Z_n(w,t) - \mathcal{Q}_n(t)| \geq t_0\big] &\leq 2\exp\Big(-\frac{nt_0^2}{2t^2}\Big). \end{aligned} \tag{31}$$

Now let us condition on the two events

$$\mathcal{A}(t,t_0) := \{|\widehat{Z}_n(w,t) - \widehat{\mathcal{Q}}_n(t)| \leq t_0\}, \quad \text{and} \quad \mathcal{A}'(t,t_0) := \{|Z_n(w,t) - \mathcal{Q}_n(t)| \leq t_0\}.$$

We then have

$$\widehat{\mathcal{R}}_K(\varepsilon_n) \overset{(a)}{\leq} \widehat{Z}_n(w,\varepsilon_n) + \frac{\varepsilon_n^2}{4e\sigma} \overset{(b)}{\leq} Z_n(w,2\varepsilon_n) + \frac{\varepsilon_n^2}{4e\sigma} \overset{(c)}{\leq} 2\mathcal{R}_{\mathbb{K}}(\varepsilon_n) + \frac{3\varepsilon_n^2}{8e\sigma} \overset{(d)}{\leq} \frac{\varepsilon_n^2}{2e\sigma},$$

where inequality (a) follows the first bound in Equation (31) with $t_0 = \frac{\varepsilon_n^2}{4e\sigma}$ and $t = \varepsilon_n^2$, inequality (b) follows from Lemma 10 with $t = \varepsilon_n$, inequality (c) follows from the second bound (31) with $t_0 = \frac{\varepsilon_n^2}{8e\sigma}$ and $t = \varepsilon_n^2$, and inequality (d) follows from the definition of ε_n. Since the events $\mathcal{A}(t,t_0)$ and $\mathcal{A}'(t,t_0)$ hold with the stated probability, the claim follows.

References

R. S. Anderssen and P. M. Prenter. A formal comparison of methods proposed for the numerical solution of first kind integral equations. *Jour. Australian Math. Soc. (Ser. B)*, 22:488–500, 1981.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

A. R. Barron, A. Cohen, W. Dahmen, and R. A. DeVore. Approximation and learning by greedy algorithms. *Annals of Statistics*, 36(1):64–94, 2008.

P. Bartlett and M. Traskin. Adaboost is consistent. *Journal of Machine Learning Research*, 8:2347–2368, 2007.

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

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *J. Complexity*, 23:52–72, 2007.

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

G. Blanchard and M. Kramer. Optimal learning rates for kernel conjugate gradient regression. In *Proceedings of the NIPS Conference*, 2010.

P. Buhlmann and B. Yu. Boosting with L^2 loss: Regression and classification. *Journal of American Statistical Association*, 98:324–340, 2003.

A. Caponetto and Y. Yao. Adaptation for regularization operators in learning theory. Technical Report CBCL Paper #265/AI Technical Report #063, Massachusetts Institute of Technology, September 2006.

A. Caponetto and Y. Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8(2):161–183, 2010.

A. Caponneto. Optimal rates for regularization operators in learning theory. Technical Report CBCL Paper #264/AI Technical Report #062, Massachusetts Institute of Technology, September 2006.

P. Drineas and M. W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

J.H. Friedman and B. Popescu. Gradient directed regularization. Technical report, Stanford University, 2004.

C. Gu. *Smoothing Spline ANOVA Models*. Springer Series in Statistics. Springer, New York, NY, 2002.

M. G. Gu and H. T. Zhu. Maximum likelihood estimation by markov chain monte carlo approximation. *J. R. Statist. Soc. B*, 63:339–355, 2001.

P. Hall and J.S. Marron. On variance estimation in nonparametric regression. *Biometrika*, 77:415–419, 1990.

W. Jiang. Process consistency for adaboost. *Annals of Statistics*, 32:13–29, 2004.

G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Jour. Math. Anal. Appl.*, 33:82–95, 1971.

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

M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

L. Mason, J. Baxter, P., and M. Frean. Boosting algorithms as gradient descent. In *Neural Information Processing Systems (NIPS)*, December 1999.

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

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209:415–446, 1909.

N. Morgan and H. Bourlard. Generalization and parameter estimation in feedforward nets: Some experiments. In *Proceedings of Neural Information Processing Systems*, 1990.

L. Orecchia and M. W. Mahoney. Implementing regularization implicitly via approximate eigenvector computation. In *ICML '11*, 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.

S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman Scientific & Technical, Harlow, UK, 1988.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151, 1981.

C. J. Stone. Additive regression and other nonparametric models. *Annals of Statistics*, 13 (2):689–705, 1985.

O. N. Strand. Theory and methods related to the singular value expansion and Landweber's iteration for integral equations of the first kind. *SIAM J. Numer. Anal.*, 11:798–825, 1974.

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

E. De Vito, S. Pereverzyev, and L. Rosasco. Adaptive kernel methods using the balancing principle. *Foundations of Computational Mathematics*, 10(4):455–479, 2010.

G. Wahba. Three topics in ill-posed problems. In M. Engl and G. Groetsch, editors, *Inverse and ill-posed problems*, pages 37–50. Academic Press, 1987.

G. Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PN, 1990.

H. L. Weinert, editor. *Reproducing Kernel Hilbert Spaces : Applications in Statistical Signal Processing*. Hutchinson Ross Publishing Co., Stroudsburg, PA, 1982.

F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *Annals of Probability*, 1(6): 1068–1070, 1973.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

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

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

T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *Annal of Statistics*, 33:1538–1579, 2005.