Constr. Approx. (2007) 26: 289–315
DOI: 10.1007/s00365-006-0663-2

CONSTRUCTIVE
APPROXIMATION

On Early Stopping in Gradient Descent Learning

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto

Abstract. In this paper we study a family of gradient descent algorithms to approximate the regression function from reproducing kernel Hilbert spaces (RKHSs), the family being characterized by a polynomial decreasing rate of step sizes (or learning rate). By solving a bias-variance trade-off we obtain an early stopping rule and some probabilistic upper bounds for the convergence of the algorithms. We also discuss the implication of these results in the context of classification where some fast convergence rates can be achieved for plug-in classifiers. Some connections are addressed with Boosting, Landweber iterations, and the online learning algorithms as stochastic approximations of the gradient descent method.

1. Introduction

In this paper we investigate the approximation by random examples of the regression function from reproducing kernel Hilbert spaces (RKHSs). We study a family of gradient descent algorithms to solve a least square problem, the family being characterized by a polynomial decreasing rate of step sizes (or learning rate).

We focus on two iteration paths in RKHSs: one is the gradient flow for expected risk minimization which depends on the unknown probability measure and is called here the *population iteration*; the other is the gradient flow for empirical risk minimization based on the sample, called here the *sample iteration*. Both paths start from the origin and, as iterations go on, leave from each other. The population iteration converges to our target, the regression function; however, the sample iteration often converges to an overfitting function. Thus, keeping the two paths close may play a role of regularization to prevent the sample iteration from an overfitting function. This exhibits a *bias-variance* phenomenon: the distance between the population iteration and the regression function is called *bias* or *approximation error*; the gap between the two paths is called *variance* or *sample error*. Stopping too early may reduce variance but enlarge bias; and stopping too late may enlarge variance though reduced bias. Solving this bias-variance trade-off leads to an early stopping rule.

In the literature such a bias-variance view has been taken, explicitly or implicitly, by boosting as a gradient descent method, where scaled convex hulls of functions are typically used instead of RKHSs. The gap between the two paths (measured by some risk functional or distance) typically grows in proportion to the radius (sum of absolute

Date received: September 21, 2005. Dates revised: February 28, 2006 and August 1, 2006. Date accepted: September 6, 2006. Communicated by Ronald A. DeVore. Online publication: March 21, 2007.
AMS classification: 62G08, 62L99, 68Q32, 68T05.
Key words and phrases: Gradient descent method, Early stopping, Regularization, Boosting, Landweber iteration, Reproducing kernel Hilbert space.

values of convex combination coefficients, or l_1 norm) of the paths and thus restricting that radius implements regularization (e.g., Lugosi and Vayatis, 2004; Blanchard, Lugosi, and Vayatis, 2003). Recently, early stopping regularization was systematically studied in the society of machine learning, see, for example, Jiang (2004) for AdaBoost, Bühlmann and Yu (2002) for L_2Boost, Zhang and Yu (2003) for Boosting with general convex loss functions, and Bickel, Ritov, and Zakai (2005) for some generalized Boosting algorithms. It is also interesting to note that Zhao and Yu (2004) introduced some backward steps which have the effect of reducing the radius. Considering the square loss function, our paper can be regarded as a sort of L_2Boost, which, roughly speaking, extends some early results in Bühlmann and Yu (2002) from Sobolev spaces with fixed designs to general RKHSs with random designs (see Chapter I in Györfi, Kohler, Krzyżak, and Walk (2002) for more discussions on random design versus fixed design).

In this paper we show by probabilistic upper bounds that under the early stopping rule above, the proposed family of algorithms converges polynomially to the regression function subject to some regularity assumption, where the constant step size algorithm is the fastest one in the family since it requires the minimal number of iterations before stopping. The rates presented in this paper are suboptimal and are expected to be improved in various aspects. We also discuss the implications of our results in the context of classification by showing that under a suitable assumption on the noise (Tsybakov, 2004) some fast convergence rates to the Bayes classifier can be achieved.

Early stopping regularization has a crucial advantage over the usual regularized least square learning algorithm (e.g., Smale and Zhou, 2005; De Vito, Rosasco, Caponnetto, Giovannini, and Odone, 2004), which is also called *ridge regression* in statistical literature or *Tikhonov regularization* in inverse problems. Early stopping does not incur the *saturation* phenomenon in the sense that the rate no longer improves when the regression function goes beyond a certain level of regularity. The saturation problem was studied intensively in inverse problems (e.g., Engl, Hanke, and Neubauer, 2000; Mathé, 2004). Our algorithms here can be regarded as a randomized discretization of the Landweber iterations in linear inverse problems.

The organization of this paper is as follows. Section 2 summarizes the main results with discussions. In Section 3 we collect more discussions on related works. In detail, Section 3.1 compares early stopping and the usual penalized least square algorithm in learning; Section 3.2 discusses the connection to boosting in view of the gradient descent method; Section 3.3 discusses the connection to the Landweber iteration in linear inverse problems; Section 3.4 discusses the connection to the online learning as a stochastic gradient method. Sections 4 and 5 contribute to the proofs. Section 4 describes some crucial decompositions for later use. Section 5 presents the proofs of the upper bounds for the sample error and the approximation error. In Section 6 we apply the main theorem to the setting of classification. Finally, Section 7 summarizes the conclusions and open problems in this paper.

2. Main Results

2.1. *Definitions and Notations*

Let the input space $X \subseteq \mathbb{R}^n$ be closed, the output space $Y = \mathbb{R}$ and $Z = X \times Y$. Given a sample $\mathbf{z} = \{(x_i, y_i) \in X \times Y : i = 1, \ldots, m\} \in Z^m$, drawn independently at random from a probability measure ρ on Z, one wants to minimize over $f \in \mathscr{H}$ the following quadratic functional:

$$(1) \qquad \mathscr{E}(f) = \int_{X \times Y} (f(x) - y)^2 \, d\rho$$

where $\mathscr{H}$ is some Hilbert space of real functions on X. In this paper we choose $\mathscr{H}$ as a *reproducing kernel Hilbert space* (RKHS), in which the gradient map takes an especially simple form.

Here we recall some basic definitions on RKHSs. Let $K : X \times X \to \mathbb{R}$ be a *Mercer kernel*, i.e., a continuous, symmetric, positive semidefinite function. Let $K_x : X \to \mathbb{R}$ be the function defined by $K_x(s) = K(x, s)$ for $x, s \in X$. Let $\mathscr{H}_K$ be the RKHS associated to a Mercer kernel K, i.e., $\mathscr{H}_K = \overline{\text{span}\{K_x : x \in X\}}$, where the closure is taken with respect to the inner product $\langle\ ,\ \rangle_K$ as the unique linear extension of $\langle K_x, K_{x'} \rangle_K = K(x, x')$. Denote by $\| \cdot \|_K$ the norm of $\mathscr{H}_K$.

The most important property of RKHSs is the *reproducing property*: for any $f \in \mathscr{H}_K$ and $x \in X$, $f(x) = \langle f, K_x \rangle_K$. The reproducing property enables one to define the *sampling operator* on $\mathscr{H}_K$. Given a set $\mathbf{x} = (x_i)_{i=1}^m \in X^m$, denote by $l_2(\mathbf{x})$ the inner product space of real functions on $\mathbf{x}$ with the inner product $\langle u, v \rangle_{l_2(\mathbf{x})} = (1/m) \sum_{x_i \in \mathbf{x}} u(x_i) v(x_i)$. As a vector space, $l_2(\mathbf{x})$ is identical to the Euclidean space $\mathbb{R}^m$. Define a *sampling operator* $S_\mathbf{x} : \mathscr{H}_K \to l_2(\mathbf{x})$ by $S_\mathbf{x}(f) = (f(x_i))_{i=1}^m = (\langle f, K_{x_i} \rangle_K)$. Its adjoint $S_\mathbf{x}^* : l_2(\mathbf{x}) \to \mathscr{H}_K$ defined by $\langle S_\mathbf{x}(f), \mathbf{y} \rangle_{l_2(\mathbf{x})} = \langle f, S_\mathbf{x}^* \mathbf{y} \rangle_K$ for $\mathbf{y} \in l_2(\mathbf{x})$, is thus $S_\mathbf{x}^*(\mathbf{y}) = (1/m) \sum_{i=1}^m y_i K_{x_i}$. Such sampling operators are used in a generalization of the Shannon Sampling Theorem (Smale and Zhou, 2004).

Our target will be the *regression function*, $f_\rho(x) = \int y \, d\rho_{Y|x}$, i.e., the conditional expectation of y with respect to x, where $\rho_{Y|x}$ is the conditional measure on Y given x. Denote by ρ_X the marginal probability measure on X and let $\mathscr{L}^2_{\rho_X}$ be the space of square integrable functions with respect to ρ_X, whose inner product (norm) is denoted by $\langle\ ,\ \rangle_\rho$ ($\|\ ,\ \|_\rho$). Due to the relation

$$\mathscr{E}(f) - \mathscr{E}(f_\rho) = \|f - f_\rho\|_\rho^2,$$

the regression function f_ρ is the minimizer of $\mathscr{E}(f)$ over $\mathscr{L}^2_{\rho_X}$.

Next we define an integral operator which plays a central role in the theory. Let $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{C}(X)$ be an integral operator defined by $(L_K f)(x') = \int K(x', x) f(x) \, d\rho_X$, where $\mathscr{C}(X)$ is the Banach space of real continuous functions on X.

Throughout the paper we assume the following.

Finiteness Assumption.

(1) Let $\kappa := \max(\sup_{x \in X} \sqrt{K(x, x)}, 1) < \infty$.
(2) There exists a constant $M \geq 0$ such that $\text{supp}(\rho) \subseteq X \times [-M, M]$.

