Early Stopping for Kernel Boosting Algorithms: A General Analysis With Localized Complexities

Yuting Wei, Fanny Yang, and Martin J. Wainwright, *Senior Member, IEEE*

***Abstract*— Early stopping of iterative algorithms is a widely used form of regularization in statistics, commonly used in conjunction with boosting and related gradient-type algorithms. Although consistency results have been established in some settings, such estimators are less well-understood than their analogues based on penalized regularization. In this paper, for a relatively broad class of loss functions and boosting algorithms (including L^2-boost, LogitBoost, and AdaBoost, among others), we exhibit a direct connection between the performance of a stopped iterate and the localized Gaussian complexity of the associated function class. This connection allows us to show that the local fixed point analysis of Gaussian or Rademacher complexities, now standard in the analysis of penalized estimators, can be used to derive optimal stopping rules. We derive such stopping rules in detail for various kernel classes and illustrate the correspondence of our theory with practice for Sobolev kernel classes.**

***Index Terms*— Boosting, kernel, early stopping, regularization, localized complexities.**

I. Introduction

WHILE non-parametric models offer great flexibility, they can also lead to overfitting, and poor generalization as a consequence. For this reason, it is well understood that procedures for fitting non-parametric models must involve some form of regularization. When models are fit via a form of empirical risk minimization, the most classical form of regularization is based on adding some type of penalty to the objective function. An alternative form of regularization is based on the principle of *early stopping*, in which an iterative algorithm is run for a pre-specified number of steps, and terminated prior to convergence.

Manuscript received March 13, 2018; revised February 23, 2019; accepted May 30, 2019. Date of publication July 9, 2019; date of current version September 13, 2019. This work was supported in part by the DOD Advanced Research Projects Agency under Grant W911NF-16-1-0552, in part by the National Science Foundation under Grant NSF-DMS-1612948, and in part by the Office of Naval Research under Grant DOD-ONR-N00014.

Y. Wei was with the Department of Statistics, University of California at Berkeley, Berkeley, CA 94720 USA. She is now with the Statistics Department, Stanford University, Stanford, CA 94305 USA. (e-mail: ytwei@berkeley.edu).

F. Yang was with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA. She is now with the Institute for Theoretical Studies, ETH Zurich, 8092 Zurich, Switzerland, and also with the Department of Statistics, Stanford University, Stanford, CA 94305 USA (e-mail: fanny-yang@berkeley.edu).

M. J. Wainwright is with the Department of Statistics, University of California at Berkeley, Berkeley, CA 94720 USA, and also with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: wainwrig@berkeley.edu).

Communicated by C. Caramanis, Associate Editor for Machine Learning.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2019.2927563

While the basic idea of early stopping is fairly old (e.g., [1], [33], [37]), recent years have witnessed renewed interests in its properties, especially in the context of boosting algorithms and neural network training (e.g., [13], [27]). Over the past decade, a line of work has yielded some theoretical insight into early stopping, including works on classification error for boosting algorithms [4], [14], [19], [25], [41], [42], L^2-boosting algorithms for regression [8], [9], and similar gradient algorithms in reproducing kernel Hilbert spaces (e.g. [11], [12], [28], [36], [41]). A number of these papers establish consistency results for particular forms of early stopping, guaranteeing that the procedure outputs a function with statistical error that converges to zero as the sample size increases. On the other hand, there are relatively few results that actually establish *rate optimality* of an early stopping procedure, meaning that the achieved error matches known statistical minimax lower bounds. To the best of our knowledge, Bühlmann and Yu [9] were the first to prove optimality for early stopping of L^2-boosting as applied to spline classes, albeit with a rule that was not computable from the data. Subsequent work by Raskutti *et al.* [28] refined this analysis of L^2-boosting for kernel classes and first established an important connection to the localized Rademacher complexity; see also the related work [10], [29], [41] with rates for particular kernel classes.

More broadly, relative to our rich and detailed understanding of regularization via penalization (e.g., see the books [18], [34], [35], [39] and papers [3], [21] for details), our understanding of early stopping regularization is not as well developed. Intuitively, early stopping should depend on the same bias-variance tradeoffs that control estimators based on penalization. In particular, for penalized estimators, it is now well-understood that complexity measures such as the *localized Gaussian width*, or its Rademacher analogue, can be used to characterize their achievable rates [3], [21], [34], [39]. Is such a general and sharp characterization also possible in the context of early stopping?

The main contribution of this paper is to answer this question in the affirmative for the early stopping of boosting algorithms as applied to various regression and classification problems involving functions in reproducing kernel Hilbert spaces (RKHS). A standard way to obtain a good estimator or classifier is through minimizing some penalized form of loss functions of which the method of kernel ridge regression [38] is a popular choice. Instead, we consider an iterative update involving the kernel that is derived from a greedy update. Borrowing tools from empirical process theory, we are able to characterize the "size" of the effective function space explored by taking T steps, and then to connect the resulting estimation error naturally to the notion of localized Gaussian width defined with respect to this effective function space. This leads to a principled analysis for a broad class of loss functions used in practice, including the loss functions that underlie the L^2-boost, LogitBoost and AdaBoost algorithms, among other procedures.

The remainder of this paper is organized as follows. In Section II, we provide background on boosting methods and reproducing kernel Hilbert spaces, and then introduce the updates studied in this paper. Section III is devoted to statements of our main results, followed by a discussion of their consequences for particular function classes in Section IV. We provide simulations that confirm the practical effectiveness of our stopping rules, and show close agreement with our theoretical predictions. In Section V, we provide the proofs of our main results, with certain more technical aspects deferred to the appendices.

II. Background and Problem Formulation

The goal of prediction is to estimate a function that maps *covariates* $x \in \mathcal{X}$ to *responses* $y \in \mathcal{Y}$. In a regression problem, the responses are typically real-valued, whereas in a classification problem, the responses take values in a finite set. In this paper, we study both regression ($\mathcal{Y} = \mathbb{R}$) and classification problems (e.g., $\mathcal{Y} = \{-1, +1\}$ in the binary case). Our primary focus is on the case of *fixed design*, in which we observe a collection of n pairs of the form $\{(x_i, Y_i)\}_{i=1}^n$, where each $x_i \in \mathcal{X}$ is a fixed covariate, whereas $Y_i \in \mathcal{Y}$ is a random response drawn independently from a distribution $\mathbb{P}_{Y|x_i}$ which depends on x_i. Later in the paper, we also discuss the consequences of our results for the case of random design, where the (X_i, Y_i) pairs are drawn in an i.i.d. fashion from the joint distribution $\mathbb{P} = \mathbb{P}_X \mathbb{P}_{Y|X}$ for some distribution $\mathbb{P}_X$ on the covariates.

In this section, we provide some necessary background on a gradient-type algorithm which is often referred to as *boosting* algorithm. We also discuss briefly about the reproducing kernel Hilbert spaces before turning to a precise formulation of the problem that is studied in this paper.

A. Boosting and Early Stopping

Consider a cost function $\phi : \mathbb{R} \times \mathbb{R} \to [0, \infty)$, where the non-negative scalar $\phi(y, \theta)$ denotes the cost associated with predicting θ when the true response is y. Some common examples of loss functions ϕ that we consider in later sections include:

- the *least-squares loss* $\phi(y, \theta) := \frac{1}{2}(y - \theta)^2$ that underlies L^2-boosting [9],
- the *logistic regression loss* $\phi(y, \theta) = \ln(1 + e^{-y\theta})$ that underlies the LogitBoost algorithm [15], [16], and
- the *exponential loss* $\phi(y, \theta) = \exp(-y\theta)$ that underlies the AdaBoost algorithm [14].

The least-squares loss is typically used for regression problems (e.g., [9], [11], [12], [28], [36], [41]), whereas the latter two losses are frequently used in the setting of binary classification (e.g., [14], [16], [25]).

Given some loss function ϕ, we define the *population cost functional* $f \mapsto \mathcal{L}(f)$ via

$$\mathcal{L}(f) := \mathbb{E}_{Y_1^n}\Big[\frac{1}{n}\sum_{i=1}^n \phi\big(Y_i, f(x_i)\big)\Big]. \tag{1}$$

Note that with the covariates $\{x_i\}_{i=1}^n$ fixed, the functional $\mathcal{L}$ is a non-random object. Given some function space $\mathscr{F}$, the optimal function[1] minimizes the population cost functional—that is

$$f^* := \arg\min_{f \in \mathscr{F}} \mathcal{L}(f). \tag{2}$$

As a standard example, when we adopt the least-squares loss $\phi(y, \theta) = \frac{1}{2}(y - \theta)^2$, the population minimizer f^* corresponds to the conditional expectation $x \mapsto \mathbb{E}[Y \mid x]$.

Since we do not have access to the population distribution of the responses however, the computation of f^* is impossible. Given our samples $\{Y_i\}_{i=1}^n$, we consider instead some procedure applied to the *empirical loss*

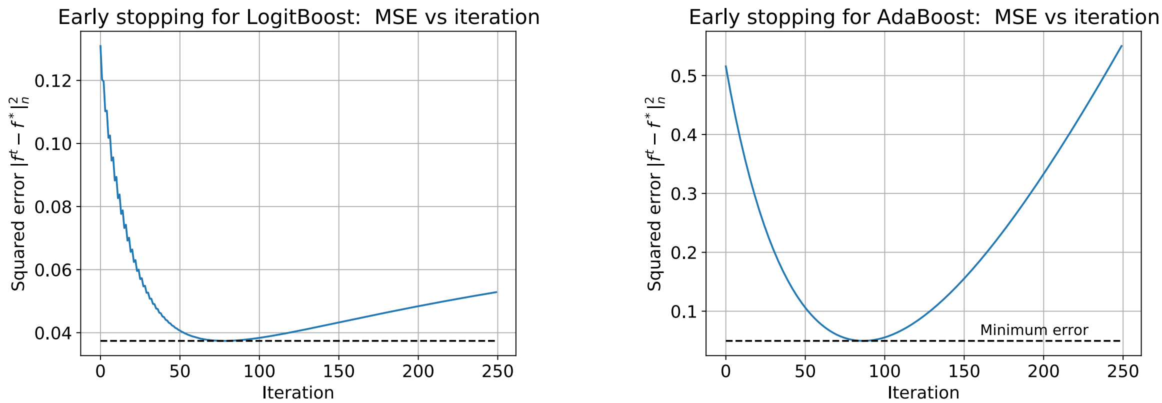
$$\mathcal{L}_n(f) := \frac{1}{n}\sum_{i=1}^n \phi(Y_i, f(x_i)), \tag{3}$$

where the population expectation has been replaced by an empirical expectation. For example, when $\mathcal{L}_n$ corresponds to the log likelihood of the samples with $\phi(Y_i, f(x_i)) = \log[\mathbb{P}(Y_i; f(x_i))]$, direct unconstrained minimization of $\mathcal{L}_n$ would yield the maximum likelihood estimator.

It is well-known that direct minimization of $\mathcal{L}_n$ over a sufficiently rich function class $\mathscr{F}$ may lead to overfitting. There are various ways to mitigate this phenomenon, among which the most classical method is to minimize the sum of the empirical loss with a penalty regularization term. Adjusting the weight on the regularization term allows for trade-off between fit to the data, and some form of regularity or smoothness in the fit. The behavior of such penalized of regularized estimation methods is now quite well understood (for instance, see the books [18], [34], [35], [39] and papers [3], [21] for more details).

In this paper, we study a form of *algorithmic regularization*, based on applying a gradient-type algorithm to $\mathcal{L}_n$ but then stopping it "early"—that is, after some fixed number of steps. Such methods are often referred to as *boosting algorithms*, since they are based on improving the fit of a function via a sequence of additive updates (e.g., see the papers [6], [7], [14], [30], [31]). Many boosting algorithms, among them AdaBoost [14], L^2-boosting [9] and LogitBoost [15], [16], can be understood as forms of functional gradient methods [16], [25]; see the survey paper [8] for further background on boosting. The way in which the number of steps is chosen is referred to as a stopping rule, and the overall procedure is referred to as *early stopping* of a boosting algorithm.

[1] Our assumptions guarantee uniqueness of f^* with regard to the design points $\{x_i\}_{i=1}^n$.



(a) (b)

Fig. 1. Plots of the squared error $\|f^t - f^*\|_n^2 = \frac{1}{n}\sum_{i=1}^n (f^t(x_i) - f^*(x_i))^2$ versus the iteration number t for (a) LogitBoost using a first-order Sobolev kernel (b) AdaBoost using the same first-order Sobolev kernel $\mathbb{K}(x, x') = 1 + \min(x, x')$ which generates a class of Lipschitz functions (splines of order one). Both plots correspond to a sample size $n = 100$.

In more detail, a broad class of boosting algorithms [25] generate a sequence $\{f^t\}_{t=0}^{\infty}$ via updates of the form

$$f^{t+1} = f^t - \alpha^t g^t \quad \text{with} \quad g^t \propto \arg\max_{\|d\|_{\mathscr{F}} \leq 1} \langle \nabla \mathcal{L}_n(f^t),\, d(x_1^n) \rangle, \tag{4}$$

where the constraint $\|d\|_{\mathscr{F}} \leq 1$ defines the unit ball in a given function class $\mathscr{F}$, and $d(x_1^n) := (d(x_1), d(x_2), \ldots, d(x_n)) \in \mathbb{R}^n$, $\nabla \mathcal{L}_n(f) \in \mathbb{R}^n$ denotes the gradient taken with respect to the vector $\big(f(x_1), \ldots, f(x_n)\big)$, and $\langle h,\, g\rangle$ is the usual inner product between vectors $h, g \in \mathbb{R}^n$. Here the scalar $\{\alpha^t\}_{t=0}^{\infty}$ is a sequence of step sizes chosen by the user. For non-decaying step sizes and a convex objective $\mathcal{L}_n$, running this procedure for an infinite number of iterations will lead to a minimizer of the empirical loss, thus causing overfitting. In order to illustrate this phenomenon, Figure 1 provides plots of the squared error $\|f^t - f^*\|_n^2 := \frac{1}{n}\sum_{i=1}^n \big(f^t(x_i) - f^*(x_i)\big)^2$ versus the iteration number, for LogitBoost in panel (a) and AdaBoost in panel (b). See Section IV-B for more details on exactly how these experiments were conducted.

In the plots in Figure 1, the dotted line indicates the minimum mean-squared error ρ_n^2 over all iterates of that particular run of the algorithm. Both plots are qualitatively similar, illustrating the existence of a "good" number of iterations to take, after which the MSE greatly increases. Hence a natural problem is to decide at what iteration T to stop such that the iterate f^T satisfies bounds of the form

$$\mathcal{L}(f^T) - \mathcal{L}(f^*) \precsim \rho_n^2 \quad \text{and} \quad \|f^T - f^*\|_n^2 \precsim \rho_n^2 \tag{5}$$

with high probability. Here $f(n) \precsim g(n)$ indicates that $f(n) \leq cg(n)$ for some universal constant $c \in (0, \infty)$. The main results of this paper provide a stopping rule T for which bounds of the form (5) do in fact hold with high probability over the randomness in the observed responses.

Moreover, as shown by our later results, under suitable regularity conditions, the expectation of the minimum squared error ρ_n^2 is proportional to the *statistical minimax risk* $\inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}[\mathcal{L}(\widehat{f}) - \mathcal{L}(f)]$, where the infimum is taken over all possible estimators $\widehat{f}$. Note that the minimax risk provides a fundamental lower bound on the performance of any estimator uniformly over the function space $\mathcal{F}$. Coupled with our stopping time guarantee (5), we are guaranteed that our estimate achieves the minimax risk up to constant factors. As a result, our bounds are unimprovable in general (see Corollary 2).

B. Reproducing Kernel Hilbert Spaces

The analysis of this paper focuses on algorithms with the update (4) when the function class $\mathscr{F}$ is a reproducing kernel Hilbert space, or RKHS for short. Here we provide a brief introduction, directing the reader to various books for more background [5], [17], [32], [38], [39]. An RKHS $\mathscr{H}$ is a function space consisting of functions mapping a domain $\mathcal{X}$ to the real line $\mathbb{R}$. Any RKHS is defined by a bivariate symmetric *kernel function* $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is required to be positive semidefinite, i.e. for any integer $N \geq 1$ and a collection of points $\{x_j\}_{j=1}^N$ in $\mathcal{X}$, the matrix $[\mathbb{K}(x_i, x_j)]_{ij} \in \mathbb{R}^{N \times N}$ is positive semidefinite.

The associated RKHS is the completion of all linear combinations $\sum_{i=1}^n \alpha_i \mathbb{K}(\cdot, x_i)$, where $n \in \mathbb{N}$, $\{x_j\}_{j=1}^n$ is some collection of points in $\mathcal{X}$, and $\{\omega_j\}_{j=1}^{\infty}$ is a real-valued sequence, with respect to the scalar product

$$\langle \sum_{i=1}^n \alpha_i \mathbb{K}(\cdot, x_i), \sum_{j=1}^m \alpha_j' \mathbb{K}(\cdot, x_j') \rangle_{\mathscr{H}} = \sum_{i,j} \alpha_i \alpha_j' \mathbb{K}(x_i, x_j').$$

For each $x \in \mathcal{X}$, the function $\mathbb{K}(\cdot, x)$ belongs to $\mathscr{H}$, and satisfies the reproducing relation

$$\langle f,\, \mathbb{K}(\cdot, x)\rangle_{\mathscr{H}} = f(x) \quad \text{for all } f \in \mathscr{H}. \tag{6}$$

Moreover, when the covariates X_i are drawn i.i.d. from a distribution $\mathbb{P}_X$ with domain $\mathcal{X}$, suppose that $\mathcal{X}$ is compact, the kernel function $\mathbb{K}$ is continuous and positive semidefinite, and satisfies the bound

$$\int_{\mathcal{X} \times \mathcal{X}} \mathbb{K}(x, x')\, d\mathbb{P}_X(x) d\mathbb{P}_X(x') < \infty.$$

Then we can invoke Mercer's theorem, which guarantees that the kernel function can be represented as

$$\mathbb{K}(x, x') = \sum_{k=1}^{\infty} \mu_k \phi_k(x)\phi_k(x'), \tag{7}$$

where $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ are the ordered *eigenvalues* of the kernel function $\mathbb{K}$ and $\{\phi_k\}_{k=1}^{\infty}$ are eigenfunctions of $\mathbb{K}$ which form an orthonormal basis of $L^2(\mathcal{X}, \mathbb{P}_X)$ with the inner product $\langle f, g\rangle := \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\mathbb{P}_X(x)$.

Throughout this paper, we assume that the kernel function is uniformly bounded, meaning that there is a constant L such that $\sup_{x\in\mathcal{X}} \mathbb{K}(x, x) \leq L$. Such a boundedness condition holds for many kernels used in practice, including the Gaussian, Laplacian, Sobolev, other types of spline kernels, as well as any trace class kernel with trigonometric eigenfunctions. By rescaling the kernel as necessary, we may assume without loss of generality that $L = 1$. As a consequence, for any function f such that $\|f\|_{\mathscr{H}} \leq r$, we have by the reproducing relation that

$$\|f\|_\infty = \sup_x \langle f, \mathbb{K}(\cdot, x)\rangle_{\mathscr{H}} \leq \|f\|_{\mathscr{H}} \sup_x \|\mathbb{K}(\cdot, x)\|_{\mathscr{H}} \leq r.$$

Given samples $\{(x_i, y_i)\}_{i=1}^n$, by the representer theorem [20], it is sufficient to restrict ourselves to the linear subspace $\mathscr{H}_n = \overline{\text{span}}\{\mathbb{K}(\cdot, x_i)\}_{i=1}^n$, for which all $f \in \mathscr{H}_n$ can be expressed as

$$f = \frac{1}{\sqrt{n}} \sum_{i=1}^n \omega_i \mathbb{K}(\cdot, x_i) \tag{8}$$

for some coefficient vector $\omega \in \mathbb{R}^n$. Among those functions which achieve the infimum in expression (1), let us define f^* as the one with the minimum Hilbert norm. This definition is equivalent to restricting f^* to be in the linear subspace $\mathscr{H}_n$.

C. Boosting in Kernel Spaces

Given a collection of n covariates $\{x_i\}_{i=1}^n$, let us define the *normalized kernel matrix* $K \in \mathbb{R}^{n\times n}$ with entries $K_{ij} = \mathbb{K}(x_i, x_j)/n$. Recall that we can restrict the minimization of $\mathcal{L}_n$ and $\mathcal{L}$ from $\mathscr{H}$ to the subspace $\mathscr{H}_n$ without loss of generality. Consequently, using the expression (8), the function value vectors $f(x_1^n) := (f(x_1), \ldots, f(x_n))$ can be written in the form $f(x_1^n) = \sqrt{n}K\omega$. Moreover, if the null space of K is empty, then there is a one-to-one correspondence between n-dimensional vectors $f(x_1^n) \in \mathbb{R}^n$ and the corresponding function $f \in \mathscr{H}_n$. In that case, minimization of an empirical loss in the subspace $\mathscr{H}_n$ essentially becomes the n-dimensional problem of fitting a response vector y over the set range(K). In the sequel, all updates will thus be performed on the function value vectors $f(x_1^n)$.

With a change of variable $d(x_1^n) = \sqrt{n}\sqrt{K}z$ we then have

$$\begin{aligned} d^t(x_1^n) &:= \underset{\substack{\|d\|_{\mathscr{H}}\leq 1\\ d\in\text{range}(K)}}{\arg\max}\ \langle \nabla\mathcal{L}_n(f^t),\ d(x_1^n)\rangle \\ &= \frac{\sqrt{n}K\nabla\mathcal{L}_n(f^t)}{\sqrt{\nabla\mathcal{L}_n(f^t)K\nabla\mathcal{L}_n(f^t)}}. \end{aligned}$$

In this paper, we study the choice $g^t = \langle\nabla\mathcal{L}_n(f^t),\ d^t(x_1^n)\rangle d^t$ in the boosting update (4), so that the function value iterates take the form

$$f^{t+1}(x_1^n) = f^t(x_1^n) - \alpha n K \nabla\mathcal{L}_n(f^t), \tag{9}$$

where $\alpha > 0$ is a constant stepsize choice. With the initialization $f^0(x_1^n) = 0$, all iterates $f^t(x_1^n)$ remain in the range space of K.

In this paper, we consider the following three error measures for an estimator $\widehat{f}$:

$$\begin{aligned} L^2(\mathbb{P}_n) \text{ norm:}\quad & \|\widehat{f} - f^*\|_n^2 = \frac{1}{n}\sum_{i=1}^n \big(\widehat{f}(x_i) - f^*(x_i)\big)^2, \\ L^2(\mathbb{P}_X) \text{ norm:}\quad & \|\widehat{f} - f^*\|_2^2 := \mathbb{E}\big(\widehat{f}(X) - f^*(X)\big)^2, \\ \text{Excess risk:}\quad & \mathcal{L}(\widehat{f}) - \mathcal{L}(f^*). \end{aligned}$$

Here the expectation defining the $L^2(\mathbb{P}_X)$-norm is taken over random covariates X that are independent of the samples (X_i, Y_i) used to form the estimate $\widehat{f}$. Our goal is to propose a stopping time T such that the averaged function $\widehat{f} = \frac{1}{T}\sum_{t=1}^T f^t$ satisfies bounds of the type (5). We begin our analysis by focusing on the empirical $L^2(\mathbb{P}_n)$ error, but as we will see in Corollary 1, bounds on the empirical error are easily transformed to bounds on the population $L^2(\mathbb{P}_X)$ error. Importantly, we exhibit such bounds with a statistical error term δ_n that is specified by the *localized Gaussian complexity* of the kernel class.

III. Main Results

We now turn to the statement of our main results, beginning with the introduction of some regularity assumptions.

A. Assumptions

Recall from our earlier set-up that we differentiate between the empirical loss function $\mathcal{L}_n$ in expression (3), and the population loss $\mathcal{L}$ in expression (1). Apart from assuming differentiability of both functions, all of our remaining conditions are imposed on the population loss. Such conditions at the population level are weaker than their analogues at the empirical level.

For a given radius $r > 0$, let us define the Hilbert ball around the optimal function f^* as

$$\mathbb{B}_{\mathscr{H}}(f^*, r) := \{f \in \mathscr{H} \mid \|f - f^*\|_{\mathscr{H}} \leq r\}. \tag{10}$$

Our analysis makes particular use of this ball defined for the radius $C_{\mathscr{H}}^2 := 2\max\{\|f^*\|_{\mathscr{H}}^2,\ 32, \sigma^2\}$ where the effective noise level σ is defined in the sequel.

We assume that the population loss is m-strongly convex and M-smooth over $\mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$, meaning that the ***m*-*M*-condition**

$$\begin{aligned} \frac{m}{2}\|f - g\|_n^2 &\leq \mathcal{L}(f) - \mathcal{L}(g) - \langle\nabla\mathcal{L}(g),\ f(x_1^n) - g(x_1^n)\rangle \\ &\leq \frac{M}{2}\|f - g\|_n^2 \end{aligned}$$

holds for all $f, g \in \mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$ and all design points $\{x_i\}_{i=1}^n$. In addition, we assume that the function ϕ is M-Lipschitz in its second argument over the interval

$$\theta \in \left[\min_{i\in[n]} f^*(x_i) - 2C_{\mathscr{H}}, \max_{i\in[n]} f^*(x_i) + 2C_{\mathscr{H}}\right].$$

To be clear, here $\nabla\mathcal{L}(g)$ denotes the vector in $\mathbb{R}^n$ obtained by taking the gradient of $\mathcal{L}$ with respect to the vector $g(x_1^n)$. It can be verified by a straightforward computation that when $\mathcal{L}$ is induced by the least-squares cost $\phi(y,\theta) = \frac{1}{2}(y-\theta)^2$, the m-M-condition holds for $m = M = 1$. The logistic and exponential loss satisfy this condition (see the supplementary material), where it is key that we have imposed the condition *only locally* on the ball $\mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$.

In addition to the least-squares cost, our theory also applies to losses $\mathcal{L}$ induced by scalar functions ϕ that satisfy the ϕ'**-boundedness** condition:

$$\max_{i=1,\dots,n} \left|\frac{\partial\phi(y,\theta)}{\partial\theta}\right|_{\theta=f(x_i)} \leq B \quad \text{for all } f \in \mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}}) \text{ and } y \in \mathcal{Y}.$$

This condition holds with $B = 1$ for the logistic loss for all $\mathcal{Y}$, and $B = \exp(2.5C_{\mathscr{H}})$ for the exponential loss for binary classification with $\mathcal{Y} = \{-1, 1\}$, using our kernel boundedness condition. Note that whenever this condition holds with some finite B, we can always rescale the scalar loss ϕ by $1/B$ so that it holds with $B = 1$, and we do so in order to simplify the statement of our results.

B. Upper Bound in Terms of Localized Gaussian Width

Our upper bounds involve a complexity measure known as the localized Gaussian width. In general, Gaussian widths are widely used to obtain risk bounds for least-squares and other types of M-estimators. In our case, we consider Gaussian complexities for "localized" sets of the form

$$\mathcal{E}_n(\delta, 1) := \Big\{f - g \mid f, g \in \mathscr{H} \quad \text{such that} \quad \|f-g\|_{\mathscr{H}} \leq 1 \quad \text{and} \quad \|f-g\|_n \leq \delta\Big\}. \tag{11}$$

The Gaussian complexity localized at scale δ is given by

$$\mathcal{G}_n\big(\mathcal{E}_n(\delta,1)\big) := \mathbb{E}\Big[\sup_{g\in\mathcal{E}_n(\delta,1)} \frac{1}{n}\sum_{i=1}^n w_i g(x_i)\Big], \tag{12}$$

where $(w_1, \dots, w_n)$ denotes an i.i.d. sequence of standard Gaussian variables.

An essential quantity in our theory is specified by a certain fixed point equation that is now standard in empirical process theory [3], [21], [28], [34], [39]. Let us define the *effective noise level*

$$\sigma := \begin{cases} \min\Big\{t \mid \max_{i=1,\dots,n} \mathbb{E}[e^{((Y_i - f^*(x_i))^2/t^2)}] < \infty\Big\} & \text{for least-squares loss} \\ 4\,(2M+1)(1+2C_{\mathscr{H}}) & \text{for } \phi'\text{-bounded losses.} \end{cases} \tag{13}$$

As a remark, for the quadratic loss function, the above assumption is equivalent to assume that each $Y_i - f^*(x_i)$ is a sub-Gaussian random variable.

The *critical radius* δ_n is the smallest positive scalar such that

$$\frac{\mathcal{G}_n(\mathcal{E}_n(\delta,1))}{\delta} \leq \frac{\delta}{\sigma}. \tag{14}$$

We note that past work on localized Rademacher and Gaussian complexity [3], [26] guarantees that there exists a unique $\delta_n > 0$ that satisfies this condition, so that our definition is sensible. In particular, one can show (more details see e.g. Wainwright [39, Lemma 13.6,]) that the left hand side of this inequality is a non-increasing function in terms of δ and the right hand side of the inequality is an increasing function of δ. Therefore there exists a smallest positive solution to the above inequality.

1) Upper Bounds on Excess Risk and Empirical $L^2(\mathbb{P}_n)$-Error: With this set-up, we are now equipped to state our main theorem. It provides high-probability bounds on the excess risk and $L^2(\mathbb{P}_n)$-error of the estimator $\bar{f}^T := \frac{1}{T}\sum_{t=1}^T f^t$ defined by averaging the T iterates of the algorithm. It applies to both the least-squares cost function, and more generally, to any loss function satisfying the m-M-condition and the ϕ'-boundedness condition.

Theorem 1. *Given a sample size n sufficiently large to ensure that $\delta_n \leq \frac{M}{m}$, suppose that we compute the sequence $\{f^t\}_{t=0}^\infty$ using the update* (9) *with initialization $f^0 = 0$ and any step size $\alpha \in (0, \min\{\frac{1}{M}, M\}]$. Then for any iteration $T \in \big\{0, 1, \dots \lfloor\frac{m}{8M\delta_n^2}\rfloor\big\}$, the averaged function estimate $\bar{f}^T$ satisfies the bounds*

$$\mathcal{L}(\bar{f}^T) - \mathcal{L}(f^*) \leq CM\Big(\frac{1}{\alpha m T} + \frac{\delta_n^2}{m^2}\Big) \tag{15a}$$

$$\text{and} \quad \|\bar{f}^T - f^*\|_n^2 \leq C\Big(\frac{1}{\alpha m T} + \frac{\delta_n^2}{m^2}\Big), \tag{15b}$$

where both inequalities hold with probability at least $1 - c_1\exp(-C_2\frac{m^2 n\delta_n^2}{\sigma^2})$.

We prove Theorem 1 in Section V-A.

A few comments about the constants in our statement: in all cases, constants of the form c_j are universal, whereas the capital C_j may depend on parameters of the joint distribution and population loss $\mathcal{L}$. In Theorem 1, we have the explicit value $C_2 = \{\frac{m^2}{\sigma^2}, 1\}$ and C^2 is proportional to the quantity $2\max\{\|f^*\|_{\mathscr{H}}^2,\ 32,\ \sigma^2\}$. While inequalities (15a) and (15b) are stated as high probability results, similar bounds for expected loss (over the response y_i, with the design fixed) can be obtained by a simple integration argument. As another remark, since the function class $\mathcal{E}_n(\delta, 1)$ does not depend on the minimizer f^*, quantity δ_n in expression (14) is also independent of f^*, and therefore our bound holds for all f^* in the unit ball of the given RKHS.

In order to gain intuition for the claims in the theorem, note that apart from factors depending on (m, M), the first term $\frac{1}{\alpha m T}$ dominates the second term $\frac{\delta_n^2}{m^2}$ whenever $T \lesssim 1/\delta_n^2$. Consequently, up to this point, taking further iterations reduces the upper bound on the error. This reduction continues until we have taken of the order $1/\delta_n^2$ many steps, at which point the upper bound is of the order δ_n^2.

More precisely, suppose that we perform the updates with step size $\alpha = \frac{m}{M}$; then, after a total number of $\tau := \frac{1}{\delta_n^2 \max\{8, M\}}$ many iterations, the extension of Theorem 1 to expectations guarantees that the mean squared error is bounded as

$$\mathbb{E}\|\bar{f}^\tau - f^*\|_n^2 \leq C' \frac{\delta_n^2}{m^2}, \tag{16}$$

where C' is another constant depending on $C_{\mathscr{H}}$. Here we have used the fact that $M \geq m$ in simplifying the expression. It is worth noting that guarantee (16) matches the best known upper bounds for kernel ridge regression (KRR)—indeed, this must be the case, since a sharp analysis of KRR is based on the same notion of localized Gaussian complexity (e.g. [2], [3]). Thus, our results establish a strong parallel between the *algorithmic regularization* of early stopping, and the *penalized regularization* of kernel ridge regression. Moreover, as will be clarified in Section III-C, under suitable regularity conditions on the RKHS, the critical squared radius δ_n^2 also acts as a lower bound for the expected risk, meaning that our upper bounds are not improvable in general.

Note that the critical radius δ_n^2 only depends on our observations $\{(x_i, y_i)\}_{i=1}^n$ through the solution of inequality (14). In many cases, it is possible to compute and/or upper bound this critical radius, so that a concrete and valid stopping rule can indeed by calculated in advance. In Section IV, we provide a number of settings in which this can be done in terms of the eigenvalues $\{\widehat{\mu}_j\}_{j=1}^n$ of the normalized kernel matrix.