Remark 2.1. The first assumption ensures that the Mercer kernel K is square-summable, whence we have the inclusion $J : L_K(\mathscr{L}^2_{\rho_X}) \hookrightarrow \mathscr{L}^2_{\rho_X}$. The composition $J \circ L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}$ is a Hilbert–Schmidt (in fact, trace-class) operator. Passing through the spectrum of $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}$, one can define $L_K^r : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}$ for $r \in \mathbb{R}$. In particular, $L_K^{1/2}$ is a Hilbert space isometry between $\mathscr{L}^2_{\rho_X}/\text{ker}(L_K)$ and $\mathscr{H}_K$, whence independent to ρ_X. The restriction $L_K|_{\mathscr{H}_K}$ induces an operator from $\mathscr{H}_K$ into $\mathscr{H}_K$. All three operators above, $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{C}(X)$, $J \circ L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}$, and $L_K|_{\mathscr{H}_K} : \mathscr{H}_K \to \mathscr{H}_K$, when their domains are clear from the context, are denoted by L_K. The first assumption leads to $\|S_{\mathbf{x}}\| = \|S^*_{\mathbf{x}}\| \leq \kappa$ and $\|L_K\| \leq \kappa^2$ for all three operators. The second assumption is met in most cases in learning, and can be relaxed to higher-order moment conditions on ρ, which is however not pursued in this paper.

Therefore, the minimization of (1) is equivalent to finding approximations of f_ρ from $\mathscr{H}_K$, a subspace of $\mathscr{L}^2_{\rho_X}$ when K is square-summable. Note that $\mathscr{H}_K$ is a closed subspace of $\mathscr{L}^2_{\rho_X}$ if and only if it is of finite dimension.

2.2. *Gradient Descent Algorithms*

First we define two iterations: sample iteration and population iteration, then we show that they are simply gradient descent algorithms with respect to proper objective functions.

Given an i.i.d. sample of size m, $\mathbf{z} \in (X \times Y)^m$, define the *sample iteration* as the sequence $(f^{\mathbf{z}}_t)_{t\in\mathbb{N}} \in \mathscr{H}_K$ by

$$f^{\mathbf{z}}_{t+1} = f^{\mathbf{z}}_t - \frac{\gamma_t}{m}\sum_{i=1}^{m}\left(f^{\mathbf{z}}_t(x_i) - y_i\right) K_{x_i}, \qquad f^{\mathbf{z}}_0 = 0, \tag{2}$$

where $\gamma_t > 0$ is the step size (or learning rate). Now define the *population iteration* as the sequence

$$f_{t+1} = f_t - \gamma_t L_K(f_t - f_\rho), \qquad f_0 = 0. \tag{3}$$

Clearly f_t is deterministic and $f^{\mathbf{z}}_t$ is an $\mathscr{H}_K$-valued random variable depending on $\mathbf{z}$. In this paper we investigate the choice of step sizes in the form of $\gamma_t = 1/[\kappa^2(t+1)^\theta]$ ($t \in \mathbb{N}$) for some $\theta \in [0, 1)$.

The following proposition shows that the algorithm (3) is a gradient descent method for minimizing (1) over $\mathscr{H}_K$ and the algorithm (2) is the gradient descent method to minimize over $\mathscr{H}_K$ the following empirical risk

$$\mathscr{E}_{\mathbf{z}}(f) := \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2. \tag{4}$$

Recall that given a real functional $V : \mathscr{H} \to \mathbb{R}$, the *Fréchet derivative* of V at f, $DV(f) : \mathscr{H} \to \mathbb{R}$ is the linear functional such that, for $g \in \mathscr{H}$,

$$\lim_{\|g\|_{\mathscr{H}}\to 0} \frac{|V(f+g) - V(f) - DV(f)(g)|}{\|g\|_{\mathscr{H}}} = 0,$$

and the gradient of V as a map $\operatorname{grad} V : \mathscr{H} \to \mathscr{H}$ is defined by

$$\langle \operatorname{grad} V(f), g\rangle_{\mathscr{H}} = DV(f)(g) \qquad \text{for all} \quad g \in \mathscr{H}.$$

Then we have the following result.

Proposition 2.2. *The gradients of* (1) *and* (4) *are the maps from* $\mathscr{H}_K$ *into* $\mathscr{H}_K$ *given by*

$$\operatorname{grad}\mathscr{E}(f) = 2L_K(f - f_\rho),$$

and

$$\operatorname{grad}\mathscr{E}_{\mathbf{z}}(f) = \frac{2}{m}\sum_{i=1}^{m}(f(x_i) - y_i)K_{x_i}.$$

Proof. Define a functional $V : \mathscr{H}_K \to \mathbb{R}$ by $V(f) = (f(x) - y)^2$. Then its *Fréchet derivative* is

$$DV(f)(g) = \langle 2(f(x) - y)K_x, g\rangle_K,$$

and thus the gradient map is $\operatorname{grad} V(f) = 2(f(x) - y)K_x$. Taking expectations, $\operatorname{grad}\mathscr{E}(f) = \mathbb{E}[\operatorname{grad} V(f)] = 2\int_{X\times Y}(f(x) - y)K_x\, d\rho = 2L_K(f - f_\rho)$ and $\operatorname{grad}\mathscr{E}_{\mathbf{z}}(f) = \hat{\mathbb{E}}[\operatorname{grad} V(f)] = (2/m)\sum_{i=1}^{m}(f(x_i) - y_i)K_{x_i}$, where $\mathbb{E}$ denotes the expectation with respect to probability measure ρ and $\hat{\mathbb{E}}$ denotes the expectation with respect to the uniform probability measure on $\mathbf{z}$, often called the *empirical measure*. ■

Soon we shall see that the population iteration f_t converges to f_ρ, while the sample iteration $f_t^{\mathbf{z}}$ does not. In most cases, $f_t^{\mathbf{z}}$ converges to an undesired overfitting solution which fits exactly the sample points but has large errors beyond them. However, via the triangle inequality

$$\|f_t^{\mathbf{z}} - f_\rho\|_\rho \le \|f_t^{\mathbf{z}} - f_t\|_\rho + \|f_t - f_\rho\|_\rho,$$

we may control $\|f_t^{\mathbf{z}} - f_\rho\|_\rho$. Here we call the gap between two iteration paths, $\|f_t^{\mathbf{z}} - f_t\|_\rho$, the *sample error* (or *variance*), and the distance, $\|f_t - f_\rho\|_\rho$, the *approximation error* (or *bias*). The theorems in the next section give upper bounds for each of them.

2.3. *Early Stopping and Probabilistic Upper Bounds*

In this section we state and discuss the main results in the paper.

First we assume some regularity property on f_ρ. Let $B_R = \{f \in \mathscr{L}^2_{\rho_X} : \|f\|_\rho \le R\}$ ($R > 0$) be the function ball in $\mathscr{L}^2_{\rho_X}$ with radius R and centered at the origin. In this paper we assume that for some $r > 0$, $f_\rho \in L_K^r(B_R)$, i.e., f_ρ lies in the image of the ball B_R under the map L_K^r. Roughly speaking, such a condition imposes a low-pass filter on f_ρ which amplifies the projections of f_ρ on the eigenvectors of $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}$ with large eigenvalues and attenuates the projections on the eigenvectors with small eigenvalues.

Main Theorem. *Suppose* $f_\rho \in L_K^r(B_R)$ *for some* $R, r > 0$. *Let* $\gamma_t = 1/[\kappa^2(t+1)^\theta]$ ($t \in \mathbb{N}$) *for some* $\theta \in [0, 1)$. *For each* $m \in \mathbb{N}$, *there is an early stopping rule* $t^* : \mathbb{N} \to \mathbb{N}$ *such that the following holds with probability at least* $1 - \delta$ ($\delta \in (0, 1)$):

(1) *if* $r > 0$, *then*

$$\|f^{\mathbf{z}}_{t^*(m)} - f_\rho\|_\rho \le C_{\rho,K,\delta}\, m^{-r/(2r+2)},$$

where $C_{\rho,K,\delta} = [4(1+\sqrt{2})M/(1-\theta)]\ \log^{1/2} 2/\delta + R(2r\kappa^2/e)^r$;

(2) *if* $r > \frac{1}{2}$, *then* $f_\rho \in \mathscr{H}_K$ *and*

$$\|f^{\mathbf{z}}_{t^*(m)} - f_\rho\|_K \leq D_{\rho,K,\delta} m^{-(r-1/2)/(2r+2)},$$

where $D_{\rho,K,\delta} = [4(1+\sqrt{2})M/\kappa(1-\theta)^{3/2}] \log^{1/2} 2/\delta + R(2(r-1/2)\kappa^2/e)^{r-1/2}$. *In both cases, the stopping rule can be chosen as*

$$t^*(m) = \lceil m^{1/(2r+2)(1-\theta)} \rceil,$$

where $\lceil x \rceil$ *denotes the smallest integer greater then or equal to* $x \in \mathbb{R}$.

Its proof will be given at the end of this section.

Remark 2.3. The first upper bound holds for all $r > 0$. In the second upper bound, $r > \frac{1}{2}$ implies $f_\rho \in \mathscr{H}_K$ as $L_K^{1/2} : \mathscr{L}^2_{\rho_X}/\ker(L_K) \to \mathscr{H}_K$ is a Hilbert space isometry. In particular, when $r \to \infty$, we approach the asymptotic rate $\|f^{\mathbf{z}}_{t^*(m)} - f_\rho\|_\rho \leq O(m^{-1/2})$ and $\|f^{\mathbf{z}}_{t^*(m)} - f_\rho\|_K \leq O(m^{-1/2})$, at a price of the constants growing exponentially with r. This happens when $\mathscr{H}_K$ is of finite dimension, e.g., when K is a polynomial kernel. Such a result improves the upper bounds for the usual regularized least square algorithm (Minh, 2005; or the Appendix by Minh, in Smale and Zhou, 2005) where the upper convergence rate is slower than $O(m^{-1/3})$ for $r > 0$ (or $O(m^{-1/4})$ for $r > \frac{1}{2}$). This fact is related to the *saturation* phenomenon in the classical studies of inverse problems (Engl, Hanke, and Neubauer, 2000). We shall come back to this point in Section 3.1.