2) Consequences for Random Design Regression: Thus far, our analysis has focused purely on the case of fixed design, in which the sequence of covariates $\{x_i\}_{i=1}^n$ is viewed as fixed. If we instead view the covariates as being sampled in an i.i.d. manner from some distribution $\mathbb{P}_X$ over $\mathcal{X}$, then the empirical error $\|\widehat{f} - f^*\|_n^2 = \frac{1}{n}\sum_{i=1}^n \big(f(x_i) - f^*(x_i)\big)^2$ of a given estimate $\widehat{f}$ is a random quantity, and it is interesting to relate it to the squared population $L^2(\mathbb{P}_X)$-norm $\|\widehat{f} - f^*\|_2^2 = \mathbb{E}\big[(\widehat{f}(X) - f^*(X))^2\big]$.

In order not to unnecessarily complicate the discussion, in this section we assume that the data is drawn from the generative model

$$y_i = f^*(x_i) + \epsilon_i, \qquad i = 1, 2, \ldots, n, \tag{17}$$

with $f^* \in \mathscr{H}$, and where we assume that the noise $\epsilon_i := y_i - f^*(x_i)$ variables are uncorrelated with x_i. Under these conditions, the function f^* is both the minimizer of the population risk when the expectation is taken over the pair (X, Y) jointly, and the minimizer of the empirical risk when the expectation is only taken with respect to the randomness in Y. In this way, our results from the case of non-random design can be generalized directly.

In order to state an upper bound on this error, we introduce a population analogue of the critical radius δ_n, which we denote by $\overline{\delta_n}$. Consider the set

$$\overline{\mathcal{E}}(\delta, 1) := \Big\{ f - g \mid f, g \in \mathscr{H},\ \|f - g\|_{\mathscr{H}} \leq 1, \quad \|f - g\|_2 \leq \delta \Big\}. \tag{18}$$

It is analogous to the previously defined set $\mathcal{E}(\delta, 1)$, except that the empirical norm $\|\cdot\|_n$ has been replaced by the population version. The population Gaussian complexity localized at scale δ is given by

$$\overline{\mathcal{G}_n}\big(\overline{\mathcal{E}}(\delta, 1)\big) := \mathbb{E}_{w,X}\Big[\sup_{g \in \overline{\mathcal{E}}(\delta,1)} \frac{1}{n}\sum_{i=1}^n w_i g(X_i)\Big], \tag{19}$$

where $\{w_i\}_{i=1}^n$ are an i.i.d. sequence of standard normal variates, and $\{X_i\}_{i=1}^n$ is a second i.i.d. sequence, independent of the normal variates, drawn according to $\mathbb{P}_X$. Finally, the population critical radius $\overline{\delta_n}$ is defined by equation (19), in which $\mathcal{G}_n$ is replaced by $\overline{\mathcal{G}_n}$.

Corollary 1. *Under the conditions of Theorem 1, suppose in addition that the sequence $\{X_i\}_{i=1}^n$ are drawn i.i.d. from distribution $\mathbb{P}_X$ and the responses $\{Y_i\}_{i=1}^n$ are generated from regression model* (17)*. If we compute the boosting updates with step size $\alpha \in (0, \min\{\frac{1}{M}, M\}]$ and initialization $f^0 = 0$, then the averaged function estimate $\bar{f}^T$ at time $T := \lfloor \frac{1}{\bar{\delta}_n^2 \max\{8, M\}} \rfloor$ satisfies the bound*

$$\mathbb{E}_X\big(\bar{f}^T(X) - f^*(X)\big)^2 \ = \ \|\bar{f}^T - f^*\|_2^2 \leq \tilde{c}\ \overline{\delta}_n^2$$

with probability at least $1 - c_1 \exp(-C_2 \frac{m^2 n \bar{\delta}_n^2}{\sigma^2})$ over the random samples.

The proof of Corollary 1 consists of two main steps:

- From standard empirical process theory bounds [3], [28], the difference between empirical risk $\|\bar{f}^T - f^*\|_n^2$ and population risk $\|\bar{f}^T - f^*\|_2^2$ can be controlled for uniformly bounded $\bar{f}^T - f^*$. Note that the latter holds when $T \leq \lfloor \frac{1}{\bar{\delta}_n^2 \max\{8, M\}} \rfloor$ by uniform boundedness of the kernel and Lemma 4 used for the proof of Theorem 1. In particular, it can be shown that
$$|\|\bar{f}^T - f^*\|_n^2 - \|\bar{f}^T - f^*\|_2^2| \leq c\overline{\delta}_n,$$
for a fixed positive constant c.
- Furthermore, one can show (e.g. Wainwright [39, Proposition 14.25]) that the empirical critical quantity δ_n is bounded by the population $\overline{\delta_n}$ up to multiplicative constant.

By combining both arguments the corollary follows. We refer the reader to the papers [3], [28] for further details on such equivalences.

It is worth comparing this guarantee with the past work of Raskutti *et al.* [28], who analyzed the kernel boosting iterates of the form (9), but with attention restricted to the special case of the least-squares loss. Their analysis was based on first decomposing the squared error into bias and variance terms, then carefully relating the combination of these terms to a particular bound on the localized Gaussian complexity (see equation (23) below). In contrast, our theory more directly analyzes the effective function class that is explored by taking T steps, so that the localized Gaussian width (19) appears more naturally. In addition, our analysis applies to a broader class of loss functions.

C. Achieving Minimax Lower Bounds

In this section, we show that the upper bound (16) for fixed design matches known minimax lower bounds on the error, so that our results are unimprovable in general. We establish this result for the class of *regular kernels*, as previously defined by Yang *et al.* [40], which includes the Gaussian and Sobolev kernels as special cases.

The class of regular kernels is defined as follows. Let $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \cdots \geq \widehat{\mu}_n \geq 0$ denote the ordered eigenvalues of the normalized kernel matrix K, and define the quantity $d_n := \text{argmin}_{j=1,\ldots,n}\{\widehat{\mu}_j \leq \delta_n^2\}$. A kernel is called *regular* whenever there is a universal constant c such that the tail sum satisfies $\sum_{j=d_n+1}^{n} \widehat{\mu}_j \leq c\, d_n \delta_n^2$. In words, the tail sum of the eigenvalues for regular kernels is roughly on the same or smaller scale as the sum of the eigenvalues bigger than δ_n^2.

For such kernels and under the Gaussian observation model ($Y_i \sim N(f^*(x_i), \sigma^2)$), Yang *et al.* [40] prove a minimax lower bound involving δ_n. In particular, they show that the minimax risk over the unit ball of the Hilbert space is lower bounded as

$$\inf_{\widehat{f}} \sup_{\|f^*\|_{\mathscr{H}} \leq 1} \mathbb{E}\|\widehat{f} - f^*\|_n^2 \geq c\delta_n^2, \tag{20}$$

for some fixed positive constant c. Comparing the lower bound (20) with upper bound (16) for our estimator $\bar{f}^T$ stopped after $O(1/\delta_n^2)$ many steps, it follows that the bounds proven in Theorem 1 are unimprovable apart from constant factors.

We now state a generalization of this minimax lower bound, one which applies to a sub-class of *generalized linear models*, or GLM for short. In these models, the conditional distribution of the observed vector $Y = (Y_1, \ldots, Y_n)$ given $\big(f^*(x_1), \ldots, f^*(x_n)\big)$ takes the form

$$\mathbb{P}_\theta(y) = \prod_{i=1}^{n} \Big[h(y_i) \exp\big(\frac{y_i f^*(x_i) - \Phi(f^*(x_i))}{s(\sigma)}\big)\Big], \tag{21}$$

where $s(\sigma)$ is a known scale factor and $\Phi : \mathbb{R} \to \mathbb{R}$ is the cumulant function of the generalized linear model. As some concrete examples:

- The linear Gaussian model is recovered by setting $s(\sigma) = \sigma^2$ and $\Phi(t) = t^2/2$.
- The logistic model for binary responses $y \in \{-1, 1\}$ is recovered by setting $s(\sigma) = 1$ and $\Phi(t) = \log(1 + \exp(t))$.

Our minimax lower bound applies to the class of GLMs for which the cumulant function Φ is differentiable and has uniformly bounded second derivative $|\Phi''| \leq L$. This class includes the linear, logistic, multinomial families, among others, but excludes (for instance) the Poisson family. Under this condition, we have the following:

Corollary 2. *Suppose that we are given i.i.d. samples* $\{y_i\}_{i=1}^n$ *from a GLM* (21) *for some function* f^* *in a regular kernel class with* $\|f^*\|_{\mathscr{H}} \leq 1$. *Then running* $T := \lfloor \frac{1}{\delta_n^2 \max\{8, M\}} \rfloor$ *iterations with step size* $\alpha \in (0, \min\{\frac{1}{M}, M\}]$ *and* $f^0 = 0$ *yields an estimate* $\bar{f}^T$ *such that*

$$\mathbb{E}\|\bar{f}^T - f^*\|_n^2 \asymp \inf_{\widehat{f}} \sup_{\|f^*\|_{\mathscr{H}} \leq 1} \mathbb{E}\|\widehat{f} - f^*\|_n^2. \tag{22}$$

Here $f(n) \asymp g(n)$ means $f(n) = cg(n)$ up to a universal constant $c \in (0, \infty)$. As always, in the minimax claim (22), the infimum is taken over all measurable functions of the input data and the expectation is taken over the randomness of the response variables $\{Y_i\}_{i=1}^n$. Since we know that $\mathbb{E}\|\bar{f}^T - f^*\|_n^2 \precsim \delta_n^2$, the way to prove bound (22) is by establishing $\inf_{\widehat{f}} \sup_{\|f^*\|_{\mathscr{H}} \leq 1} \mathbb{E}\|\widehat{f} - f^*\|_n^2 \succsim \delta_n^2$. See Section V-B for the proof of this result.

At a high level, the statement in Corollary 2 shows that early stopping prevents us from overfitting to the data; in particular, using the stopping time T yields an estimate that attains the optimal balance between bias and variance. Note that for the special case of the squared loss, related works such as [11], [24] have proved bounds which adapt to the smoothness and regularity of f^* for an early stopping time (which also depends on the regularity parameter). The novelty in our work is that our general unified framework which connects penalized and algorithmic regularization, can be used to compute bounds for a large variety of loss functions.

IV. Consequences for Various Kernel Classes

In this section, we apply Theorem 1 to derive some concrete rates for different kernel spaces for the random design case and then illustrate them with some numerical experiments. It is known that the complexity of an RKHS in association with a distribution over the covariates $\mathbb{P}_X$ can be characterized by the decay rate (7) of the eigenvalues of the kernel operator (also called eigen-decay). The representation power of a kernel class is directly correlated with the eigen-decay: the faster the decay, the smaller the function class.

A. Theoretical Predictions as a Function of Decay

In this section, let us consider two forms of decay of kernel eigenvalues μ_i defined in equation (7).

- **γ-exponential decay**: For some $\gamma > 0$, the eigenvalues satisfy a decay condition of the form $\mu_j \leq c_1 \exp(-c_2\, j^\gamma)$, where c_1, c_2 are universal constants. Examples of kernels in this class include the Gaussian kernel, which for the Lebesgue measure satisfies such a bound with $\gamma = 2$ (over the real line) or $\gamma = 1$ (on a compact domain).
- **β-polynomial decay**: For some $\beta > 1/2$, the eigenvalues satisfy a decay condition of the form $\mu_j \leq c_1\, j^{-2\beta}$, where c_1 is a universal constant. Examples of kernels in this class include the k^{th}-order Sobolev spaces for some fixed integer $k \geq 1$ with Lebesgue measure on a bounded domain. We consider Sobolev spaces that consist of functions that have k^{th}-order weak derivatives $f^{(k)}$ being Lebesgue integrable and $f(0) = f^{(1)}(0) = \cdots = f^{(k-1)}(0) = 0$. For such classes, the β-polynomial decay condition holds with $\beta = k$.

Given eigendecay conditions of these types, it is possible to compute an upper bound on the critical radius $\overline{\delta_n}$ by upper bounding the localized Gaussian complexity by a function of the eigenvalues. In particular, we have the following lemma:

Lemma 1. *Letting $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ denote the ordered eigenvalues of the kernel operator, it holds that*

$$\overline{\mathcal{G}}_n(\overline{\mathcal{E}}(\delta, 1)) \leq \sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^{\infty} \min\{\delta^2, \mu_j\}}. \tag{23}$$

The proof can be found in Section E. As a direct consequence, the smallest δ which satisfies the following inequality

$$\sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^{\infty} \min\{\delta^2, \mu_j\}} \;\leq\; \frac{\delta^2}{\sigma} \tag{24}$$

is an upper bound on the critical radius $\overline{\delta}_n$. We can now see that $\overline{\delta}_n$ is directly related to the representation power of a kernel class by means of the eigenvalues of the corresponding kernel operator.

Consequently, we can show that for γ -exponentially decaying kernels, we have $\overline{\delta}_n^2 \precsim \frac{(\log n)^{1/\gamma}}{n}$, whereas for β-polynomial kernels, we have $\overline{\delta}_n^2 \precsim n^{-\frac{2\beta}{2\beta+1}}$, where $\precsim$ denotes an inequality holding with a universal constant pre-factor. Combining with our Theorem 1, we obtain the following result:

Corollary 3 (Bounds based on eigendecay)**.** *Under the conditions of Corollary 1 and for $\{x_i\}_{i=1}^n$ drawn i.i.d. from $\mathbb{P}_X$, the following bounds on the expected empirical errors over the outputs hold with probability at least $1 - e^{-n\overline{\delta}_n}$*

(a) For kernels with γ -exponential eigen-decay, we have

$$\mathbb{E}\|\bar{f}^T - f^*\|_n^2 \leq c\, \frac{\log^{1/\gamma} n}{n}$$

at $T \asymp \frac{n}{\log^{1/\gamma} n}$ steps.

(b) For kernels with β-polynomial eigen-decay, we have

$$\mathbb{E}\|\bar{f}^T - f^*\|_n^2 \leq c\, n^{-2\beta/(2\beta+1)}$$

at $T \asymp n^{2\beta/(2\beta+1)}$ steps.

In particular, these bounds hold for LogitBoost and AdaBoost. See Section V-C for the proof of Corollary 3. These bounds directly carry over to bounds on the excess risk $\mathcal{L}(f^T) - \mathcal{L}(f^*)$ as outlined in the proof of Theorem 1 and also hold for the population errors by invoking Corollary 1.

To the best of our knowledge, our results are the first to show non-asymptotic and minimax optimal rates for the $\|\cdot\|_n^2$-error when applying early stopping to either the LogitBoost or AdaBoost updates, along with an explicit dependence of the stopping rule on n. Our results yield analogous guarantees for L^2-boosting, as have been established in past work [28]. Note that we observe a similar trade-off between computational efficiency and statistical accuracy as in the case of kernel least-squares regression [28], [41]: although larger kernel classes (e.g. Sobolev classes) yield higher estimation errors, boosting updates reach the optimum faster than for a smaller kernel class (e.g. Gaussian kernels).

It is worth noting that whiel Lemma 1 is stated in terms of the kernel operator eigenvalues, an analogous bound also holds in terms of the eigenvalues of the kernel matrix. The proof of this claim is similar; see Lemma 13.22 in the book [39] for details. Since the empirical and population critical quantities are close up to constants ([39, Lemma 13.6,]), this connection provides an explicit way to compute the optimal stopping rule T up to a constant factor. It does entail computing (a subset of) the eigenvalues of the n-dimensional kernel matrix. In applications where there are multiple estimation problems involving the same set of design points, this cost could be amortized. In other settings, this cost could be significant, and it would be interesting to devise algorithms that approximate the kernel eigenvalues to sufficient accuracy. We leave this as a direction for future work.

B. Numerical Experiments

We now describe some numerical experiments that provide illustrative confirmations of our theoretical predictions. While we have applied our methods to various kernel classes, in this section, we present numerical results for the first-order Sobolev kernel as two typical examples for exponential and polynomial eigen-decay kernel classes.

Let us start with the first-order Sobolev space of Lipschitz functions on the unit interval $[0, 1]$, defined by the kernel $\mathbb{K}(x, x') = 1 + \min(x, x')$, and with the design points $\{x_i\}_{i=1}^n$ set equidistantly over $[0, 1]$. Note that the equidistant design yields β-polynomial decay of the eigenvalues of K with $\beta = 1$ as in the case when x_i are drawn i.i.d. from the uniform measure on $[0, 1]$. Consequently we have that $\delta_n^2 \asymp n^{-2/3}$. Accordingly, our theory predicts that the stopping time $T = (cn)^{2/3}$ should lead to an estimate $\bar{f}^T$ such that $\|\bar{f}^T - f^*\|_n^2 \precsim n^{-2/3}$.

In our experiments for L^2-Boost, we sampled Y_i according to $Y_i = f^*(x_i) + w_i$ with $w_i \sim \mathcal{N}(0, 0.5)$, which corresponds to the probability distribution $\mathbb{P}(Y \mid x_i) = \mathcal{N}(f^*(x_i); 0.5)$, where $f^*(x) = |x - \frac{1}{2}| - \frac{1}{4}$ is defined on the unit interval $[0, 1]$. By construction, the function f^* belongs to the first-order Sobolev space with $\|f^*\|_{\mathscr{H}} = 1$. For LogitBoost, we sampled Y_i according to $\mathrm{Bin}(p(x_i), 5)$ where $p(x) = \frac{\exp(f^*(x))}{1+\exp(f^*(x))}$. In all cases, we fixed the initialization $f^0 = 0$, and ran the updates (9) for L^2-Boost and LogitBoost with the constant step size $\alpha = 0.75$. We compared various stopping rules to the *oracle rule* G, meaning the procedure that examines all iterates $\{f^t\}$, and chooses the stopping time $G = \arg\min_{t \geq 1} \|f^t - f^*\|_n^2$ that yields the minimum prediction error. Of course, this procedure is unimplementable in practice, but it serves as a convenient lower bound with which to compare.

Each panel in Figure 2 shows plots of the mean-squared error $\|\bar{f}^T - f^*\|_n^2$ versus the sample size n, with each point corresponding to an average over 40 trials, for four different stopping rules. Error bars correspond to the standard errors computed from our simulations. The top two panels ((a) and (b)) show results for L^2-Boost, whereas the bottom two panels ((c) and (d)) show results for LogitBoost. The left two panels ((a) and (c)) gives plots on a linear-linear scale, whereas the corresponding right two panels ((b) and (d)) replots the same data on a log-log scale. The blue-solid curve corresponds to the performance of the oracle stopping

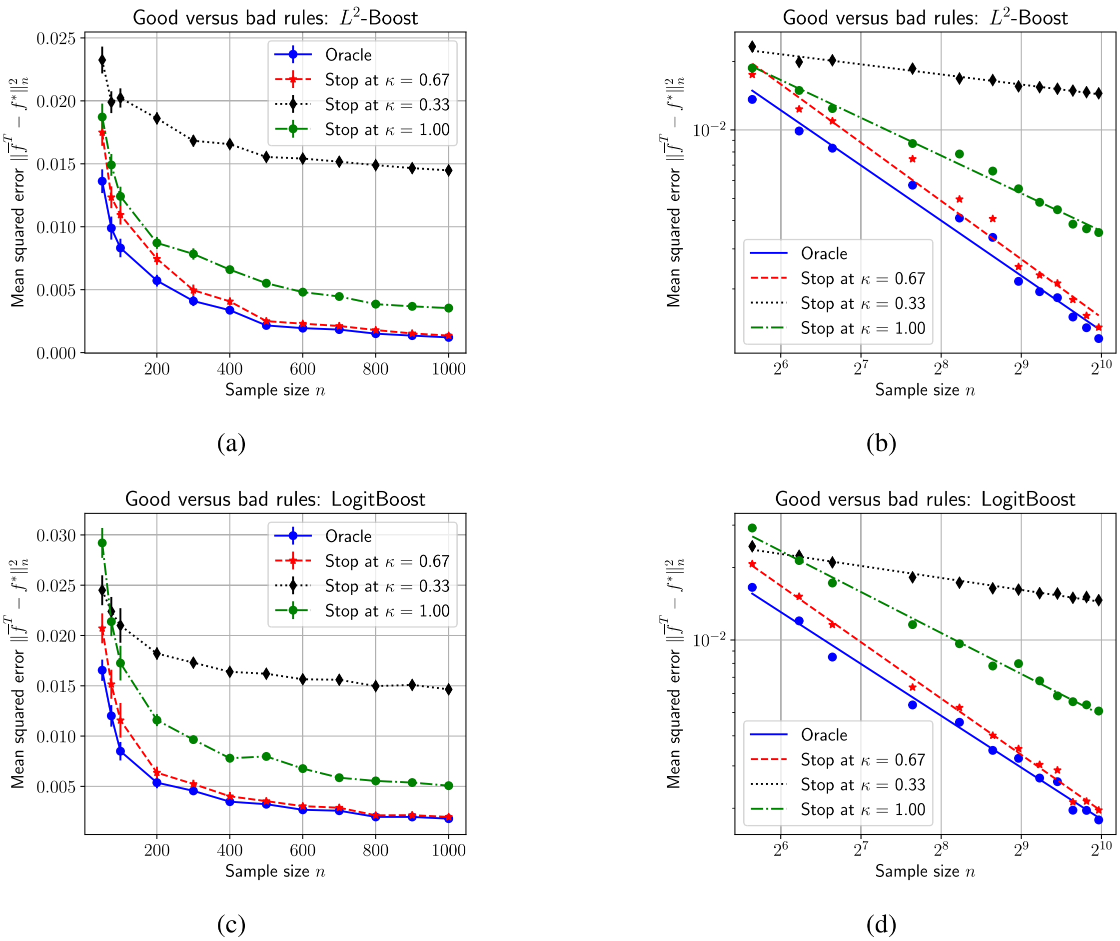


Fig. 2. Plots of the mean-squared error versus the sample size n for various types of stopping rules. Panel (a) shows plots on a linear-linear scale, whereas panel (b) shows plots on a log-log scale. The blue-solid curve corresponds the *oracle rule* of stopping the algorithm at the time T with the minimum error; it is an oracle procedure, since we cannot actually determine the point of minimum error in practice. The remaining three curves show the behavior of three different practical stopping rules, all based on stopping at the time $T = (7n)^\kappa$ for different choices of exponent κ. The performance of the rules given by $\kappa \in \{1/3, 2/3, 1\}$ are plotted in black-dotted, red-dashed, and green-dot-dashed curves, respectively. The choice $\kappa = 0.67$ is the theoretically optimal choice. On the log-log plot, we see linear fits with slopes -0.80 for the oracle rule versus -0.85 for the rule with $\kappa = 2/3$ in the case of L^2-boost, and -0.71 for the oracle rule versus -0.77 for the rule with $\kappa = 2/3$ in the case of LogitBoost.

time $T = G$ described above; the other three curves show the performance of rules of the form $T = (7n)^\kappa$ for different choices of κ. Note that the behavior of the stopping rules for L^2-Boost and LogitBoost are qualitatively similar. In both cases, the theoretically derived stopping rule $T = (7n)^\kappa$ with $\kappa^* = 2/3 \approx 0.67$, while slightly worse than the oracle, tracks its performance closely. We also performed simulations for some "bad" stopping rules, in particular for an exponent κ *not equal* to the theoretically optimal choice $2/3$; the performance of these bad rules is indicated in the green-dot-dashed and black-dotted curves. The log scale plots in Figure 2, namely panels (b) and (d), show that with rules defined by $\kappa \in \{1/3, 1\}$, the resulting performance is indeed substantially worse, with the difference in slope even suggesting a different scaling of the error with the number of observations n. Recalling our discussion for Figure 1, this phenomenon likely occurs due to underfitting and overfitting effects. These qualitative shifts are consistent with our theory. As regards to the choice of pre-factor 7 in our stopping rule, this was made on a heuristic basis, mainly to ensure that we had reasonable rules for all choices of κ. Our main goal here was to illustrate the correspondence between our theoretical predictions and behavior in simulation.

V. Proof of Main Results

We now turn to the proofs of our main results, with the bulk of the technical details deferred to the appendices. We begin by recalling some notation from Section II-C. We denote the vector of function values of a function $f \in \mathscr{H}$ evaluated at $(x_1, x_2, \ldots, x_n)$ as

$$\theta_f := f(x_1^n) = (f(x_1), f(x_2), \ldots f(x_n)) \in \mathbb{R}^n,$$

where we omit the subscript f when it is clear from the context. As mentioned in the main text, updates on the function value vectors $\theta^t \in \mathbb{R}^n$ correspond uniquely to updates of the functions $f^t \in \mathscr{H}$ (see a complete discussion in Section A). In the following we repeatedly abuse notation by defining the Hilbert norm and empirical norm on vectors in $\Delta \in \text{range}(K)$ as

$$\|\Delta\|_{\mathscr{H}}^2 = \frac{1}{n}\Delta^T K^\dagger \Delta \quad \text{and} \quad \|\Delta\|_n^2 = \frac{1}{n}\|\Delta\|_2^2,$$

where $K^\dagger$ is the pseudoinverse of K. We also use $\mathbb{B}_{\mathscr{H}}(\theta, r)$ to denote the ball with respect to the $\|\cdot\|_{\mathscr{H}}$-norm in range(K).

A. Proof of Theorem 1

On a high-level, the key to the proof is the fact that the Hilbert norm is bounded for all iterations t such that $t \le \frac{m}{8M\delta_n^2}$. Doing so involves combining arguments from convex optimization with arguments from empirical process theory, with the latter being required to relate the derivatives of the sample and population-based objective functions.

The proof of our main theorem is based on a sequence of lemmas, all of which are stated with the assumptions of Theorem 1 in force. The first lemma establishes a bound on the empirical norm $\|\cdot\|_n$ of the error $\Delta^{t+1} := \theta^{t+1} - \theta^*$, provided that its Hilbert norm is suitably controlled.

Lemma 2. *For any stepsize $\alpha \in (0, \frac{1}{M}]$ and any iteration t we have*

$$\frac{m}{2}\|\Delta^{t+1}\|_n^2 \le \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\} + \langle \nabla\mathcal{L}(\theta^* + \Delta^t) - \nabla\mathcal{L}_n(\theta^* + \Delta^t),\ \Delta^{t+1}\rangle. \quad (26)$$

See Appendix B for the proof of this claim.

The second term on the right-hand side of the bound (26) involves the difference between the population and empirical gradient operators. Since this difference is being evaluated at the random points Δ^t and Δ^{t+1}, the following lemma establishes a form of uniform control on this term.

Let us define the set

$$\mathbb{S} := \Big\{\Delta, \widetilde{\Delta} \in \mathbb{R}^n \mid \|\Delta\|_{\mathscr{H}} \ge 1, \text{ and } \Delta,\ \widetilde{\Delta} \in \mathbb{B}_{\mathscr{H}}(0, 2C_{\mathscr{H}})\Big\},$$

and consider the uniform bound

$$\langle \nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}),\ \Delta\rangle \le 2\delta_n\|\Delta\|_n + 2\delta_n^2\|\Delta\|_{\mathscr{H}} + \frac{m}{c_3}\|\Delta\|_n^2 \quad \text{for all } \Delta, \widetilde{\Delta} \in \mathbb{S}. \quad (27)$$

Lemma 3. *Let $\mathcal{E}$ be the event that bound (27) holds. There are universal constants (c_1, c_2) such that $\mathbb{P}[\mathcal{E}] \ge 1 - c_1\exp(-c_2\frac{m^2 n\delta_n^2}{\sigma^2})$.*

See Appendix C for the proof of this claim.

Note that Lemma 2 applies only to error iterates with a bounded Hilbert norm. Our last lemma provides this control for some number of iterations:

Lemma 4. *There are constants (C_1, C_2) independent of n such that for any step size $\alpha \in \big(0, \min\{M, \frac{1}{M}\}\big]$, we have*

$$\|\Delta^t\|_{\mathscr{H}} \le C_{\mathscr{H}} \quad \textit{for all iterations } t \le \frac{m}{8M\delta_n^2} \quad (28)$$

with probability at least $1 - C_1\exp(-C_2 n\delta_n^2)$, where $C_2 = \max\{\frac{m^2}{\sigma^2}, 1\}$.

See Appendix D for the proof of this lemma; note that this proof also uses Lemma 3.

Taking these lemmas as given, we now complete the proof of the theorem. We first condition on the event $\mathcal{E}$ from Lemma 3, so that we may apply the bound (27). We then fix some iterate t such that $t < \frac{m}{8M\delta_n^2} - 1$, and condition on the event that the bound (28) in Lemma 4 holds, so that we are guaranteed that $\|\Delta^{t+1}\|_{\mathscr{H}} \le C_{\mathscr{H}}$. We then split the analysis into two cases:

a) Case 1: First, suppose that $\|\Delta^{t+1}\|_n \le \delta_n C_{\mathscr{H}}$. In this case, inequality (15b) holds directly.

b) Case 2: Otherwise, we may assume that $\|\Delta^{t+1}\|_n > \delta_n\|\Delta^{t+1}\|_{\mathscr{H}}$. Applying bound (27) with the choice $(\widetilde{\Delta}, \Delta) = (\Delta^t, \Delta^{t+1})$ yields

$$\langle \nabla\mathcal{L}(\theta^* + \Delta^t) - \nabla\mathcal{L}_n(\theta^* + \Delta^t),\ \Delta^{t+1}\rangle \le 4\delta_n\|\Delta^{t+1}\|_n + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2. \quad (29)$$

Substituting inequality (29) back into equation (26) yields

$$\frac{m}{2}\|\Delta^{t+1}\|_n^2 \le \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\} + 4\delta_n\|\Delta^{t+1}\|_n + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2.$$

Re-arranging terms yields the bound

$$\gamma m\|\Delta^{t+1}\|_n^2 \le D^t + 4\delta_n\|\Delta^{t+1}\|_n, \quad (30)$$

where we have introduced the shorthand notation $D^t := \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\}$, as well as $\gamma = \frac{1}{2} - \frac{1}{c_3}$

Equation (30) defines a quadratic inequality with respect to $\|\Delta^{t+1}\|_n$; solving it and making use of the inequality $(a + b)^2 \le 2a^2 + 2\,b^2$ yields the bound

$$\|\Delta^{t+1}\|_n^2 \le \frac{c\delta_n^2}{\gamma^2 m^2} + \frac{2D^t}{\gamma m}, \quad (31)$$

for some universal constant c. By telescoping inequality (31), we find that

$$\frac{1}{T}\sum_{t=1}^{T}\|\Delta^t\|_n^2 \le \frac{c\delta_n^2}{\gamma^2 m^2} + \frac{1}{T}\sum_{t=1}^{T}\frac{2D^t}{\gamma m} \quad (32)$$

$$\le \frac{c\delta_n^2}{\gamma^2 m^2} + \frac{1}{\alpha\gamma mT}[\|\Delta^0\|_{\mathscr{H}}^2 - \|\Delta^T\|_{\mathscr{H}}^2]. \quad (33)$$

By Jensen's inequality, we have

$$\|\bar{f}^T - f^*\|_n^2 = \|\frac{1}{T}\sum_{t=1}^{T}\Delta^t\|_n^2 \ \le\ \frac{1}{T}\sum_{t=1}^{T}\|\Delta^t\|_n^2,$$

so that inequality (15b) follows from the bound (32).

On the other hand, by the smoothness assumption, we have

$$\mathcal{L}(\bar{f}^T) - \mathcal{L}(f^*) \le \frac{M}{2}\|\bar{f}^T - f^*\|_n^2,$$

from which inequality (15a) follows.

B. Proof of Corollary 2

We first provide a proof outline before proceeding with the details. Our proof is based on a standard application of Fano's inequality, together with a random packing argument. Fano's inequality helps reduce proving lower bounds to finding an upper bound on the Kullback-Leibler divergence between distributions associated to the packing set. Then we control the Kullback-Leibler divergence between the GLM (21) through the properties of their cumulant functions. See Chapter 15 in the book [39] for more background on arguments of this type.

Our proof builds on and generalizes Theorem 1 in Yang et al. [40]. By definition of the transformed vector $\theta = DU\alpha$ with $K = U^T DU$, we have for any estimator $\widehat{f} = \sqrt{n}U^T\theta$ that $\|\widehat{f} - f^*\|_n^2 = \|\widehat{\theta} - \theta^*\|_2^2$. Therefore our goal is to lower bound the Euclidean error $\|\widehat{\theta} - \theta^*\|_2$ of any estimator of θ^*. Borrowing Lemma 4 in Yang *et al.* [40], there exists $\delta/2$-packing of the set $B = \{\theta \in \mathbb{R}^n \mid \|D^{-1/2}\theta\|_2 \leq 1\}$ of cardinality $M = e^{d_n/64}$ with $d_n := \arg\min_{j=1,\ldots,n}\{\widehat{\mu}_j \leq \delta_n^2\}$. The packing actually belongs to the set

$$\mathcal{E}(\delta) := \Big\{\theta \in \mathbb{R}^n \mid \sum_{j=1}^n \frac{\theta_j^2}{\min\{\delta^2, \widehat{\mu}_j\}} \leq 1\Big\},$$

which is a subset of B. Let us denote the packing set by $\{\theta^1, \ldots, \theta^M\}$. Since $\theta \in \mathcal{E}(\delta)$, by simple calculation, we have $\|\theta^i\|_2 \leq \delta$.

By considering the random ensemble of regression problems in which we first draw an index Z at random from the index set $[M]$ and then condition on $Z = z$, we observe n i.i.d samples $y_1^n := \{y_1, \ldots, y_n\}$ from $\mathbb{P}_{\theta^z}$, Fano's inequality implies that

$$\mathbb{P}(\|\widehat{\theta} - \theta^*\|_2 \geq \frac{\delta^2}{4}) \geq 1 - \frac{I(y_1^n; Z) + \log 2}{\log M}.$$

where $I(y_1^n; Z)$ is the mutual information between the samples Y and the random index Z.