Remark 2.4. Here we address the optimality issue about the convergence rates. Some minimax lower rates (De Vore, Kerkyacharian, Picard, and Temlyakov, 2004, Temlyakov, 2004) and individual lower rates (Caponnetto and De Vito, 2005) suggest that, for $r > 0$ the $\mathscr{L}^2_{\rho_X}$-convergence rate $O(m^{-r/(2r+1)})$ is the optimal kernel-independent rate, in the sense that the rate is independent of the decaying rate of the eigenvalues of L_K. In this sense, the Main Theorem has only suboptimal rates. To the authors' knowledge, recently there have been some improvements in several restricted settings. In fixed designs with constant step sizes, Bissantz, Hohage, Munk, and Ruymgaart (2006) show that the optimal rate can be achieved. In random designs, if $f_\rho \in \mathscr{H}_K$ (i.e., $r \geq \frac{1}{2}$) Bauer, Pereverzev, and Rosasco (2006) show that the optimal rates can be achieved. The difficulty in random designs is due to that the sample iteration and population iteration use distinct operators which cannot be diagonalized simultaneously as can be done in fixed designs. In the latter work this difficulty was overcome by a crucial notion of *operator monotonicity* (Mathé and Pereverzev, 2002). However, in their work, a different bias-variance decomposition approach was used which requires $f_\rho \in \mathscr{H}_K$ and thus cannot be applied to the setting of $0 < r < \frac{1}{2}$. Recently, Caponnetto (2006) suggests that in semisupervised learning, by exploiting unlabeled data, one may get the optimal upper rates when $r \in (0, \frac{1}{2})$. Therefore, it is still an open problem of how to achieve the optimal rates for $r \in (0, \frac{1}{2})$.

Remark 2.5. Roughly speaking, the Main Theorem extends the convergence of L_2 Boost (Bühlmann and Yu, 2002) in Sobolev spaces with fixed designs to general RKHSs with random designs. For example, let $X = S^d$ be the d-sphere and let ρ_X be the uniform measure on S^d, then following Wahba (1990) one can take a Sobolev space $W_d(S^d)$ as an RKHS $\mathscr{H}_K$, such that the associated integral L_K has eigenvalues $\lambda_n \sim n^{-1}$. Then $f_\rho \in L_K^{s/d}(B_R)$ implies that $f_\rho \in W_s(S^d)$. The Main Theorem gives an upper $\mathscr{L}^2_{\rho_X}$-convergence rate $O(m^{-s/(2s+2d)})$ for $r > 0$, which is suboptimal due to the reason stated in Remark 2.4.

Now we discuss a direct application of the Main Theorem to the setting of classification. Notice that when $Y = \{-1, 1\}$, algorithm (2) may provide a classification rule sign $f^{\mathbf{z}}_{t^*}$, often called *plug-in classifier* in literature. Hence we may consider such a rule as an approximation of the Bayes rule, sign f_ρ. The following result gives an upper bound on the distance $\|$ sign $f^{\mathbf{z}}_t -$ sign $f_\rho\|_\rho$.

Theorem 2.6. *Assume the same condition as the Main Theorem. Moreover, suppose* $Y = \{-1, 1\}$ *and Tsybakov's noise condition*

$$\rho_X(\{x \in X : |f_\rho(x)| \leq t\}) \leq B_q t^q, \qquad \forall t > 0, \tag{5}$$

for some $q \in [0, \infty]$ *and* $B_q \geq 0$. *Then*:

(1) *if* $r > 0$, *the following holds with probability at least* $1 - \delta$ $(\delta \in (0, 1))$,

$$\|\text{sign } f^{\mathbf{z}}_{t^*(m)} - \text{sign } f_\rho\|_\rho \leq C_1 m^{-\alpha r/2(r+1)(2-\alpha)},$$

where $\alpha = q/(q+1)$ *and* $C_1 = [16(1+\sqrt{2})(B_q+1)M/(1-\theta)]\ \log^{1/2} 2/\delta + 4(B_q+1)R(2r\kappa^2/e)^r$;

(2) *if* $r > \frac{1}{2}$, *the following holds with probability at least* $1 - \delta$ $(\delta \in (0, 1))$,

$$\|\text{sign } f^{\mathbf{z}}_{t^*(m)} - \text{sign } f_\rho\|_\rho \leq C_2 m^{-q(r-1/2)/(4r+4)},$$

where $C_2 = 2\sqrt{\kappa B_q D_{\rho,K,\delta}}$;

(3) *if* $r > \frac{1}{2}$, *and the hard margin condition holds*

$$\rho_X(\{x \in X : |f_\rho(x)| \leq \gamma\}) = 0,$$

the following upper bound holds:

$$\mathbb{E}_{\mathbf{z}\in Z^m}[\|\text{sign } f^{\mathbf{z}}_{t^*(m)} - \text{sign } f_\rho\|_\rho] \leq C_3 e^{-C_4 \gamma m^{(r-1/2)/(2r+2)}},$$

where $C_3 = 2^{-1-[\kappa R(1-\theta)^{3/2}/4M(1+\sqrt{2})]\cdot(2(r-1/2)\kappa^2/e)^{r-1/2}}$ *and* $C_4 = [(1-\theta)^{3/2}\log 2]/4M(1+\sqrt{2})$.

The proof, together with a detailed introduction to the background, is given in Section 6.

Remark 2.7. The first bound holds for all $r > 0$, with an implication that as $\alpha = 1$ (e.g., f_ρ has a hard margin) and $r \to \infty$ (e.g., $\mathscr{H}_K$ is of finite dimension), the convergence rate may approach $O(1/\sqrt{m})$ arbitrarily. The second upper bound holds only for $r > \frac{1}{2}$

(whence $f_\rho \in \mathscr{H}_K$), which however implies that as $q \to \infty$ (i.e., $\alpha \to 1$), one can achieve an arbitrarily fast polynomial convergence rate. The third upper bound goes even further, which says that for $r > \frac{1}{2}$ and a hard margin assumption, an exponential convergence rate can be achieved for the mean error. For more types of exponential rates for plug-in classifiers, see Audibert and Tsybokov (2005) and references therein.

Remark 2.8. Consider the Bayes consistency. Define the risk of f by

$$R(f) = \rho_Z(\{(x, y) \in Z \mid \operatorname{sign} f(x) \neq y\}),$$

and let $R(f_\rho)$ be the *Bayes risk*. Using the comparison results in Proposition 6.2, we may obtain similar upper bounds for the error $R(f^{\mathbf{z}}_{t^*(m)}) - R(f_\rho)$.

Next we present upper bounds for the sample error and the approximation error, respectively, which are used to prove the Main Theorem. To improve the Main Theorem, it is crucial to get sharper results on the sample error rate.

Theorem 2.9 (Sample Error). *With probability at least* $1 - \delta$ ($\delta \in (0, 1)$) *there holds, for all* $t \in \mathbb{N}$,

$$\|f^{\mathbf{z}}_t - f_t\|_\rho \le C_5 \frac{t^{1-\theta}}{\sqrt{m}},$$

where $C_5 = [2(1 + \sqrt{2})M/(1 - \theta)]\ \log^{1/2} 2/\delta$; *and*

$$\|f^{\mathbf{z}}_t - f_t\|_K \le C_6 \sqrt{\frac{t^{3(1-\theta)}}{m}},$$

where $C_6 = [2(1 + \sqrt{2})M/\kappa(1 - \theta)^{3/2}]\ \log^{1/2} 2/\delta$.

Theorem 2.10 (Approximation Error). *Suppose* $f_\rho \in L^r_K(B_R)$ *for some* $R, r > 0$ *and* $f_0 = 0$. *Then, for all* $t \in \mathbb{N}$,

$$\|f_t - f_\rho\|_\rho \le C_7 t^{-r(1-\theta)},$$

where $C_7 = R(2r\kappa^2/e)^r$; *and if, moreover,* $r > \frac{1}{2}$, *then* $f_\rho \in \mathscr{H}_K$ *and*

$$\|f_t - f_\rho\|_K \le C_8 t^{-(r-1/2)(1-\theta)},$$

where $C_8 = R(2(r - 1/2)\kappa^2/e)^{r-1/2}$.

Their proofs are given in Section 5.

Remark 2.11. It can be seen that the population iteration f_t converges to f_ρ, while the gap between the population iteration and sample iteration (i.e., the sample error) expands simultaneously. The step size γ_t affects the rates of both. When γ_t shrinks faster (larger θ), the approximation error (bias) drops slower, while the sample error (variance) grows slower.

Finally, combining these upper bounds, we obtain an immediate proof of the Main Theorem by solving a bias-variance trade-off.

Proof of the Main Theorem. Combining Theorems 2.9 and 2.10, we have

$$\|f_t^{\mathbf{z}} - f_\rho\|_\rho \leq C_5 \frac{t^{1-\theta}}{\sqrt{m}} + C_7 t^{-r(1-\theta)}.$$

Let $t^*(m) = \lceil m^\alpha \rceil$, the smallest integer greater than or equal to m^α for some $\alpha > 0$. Minimizing the right-hand side over $\alpha > 0$ we arrive at the linear equation

$$\alpha(1-\theta) - \tfrac{1}{2} = -\alpha r(1-\theta)$$

whose solution is $\alpha = 1/(2r+2)(1-\theta)$.

Assume for some $\beta \in [1, 2]$ such that $m^\alpha \leq t^*(m) = \beta m^\alpha \leq m^\alpha + 1 \leq 2m^\alpha$. Then

$$\|f_{t^*}^{\mathbf{z}} - f_\rho\|_\rho \leq (\beta^{1-\theta} C_5 + \beta^{-r(1-\theta)} C_7) m^{-r/(2r+2)} \leq (2C_5 + C_7) m^{-r/(2r+2)}.$$

Essentially the same reasoning leads to the second bound. ■

3. Discussions on Related Work

In this section we provide more discussions on the comparison between early stopping and Tikhonov regularization used in the usual regularized least square algorithm, Boosting in the gradient descent view, Landweber iterations to solve linear equations, and online learning algorithms as stochastic approximations of the gradient descent method.