So it is only left for us to control the mutual information $I(y_1^n; Z)$. Using the mixture representation $\bar{\mathbb{P}} = \frac{1}{M}\sum_{i=1}^M \mathbb{P}_{\theta^i}$ and the convexity of the Kullback–Leibler divergence, we have

$$I(y_1^n; Z) = \frac{1}{M}\sum_{j=1}^M \|\mathbb{P}_{\theta^j},\ \bar{\mathbb{P}}\|_{\mathrm{KL}} \leq \frac{1}{M^2}\sum_{i,j} \|\mathbb{P}_{\theta^i},\ \mathbb{P}_{\theta^j}\|_{\mathrm{KL}}.$$

In order to proceed, we claim that

$$\|\mathbb{P}_\theta(y),\ \mathbb{P}_{\theta'}(y)\|_{\mathrm{KL}} \leq \frac{nL\|\theta - \theta'\|_2^2}{s(\sigma)}. \tag{34}$$

Taking this claim as given for the moment, since each $\|\theta^i\|_2 \leq \delta$, triangle inequality yields $\|\theta^i - \theta^j\|_2 \leq 2\delta$ for all $i \neq j$. It is therefore guaranteed that

$$I(y_1^n; Z) \leq \frac{4nL\delta^2}{s(\sigma)}.$$

Therefore, similar to Yang *et al.* [40], following the fact that the kernel is regular and hence $s(\sigma)d_n \geq cn\delta_n^2$, any estimator $\widehat{f}$ has prediction error lower bounded as

$$\sup_{\|f^*\|_{\mathscr{H}} \leq 1} \mathbb{E}\|\widehat{f} - f^*\|_n^2 \geq c_l\delta_n^2.$$

This lower bound, in conjunction with the upper bound from Theorem 1, yields the statement of Corollary 2.

Proof of inequality (34)*:* Recall that we define the transformed parameter $\theta = DU\alpha$ with $K = U^T DU$, and any estimator $\widehat{f} = \sqrt{n}U^T\theta$. Let us write $U = [u_1, u_2, \ldots, u_n]$. Direct calculation of the KL divergence yields

$$\begin{aligned}\|\mathbb{P}_\theta(y),\ \mathbb{P}_{\theta'}(y)\|_{\mathrm{KL}} &= \int \log(\frac{\mathbb{P}_\theta(y)}{\mathbb{P}_{\theta'}(y)})\mathbb{P}_\theta(y)dy\\ &= \frac{1}{s(\sigma)}\sum_{i=1}^n \Phi(\sqrt{n}\langle u_i,\ \theta'\rangle) - \Phi(\sqrt{n}\langle u_i,\ \theta\rangle)\\ &\quad + \frac{\sqrt{n}}{s(\sigma)}\int\sum_{i=1}^n \big[y_i\langle u_i,\ \theta - \theta'\rangle\big]\mathbb{P}_\theta dy.\end{aligned} \tag{35}$$

To further control the right hand side of expression (35), we concentrate on expressing $\int\sum_{i=1}^n y_i u_i \mathbb{P}_\theta dy$ differently. Leibniz's rule allow us to inter-change the order of integral and derivative, so that

$$\int \frac{dP_\theta}{d\theta}dy = \frac{d}{d\theta}\int P_\theta dy = 0. \tag{36}$$

Observe that

$$\int \frac{dP_\theta}{d\theta}dy = \frac{\sqrt{n}}{s(\sigma)}\int P_\theta \cdot \sum_{i=1}^n u_i\big(y_i - \Phi'(\sqrt{n}\langle u_i,\ \theta'\rangle)\big)dy$$

so that equality (36) yields

$$\int\sum_{i=1}^n y_i u_i \mathbb{P}_\theta dy = \sum_{i=1}^n u_i\Phi'(\sqrt{n}\langle u_i,\ \theta\rangle).$$

Combining the above inequality with expression (35), the KL divergence between two generalized linear models $\mathbb{P}_\theta, \mathbb{P}_{\theta'}$ can thus be written as

$$\begin{aligned}\|\mathbb{P}_\theta(y),\ \mathbb{P}_{\theta'}(y)\|_{\mathrm{KL}} &= \frac{1}{s(\sigma)}\sum_{i=1}^n \Phi(\sqrt{n}\langle u_i,\ \theta'\rangle)\\ &-\Phi(\sqrt{n}\langle u_i,\ \theta\rangle) - \sqrt{n}\langle u_i,\ \theta' - \theta\rangle\Phi'(\sqrt{n}\langle u_i,\ \theta\rangle).\end{aligned} \tag{37}$$

Together with the fact that

$$\begin{aligned}|\Phi(\sqrt{n}\langle u_i,\ \theta'\rangle) &- \Phi(\sqrt{n}\langle u_i,\ \theta\rangle)\\ &- \sqrt{n}\langle u_i,\ \theta' - \theta\rangle\Phi'(\sqrt{n}\langle u_i,\ \theta\rangle)| \leq nL\|\theta - \theta'\|_2^2.\end{aligned}$$

which follows by assumption on Φ having a uniformly bounded second derivative. Combining the above inequality with inequality (37) establishes our claim (34).

C. Proof of Corollary 3

Throughout the proof, we condition on the event

$$\mathcal{E} = \big\{\frac{\overline{\delta_n}}{4} \leq \delta_n \leq 3\overline{\delta_n}\big\}$$

which holds with probability at least $1 - c_1 e^{-c_2 n\overline{\delta_n}}$ for some universal constants c_1, c_2 (see Proposition 14.1 in the book [39]). The expectations are with respect to the outputs $Y_1, \ldots, Y_n$. When the conditions in Theorem 1 are satisfied, it then follows from the extension (16) of the theorem that

$$\mathbb{E}\|\bar{f}^T - f^*\|_n^2 \leq C'\frac{\delta_n^2}{m^2} \leq C''\frac{\overline{\delta_n^2}}{m^2} \ \text{ at } T \asymp \tfrac{1}{\delta_n^2} \text{ steps.} \tag{38}$$

In order to invoke Theorem 1 for the particular cases of LogitBoost and AdaBoost, we need to verify the conditions, i.e. that the m-M-condition and ϕ'-boundedness conditions hold for the respective loss function over the ball $\mathbb{B}_{\mathscr{H}}(\theta^*, 2C_{\mathscr{H}})$. The following lemma provides such a guarantee:

Lemma 5. *With $D := C_{\mathscr{H}} + \|\theta^*\|_{\mathscr{H}}$, the logistic regression cost function satisfies the m-M-condition with parameters*

$$m = \frac{1}{e^{-D} + e^{D} + 2}, \quad M = \frac{1}{4}, \quad \text{and} \quad B = 1.$$

The AdaBoost cost function satisfies the m-M-condition with parameters

$$m = e^{-D}, \quad M = e^{D}, \quad \text{and} \quad B = e^{D}.$$

In particular, the conditions hold for any sequence $\{x_i\}_{i=1}^n$ in the domain $\mathcal{X}$.

See Appendix F for the proof of Lemma 5.

It follows from Lemma 1 that is suffices to compute the function

$$\mathcal{R}(\delta) := \sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{\infty} \min\{\delta^2, \mu_j\}} \tag{39}$$

to obtain an upper bound for the critical radius $\overline{\delta}_n$.

c) γ-exponential decay: Suppose that the kernel operator eigenvalues satisfy a decay condition of the form $\mu_j \leq c_1 \exp(-c_2\ j^\gamma)$, where c_1, c_2 are universal constants. Then the function $\mathcal{R}$ from equation (39) can be upper bounded as

$$\mathcal{R}(\delta) = \sqrt{\frac{1}{n}\sum_{i=1}^{\infty} \min\{\delta^2, \mu_j\}}$$
$$\leq \sqrt{\frac{1}{n}}\sqrt{k\delta^2 + \sum_{j=k+1}^{\infty} c_1 e^{-c_2\ j^2}},$$

and $\mathcal{R}(\delta) \geq \sqrt{k\delta^2}$ where k is the smallest integer such that $c_1 \exp(-c_2\, k^\gamma) < \delta^2$. Some algebra then shows that the critical radius scales as $\overline{\delta}_n^2 \asymp \frac{n}{\log(n)^{1/\gamma}\sigma^2}$.

Consequently, if we take $T \asymp \frac{\log(n)^{1/\gamma}\sigma^2}{n}$ steps, then Theorem 1 guarantees that the averaged estimator $\bar{\theta}^T$ satisfies the bound

$$\|\bar{\theta}^T - \theta^*\|_n^2 \lesssim \left(\frac{1}{\alpha m} + \frac{1}{m^2}\right)\frac{\log^{1/\gamma} n}{n}\sigma^2,$$

with probability $1 - c_1 \exp(-c_2 m^2 \log^{1/\gamma} n)$.

d) β-polynomial decay: Now suppose that the kernel eigenvalues satisfy a decay condition of the form $\mu_j \leq c_1\ j^{-2\beta}$ for some $\beta > 1/2$ and constant c_1. In this case, a direct calculation yields the bounds

$$\mathcal{R}(\delta) \leq \sqrt{k\delta^2 + c_2\sum_{j=k+1}^{\infty} j^{-2}} \quad \text{and} \quad \mathcal{R}(\delta) \geq \sqrt{k\delta^2},$$

where k is the smallest integer such that $c_2\ k^{-2} < \delta^2$. Combined with upper bound

$$c_2 \sum_{j=k+1}^{n} j^{-2} \leq c_2 \int_{k+1} j^{-2} \leq k\delta^2,$$

we find that the critical radius scales as $\overline{\delta}_n^2 \asymp n^{-2\beta/(1+2\beta)}$.

Consequently, if we take $T \asymp n^{-2\beta/(1+2\beta)}$ many steps, then Theorem 1 guarantees that the averaged estimator $\bar{\theta}^T$ satisfies the bound

$$\|\bar{\theta}^T - \theta^*\|_n^2 \leq \left(\frac{1}{\alpha m} + \frac{1}{m^2}\right)\left(\frac{\sigma^2}{n}\right)^{2\beta/(2\beta+1)},$$

with probability at least $1 - c_1 \exp(-c_2 m^2 (\frac{n}{\sigma^2})^{1/(2\beta+1)})$.

VI. Discussion

In this paper, we have proven non-asymptotic bounds for early stopping of kernel boosting for a relatively broad class of loss functions. These bounds allowed us to propose simple stopping rules which, for the class of regular kernel functions [40], yield minimax optimal rates of estimation. Although the connection between early stopping and regularization has long been studied and explored in the theoretical literature and applications alike, to the best of our knowledge, this paper is the first one to establish a general relationship between the statistical optimality of stopped iterates and the localized Gaussian complexity. This connection is important, because this localized Gaussian complexity measure, as well as its Rademacher analogue, are now well-understood to play a central role in controlling the behavior of estimators based on regularization [3], [21], [34], [39].

There are various open questions suggested by our results. The stopping rules in this paper depend on the eigenvalues of the empirical kernel matrix; for this reason, they are data-dependent and computable given the data. However, in practice, it would be desirable to avoid the cost of computing all the empirical eigenvalues. Can fast approximation techniques for kernels be used to approximately compute our optimal stopping rules? Second, our current theoretical results apply to the averaged estimator $\bar{f}^T$. We strongly suspect that the same results apply to the stopped estimator f^T, but some new ingredients are required to extend our proofs.

Appendix

A. Notation and Three Equivalent Iteration Forms

Before directly diving into the proofs, let us first introduce some shorthand notation and provide more details on RKHS that are relevant for developing our main results.

Recalling the discussion in Section II-C, we denote the vector of function values of a function $f \in \mathscr{H}$ evaluated at $(x_1, x_2, \ldots, x_n)$ as

$$\theta_f := f(x_1^n) = (f(x_1), f(x_2), \ldots f(x_n)) \in \mathbb{R}^n,$$

where we omit the subscript f when it is clear from the context. As mentioned in the main text, updates on the function value vectors $\theta^t \in \mathbb{R}^n$ correspond uniquely to updates of the functions $f^t \in \mathscr{H}_n$. Therefore updates on f^t which is written as

$$f^{t+1} = f^t - \alpha \nabla \mathcal{L}_n(f^t), \quad \text{with } f^0 = 0, \tag{40a}$$

can be written as updates on the function value

$$\theta^{t+1} = \theta^t - n\alpha K \nabla \tilde{\mathcal{L}}_n(\theta^t), \quad \text{with } \theta^0 = 0, \tag{40b}$$

where

$$\mathcal{L}_n(f) = \tilde{\mathcal{L}_n}(\theta^t) = \frac{1}{n}\sum_{i=1}^n \phi(Y_i, \theta_i).$$

Denote $K^\dagger$ as the pseudoinverse of K, in our proof, we also use the linear transformation

$$z := n^{-1/2}(K^\dagger)^{1/2}\theta \iff \theta = \sqrt{n}K^{1/2}z.$$

as well as the new function $\mathcal{J}_n(z) := \tilde{\mathcal{L}_n}(\sqrt{n}\sqrt{K}z)$ and its population equivalent $\mathcal{J}(z) := \mathbb{E}\mathcal{J}_n(z)$. Ordinary gradient descent on $\mathcal{J}_n$ with stepsize α takes the form

$$\begin{aligned} z^{t+1} &= z^t - \alpha\nabla\mathcal{J}_n(z^t) \\ &= z^t - \alpha\sqrt{n}\sqrt{K}\nabla\tilde{\mathcal{L}_n}(\sqrt{n}\sqrt{K}z^t), \end{aligned} \tag{40c}$$

with $z^0 = 0$. If we transform this update on z back to an equivalent one on θ by multiplying both sides by $\sqrt{n}\sqrt{K}$, we see that ordinary gradient descent on $\mathcal{J}_n$ is equivalent to the kernel boosting update $\theta^{t+1} = \theta^t - \alpha nK\nabla\tilde{\mathcal{L}_n}(\theta^t)$.

Therefore in our context, these three forms (40a)-(40c) are equivalent to each other, and we interchange between them from time to time for convenience of the proof. Also note that we abuse notation and write $\mathcal{L}_n$ for both $\mathcal{L}_n$ and $\tilde{\mathcal{L}_n}$ as it is clear from the argument which loss we refer to.

B. Proof of Lemma 2

Recalling that $K^\dagger$ denotes the pseudoinverse of K(see Section A), our proof is based on the linear transformation

$$z := n^{-1/2}(K^\dagger)^{1/2}\theta \iff \theta = \sqrt{n}K^{1/2}z.$$

and iterates (40c) where

$$\begin{aligned} z^{t+1} &= z^t - \alpha\nabla\mathcal{J}_n(z^t) \\ & z^t - \alpha\sqrt{n}\sqrt{K}\nabla\mathcal{L}_n(\sqrt{n}\sqrt{K}z^t). \end{aligned}$$

Our goal is to analyze the behavior of the update (40c) in terms of the population cost $\mathcal{J}(z^t)$. Thus, our problem is one of analyzing a noisy form of gradient descent on the function $\mathcal{J}$, where the noise is induced by the difference between the empirical gradient operator $\nabla\mathcal{J}_n$ and the population gradient operator $\nabla\mathcal{J}$.

We show in the first part of the proof of Lemma 4 (without having to use the statement of this Lemma 2), that for all $t \le \frac{m}{8M\delta_n^2}$ we have $\|\Delta^t\|_{\mathscr{H}} \le 2C_{\mathscr{H}}$. Therefore we can readily assume that the $\mathcal{L}$ is M-smooth. Since the kernel matrix K has been normalized to have largest eigenvalue at most one, the function $\mathcal{J}$ is also M-smooth, whence

$$\mathcal{J}(z^{t+1}) \le \mathcal{J}(z^t) + \langle\nabla\mathcal{J}(z^t),\, d^t\rangle + \frac{M}{2}\|d^t\|_2^2, \quad \text{where} \quad d^t := z^{t+1} - z^t = -\alpha\nabla\mathcal{J}_n(z^t).$$

Morever, since the function $\mathcal{J}$ is convex, we have $\mathcal{J}(z^*) \ge \mathcal{J}(z^t) + \langle\nabla\mathcal{J}(z^t),\, z^* - z^t\rangle$, whence

$$\begin{aligned} \mathcal{J}(z^{t+1}) - \mathcal{J}(z^*) &\le \langle\nabla\mathcal{J}(z^t),\, d^t + z^t - z^*\rangle + \frac{M}{2}\|d^t\|_2^2 \\ &= \langle\nabla\mathcal{J}(z^t),\, z^{t+1} - z^*\rangle + \frac{M}{2}\|d^t\|_2^2. \end{aligned} \tag{41}$$

Now define the difference of the squared errors

$$V^t := \frac{1}{2}\Big\{\|z^t - z^*\|_2^2 - \|z^{t+1} - z^*\|_2^2\Big\}.$$

By some simple algebra, we have

$$\begin{aligned} V^t &= \frac{1}{2}\Big\{\|z^t - z^*\|_2^2 - \|d^t + z^t - z^*\|_2^2\Big\} \\ &= -\langle d^t,\, z^t - z^*\rangle - \frac{1}{2}\|d^t\|_2^2 \\ &= -\langle d^t,\, -d^t + z^{t+1} - z^*\rangle - \frac{1}{2}\|d^t\|_2^2 \\ &= -\langle d^t,\, z^{t+1} - z^*\rangle + \frac{1}{2}\|d^t\|_2^2. \end{aligned}$$

Substituting back into equation (41) yields

$$\begin{aligned} \mathcal{J}(z^{t+1}) - \mathcal{J}(z^*) &\le \frac{1}{\alpha}V^t + \langle\nabla\mathcal{J}(z^t) + \frac{d^t}{\alpha},\, z^{t+1} - z^*\rangle \\ &= \frac{1}{\alpha}V^t + \langle\nabla\mathcal{J}(z^t) - \nabla\mathcal{J}_n(z^t),\, z^{t+1} - z^*\rangle, \end{aligned}$$

where we have used the fact that $\frac{1}{\alpha} \ge M$ by our choice of stepsize α.

Finally, we transform back to the original variables $\theta = \sqrt{n}\sqrt{K}z$, using the relation $\nabla\mathcal{J}(z) = \sqrt{n}\sqrt{K}\nabla\mathcal{L}(\theta)$, so as to obtain the bound

$$\begin{aligned} \mathcal{L}(\theta^{t+1}) - \mathcal{L}(\theta^*) \le \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\} \\ + \langle\nabla\mathcal{L}(\theta^t) - \nabla\mathcal{L}_n(\theta^t),\, \theta^{t+1} - \theta^*\rangle. \end{aligned}$$

Note that the optimality of θ^* implies that $\nabla\mathcal{L}(\theta^*) = 0$. Combined with m-strong convexity, we are guaranteed that $\frac{m}{2}\|\Delta^{t+1}\|_n^2 \le \mathcal{L}(\theta^{t+1}) - \mathcal{L}(\theta^*)$, and hence

$$\begin{aligned} \frac{m}{2}\|\Delta^{t+1}\|_n^2 \le \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\} \\ + \langle\nabla\mathcal{L}(\theta^* + \Delta^t) - \nabla\mathcal{L}_n(\theta^* + \Delta^t),\, \Delta^{t+1}\rangle, \end{aligned}$$

as claimed.

C. Proof of Lemma 3

We split our proof into two cases, depending on whether we are dealing with the least-squares loss $\phi(y, \theta) = \frac{1}{2}(y - \theta)^2$, or a classification loss with uniformly bounded gradient ($\|\phi'\|_\infty \le 1$).

1) Least-Squares Case: The least-squares loss is m-strongly convex with $m = M = 1$. Moreover, the difference between the population and empirical gradients can be written as $\nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}) = \frac{\sigma}{n}(w_1, \ldots, w_n)$, where the random variables $\{w_i\}_{i=1}^n$ are i.i.d. and sub-Gaussian with parameter 1. Consequently, we have

$$|\langle\nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}),\, \Delta\rangle| = \left|\frac{\sigma}{n}\sum_{i=1}^n w_i\,\Delta(x_i)\right|.$$

Under these conditions, one can show (see [39] Lemma 13.4.) that

$$\left|\frac{\sigma}{n}\sum_{i=1}^n w_i\,\Delta(x_i)\right| \le 2\delta_n\|\Delta\|_n + 2\delta_n^2\|\Delta\|_{\mathscr{H}} + \frac{1}{16}\|\Delta\|_n^2, \tag{42}$$

which implies that Lemma 3 holds with $c_3 = 16$.

2) Gradient-Bounded ϕ-Functions: We now turn to the proof of Lemma 3 for gradient bounded ϕ-functions. First, we claim that it suffices to prove the bound (27) for functions $g \in \partial\mathscr{H}$ and $\|g\|_{\mathscr{H}} = 1$ where $\partial\mathscr{H} := \{f - g \mid f, g \in \mathscr{H}\}$. Indeed, suppose that it holds for all such functions, and that we are given a function Δ with $\|\Delta\|_{\mathscr{H}} > 1$. By assumption, we can apply the inequality (27) to the new function $g := \Delta/\|\Delta\|_{\mathscr{H}}$, which belongs to $\partial\mathscr{H}$ by nature of the subspace $\mathscr{H} = \overline{\text{span}}\{\mathbb{K}(\cdot, x_i)\}_{i=1}^n$.

Applying the bound (27) to g and then multiplying both sides by $\|\Delta\|_{\mathscr{H}}$, we obtain

$$\begin{aligned}\langle \nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}),\ \Delta\rangle &\le 2\delta_n\|\Delta\|_n + 2\delta_n^2\|\Delta\|_{\mathscr{H}} + \frac{m}{c_3}\frac{\|\Delta\|_n^2}{\|\Delta\|_{\mathscr{H}}} \\ &\le 2\delta_n\|\Delta\|_n + 2\delta_n^2\|\Delta\|_{\mathscr{H}} + \frac{m}{c_3}\|\Delta\|_n^2,\end{aligned}$$

where the second inequality uses the fact that $\|\Delta\|_{\mathscr{H}} > 1$ by assumption.

In order to establish the bound (27) for functions with $\|g\|_{\mathscr{H}} = 1$, we first prove it uniformly over the set $\{g \mid \|g\|_{\mathscr{H}} = 1,\ \ \|g\|_n \le t\}$, where $t > 1$ is a fixed radius (of course, we restrict our attention to those radii t for which this set is non-empty.) We then extend the argument to one that is also uniform over the choice of t by a "peeling" argument.

Define the random variable

$$\mathcal{Z}_n(t) := \sup_{\Delta, \widetilde{\Delta} \in \mathcal{E}(t,1)} \langle \nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}),\ \Delta\rangle. \tag{43}$$

The following two lemmas, respectively, bound the mean of this random variable, and its deviations above the mean:

Lemma 6. *For any $t > 0$, the mean is upper bounded as*

$$\mathbb{E}\mathcal{Z}_n(t) \le \sigma\mathcal{G}_n(\mathcal{E}(t, 1)), \tag{44}$$

where $\sigma := 2M + 4C_{\mathscr{H}}$.

Lemma 7. *There are universal constants (c_1, c_2) such that*

$$\mathbb{P}\Big[\mathcal{Z}_n(t) \ge \mathbb{E}\mathcal{Z}_n(t) + \alpha\Big] \le c_1 \exp\Big(-\frac{c_2 n\alpha^2}{t^2}\Big). \tag{45}$$

See Appendices C.3 and C.4 for the proofs of these two claims.

Equipped with Lemmas 6 and 7, we now prove inequality (27). We divide our argument into two cases:

a) Case $t = \delta_n$: We first prove inequality (27) for $t = \delta_n$. From Lemma 6, we have

$$\mathbb{E}\mathcal{Z}_n(\delta_n) \le \sigma\mathcal{G}_n(\mathcal{E}(\delta_n, 1)) \overset{(i)}{\le} \delta_n^2, \tag{46}$$

where inequality (i) follows from the definition of δ_n in inequality (14). Setting $\alpha = \delta_n^2$ in expression (45) yields

$$\mathbb{P}\Big[\mathcal{Z}_n(\delta_n) \ge 2\delta_n^2\Big] \le c_1 \exp\left(-c_2 n\delta_n^2\right), \tag{47}$$

which establishes the claim for $t = \delta_n$.

b) Case $t > \delta_n$: On the other hand, for any $t > \delta_n$, we have

$$\mathbb{E}\mathcal{Z}_n(t) \overset{(i)}{\le} \sigma\mathcal{G}_n(\mathcal{E}(t, 1)) \overset{(ii)}{\le} t\sigma\frac{\mathcal{G}_n(\mathcal{E}(t, 1))}{t} \le t\delta_n,$$

where step (i) follows from Lemma 6, and step (ii) follows because the function $u \mapsto \frac{\mathcal{G}_n(\mathcal{E}(u,1))}{u}$ is non-increasing on the positive real line. (This non-increasing property is a direct consequence of the star-shaped nature of $\partial\mathscr{H}$.) Finally, using this upper bound on expression $\mathbb{E}\mathcal{Z}_n(\delta_n)$ and setting $\alpha = t^2m/(4c_3)$ in the tail bound (45) yields

$$\mathbb{P}\Big[\mathcal{Z}_n(t) \ge t\delta_n + \frac{t^2 m}{4c_3}\Big] \le c_1 \exp\left(-c_2 n m^2 t^2\right). \tag{48}$$

Note that the precise values of the universal constants c_2 may change from line to line throughout this section.

c) Peeling argument: Equipped with the tail bounds (47) and (48), we are now ready to complete the peeling argument. Let $\mathcal{A}$ denote the event that the bound (27) is violated for some function $g \in \partial\mathscr{H}$ with $\|g\|_{\mathscr{H}} = 1$. For real numbers $0 \le a < b$, let $\mathcal{A}(a, b)$ denote the event that it is violated for some function such that $\|g\|_n \in [a, b]$, and $\|g\|_{\mathscr{H}} = 1$. For $k = 0, 1, 2, \ldots$, define $t_k = 2^k\delta_n$. We then have the decomposition $\mathcal{E} = (0, t_0) \cup (\bigcup_{k=0}^{\infty} \mathcal{A}(t_k, t_{k+1}))$ and hence by union bound,

$$\mathbb{P}[\mathcal{E}] \le \mathbb{P}[\mathcal{A}(0, \delta_n)] + \sum_{k=1}^{\infty} \mathbb{P}[\mathcal{A}(t_k, t_{k+1})]. \tag{49}$$

From the bound (47), we have $\mathbb{P}[\mathcal{A}(0, \delta_n)] \le c_1 \exp\left(-c_2\ n\delta_n^2\right)$. On the other hand, suppose that $\mathcal{A}(t_k, t_{k+1})$ holds, meaning that there exists some function g with $\|g\|_{\mathscr{H}} = 1$ and $\|g\|_n \in [t_k, t_{k+1}]$ such that

$$\begin{aligned}\langle \nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}),\ g\rangle &> 2\delta_n\|g\|_n + 2\delta_n^2 + \frac{m}{c_3}\|g\|_n^2 \\ &\overset{(i)}{\ge} 2\delta_n t_k + 2\delta_n^2 + \frac{m}{c_3}t_k^2 \\ &\overset{(ii)}{\ge} \delta_n t_{k+1} + 2\delta_n^2 + \frac{m}{4c_3}t_{k+1}^2,\end{aligned}$$

where step (i) uses the $\|g\|_n \ge t_k$ and step (ii) uses the fact that $t_{k+1} = 2t_k$. This lower bound implies that $\mathcal{Z}_n(t_{k+1}) > t_{k+1}\delta_n + \frac{t_{k+1}^2 m}{4c_3}$ and applying the tail bound (48) yields

$$\begin{aligned}\mathbb{P}(\mathcal{A}(t_k, t_{k+1})) &\le \mathbb{P}(\mathcal{Z}_n(t_{k+1}) > t_{k+1}\delta_n + \frac{t_{k+1}^2 m}{4\ c_3}) \\ &\le \exp\left(-c_2 n m^2 2^{2k+2}\delta_n^2\right).\end{aligned}$$

Substituting this inequality and our earlier bound (47) into equation (49) yields

$$\mathbb{P}(\mathcal{E}) \le c_1 \exp(-c_2 n m^2 \delta_n^2),$$

where the reader should recall that the precise values of universal constants may change from line-to-line. This concludes the proof of Lemma 3.

3) Proof of Lemma 6: Recalling the definitions (1) and (3) of $\mathcal{L}$ and $\mathcal{L}_n$, we can write

$$\mathcal{Z}_n(t) = \sup_{\Delta,\widetilde{\Delta}\in\mathcal{E}(t,1)} \frac{1}{n}\sum_{i=1}^{n}(\phi'(y_i,\theta_i^*+\widetilde{\Delta}_i) - \mathbb{E}\phi'(y_i,\theta_i^*+\widetilde{\Delta}_i))\Delta_i$$

Note that the vectors Δ and $\widetilde{\Delta}$ contain function values of the form $f(x_i) - f^*(x_i)$ for functions $f \in \mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$. Recall that the kernel function is bounded uniformly by one. Consequently, for any function $f \in \mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$, we have

$$\begin{aligned}|f(x) - f^*(x)| &= |\langle f - f^*, \mathbb{K}(\cdot,x)\rangle_{\mathscr{H}}| \\ &\le \|f - f^*\|_{\mathscr{H}}\|\mathbb{K}(\cdot,x)\|_{\mathscr{H}} \le 2C_{\mathscr{H}}.\end{aligned}$$

Thus, we can restrict our attention to vectors $\Delta, \widetilde{\Delta}$ with $\|\Delta\|_\infty, \|\widetilde{\Delta}\|_\infty \le 2C_{\mathscr{H}}$ from hereonwards.

Letting $\{\varepsilon_i\}_{i=1}^n$ denote an i.i.d. sequence of Rademacher variables, define the symmetrized variable

$$\tilde{\mathcal{Z}}_n(t) := \sup_{\Delta,\widetilde{\Delta}\in\mathcal{E}(t,1)} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\phi'(y_i,\theta_i^*+\widetilde{\Delta}_i)\ \Delta_i. \tag{50}$$

By a standard symmetrization argument [35], we have $\mathbb{E}_y[\mathcal{Z}_n(t)] \le 2\mathbb{E}_{y,\epsilon}[\tilde{\mathcal{Z}}_n(t)]$. Moreover, since

$$\phi'(y_i,\theta_i^*+\widetilde{\Delta}_i)\ \Delta_i \le \frac{1}{2}\Big(\phi'(y_i,\theta_i^*+\widetilde{\Delta}_i)\Big)^2 + \frac{1}{2}\Delta_i^2$$

we have

$$\begin{aligned}\mathbb{E}\mathcal{Z}_n(t) &\le \mathbb{E}\sup_{\widetilde{\Delta}\in\mathcal{E}(t,1)}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big(\phi'(y_i,\theta_i^*+\widetilde{\Delta}_i)\big)^2 \\ &\quad + \mathbb{E}\sup_{\Delta\in\mathcal{E}(t,1)}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta_i^2 \\ &\le 2\,\underbrace{\mathbb{E}\sup_{\widetilde{\Delta}\in\mathcal{E}(t,1)}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\phi'(y_i,\theta_i^*+\widetilde{\Delta}_i)}_{T_1} \\ &\quad + 4C_{\mathscr{H}}\underbrace{\mathbb{E}\sup_{\Delta\in\mathcal{E}(t,1)}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta_i}_{T_2},\end{aligned}$$

where the second inequality follows by applying the Rademacher contraction inequality [23], using the fact that $\|\phi'\|_\infty \le 1$ for the first term, and $\|\Delta\|_\infty \le 2C_{\mathscr{H}}$ for the second term.

Focusing first on the term T_1, since $\mathbb{E}[\varepsilon_i\phi'(y_i,\theta_i^*)] = 0$, we have

$$\begin{aligned}T_1 &= \mathbb{E}\sup_{\widetilde{\Delta}\in\mathcal{E}(t,1)}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\underbrace{\Big(\phi'(y_i,\theta_i^*+\widetilde{\Delta}_i) - \phi'(y_i;\theta_i^*)\Big)}_{\varphi_i(\widetilde{\Delta}_i)} \\ &\overset{(i)}{\le} M\mathbb{E}\sup_{\widetilde{\Delta}\in\mathcal{E}(t,1)}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\widetilde{\Delta}_i \\ &\overset{(ii)}{\le} \sqrt{\frac{\pi}{2}}M\mathcal{G}_n(\mathcal{E}(t,1)),\end{aligned}$$

where step (i) follows since each function φ_i is M-Lipschitz by assumption; and step (ii) follows since the Gaussian complexity upper bounds the Rademacher complexity up to a factor of $\sqrt{\frac{\pi}{2}}$. Similarly, we have

$$T_2 \le \sqrt{\frac{\pi}{2}}\,\mathcal{G}_n(\mathcal{E}(t,1)),$$

and putting together the pieces yields the claim.