3.1. *Early Stopping versus Penalized Least Square*

Recently, the following penalized least square algorithm gained an extensive study in learning (e.g., Cucker and Smale, 2002; Smale and Zhou, 2005),

$$f_\lambda = \arg \min_{f \in \mathscr{H}_K} \mathscr{E}(f) + \lambda \|f\|_K^2,$$

where it can be shown that

$$f_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho.$$

On the other hand, for constant step sizes $\gamma_t = \gamma_0 = 1$ (assuming $\kappa = 1$), the population iteration (3) becomes

$$f_t = \sum_{i=0}^{t-1} (I - L_K)^i L_K f_\rho = \sum_{i=0}^{t-1} (I - L_K)^i (I - (I - L_K)) f_\rho = (I - (I - L_K)^t) f_\rho.$$

Both f_λ and f_t can be regarded as low-pass filters on f_ρ, which tends to project f_ρ to the eigenfunctions corresponding to large eigenvalues. The applications of early stopping regularization in statistical estimation can at least be traced back to Wahba (1987). Moreover, it is shown in Fleming (1990) that if the L_K is a finite rank operator (matrix) and if the step size γ_t is taken to be a finite rank operator (matrix), then there is a one-to-one correspondence between the two regularization methods.

On the other hand, there are also significant differences between f_t and f_λ. First of all, f_t has a better approximation ability than f_λ. In fact, it can be shown (Minh, 2005; or the Appendix by Minh, in Smale and Zhou, 2005) that, if $f_\rho \in L_K^r(B_R)$ for some $r > 0$,

$$\|f_\lambda - f_\rho\|_\rho \leq O(\lambda^{\min(r,1)}),$$

and for $r > \frac{1}{2}$,

$$\|f_\lambda - f_\rho\|_K \leq O(\lambda^{\min(r-1/2,1)}).$$

We can see for large r, the upper bound cannot go faster than λ (as $\lambda \to 0$) in Tikhonov regularization. On the other hand, taking $\theta = 0$ in Theorem 2.10 we have that, for $r > 0$,

$$\|f_t - f_\rho\|_\rho \leq O(t^{-r}),$$

and for $r > \frac{1}{2}$,

$$\|f_t - f_\rho\|_K \leq O(t^{-(r-1/2)}).$$

We may roughly regard such a relationship between regularization parameters, $\lambda \sim 1/t$, where f_t has faster approximation rates than f_λ for large r. In particular, this leads to an optimal convergence rate $O(m^{-1/2})$ in the Main Theorem as $r \to \infty$, in contrast for the usual penalized least square the best known upper convergence rates in $\mathscr{L}^2_{\rho_X}$ and $\mathscr{H}_K$ are $O(m^{-1/3})$ (for $r \geq 1$) and $O(m^{-1/4})$ (for $r \geq \frac{3}{2}$), respectively. This phenomenon is studied as the *saturation* of regularizations in classical inverse problems (e.g., Engl, Hanke, and Neubauer, 2000).

On numerical aspects, the computational cost of Tikhonov regularization essentially needs inverting a matrix which is of $O(m^3)$ floating point operations, where early stopping regularization needs $O(t^* m^2)$, where t^* is the early stopping time. Thus, for those kernel matrices with special structures, where a few iterations are sufficient to provide a good approximation (i.e., $t^* \ll m$), early stopping regularization is favored. For those very ill-conditioned kernel matrices, conjugate gradient descent methods or more sophisticate iteration methods (Hanke, 1995; Ong, 2005) are suggested to achieve faster numerical convergence.

3.2. *Perspectives on Boosting*

The notion of boosting was originally proposed as the question as to whether a "weak" learning algorithm which performs just slightly better than random guessing can be "boosted" into a "strong" learning algorithm of high accuracy (Valiant, 1984; or see the review by Schapire, 2002 or Dietterich, 1997). For example, AdaBoost (Freund and Schapire, 1997) is claimed to be one of the "best off-shelf" machine learning algorithms.

Although running long enough AdaBoost will eventually overfit, during the process it exhibits resistance against overfitting. This phenomenon suggests that it might be the dynamical process of boosting which accounts for regularization. Note that there are two dynamical systems in AdaBoost: one is the evolution of the empirical distributions on the sample, and the other is the evolution in hypothesis spaces. Thus one may study both dynamical systems, or either one. For example, studies on both lead to game theory (e.g., Breiman, 1999; Freund and Schapire, 1999; Schapire, 2001; Stoltz and Lugosi, 2004), the first has been seen in Rudin, Daubechies, and Schapire (2004) and the second leads to the functional gradient descent view with general convex loss functions (e.g., Breiman, 1999; Friedman, Hastie, and Tibshirani, 2000; Mason, Baxter, Bartlett, and Frean, 2000; Friedman, 2001), where this paper is also included.

In view of gradient descent with L_2 loss, our algorithms can also be regarded as a boosting procedure, L_2Boost (Bühlmann and Yu, 2002). The "weak learners" here are the functions K_{x_i} $(i = 1, \ldots, m)$, where $x_i \in X$ is an example. Such functions can be regarded as generalizations of the *sinc* function in the Shannon Sampling Theorem (Smale and Zhou, 2004). A rough comparison on the convergence rates with Bühlmann and Yu (2002) is in Remark 2.5.

The treatment in this paper adopts the same bias-variance decomposition as in other consistency studies on boosting (e.g., Jiang, 2004; Breiman, 2004; Lugosi and Vayatis, 2004; Zhang and Yu, 2003). However, we did not use VC-dimension or Rademacher complexity to bound the sample error. Instead, we benefit from the linear structure in RKHS by exploiting concentration inequalities for random Hilbert–Schmidt operators and vectors to derive the uniform convergence and its rates, which simplifies the analysis. The idea that norm convergence of operators leading to uniform convergence of sequences is, in fact, not new in the literature, e.g., Yosida and Kakutani (1941) or see the comments in Peskir (2000).

3.3. *Perspectives on Landweber Iterations*

In this subsection we show that the population iteration (3) can be regarded as the Landweber iteration for a specific linear operator equation and the sample iteration (2) is a discretization of such an algorithm by random samples. We also point out one difference in learning and inverse problem formulation, where one should be careful in applying the results in inverse problems to learning.

Consider the following linear operator equation:

$$I_K f = f_\rho, \tag{6}$$

where the linear map $I_K : \mathscr{H}_K \hookrightarrow \mathscr{L}^2_{\rho_X}$ is a continuous inclusion. Without loss of generality, assume that ρ_X is strictly positive on X, which makes I_K injective, i.e., an embedding. This is an ill-posed problem as in general $f_\rho \notin \mathscr{H}_K$. However, if $f_\rho \in \mathscr{H}_K \oplus \overline{\mathscr{H}_K}^{\perp}$, then the following normal equation has a solution:

$$I_K^* I_K f = I_K^* f_\rho, \tag{7}$$

where the adjoint of I_K, $I_K^* : \mathscr{L}^2_{\rho_X} \to \mathscr{H}_K$ is simply the operator $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{H}_K$. Note that $I_K^* I_K = L_K : \mathscr{H}_K \to \mathscr{H}_K$ and $I_K I_K^* = L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}$. Thus the normal equation is simply

$$L_K f = L_K f_\rho. \tag{8}$$

Actually, the first L_K is $L_K : \mathscr{H}_K \to \mathscr{H}_K$ and the second L_K is $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{H}_K$. In this way one can see that the population iteration (3) with the choice $\gamma_t = 1/\kappa^2$ is the Landweber iteration (Engl, Hanke, and Neubauer, 2000) to solve (6).

Consider the following discretized version of (6),

$$S_{\mathbf{x}} f = \mathbf{y}, \tag{9}$$

whose normal equation is

$$S_{\mathbf{x}}^* S_{\mathbf{x}} f = S_{\mathbf{x}}^* \mathbf{y}, \tag{10}$$

where $S_{\mathbf{x}} : \mathscr{H}_K \to l_2(\mathbf{x})$ is the sampling operator. In this way, the sample iteration (2) can be rewritten as

$$f_{t+1}^{\mathbf{z}} = f_t^{\mathbf{z}} - \gamma_t (S_{\mathbf{x}}^* S_{\mathbf{x}} f_t^{\mathbf{z}} - S_{\mathbf{x}}^* \mathbf{y}), \tag{11}$$

and with the choice $\gamma_t = 1/\kappa^2$, it is the Landweber iteration to solve (9). In *fixed designs*, $\mathbf{x}$ is not randomly sampled from ρ_X but fixed in advance such that $L_K = S_{\mathbf{x}}^* S_{\mathbf{x}}$, where only the output noise on $\mathbf{y}$ is considered.

It should be noted that the setting of learning goes slightly beyond the classical setting of inverse problems.

In classical inverse problems we require $f_\rho \in \mathscr{H}_K \oplus \overline{\mathscr{H}_K}^{\perp}$ (if $f_\rho \in L_K^r(B_R)$, this requirement becomes $f_\rho \in \mathscr{H}_K$) for the existence of a solution of normal equation (7) or, equivalently, we require $P_K f_\rho \in \mathscr{H}_K$ where $P_K : \mathscr{L}^2_{\rho_X} \to \overline{\mathscr{H}_K}$ is the projection from $\mathscr{L}^2_{\rho_X}$ onto the closure of $\mathscr{H}_K$ in $\mathscr{L}^2_{\rho_X}$. In this case, one can define $f_\rho^\dagger = P_K f_\rho$ as the generalized inverse of f_ρ, i.e., the unique minimal norm least square solution of (6) in $\mathscr{H}_K$, and study the convergence

$$\|f_t - f_\rho^\dagger\|_K \to 0$$

under the assumption $f_\rho^\dagger = (I_K^* I_K)^r g$ for some $\|g\|_K \leq R$.

However, in learning, one typically has $f_\rho \notin \mathscr{H}_K \oplus \overline{\mathscr{H}_K}^{\perp}$ or $P_K f_\rho \notin \mathscr{H}_K$, whence $f_\rho^\dagger$ does not exist. For example, choose a Gaussian kernel $K(x, x') = e^{-a\|x-x'\|^2}$ $(a > 0)$ such that $\mathscr{H}_K$ is dense in $\mathscr{L}^2_{\rho_X}$ and $f_\rho \notin \mathscr{H}_K$. In this case, f_t diverges in $\mathscr{H}_K$ but converges in $\mathscr{L}^2_{\rho_X}$,

$$\|f_t - P_K f_\rho\|_\rho \to 0$$

under the assumption that $f_\rho = (I_K I_K^*)^r g$ for some $\|g\|_\rho \leq R$.

For a broader discussion on learning versus inverse problems, see De Vito, Rosasco, Caponnetto, Giovannini, and Odone (2004).

3.4. *Perspectives on Online Learning*

An online learning algorithm was suggested in Smale and Yao (2005) as stochastic approximations of the gradient descent method for the following penalized least square problem:

$$\min_{f \in \mathscr{H}_K} \mathscr{E}(f) + \lambda \|f\|_K^2, \qquad \lambda \geq 0.$$

To be precise, the algorithm returns a sequence $(f_t)_{t \in \mathbb{N}}$ defined by

$$\hat{f}_t = \hat{f}_{t-1} - \gamma_t [(\hat{f}_{t-1}(x_t) - y_t) K_{x_t} + \lambda \hat{f}_{t-1}] \qquad \text{for some} \quad \hat{f}_0 \in \mathscr{H}_K, \tag{12}$$

where $\hat{f}_t$ depends on $z_t = (x_t, y_t)$ and $\hat{f}_{t-1}$ which only relies on the previous examples $\mathbf{z}_{t-1} = (x_i, y_i)_{1 \leq i \leq t-1}$. Note that, in this paper, the sample $\mathbf{z} \in Z^m$ is fixed during the iterations and the regularization parameter $\lambda = 0$ is replaced by some early stopping rule.

Shrinking the step size γ_t plays different roles in this paper and in online learning algorithms. In this paper it might only affect the stopping time, but not the rate of convergence. In fact, in the Main Theorem, the constant step size, i.e., $\theta = 0$, leads to the fastest algorithm in the family in the sense that the algorithm requires the minimal number of iterations before stopping, while all the algorithms are guaranteed with the same upper convergence rates. Therefore, shrinking the step size chosen in this paper does not contribute to the regularizations. However, in online learning, it significantly affects the regularization. In fact, to ensure the convergence of online learning algorithm (12), one needs to shrink step sizes $\gamma_t \to 0$; but the shrinkage can't go too fast: $\sum_t \gamma_t = \infty$ is necessary to "forget" the initial error (e.g., see Yao (2005)). This phenomenon might suggest further investigations on the roles of restricting step sizes as regularization in various settings.

A close connection can be seen from the structural decomposition in Proposition 4.3, which gives

$$f_t^{\mathbf{z}} - f_t = r_t(L_K)(f_0^{\mathbf{z}} - f_0) + \sum_{k=0}^{t-1} \gamma_k \pi_{k+1}^t (L_K) \chi_k,$$

where $\chi_k = (L_K - S_{\mathbf{x}}^* S_{\mathbf{x}}) f_k^{\mathbf{z}} + S_{\mathbf{x}}^* \mathbf{y} - L_K f_\rho$. In Yao (2005), the martingale decomposition gives

$$\hat{f}_t - f_\lambda = r_t(L_K + \lambda I)(\hat{f}_0 - f_\lambda) + \sum_{k=0}^{t-1} \gamma_k \pi_{k+1}^t (L_K + \lambda I) \hat{\chi}_k,$$

where $\hat{\chi}_k = (L_K - S_{x_{k+1}}^* S_{x_{k+1}}) \hat{f}_k + S_{x_{k+1}}^* (y_{k+1}) - L_K f_\rho$. The key difference lies in the fact that since $\hat{f}_k$ only depends on historical examples $z_1, \ldots, z_k$, $\hat{\chi}_k$ is a martingale sequence with $\mathbb{E}[\hat{\chi}_k | z_1, \ldots, z_k] = 0$. However, χ_k loses this feature since in iterations every $f_k^{\mathbf{z}}$ depends on the whole sample $\mathbf{z}$.

4. Some Function Decompositions

The next two sections are devoted to the proof of the upper bounds on sample error and approximation error, i.e., Theorems 2.9 and 2.10. In this section we provides some decompositions for f_t, $f_t^{\mathbf{z}}$, and $f_t^{\mathbf{z}} - f_t$, which are crucial to estimating the sample error in Section 5.

4.1. *Regularization and Residue Polynomials*

Before studying the sample error, we define some polynomials which will be used to represent the decomposition in a neat way.

For $x \in \mathbb{R}$, define a polynomial of degree $t - k + 1$,

$$\pi_k^t(x) = \begin{cases} \prod_{i=k}^{t} (1 - \gamma_i x), & k \le t, \\ 1, & k > t. \end{cases} \tag{13}$$

An important property about π_k^t is that by the telescope sum

$$\begin{aligned} \sum_{k=\tau}^{t-1} \gamma_k x \pi_{k+1}^{t-1}(x) &= \sum_{k=\tau}^{t-1} (1 - (1 - \gamma_k x)) \pi_{k+1}^{t-1}(x) \\ &= \sum_{k=\tau}^{t-1} (\pi_{k+1}^{t-1}(x) - \pi_k^{t-1}(x)) = 1 - \pi_\tau^{t-1}(x). \end{aligned} \tag{14}$$

This property motivates the definition of two important polynomials: define the *regularization polynomial*

$$g_t(x) = \sum_{k=0}^{t-1} \gamma_k \pi_{k+1}^{t-1}(x), \tag{15}$$

and the *residue polynomial*

$$r_t(x) = 1 - x g_t(x) = \pi_0^{t-1}(x). \tag{16}$$

Given a polynomial $p(x) = a_0 + a_1 x + \cdots + a_n x^n$ and a self-adjoint operator T, we write $p(T)$ for the operator $a_0 I + a_1 T + \cdots + a_n T^n$.

Lemma 4.1. *Let T be a compact self-adjoint operator. Suppose $0 \le \gamma_t \le 1/\|T\|$ for all $t \in \mathbb{N}$. Then*:

(1) $\|\pi_k^t(T)\| \le 1$;
(2) $\|g_t(T)\| \le \sum_{k=0}^{t-1} \gamma_k$;
(3) $\|r_t(T)\| \le 1$.

Proof. The results follow from the spectral decomposition of T (see, e.g., Engl, Hanke, and Neubauer (2000)) and the following estimates: suppose $0 \le \gamma_t x \le 1$ for all $t \in \mathbb{N}$, then:

(A) $|\pi_k^t(x)| \le 1$;
(B) $|g_t(x)| \le \sum_{k=0}^{t-1} \gamma_k$;
(C) $|r_t(x)| \le 1$.

These bounds are tight since $\pi_k^t(0) = r_t(0) = 1$ and $g_t(0) = \sum_{k=0}^{t-1} \gamma_k$. ■

4.2. *Some Decompositions*

The following proposition gives explicit representations of f_t and $f_t^{\mathbf{z}}$.

Proposition 4.2. *For all* $t \in \mathbb{N}$,

(1) $f_t = r_t(L_K) f_0 + g_t(L_K) L_K f_\rho$;
(2) $f_t^{\mathbf{z}} = r_t(S_{\mathbf{x}}^* S_{\mathbf{x}}) f_0^{\mathbf{z}} + g_t(S_{\mathbf{x}}^* S_{\mathbf{x}}) S_{\mathbf{x}}^* \mathbf{y}$.

Proof. The first identity follows from induction on (3) and the second follows from induction on (11). ■

Define the *remainder* at time t to be $r_t = f_t^{\mathbf{z}} - f_t$. The following proposition gives a decomposition of remainder which is crucial in the upper bound for the sample error.

Proposition 4.3 (Remainder Decomposition). *For all* $t \in \mathbb{N}$,

$$f_t^{\mathbf{z}} - f_t = r_t(L_K)(f_0^{\mathbf{z}} - f_0) + \sum_{k=0}^{t-1} \gamma_k \pi_{k+1}^t (L_K) \chi_k,$$

where $\chi_k = (L_K - S_{\mathbf{x}}^* S_{\mathbf{x}}) f_k^{\mathbf{z}} + S_{\mathbf{x}}^* \mathbf{y} - L_K f_\rho$.

Proof. We use a new representation of $f_t^{\mathbf{z}}$, other than Proposition4.2(2),

$$f_{t+1}^{\mathbf{z}} = f_t^{\mathbf{z}} - \gamma_t (S_{\mathbf{x}}^* S_{\mathbf{x}} f_t^{\mathbf{z}} - S_{\mathbf{x}}^* \mathbf{y}) = (1 - \gamma_t L_K) f_t^{\mathbf{z}} + \gamma_t [(L_K - S_{\mathbf{x}}^* S_{\mathbf{x}}) f_t^{\mathbf{z}} + S_{\mathbf{x}}^* \mathbf{y}].$$

By induction on $t \in \mathbb{N}$, we reach

$$f_t^{\mathbf{z}} = \pi_0^{t-1}(L_K) f_0^{\mathbf{z}} + \sum_{k=0}^{t-1} \gamma_k \pi_{k+1}^{t-1}(L_K)((L_K - S_{\mathbf{x}}^* S_{\mathbf{x}}) f_k^{\mathbf{z}} + S_{\mathbf{x}}^* \mathbf{y}).$$

Subtracting on both sides of Proposition 4.2(1), we obtain the result. ■

Some useful upper bounds are collected in the following proposition.

Proposition 4.4. *Assume that* $f_0 = f_0^{\mathbf{z}} = 0$. *Then, for all* $t \in \mathbb{N}$,

(1) $\|f_t\|_K \le \sqrt{\sum_{k=0}^{t-1} \gamma_k} \|f_\rho\|_\rho$;
(2) $\|f_t\|_\rho \le \|f_\rho\|_\rho$;
(3) $\|f_t^{\mathbf{z}}\|_K \le M \sqrt{\sum_{k=0}^{t-1} \gamma_k}$;
(4) $\|f_t^{\mathbf{z}} - f_t\|_K \le (\sum_{k=0}^{t-1} \gamma_k) \sup_{1 \le k \le t-1} \|\chi_k\|_K$;
(5) $\|f_t^{\mathbf{z}} - f_t\|_{\mathscr{L}^2_{\rho_X}} \le \sqrt{\sum_{k=0}^{t-1} \gamma_k} \sup_{1 \le k \le t-1} \|\chi_k\|_K$.