4) Proof of Lemma 7: Recall the definition (50) of the symmetrized variable $\tilde{\mathcal{Z}}_n$. By a standard symmetrization argument [35], there are universal constants c_1, c_2 such that

$$\mathbb{P}\Big[\mathcal{Z}_n(t) \ge \mathbb{E}\mathcal{Z}_n[t] + c_1\alpha\Big] \le c_2\mathbb{P}\Big[\tilde{\mathcal{Z}}_n(t) \ge \mathbb{E}\tilde{\mathcal{Z}}_n[t] + \alpha\Big].$$

Since $\{\varepsilon_i\}_{i=1}^n$ are $\{y_i\}_{i=1}^n$ are independent, we can study $\tilde{\mathcal{Z}}_n(t)$ conditionally on $\{y_i\}_{i=1}^n$. Viewed as a function of $\{\varepsilon_i\}_{i=1}^n$, the function $\tilde{\mathcal{Z}}_n(t)$ is convex and Lipschitz with respect to the Euclidean norm with parameter

$$L^2 := \sup_{\Delta,\widetilde{\Delta}\in\mathcal{E}(t,1)}\frac{1}{n^2}\sum_{i=1}^{n}\Big(\phi'(y_i,\theta_i^*+\widetilde{\Delta}_i)\ \Delta_i\Big)^2 \ \le\ \frac{t^2}{n},$$

where we have used the facts that $\|\phi'\|_\infty \le 1$ and $\|\Delta\|_n \le t$. By Ledoux's concentration for convex and Lipschitz functions [22], we have

$$\mathbb{P}\Big[\tilde{\mathcal{Z}}_n(t) \ge \mathbb{E}\tilde{\mathcal{Z}}_n[t] + \alpha \ \mid\ \{y_i\}_{i=1}^n\Big] \le c_3\exp\Big(-c_4\frac{n\alpha^2}{t^2}\Big).$$

Since the right-hand side does not involve $\{y_i\}_{i=1}^n$, the same bound holds unconditionally over the randomness in both the Rademacher variables and the sequence $\{y_i\}_{i=1}^n$. Consequently, the claimed bound (45) follows, with suitable redefinitions of the universal constants.

D. Proof of Lemma 4

We first require an auxiliary lemma, which we state and prove in the following section. We then prove Lemma 4 in Section D.2.

1) An Auxiliary Lemma: The following result relates the Hilbert norm of the error to the difference between the empirical and population gradients:

Lemma 8. *For any convex and differentiable loss function $\mathcal{L}$, the kernel boosting error $\Delta^{t+1} := \theta^{t+1} - \theta^*$ satisfies the bound*

$$\begin{aligned}\|\Delta^{t+1}\|_{\mathscr{H}}^2 &\le \|\Delta^t\|_{\mathscr{H}}\|\Delta^{t+1}\|_{\mathscr{H}} \\ &\quad + \alpha\langle\nabla\mathcal{L}(\theta^*+\Delta^t) - \nabla\mathcal{L}_n(\theta^*+\Delta^t),\ \Delta^{t+1}\rangle.\end{aligned} \tag{51}$$

Proof: Recall that $\|\Delta^t\|_{\mathscr{H}}^2 = \|\theta^t - \theta^*\|_{\mathscr{H}}^2 = \|z^t - z^*\|_2^2$ by definition of the Hilbert norm. Let us define the population update operator G on the population function $\mathcal{J}$ and the empirical update operator G_n on $\mathcal{J}_n$ as

$$\begin{aligned}G(z^t) &:= z^t - \alpha\nabla\mathcal{J}(\sqrt{n}\sqrt{K}z^t), \\ \text{and}\quad z^{t+1} &:= G_n(z^t) = z^t - \alpha\nabla\mathcal{J}_n(\sqrt{n}\sqrt{K}z^t).\end{aligned} \tag{52}$$

Since $\mathcal{J}$ is convex and smooth, it follows from standard arguments in convex optimization that G is a non-expansive operator—viz.

$$\|G(x) - G(y)\|_2 \leq \|x - y\|_2 \quad \text{for all } x, y \in \mathcal{C}. \tag{53}$$

In addition, we note that the vector z^* is a fixed point of G—that is, $G(z^*) = z^*$. From these ingredients, we have

$$\begin{aligned}
&\|\Delta^{t+1}\|_{\mathcal{H}}^2 \\
&= \langle z^{t+1} - z^*,\ G_n(z^t) - G(z^t) + G(z^t) - z^* \rangle \\
&\overset{(i)}{\leq} \|z^{t+1} - z^*\|_2 \|G(z^t) - G(z^*)\|_2 \\
&\quad + \alpha \langle \sqrt{n}\sqrt{K}[\nabla \mathcal{L}(\theta^* + \Delta^t) - \nabla \mathcal{L}_n(\theta^* + \Delta^t)],\ z^{t+1} - z^* \rangle \\
&\overset{(ii)}{\leq} \|\Delta^{t+1}\|_{\mathcal{H}} \|\Delta^t\|_{\mathcal{H}} \\
&\quad + \alpha \langle \nabla \mathcal{L}(\theta^* + \Delta^t) - \nabla \mathcal{L}_n(\theta^* + \Delta^t),\ \Delta^{t+1} \rangle
\end{aligned}$$

where step (i) follows by applying the Cauchy-Schwarz to control the inner product, and step (ii) follows since $\Delta^{t+1} = \sqrt{n}\sqrt{K}(z^{t+1} - z^*)$, and the square root kernel matrix $\sqrt{K}$ is symmetric. ■

2) Proof of Lemma 4: We now prove Lemma 4. The argument makes use of Lemmas 2 and 3 combined with Lemma 8.

In order to prove inequality (28), we follow an inductive argument. Instead of proving (28) directly, we prove a slightly stronger relation which implies it, namely

$$\max\{1, \|\Delta^t\|_{\mathcal{H}}^2\} \leq \max\{1, \|\Delta^0\|_{\mathcal{H}}^2\} + t\delta_n^2 \frac{4M}{\widetilde{\gamma} m}. \tag{54}$$

Here $\widetilde{\gamma}$ and c_3 are constants linked by the relation

$$\widetilde{\gamma} := \frac{1}{32} - \frac{1}{4c_3} = 1/C_{\mathcal{H}}^2. \tag{55}$$

We claim that it suffices to prove that the error iterates Δ^{t+1} satisfy the inequality (54). Indeed, if we take inequality (54) as given, then we have

$$\|\Delta^t\|_{\mathcal{H}}^2 \leq \max\{1, \|\Delta^0\|_{\mathcal{H}}^2\} + \frac{1}{2\widetilde{\gamma}} \leq C_{\mathcal{H}}^2,$$

where we used the definition $C_{\mathcal{H}}^2 = 2\max\{\|\theta^*\|_{\mathcal{H}}^2,\ 32\}$. Thus, it suffices to focus our attention on proving inequality (54).

For $t = 0$, it is trivially true. Now let us assume inequality (54) holds for some $t \leq \frac{m}{8M\delta_n^2}$, and then prove that it also holds for step $t + 1$.

If $\|\Delta^{t+1}\|_{\mathcal{H}} < 1$, then inequality (54) follows directly. Therefore, we can assume without loss of generality that $\|\Delta^{t+1}\|_{\mathcal{H}} \geq 1$.

We break down the proof of this induction into two steps:

- First, we show that $\|\Delta^{t+1}\|_{\mathcal{H}} \leq 2C_{\mathcal{H}}$ so that Lemma 3 is applicable.
- Second, we show that the bound (54) holds and thus in fact $\|\Delta^{t+1}\|_{\mathcal{H}} \leq C_{\mathcal{H}}$.

Throughout the proof, we condition on the event $\mathcal{E}$ and $\mathcal{E}_0 := \{\frac{1}{\sqrt{n}}\|y - \mathbb{E}[y \mid x]\|_2 \leq \sqrt{2}\sigma\}$. Lemma 3 guarantees that $\mathbb{P}(\mathcal{E}^c) \leq c_1 \exp(-c_2 \frac{m^2 n\delta_n^2}{\sigma^2})$ whereas $\mathbb{P}(\mathcal{E}_0) \geq 1 - e^{-n}$ follows from the fact that Y^2 is sub-exponential with parameter $\sigma^2 n$ and applying Hoeffding's inequality. Putting things together yields an upper bound on the probability of the complementary event, namely

$$\mathbb{P}(\mathcal{E}^c \cup \mathcal{E}_0^c) \leq 2c_1 \exp(-C_2 n\delta_n^2)$$

with $C_2 = \max\{\frac{m^2}{\sigma^2}, 1\}$.

a) Showing that $\|\Delta^{t+1}\|_{\mathcal{H}} \leq 2C_{\mathcal{H}}$: In this step, we assume that inequality (54) holds at step t, and show that $\|\Delta^{t+1}\|_{\mathcal{H}} \leq 2C_{\mathcal{H}}$. Recalling that $z := \frac{(K^{\dagger})^{1/2}}{\sqrt{n}}\theta$, our update can be written as

$$\begin{aligned}
z^{t+1} - z^* = z^t - \alpha\sqrt{n}\sqrt{K}\nabla \mathcal{L}(\theta^t) - z^* \\
+ \alpha\sqrt{n}\sqrt{K}(\nabla \mathcal{L}_n(\theta^t) - \nabla \mathcal{L}(\theta^t)).
\end{aligned}$$

Applying the triangle inequality yields the bound

$$\begin{aligned}
\|z^{t+1} - z^*\|_2 \leq \|\underbrace{z^t - \alpha\sqrt{n}\sqrt{K}\nabla \mathcal{L}(\theta^t)}_{G(z^t)} - z^*\|_2 \\
+ \|\alpha\sqrt{n}\sqrt{K}(\nabla \mathcal{L}_n(\theta^t) - \nabla \mathcal{L}(\theta^t))\|_2
\end{aligned}$$

where the population update operator G was previously defined (52), and observed to be non-expansive (53). From this non-expansiveness, we find that

$$\begin{aligned}
\|z^{t+1} - z^*\|_2 \leq \|z^t - z^*\|_2 \\
+ \|\alpha\sqrt{n}\sqrt{K}(\nabla \mathcal{L}_n(\theta^t) - \nabla \mathcal{L}(\theta^t))\|_2,
\end{aligned}$$

Note that the ℓ_2 norm of z corresponds to the Hilbert norm of θ. This implies

$$\begin{aligned}
\|\Delta^{t+1}\|_{\mathcal{H}} \leq \|\Delta^t\|_{\mathcal{H}} + \\
\underbrace{\|\alpha\sqrt{n}\sqrt{K}(\nabla \mathcal{L}_n(\theta^t) - \nabla \mathcal{L}(\theta^t))\|_2}_{:=T}
\end{aligned}$$

Observe that because of uniform boundedness of the kernel by one, the quantity T can be bounded as

$$T \leq \alpha\sqrt{n}\|\nabla \mathcal{L}_n(\theta^t) - \nabla \mathcal{L}(\theta^t))\|_2 = \alpha\sqrt{n}\frac{1}{n}\|v - \mathbb{E}v\|_2,$$

where we have defined the vector $v \in \mathbb{R}^n$ with coordinates $v_i := \phi'(y_i, \theta_i^t)$. For functions ϕ satisfying the gradient boundedness and $m - M$ condition, since $\theta^t \in \mathbb{B}_{\mathcal{H}}(\theta^*, C_{\mathcal{H}})$, each coordinate of the vectors v and $\mathbb{E}v$ is bounded by 1 in absolute value. We consequently have

$$T \leq \alpha \leq C_{\mathcal{H}},$$

where we have used the fact that $\alpha \leq m/M < 1 \leq \frac{C_{\mathcal{H}}}{2}$. For least-squares ϕ we instead have

$$T \leq \alpha\frac{\sqrt{n}}{n}\|y - \mathbb{E}[y \mid x]\|_2 =: \frac{\alpha}{\sqrt{n}}Y \leq \sqrt{2}\sigma \leq C_{\mathcal{H}}$$

conditioned on the event $\mathcal{E}_0 := \{\frac{1}{\sqrt{n}}\|y - \mathbb{E}[y \mid x]\|_2 \leq \sqrt{2}\sigma\}$. Since Y^2 is sub-exponential with parameter $\sigma^2 n$ it follows by Hoeffding's inequality that $\mathbb{P}(\mathcal{E}_0) \geq 1 - e^{-n}$. Putting together the pieces yields that $\|\Delta^{t+1}\|_{\mathcal{H}} \leq 2C_{\mathcal{H}}$, as claimed.

b) Completing the induction step: We are now ready to complete the induction step for proving inequality (54) using Lemma 2 and Lemma 3 since $\|\Delta^{t+1}\|_{\mathscr{H}} \geq 1$. We split the argument into two cases separately depending on whether or not $\|\Delta^{t+1}\|_{\mathscr{H}}\delta_n \geq \|\Delta^{t+1}\|_n$. In general we can assume that $\|\Delta^{t+1}\|_{\mathscr{H}} > \|\Delta^t\|_{\mathscr{H}}$, otherwise the induction inequality (54) is satisfied trivially.

c) Case 1: When $\|\Delta^{t+1}\|_{\mathscr{H}}\delta_n \geq \|\Delta^{t+1}\|_n$, inequality (27) implies that

$$\langle \nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}),\ \Delta^{t+1}\rangle \leq 4\delta_n^2\|\Delta^{t+1}\|_{\mathscr{H}} + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2, \tag{56}$$

Combining Lemma 8 and inequality (56), we obtain

$$\|\Delta^{t+1}\|_{\mathscr{H}}^2 \leq \|\Delta^t\|_{\mathscr{H}}\|\Delta^{t+1}\|_{\mathscr{H}} + 4\alpha\delta_n^2\|\Delta^{t+1}\|_{\mathscr{H}} + \alpha\frac{m}{c_3}\|\Delta^{t+1}\|_n^2$$
$$\implies \|\Delta^{t+1}\|_{\mathscr{H}} \leq \frac{1}{1-\alpha\delta_n^2\frac{m}{c_3}}\big[\|\Delta^t\|_{\mathscr{H}} + 4\alpha\delta_n^2\big], \tag{57}$$

where the last inequality uses the fact that $\|\Delta^{t+1}\|_n \leq \delta_n\|\Delta^{t+1}\|_{\mathscr{H}}$.

d) Case 2: When $\|\Delta^{t+1}\|_{\mathscr{H}}\delta_n < \|\Delta^{t+1}\|_n$, we use our assumption $\|\Delta^{t+1}\|_{\mathscr{H}} \geq \|\Delta^t\|_{\mathscr{H}}$ together with Lemma 8 and inequality (27). Collectively, they guarantee that

$$\begin{aligned}\|\Delta^{t+1}\|_{\mathscr{H}}^2 &\leq \|\Delta^t\|_{\mathscr{H}}^2 + 2\alpha\langle\nabla\mathcal{L}(\theta^*+\Delta^t) - \nabla\mathcal{L}_n(\theta^*+\Delta^t),\ \Delta^{t+1}\rangle\\ &\leq \|\Delta^t\|_{\mathscr{H}}^2 + 8\alpha\delta_n\|\Delta^{t+1}\|_n + 2\alpha\frac{m}{c_3}\|\Delta^{t+1}\|_n^2.\end{aligned}$$

Using the elementary inequality $2ab \leq a^2 + b^2$, we find that

$$\begin{aligned}\|\Delta^{t+1}\|_{\mathscr{H}}^2 &\leq \|\Delta^t\|_{\mathscr{H}}^2 + 8\alpha\left[m\widetilde{\gamma}\|\Delta^{t+1}\|_n^2 + \frac{1}{4\widetilde{\gamma}m}\delta_n^2\right] + 2\alpha\frac{m}{c_3}\|\Delta^{t+1}\|_n^2\\ &\leq \|\Delta^t\|_{\mathscr{H}}^2 + \alpha\frac{m}{4}\|\Delta^{t+1}\|_n^2 + \frac{2\alpha\delta_n^2}{\widetilde{\gamma}m},\end{aligned} \tag{58}$$

where in the final step, we plug in the constants $\widetilde{\gamma}, c_3$ which satisfy equation (55).

Now Lemma 2 implies that

$$\begin{aligned}\frac{m}{2}\|\Delta^{t+1}\|_n^2 &\leq D^t + 4\|\Delta^{t+1}\|_n\delta_n + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2\\ &\overset{(i)}{\leq} D^t + 4\left[\widetilde{\gamma}m\|\Delta^{t+1}\|_n^2 + \frac{1}{4\widetilde{\gamma}m}\delta_n^2\right] + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2,\end{aligned}$$

where step (i) again uses $2ab \leq a^2 + b^2$. Thus, we have $\frac{m}{4}\|\Delta^{t+1}\|_n^2 \leq D^t + \frac{1}{\widetilde{\gamma}m}\delta_n^2$. Together with expression (58), we find that

$$\begin{aligned}&\|\Delta^{t+1}\|_{\mathscr{H}}^2 \leq \|\Delta^t\|_{\mathscr{H}}^2 + \frac{1}{2}(\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2) + \frac{4\alpha}{\widetilde{\gamma}m}\delta_n^2\\ &\implies \|\Delta^{t+1}\|_{\mathscr{H}}^2 \leq \|\Delta^t\|_{\mathscr{H}}^2 + \frac{4\alpha}{\widetilde{\gamma}m}\delta_n^2.\end{aligned}$$

e) Combining the pieces: By combining the two previous cases, we arrive at the bound

$$\max\left\{1, \|\Delta^{t+1}\|_{\mathscr{H}}^2\right\} \leq \max\left\{1, \kappa^2(\|\Delta^t\|_{\mathscr{H}} + 4\alpha\delta_n^2)^2, \|\Delta^t\|_{\mathscr{H}}^2 + \frac{4M}{\widetilde{\gamma}m}\delta_n^2\right\}, \tag{59}$$

where $\kappa := \frac{1}{(1-\alpha\delta_n^2\frac{m}{c_3})}$ and we used the inequality $\alpha \leq \min\{\frac{1}{M}, M\}$.