Proof. Throughout the proof we repeatedly use Corollary 4.1 and the isometry $L_K^{1/2} : \mathscr{L}^2_{\rho_X}/\ker(L_K) \to \mathscr{H}_K$.

The first three parts are based on Proposition 4.2 with $f_0 = f_0^{\mathbf{z}} = 0$,

$$f_t = g_t(L_K)L_K f_\rho, \qquad \text{and} \qquad f_t^{\mathbf{z}} = g_t(S_{\mathbf{x}}^* S_{\mathbf{x}}) S_{\mathbf{x}}^* \mathbf{y}.$$

(1) Note that

$$\|f_t\|_K^2 \ = \ \langle g_t(L_K)L_K f_\rho, g_t(L_K)L_K f_\rho \rangle_K = \langle L_K^{1/2} f_\rho, [g_t(L_K)L_K] g_t(L_K) L_K^{1/2} f_\rho \rangle_K,$$

where, using $r_t(\lambda) = 1 - \lambda g_t(\lambda)$,

$$\begin{aligned} \text{r.h.s.} \ &= \ \langle L_K^{1/2} f_\rho, (I - r_t(L_K)) g_t(L_K) L_K^{1/2} f_\rho \rangle_K \\ &\le \ \|g_t(L_K)\| \|L_K^{1/2} f_\rho\|_K^2 = \sum_{k=0}^{t-1} \gamma_k \|f_\rho\|_\rho^2. \end{aligned}$$

Taking the square root gives the result.

(2) Note that $\|f_t\|_\rho^2 = \|L_K^{1/2} f_t\|_K^2$, whence

$$\|f_t\|_\rho \ = \ \|L_K^{1/2} g_t(L_K) L_K f_\rho\|_K = \|(I - r_t(L_K)) L_K^{1/2} f_\rho\|_K \le \|L_K^{1/2} f_\rho\|_K^2 = \|f_\rho\|_\rho^2.$$

(3) Let G be the $m \times m$ Grammian matrix $G_{ij} = (1/m) K(x_i, x_j)$. Clearly, $G = S_{\mathbf{x}} S_{\mathbf{x}}^*$.

$$\begin{aligned} \|f_t^{\mathbf{z}}\|_K^2 \ &= \ \langle g_t(S_{\mathbf{x}}^* S_{\mathbf{x}}) S_{\mathbf{x}}^* \mathbf{y}, g_t(S_{\mathbf{x}}^* S_{\mathbf{x}}) S_{\mathbf{x}}^* \mathbf{y} \rangle_K = \langle g_t(G)\mathbf{y}, g_t(G) G \mathbf{y} \rangle_m \\ &= \ \langle g_t(G)\mathbf{y}, (I - r_t(G))\mathbf{y} \rangle_m \le \|g_t(G)\| \|\mathbf{y}\|_m^2 \le M^2 \sum_{k=0}^{t-1} \gamma_k. \end{aligned}$$

The next two parts are based on Proposition 4.3 with zero initial conditions,

$$f_t^{\mathbf{z}} - f_t = \sum_{k=0}^{t-1} \gamma_k \pi_{k+1}^t (L_K) \chi_k.$$

(4) $\|f_t^{\mathbf{z}} - f_t\|_K \le \left(\sum_{k=0}^{t-1} \gamma_k \|\pi_{k+1}^t(L_K)\| \right) \sup_{1 \le k \le t-1} \|\chi_k\|_K \le \left(\sum_{k=0}^{t-1} \gamma_k \right) \sup_{1 \le k \le t-1} \|\chi_k\|_K.$

(5) Note that $\|f_t^{\mathbf{z}} - f_t\|_\rho^2 = \|L_K^{1/2}(f_t^{\mathbf{z}} - f_t)\|_K^2$, whence similar to part (4),

$$\begin{aligned} \|f_t^{\mathbf{z}} - f_t\|_\rho^2 \ &= \ \left\langle L_K^{1/2} \sum_{k=0}^{t-1} \gamma_k \pi_{k+1}^t (L_K) \chi_k, L_K^{1/2} \sum_{k=0}^{t-1} \gamma_k \pi_{k+1}^t (L_K) \chi_k \right\rangle \\ &\le \ \|r_t(L_K)\| \left(\sum_{k=0}^{t-1} \gamma_k \|\pi_{k+1}^t (L_K)\| \right) \left(\sup_{1 \le k \le t-1} \|\chi_k\|_K \right)^2, \\ &\le \ \left(\sum_{k=0}^{t-1} \gamma_k \right) \left(\sup_{1 \le k \le t-1} \|\chi_k\|_K \right)^2. \end{aligned}$$

The result follows by taking the square root. ■

5. Bounds for Sample Error and Approximation Error

In this section we present the proofs of Theorems 2.9 and 2.10. Before that, we present some concentration inequalities which are used in the probabilistic upper bound of sample error.

5.1. *Concentration of Random Hilbert–Schmidt Operators and Vectors*

Recall that a bounded linear operator T is called a *Hilbert–Schmidt operator* if $T = T^*$ and $\mathrm{tr}(T^2) < \infty$. The set of Hilbert–Schmidt operators contains all finite-rank self-adjoint operators and are contained in the set of compact operators. Given two Hilbert–Schmidt operators $S, T : \mathscr{H} \to \mathscr{H}$, we can define the inner product $\langle S, T\rangle_{HS} = \mathrm{tr}(S^*T)$ and whence the norm $\|S\|_{HS} = \sqrt{\langle S, S\rangle_{HS}}$. The completion with respect to this norm gives a Hilbert space consisting of Hilbert–Schmidt operators. Therefore, we can apply concentration inequalities in Hilbert spaces to study the random operators in this space. In this paper we use the following result due to Iosif Pinelis (Pinelis, 1992).

Lemma 5.1 (Pinelis–Hoeffding). *Let $(\xi_i)_{i\in\mathbb{N}} \in \mathscr{H}$ be an independent random sequence of zero means in a Hilbert space $\mathscr{H}$ such that for all i almost surely $\|\xi_i\| \le c_i < \infty$. Then, for all $t \in \mathbb{N}$,*

$$\mathbf{Prob}\left\{\left\|\sum_{i=1}^{m}\xi_i\right\| \ge \varepsilon\right\} \le 2\exp\left\{-\frac{\varepsilon^2}{2\sum_{i=1}^{m} c_i^2}\right\}.$$

Note that $S_{\mathbf{x}}^*S_{\mathbf{x}} : \mathscr{H}_K \to \mathscr{H}_K$ is a random Hilbert–Schmidt operator whose expectation is $L_K : \mathscr{H}_K \to \mathscr{H}_K$. The following proposition collects some useful bounds.

Proposition 5.2.

(1) $\|L_K\|_{HS} = \sqrt{\mathrm{tr}(L_K^2)} \le \kappa^2$;
(2) $\mathrm{tr}(S_x^*S_x) \le \kappa^2$ *for* $x \in X$;
(3) $\mathrm{tr}(S_{\mathbf{x}}^*S_{\mathbf{x}}) \le \kappa^2$ *for* $\mathbf{x} \in X^m$;
(4) $\|S_{\mathbf{x}}^*S_{\mathbf{x}}\|_{HS} = \sqrt{\mathrm{tr}((S_{\mathbf{x}}^*S_{\mathbf{x}})^2)} \le \kappa^2$;
(5) $\|S_x^*S_x - L_K\|_{HS} \le \sqrt{2}\kappa^2$.

Proof. (1) This follows from Corollary 3 in Section 2, Chapter III of Cucker and Smale (2002).

(2) Since $S_x^*S_x$ is a rank-one operator, then $\mathrm{tr}(S_x^*S_x) \le \|S_x^*S_x\| \le \kappa^2$.

(3) By $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$,

$$\mathrm{tr}(S_{\mathbf{x}}^*S_{\mathbf{x}}) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{tr}(S_{x_i}^*S_{x_i}) \le \kappa^2.$$

(4) Similarly,

$$\mathrm{tr}((S_{\mathbf{x}}^*S_{\mathbf{x}})^2) = \frac{1}{m^2}\sum_{i,j=1}^{m}\mathrm{tr}(S_{x_i}^*S_{x_i}S_{x_j}^*S_{x_j}) \le \kappa^4,$$

where $\mathrm{tr}(S_{x_i}^*S_{x_i}S_{x_j}^*S_{x_j}) = k(x_i, x_j)\,\mathrm{tr}(S_{x_i}^*S_{x_j}) \le \kappa^4$.

(5) By definition,

$$\begin{aligned}\|S_x^*S_x - L_K\|_{HS}^2 &= \|S_x^*S_x\|_{HS}^2 + \|L_K\|_{HS}^2 - 2\,\mathrm{tr}(S_x^*S_xL_K)\\ &\le \|S_x^*S_x\|_{HS}^2 + \|L_K\|_{HS}^2 \le 2\kappa^4\end{aligned}$$

since $\mathrm{tr}(S_x^*S_xL_K) = \sum_i \mu_i\varphi_i(x)^2 = K(x,x) \ge 0$, where (μ_i, φ_i) is an eigensystem of L_K. ■

Let $\xi_i = S_{x_i}^*S_{x_i} - L_K$. Setting $c_i = \sqrt{2}\kappa^2$ and $\varepsilon = m\varepsilon$, by Lemma 5.1 we obtain

Proposition 5.3.

$$\mathbf{Prob}\{\|S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K\|_{HS} \ge \varepsilon\} \le 2\exp\left\{-\frac{m\varepsilon^2}{4\kappa^4}\right\}.$$

Therefore with probability at least $1-\delta$ $(\delta \in (0,1))$,

$$\|S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K\| \le \left\|S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K\right\|_{HS} \le \frac{2\kappa^2}{\sqrt{m}}\log^{1/2}\frac{2}{\delta}.$$

Note that $S_{\mathbf{x}}^*\mathbf{y} = (1/m)\sum_{i=1}^m y_iK_{x_i}$ is a random vector in $\mathscr{H}_K$ with expectation $\mathbb{E}[S_{\mathbf{x}}^*\mathbf{y}] = L_Kf_\rho$. Moreover, $\|S_{\mathbf{x}}^*\mathbf{y}\| \le \|S_{\mathbf{x}}^*\|\|\mathbf{y}\| \le \kappa M$ and $\|L_Kf_\rho\| \le \kappa M$. Thus, setting $\xi_i = S_{\mathbf{x}}^*\mathbf{y} - L_Kf_\rho$, $c_i = 2\kappa M$, and $\varepsilon = m\varepsilon$, by Lemma 5.1 we obtain

Proposition 5.4.

$$\mathbf{Prob}\{\|S_{\mathbf{x}}^*\mathbf{y} - L_Kf_\rho\|_K \ge \varepsilon\} \le 2\exp\left\{-\frac{m\varepsilon^2}{8\kappa^2M^2}\right\}.$$

Therefore with probability at least $1-\delta$ $(\delta \in (0,1))$,

$$\left\|S_{\mathbf{x}}^*\mathbf{y} - L_Kf_\rho\right\|_K \le \frac{2\sqrt{2}\kappa M}{\sqrt{m}}\log^{1/2}\frac{2}{\delta}.$$

Concentration results of this kind were first obtained by De Vito, Rosasco, Caponnetto, Giovannini, and Odone (2004).

5.2. *A Probabilistic Upper Bound for Sample Error*

Before the formal proof, we present a proposition which gives a probabilistic upper bound on the random variable $\chi_t = (L_K - S_{\mathbf{x}}^*S_{\mathbf{x}})f_t^{\mathbf{z}} + S_{\mathbf{x}}^*\mathbf{y} - L_Kf_\rho$ using the concentration inequalities in Propositions 5.3 and 5.4.

Proposition 5.5. *With probability at least* $1-\delta$ $(\delta \in (0,1))$ *there holds, for all* $t \in \mathbb{N}$,

$$\sup_{1\le k\le t-1}\|\chi_k\|_K \le \frac{2(1+\sqrt{2})\kappa M}{\sqrt{1-\theta}}\log^{1/2}\frac{2}{\delta}\sqrt{\frac{t^{1-\theta}}{m}}.$$

Proof. Note that

$$\begin{aligned}\sup_{1\le k\le t}\|\chi_k\|_K &\le \|L_K - S_{\mathbf{x}}^* S_{\mathbf{x}}\| \sup_{1\le k\le t-1}\|f_k^{\mathbf{z}}\|_K + \|S_{\mathbf{x}}^*\mathbf{y} - L_K f_\rho\|_K \\ &\le M\sqrt{\sum_{k=0}^{t-1}\gamma_k}\|L_K - S_{\mathbf{x}}^* S_{\mathbf{x}}\| + \|S_{\mathbf{x}}^*\mathbf{y} - L_K f_\rho\|_K.\end{aligned}$$

Using $\sum_{k=0}^{t-1}\gamma_k \le [1/\kappa^2(1-\theta)]t^{1-\theta}$ and Propositions 5.3 and 5.4, we have

$$\begin{aligned}M\sqrt{\sum_{k=0}^{t-1}\gamma_k}\|L_K - S_{\mathbf{x}}^* S_{\mathbf{x}}\| &\le \frac{2\kappa^2 M}{\sqrt{m}}\log^{1/2}\frac{2}{\delta}\cdot\frac{1}{\kappa\sqrt{1-\theta}}t^{(1-\theta)/2} \\ &\le \frac{2\kappa M}{\sqrt{1-\theta}}\log^{1/2}\frac{2}{\delta}\sqrt{\frac{t^{1-\theta}}{m}},\end{aligned}$$

and

$$\|S_{\mathbf{x}}^*\mathbf{y} - L_K f_\rho\|_K \le \frac{2\sqrt{2}\kappa M}{\sqrt{m}}\log^{1/2}\frac{2}{\delta}.$$

Adding them together, and noticing that $1 \le \sqrt{t^{(1-\theta)}/(1-\theta)}$, we obtain the result. ■

Now we are in a position to prove Theorem 2.9.

Proof of Theorem 2.9. From Propositions 4.4(5) and 5.5, and $\sum_{k=0}^{t-1}\gamma_k \le [1/\kappa^2(1-\theta)]t^{1-\theta}$,

$$\begin{aligned}\|f_t^{\mathbf{z}} - f_t\|_\rho &\le \sqrt{\sum_{k=0}^{t-1}\gamma_k}\sup_{1\le k\le t-1}\|\chi_k\|_K \\ &\le \frac{1}{\kappa\sqrt{1-\theta}}t^{(1-\theta)/2}\frac{2(1+\sqrt{2})\kappa M}{\sqrt{1-\theta}}\log^{1/2}\frac{2}{\delta}\sqrt{\frac{t^{1-\theta}}{m}} \\ &\le \frac{2(1+\sqrt{2})M}{1-\theta}\log^{1/2}\frac{2}{\delta}\cdot\frac{t^{1-\theta}}{\sqrt{m}},\end{aligned}$$

which gives the first bound.

Similarly, replacing Proposition 4.4(5) by 4.4(4), we have

$$\begin{aligned}\|f_t^{\mathbf{z}} - f_t\|_K &\le \sum_{k=0}^{t-1}\gamma_k\sup_{1\le k\le t-1}\|\chi_k\|_K \\ &\le \frac{1}{\kappa^2(1-\theta)}t^{1-\theta}\frac{2(1+\sqrt{2})\kappa M}{\sqrt{1-\theta}}\log^{1/2}\frac{2}{\delta}\sqrt{\frac{t^{1-\theta}}{m}} \\ &\le \frac{2(1+\sqrt{2})M}{\kappa(1-\theta)^{3/2}}\log^{1/2}\frac{2}{\delta}\cdot\frac{t^{(3/2)(1-\theta)}}{\sqrt{m}},\end{aligned}$$

which gives the second bound. ■

5.3. *A Deterministic Upper Bound for Approximation Error*

The following is the proof of Theorem 2.10 using a similar technique to Engl, Hanke, and Neubauer (2000).

Proof of Theorem 2.10. Let $f_\rho = L_K^r g$ with $\|g\|_\rho \leq R$. By Proposition 4.2, with $f_0 = 0$,

$$f_t - f_\rho = g_t(L_K)L_K f_\rho - f_\rho = -r_t(L_K)f_\rho,$$

whence

$$\|f_t - f_\rho\|_\rho = \|r_t(L_K)L_K^r g\|_\rho \leq R\|L_K^r r_t(L_K)\|.$$

where, with eigenvalues $(\mu_j)_{j\in\mathbb{N}}$ for L_K,

$$\begin{aligned}\|L_K^r r_t(L_K)\| &\leq \sup_j \lambda_j^r \prod_{i=0}^{t-1}(1-\gamma_i\mu_j) = \sup_j \exp\left\{\sum_{i=0}^{t-1}\log(1-\gamma_i\mu_j) + r\log\mu_j\right\}\\ &\leq \sup_j \exp\left\{-\sum_{i=0}^{t-1}\gamma_i\mu_j + r\log\mu_j\right\},\\ &\quad \text{where} \quad \log(1+x) \leq x \text{ for } x > -1.\end{aligned}$$

But the function

$$f(x) = -\sum_i \gamma_i x + r\log x, \qquad x > 0,$$

is maximized at $x^* = r/(\sum_i \gamma_i)$ with $f(x^*) = -r + r\log r - r\log\sum_i\gamma_i$. Taking $\gamma_t = 1/[\kappa^2(t+1)^\theta]$, we obtain

$$\|L_K^r r_t(L_K)\| \leq (r/e)^r\left(\sum_{i=0}^{t-1}\gamma_i\right)^{-r} \leq \left(\frac{2r\kappa^2}{e}\right)^r t^{-r(1-\theta)},$$

using $\sum_{k=0}^{t-1}\gamma_k \geq (1/2\kappa^2)t^{1-\theta}$.

For the case of $r > \frac{1}{2}$, $f_\rho \in \mathscr{H}_K$, and by the isomorphism $L_K^{1/2} : \mathscr{L}^2_{\rho_X}/\ker(L_K) \to \mathscr{H}_K$,

$$\|f_t - f_\rho\|_K = \|L_K^{-1/2}(f_t - f_\rho)\|_\rho = \|L_K^{r-1/2}r_t(L_K)g\|_\rho \leq R\|L_K^{r-1/2}r_t(L_K)\|.$$

Replacing r by $r - \frac{1}{2}$ above leads to the second bound. ■

6. Early Stopping in Classification

In this section we apply the Main Theorem to classifications and give a proof of Theorem 2.6. The formal proof is presented at the end of this section and before that we provide some background. For simplicity, we only use Tsybakov's noise condition to derive the convergence rates. Our results can be extended to incorporate the geometric noise condition introduced by Steinwart and Scovel (2005), which is however not pursued in this paper.

First recall different error measures for binary classification problems and then collect some results on the relation between them. In this section let $Y = \{\pm 1\}$. Define the *misclassification set*

$$X_f := \{x \in X \mid \operatorname{sign} f \neq \operatorname{sign} f_\rho\}.$$

For classification problems, the following error measure is proposed in Smale and Zhou (2005),

$$\|\operatorname{sign} f - \operatorname{sign} f_\rho\|_\rho$$

which is equivalent to the probability of misclassification by f,

$$\|\operatorname{sign} f - \operatorname{sign} f_\rho\|_\rho^2 = 4\rho_X(X_f). \tag{17}$$

More often in the literature, the following *misclassification risk* is used

$$R(f) = \rho_Z(\{(x, y) \in Z \mid \operatorname{sign} f(x) \neq y\}),$$

which is minimized at the so-called *Bayes rule*, sign f_ρ. It is easy to check that

$$R(f) - R(f_\rho) = \int_{X_f} |f_\rho(x)|\, d\rho_X(x). \tag{18}$$

6.1. *Tsybakov's Noise Condition*

The Tsybokov Noise Condition characterizes the regularity of the regression function f_ρ when crossing its zero level set.