Now it is only left for us to show that with the constant c_3 chosen such that $\widetilde{\gamma} = \frac{1}{32} - \frac{1}{4c_3} = 1/C_{\mathscr{H}}^2$, we have

$$\kappa^2(\|\Delta^t\|_{\mathscr{H}} + 4\alpha\delta_n^2)^2 \leq \|\Delta^t\|_{\mathscr{H}}^2 + \frac{4M}{\widetilde{\gamma}m}\delta_n^2.$$

Define the function $f : (0, C_{\mathscr{H}}] \to \mathbb{R}$ via

$$f(\xi) := \kappa^2(\xi + 4\alpha\delta_n^2)^2 - \xi^2 - \frac{4M}{\widetilde{\gamma}m}\delta_n^2.$$

Since $\kappa \geq 1$, in order to conclude that $f(\xi) < 0$ for all $\xi \in (0, C_{\mathscr{H}}]$, it suffices to show that $\operatorname{argmin}_{x\in\mathbb{R}} f(x) < 0$ and $f(C_{\mathscr{H}}) < 0$.

The former is obtained by basic algebra and follows directly from $\kappa \geq 1$. For the latter, since $\widetilde{\gamma} = \frac{1}{32} - \frac{1}{4c_3} = 1/C_{\mathscr{H}}^2$, $\alpha < \frac{1}{M}$ and $\delta_n^2 \leq \frac{M^2}{m^2}$ it thus suffices to show

$$\frac{1}{(1-\frac{M}{8m})^2} \leq \frac{4M}{m} + 1$$

Since $(4x+1)(1-\frac{x}{8})^2 \geq 1$ for all $x \leq 1$ and $\frac{m}{M} \leq 1$, we conclude that $f(C_{\mathscr{H}}) < 0$.

Now that we have established $\max\{1, \|\Delta^{t+1}\|_{\mathscr{H}}^2\} \leq \max\{1, \|\Delta^t\|_{\mathscr{H}}^2\} + \frac{4M}{\widetilde{\gamma}m}\delta_n^2$, the induction step (54) follows, which completes the proof of Lemma 4.

E. Proof of Lemma 1

Recall that the eigenfunctions $\{\phi_k\}_{k=1}^\infty$ associated with a kernel operator form an orthonormal basis of $L^2(\mathcal{X}, \mathbb{P}_X)$ with the inner product $\langle f, g\rangle := \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\mathbb{P}_X(x)$. Every function $f \in \mathscr{H}$ induced by the kernel can then be written as

$$f(x) = \sum_{j=1}^\infty \beta_j\phi_j(x)$$

with $\beta \in \ell^2(\mathbb{N})$ (i.e. $\sum_{j=1}^\infty \beta_j^2 < \infty$) and $\sum_{j=1}^\infty \frac{\beta_j^2}{\mu_j} < \infty$. It can be shown (see e.g. [39]) that the corresponding inner product of the function space $\mathscr{H}$ reads

$$\langle f, g\rangle_{\mathscr{H}} = \sum_{j=1}^\infty \frac{\langle f, \phi_j\rangle\langle g, \phi_j\rangle}{\mu_j}$$

and thus for f as defined in Equation (E) we have $\|f\|_{\mathscr{H}}^2 = \sum_{j=1}^\infty \frac{\beta_j^2}{\mu_j}$ as well as $\|f\|_2^2 = \sum_{j=1}^\infty \beta_j^2$. The localized population Gaussian complexity is defined as

$$\overline{\mathcal{G}}_n(\overline{\mathcal{E}}(\delta,1)) = \mathbb{E}_X\mathbb{E}_w \sup_{\|f\|_2\leq\delta, \|f\|_H\leq 1}\Big|\frac{1}{n}\sum_{i=1}^n w_i f(x_i)\Big|.$$

Rewriting the objective and defining $\tilde{w}_j = \sum_{i=1}^n w_i \phi_j(x_i)$, the inner constrained maximization can be upper bounded as follows

$$\sup_{\substack{\sum_{j=1}^\infty \beta_j^2 \leq \delta^2 \\ \sum_{j=1}^\infty \frac{\beta_j^2}{\mu_j} \leq 1}} \sum_{j=1}^\infty \tilde{w}_j \beta_j \leq \sup_{\sum_{j=1}^\infty \eta_j \beta_j^2 \leq 2} \sum_{j=1}^\infty \tilde{w}_j \beta_j$$

$$= \sup_{\sum_{j=1}^\infty \gamma_j^2 \leq 1} \sum_{j=1}^\infty \sqrt{\frac{2}{\eta_j}} \tilde{w}_j \gamma_j,$$

where $\eta_j = \max\{\delta^{-2}, \mu_j^{-1}\}$. Now using Hölders inequality we obtain

$$\sup_{\sum_{j=1}^\infty \beta_j^2 \leq \delta^2, \sum_{j=1}^\infty \frac{\beta_j^2}{\mu_j} \leq 1} \sum_{j=1}^\infty \tilde{w}_j \beta_j \leq \sqrt{\sum_{j=1}^\infty \frac{2\tilde{w}_j^2}{\eta_j}}.$$

Furthermore, using Jensen's inequality we have

$$\frac{1}{n}\mathbb{E}_X \mathbb{E}_w \sqrt{\sum_{j=1}^\infty \frac{2\tilde{w}_j^2}{\eta_j}} \leq \sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^\infty \frac{\mathbb{E}_{X,w} \tilde{w}_j^2}{\eta_j}}$$

$$\leq \sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^\infty \frac{1}{\eta_j}},$$

which concludes the proof.

F. Proof of Lemma 5

Recall that the LogitBoost algorithm is based on logistic loss $\phi(y,\theta) = \ln(1 + e^{-y\theta})$, whereas the AdaBoost algorithm is based on the exponential loss $\phi(y,\theta) = \exp(-y\theta)$. We now verify the m-M-condition for these two losses with the corresponding parameters specified in Lemma 5.

1) m-M-condition for Logistic Loss: The first and second derivatives are given by

$$\frac{\partial \phi(y,\theta)}{\partial \theta} = \frac{-y e^{-y\theta}}{1 + e^{-y\theta}}, \qquad \text{and}$$
$$\frac{\partial^2 \phi(y,\theta)}{(\partial \theta)^2} = \frac{y^2}{(e^{-y\theta/2} + e^{y\theta/2})^2}.$$

It is easy to check that $|\frac{\partial \phi(y,\theta)}{\partial \theta}|$ is uniformly bounded by $B = 1$.

Turning to the second derivative, recalling that $y \in \{-1, +1\}$, it is straightforward to show that

$$\max_{y \in \{-1,+1\}} \sup_\theta \frac{y^2}{(e^{-y\theta/2} + e^{y\theta/2})^2} \leq \frac{1}{4},$$

which implies that $\frac{\partial \phi(y,\theta)}{\partial \theta}$ is a 1/4-Lipschitz function of θ, i.e. with $M = 1/4$.

Our final step is to compute a value for m by deriving a uniform lower bound on the Hessian. For this step, we need to exploit the fact that $\theta = f(x)$ must arise from a function f such that $\|f\|_{\mathscr{H}} \leq D := C_{\mathscr{H}} + \|\theta^*\|_{\mathscr{H}}$. Since $\sup_x \mathbb{K}(x,x) \leq 1$ by assumption, the reproducing relation for RKHS then implies that $|f(x)| \leq D$. Combining this inequality with the fact that $y \in \{-1, 1\}$, it suffices to lower bound the quantity

$$\min_{y \in \{-1,+1\}} \min_{|\theta| \leq D} \left| \frac{\partial^2 \phi(y,\theta)}{(\partial \theta)^2} \right|$$
$$= \min_{|y| \leq 1} \min_{|\theta| \leq D} \frac{y^2}{(e^{-y\theta/2} + e^{y\theta/2})^2} \geq \underbrace{\frac{1}{e^{-D} + e^{D} + 2}}_{m},$$

which completes the proof for the logistic loss.

2) m-M-condition for AdaBoost: The AdaBoost algorithm is based on the cost function $\phi(y,\theta) = e^{-y\theta}$, which has first and second derivatives (with respect to its second argument) given by

$$\frac{\partial \phi(y,\theta)}{\partial \theta} = -y e^{-y\theta}, \qquad \text{and} \qquad \frac{\partial^2 \phi(y,\theta)}{(\partial \theta)^2} = e^{-y\theta}.$$

As in the preceding argument for logistic loss, we have the bound $|y| \leq 1$ and $|\theta| \leq D$. By inspection, the absolute value of the first derivative is uniformly bounded $B := e^D$, whereas the second derivative always lies in the interval $[m, M]$ with $M := e^D$ and $m := e^{-D}$, as claimed.

ACKNOWLEDGEMENT

Yuting Wei and Fanny Yang contributed equally to this work.

REFERENCES

[1] R. S. Anderssen and P. M. Prenter, "A formal comparison of methods proposed for the numerical solution of first kind integral equations," *ANZIAM J.*, vol. 22, no. 4, pp. 488–500, 1981.

[2] P. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Nov. 2002.

[3] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local rademacher complexities," *Ann. Statist.*, vol. 33, no. 4, pp. 1497–1537, 2005.

[4] P. L. Bartlett and M. Traskin, "Adaboost is consistent," *J. Mach. Learn. Res.*, vol. 8, pp. 2347–2368, Oct. 2007.

[5] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Norwell, MA, USA: Kluwer Academic, 2004.

[6] L. Breiman, "Prediction games and arcing algorithms," *Neural Comput.*, vol. 11, no. 7, pp. 1493–1517, Oct. 1999.

[7] L. Breiman, "Arcing classifier (with discussion and a rejoinder by the author)," *Ann. Statist.*, vol. 26, no. 3, pp. 801–849, 1998.

[8] P. Bühlmann and T. Hothorn, "Boosting algorithms: Regularization, prediction and model fitting," *Stat. Sci.*, vol. 22, no. 4, pp. 477–505, 2007.

[9] P. Bühlmann and B. Yu, "Boosting with L_2 loss: Regression and classification," *J. Amer. Stat. Assoc.*, vol. 98, no. 462, pp. 324–339, 2003.

[10] R. Camoriano, T. Angles, A. Rudi, and L. Rosasco, "NYTRO: When subsampling meets early stopping," in *Proc. 19th Int. Conf., Artif. Intell. Statist.*, 2016, pp. 1403–1411.

[11] A. Caponnetto, "Optimal rates for regularization operators in learning theory," Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. CBCL Paper 264/AI Tech. Rep. 062, Sep. 2006.

[12] A. Caponnetto and Y. Yao, "Adaptation for regularization operators in learning theory," Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. CBCL Paper 265/AI Tech. Rep. 063, Sep. 2006.

[13] R. Caruana, S. Lawrence, and C. L. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 402–408.

[14] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[15] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors)," *Ann. Statist.*, vol. 28, no. 2, pp. 337–407, 2000.

[16] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[17] C. Gu, *Smoothing Spline ANOVA Models* (Springer Series in Statistics). New York, NY, USA: Springer, 2002.
[18] L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression* (Springer Series in Statistics). New York, NY, USA: Springer-Verlag, 2002.
[19] W. Jiang, "Process consistency for adaboost," *Ann. Statist.*, vol. 21, no. 1, pp. 13–29, 2004.
[20] G. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *J. Math. Anal. Appl.*, vol. 33, pp. 82–95, Sep. 1971.
[21] V. Koltchinskii, "Local rademacher complexities and oracle inequalities in risk minimization," *Ann. Statist.*, vol. 34, no. 6, pp. 2593–2656, 2006.
[22] M. Ledoux, *The Concentration of Measure Phenomenon* (Mathematical Surveys and Monographs). Providence, RI, USA: American Mathematical Society, 2001.
[23] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. New York, NY, USA: Springer-Verlag, 1991.
[24] J. Lin, A. Rudi, L. Rosasco, and V. Cevher, "Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces," *Appl. Comput. Harmon. Anal.*, to be published. doi: 10.1016/j.acha.2018.09.009.
[25] L. Mason, J. Baxter, P. L. Bartlett, and M. R. Frean, "Boosting algorithms as gradient descent," in *Adv. Neural Inf. Process. Syst.*, vol. 12, 1999, pp. 512–518.
[26] S. Mendelson, "Geometric parameters of kernel machines," in *Proc. Conf. Learn. Theory (COLT)*, 2002, pp. 29–43.
[27] L. Prechelt, "Early stopping-but when?" in *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 1998, pp. 55–69.
[28] G. Raskutti, M. J. Wainwright, and B. Yu, "Early stopping and non-parametric regression: An optimal data-dependent stopping rule," *J. Mach. Learn. Res.*, vol. 15, pp. 335–366, Jan. 2014.
[29] L. Rosasco and S. Villa, "Learning with incremental iterative regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1630–1638.
[30] R. E. Schapire, "The strength of weak learnability," *Mach. Learn.*, vol. 5, no. 2, pp. 197–227, 1990.
[31] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Nonlinear Estimation and Classification*. New York, NY, USA: Springer, 2003, pp. 149–171.
[32] B. Schölkopf and A. Smola, *Learning With Kernels*. Cambridge, MA, USA: MIT Press, 2002.
[33] O. N. Strand, "Theory and methods related to the singular-function expansion and Landweber's iteration for integral equations of the first kind," *SIAM J. Numer. Anal.*, vol. 11, no. 4, pp. 798–825, 1974.
[34] S. van de Geer, *Empirical Processes in M-Estimation*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[35] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*. New York, NY, USA: Springer-Verlag, 1996.
[36] E. De Vito, S. Pereverzyev, and L. Rosasco, "Adaptive kernel methods using the balancing principle," *Found. Comput. Math.*, vol. 10, no. 4, pp. 455–479, 2010.
[37] G. Wahba, "Three topics in ill-posed problems," in *Inverse and Ill-Posed Problems*, H. W. Engl and G. W. Groetsch, Eds. New York, NY, USA: Academic, 1987, pp. 37–50.
[38] G. Wahba, "Spline models for observational data," in *Proc. CBMS-NSF Regional Conf. Ser. Appl. Math.* Philadelphia, PA, USA: SIAM, 1990, pp. 1–169.
[39] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2019.
[40] Y. Yang, M. Pilanci, and M. J. Wainwright, "Randomized sketches for kernels: Fast and optimal nonparametric regression," *Ann. Statist.*, vol. 45, no. 3, pp. 991–1023, 2017.
[41] Y. Yao, L. Rosasco, and A. Caponnetto, "On early stopping in gradient descent learning," *Construct. Approx.*, vol. 26, no. 2, pp. 289–315, 2007.
[42] T. Zhang and B. Yu, "Boosting with early stopping: Convergence and consistency," *Ann. Statist.*, vol. 33, no. 4, pp. 1538–1579, Aug. 2005.

Yuting Wei Yuting Wei is currently a Stein Fellow in Statistics Department at Stanford University. She received a Ph.D. degree in Statistics from University of California, Berkeley in 2018 and a B.S. in Statistics, a B.A. in Economics both from Peking University in China. Yuting was the recipient of the 2018 Erich L. Lehmann Citation from the Berkeley statistics department for an outstanding Ph.D. dissertation in theoretical statistics. Her research interests include non-parametric statistics, high dimensional statistical inference, optimization, and statistical information theory.

Fanny Yang Fanny Yang is a Junior Fellow at the Institute of Theoretical Studies at ETH Zurich and holds a joint appointment as postdoctoral scholar Department of Statistics and Department of Computer Sciences at Stanford University. She received a Bachelor's degree in Electrical Engineering from Karlsruhe Institute of Technology, Germany, a M. Sc. degree from the Technical University, Munich and a Ph.D. degree in EECS from University of California, Berkeley. She was the recipient of the Best Student Paper Award at the International Conference on Sampling Theory and Applications 2013, Berkeley Departmental Fellowship 2013/14 and Samuel Silver Memorial Scholarship Award 2015. Her research interests include non-parametric statistics, theory for neural networks, statistical and computational trade-offs in machine learning and high-dimensional statistics.

Martin J. Wainwright is Chancellor's Professor at the University of California at Berkeley, with a joint appointment between the Department of Statistics and the Department of Electrical Engineering and Computer Sciences (EECS). He received a Bachelor's degree in Mathematics from University of Waterloo, Canada, and Ph.D. degree in EECS from Massachusetts Institute of Technology (MIT). His research interests include high-dimensional statistics, information theory, statistical machine learning, and optimization theory. He has been awarded an Alfred P. Sloan Foundation Fellowship (2005), Best Paper Awards from the IEEE Signal Processing Society (2008), and IEEE Communications Society (2010); the Joint Paper Prize (2012) from IEEE Information Theory and Communication Societies; a Medallion Lectureship (2013) from the Institute of Mathematical Statistics; a Section Lecturer at the International Congress of Mathematicians (2014); the COPSS Presidents' Award (2014) from the Joint Statistical Societies; and the Blackwell Lecturer (2016) from the Institute of Mathematical Statistics.