Define the *Tsybakov function* $T_\rho : [0, 1] \to [0, 1]$ by

$$T_\rho(s) = \rho_X(\{x \in X : f_\rho(x) \in [-s, s]\}), \tag{19}$$

which characterizes the probability of the level sets of f_ρ within $[-s, s]$. The following *Tsybakov noise condition* (Tsyvakov, 2004), for some $q \in [0, \infty]$,

$$T_\rho(s) \leq B_q s^q, \qquad \forall s \in [0, 1], \tag{20}$$

characterizes the decay rate of $T_\rho(s)$. In particular, when T_ρ vanishes at a neighborhood of 0 (i.e., $T_\rho(s) = 0$ when $s \leq \varepsilon$ for some $\varepsilon > 0$), indicating a nonzero hard margin, we have $q = \infty$.

The following equivalent condition is useful (see Tsyvakov, 2004; or Bousquet, Boucheron, and Lugosi, 2004).

Lemma 6.1. *Tsybakov's condition* (20) *is equivalent*[1] *to that, for all* $f \in \mathscr{L}^2_{\rho_X}$,

$$\rho_X(X_f) \leq c_\alpha (R(f) - R(f_\rho))^\alpha, \tag{21}$$

[1] The uniform condition, for all $f \in \mathscr{L}^2_{\rho_X}$, is crucial for the direction (21) ⇒ (20) as shown in the proof. If we replace it by $f \in \mathscr{H}_K$, the two conditions are not equivalent. However, the proof of Theorem 2.6, or Proposition 6.2(5), only requires the direction (20) ⇒ (21).

where

$$\alpha = \frac{q}{q+1} \in [0, 1] \tag{22}$$

and $c_\alpha = B_q + 1 \geq 1$.

Proof. (20) ⇒ (21) Recalling (18) we have the following chains of inequalities:

$$\begin{aligned} R(f) - R(f_\rho) &\geq \int_{X_f} |f_\rho(x)| \chi_{|f_\rho(x)|>t} \, d\rho_X \geq t \int_{X_f} \chi_{|f_\rho(x)|>t} \, d\rho_X \\ &= t \left[\int_X \chi_{|f_\rho(x)|>t} \, d\rho_X - \int_{X/X_f} \chi_{|f_\rho(x)|>t} \, d\rho_X \right] \\ &\geq t[(1 - B_q t^q) - \rho_X(X \backslash X_f)] = t(\rho_X(X_f) - B_q t^q). \end{aligned}$$

The proof follows by taking

$$t = \left(\frac{1}{B_q + 1} \rho_X(X_f) \right)^{1/q}$$

and setting α as in (22).

(21) ⇒ (20) Define, for $s > 0$,

$$X_s = \{x \in X : |f_\rho(x)| \leq s\}.$$

Choose a $f \in \mathscr{L}^2_{\rho_X}$ such that sign f = sign f_ρ on $X \backslash X_s$ and otherwise sign $f \neq$ sign f_ρ, then $X_f = X_s$. Therefore,

$$\begin{aligned} \rho_X(X_f) &= \rho_X(X_s) \leq c_\alpha (R(f) - R(f_\rho))^\alpha \leq c_\alpha \left(\int_{X_s} |f_\rho(x)| \, d\rho_X \right)^\alpha \\ &\leq c_\alpha t^\alpha \rho_X(X_s)^\alpha = c_\alpha t^\alpha \rho_X(X_f)^\alpha, \end{aligned}$$

whence $\rho_X(X_f) \leq c_\alpha^{1/(1-\alpha)} t^{\alpha/(1-\alpha)}$ which recovers (20) with $q = \alpha/(1-\alpha)$ and $B_q = c_\alpha^{1/(1-\alpha)}$. ■

6.2. *Comparison Results and Proof of Theorem* 2.6

Now recall several results relating the different error measures introduced above.

Proposition 6.2. *Let f be some function in $\mathscr{L}^2_{\rho_X}$. The following inequalities hold*:

(1) $R(f) - R(f_\rho) \leq \|f - f_\rho\|_\rho$;

(2) *If* (21) *hold*, *then* $R(f) - R(f_\rho) \leq 4c_\alpha \|f - f_\rho\|_\rho^{2/(2-\alpha)}$;

(3) $R(f) - R(f_\rho) \leq \frac{1}{2} \|f_\rho\|_\rho \|\text{sign}\, f - \text{sign}\, f_\rho\|_\rho$;

(4) $R(f) - R(f_\rho) \leq \frac{1}{4} \|f - f_\rho\|_\infty \|\text{sign}\, f - \text{sign}\, f_\rho\|_\rho^2$;

(5) $\|\text{sign}\, f - \text{sign}\, f_\rho\|_\rho^2 \leq 4T(\|f - f_\rho\|_\infty)$;

(6) *If* (21) *hold*, *then* $\|\text{sign}\, f - \text{sign}\, f_\rho\|_\rho \leq 4c_\alpha \|f - f_\rho\|_\rho^{\alpha/(2-\alpha)}$.

Remark 6.3. Part (4) was used in Smale and Zhou (2005) by applying bounds on $\|f - f_\rho\|_K$ to estimate $\|f - f_\rho\|_\infty$. Due to the square on the left-hand side, this loses a power of $\frac{1}{2}$ in the asymptotic rate. But turning to the weaker norm $\|f - f_\rho\|_\rho$, part (5) remedies this problem without losing the rate.

Proof. (1) The proof is straightforward by noting that

$$|f_\rho(x)| \le |f(x) - f_\rho(x)| \tag{23}$$

when $x \in X_f$. In fact, from (18),

$$R(f) - R(f_\rho) \le \int_{X_f} |f(x) - f_\rho(x)| \le \|f - f_\rho\|_\rho.$$

(2) The inequality is a special case of Theorem 10 in Bartlett, Jordan, and McAuliffe (2003). Here we give the proof for completeness. If we further develop (18) we get

$$R(f) - R(f_\rho) = \int_{X_f} |f_\rho(x)|\chi_{|f_\rho(x)|\le t}\, d\rho_X(x) + \int_{X_f} |f_\rho(x)|\chi_{|f_\rho(x)|>t}\, d\rho_X(x),$$

where for $|f_\rho(x)| > t$, $|f_\rho(x)| = |f_\rho(x)|^2/|f_\rho(x)| < (1/t)|f_\rho(x)|^2$. Then, by conditions (21) and (23), we have

$$\begin{aligned} R(f) - R(f_\rho) &\le t\rho_X(X_f) + \frac{1}{t}\int_{X_f} |f_\rho(x)|_\rho^2\, d\rho_X(x) \\ &\le tc_\alpha(R(f) - R(f_\rho))^\alpha + \frac{1}{t}\|f - f_\rho\|_\rho^2. \end{aligned}$$

The result follows by taking $t = (1/2c_\alpha)(R(f) - R(f_\rho))^{1-\alpha}$ and $(4c_\alpha)^{1/(2-\alpha)} \le 4c_\alpha$ as $\alpha \in [0, 1]$ and $c_\alpha \ge 1$.

(3) From (18), simply using the Schwartz Inequality we have

$$\begin{aligned} R(f) - R(f_\rho) &= \tfrac{1}{2}\int_X f_\rho(x)(\operatorname{sign} f(x) - \operatorname{sign} f_\rho(x))\, d\rho_X(x) \\ &\le \tfrac{1}{2}\|f_\rho\|_\rho\|\operatorname{sign} f - \operatorname{sign} f_\rho\|_\rho. \end{aligned}$$

(4) Similarly to part (3), noting that for $x \in X_f$, $|f_\rho(x)| \le |f(x) - f_\rho(x)|$, by (18),

$$\begin{aligned} R(f) - R(f_\rho) &\le \tfrac{1}{4}\int_X |f(x) - f_\rho(x)| \cdot |\operatorname{sign} f(x) - \operatorname{sign} f_\rho(x)|^2\, d\rho_X(x) \\ &\le \tfrac{1}{4}\|f - f_\rho\|_\infty\|\operatorname{sign} f - \operatorname{sign} f_\rho\|_\rho, \end{aligned}$$

which gives the result.

(5) See Proposition 2 in Smale and Zhou (2005).

(6) The proof follows from (17) by plugging in (21) and part (2). ■

The following result provides an upper bound of the average $\rho_X(X_f)$ over the training samples by probability of a large deviation in supreme norm, when a hard margin exists. This bound is crucial to obtain exponential rates in Theorem 2.6(3) for plug-in classifiers. It is an adapted form from Lemma 3.6 in Audibert and Tsybokov (2005).

Proposition 6.4. *Assume that*

$$\rho_X(x \in X : |f_\rho(x)| \leq \gamma) = 0.$$

Then, for any mapping $f_{\mathbf{z}} : Z^m \to \mathscr{H}_K$, *the following holds*:

$$\mathbb{E}_{\mathbf{z}\in Z^m}[\rho_X(X_{f_{\mathbf{z}}})] \leq \mathbf{Prob}\{\mathbf{z} \in Z^m : \|f_{\mathbf{z}} - f_\rho\|_\infty > \gamma\}.$$

Proof. For any $x \in X_{f_{\mathbf{z}}}$, note that

$$\gamma < |f_\rho(x)| \leq |f_{\mathbf{z}}(x) - f_\rho(x)| \leq \|f_{\mathbf{z}} - f_\rho\|_\infty.$$

Hence $\rho_X(X_{f_{\mathbf{z}}}) \leq \rho_X(|f_{\mathbf{z}}(x) - f_\rho(x)| > \gamma)$ and

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\in Z^m}[\rho_X(|f_{\mathbf{z}}(x) - f_\rho(x)| > \gamma)] &= \mathbb{E}_{\mathbf{z}\in Z^m}[\mathbb{E}_{x\in X}[\mathbf{1}_{\{|f_{\mathbf{z}}(x)-f_\rho(x)|>\gamma\}}]] \\ &\leq \mathbb{E}_{\mathbf{z}\in Z^m}[\mathbf{1}_{\{\|f_{\mathbf{z}}-f_\rho\|_\infty>\gamma\}}\end{aligned}$$

which equals the right-hand side. ■

Now we are ready to give the proof of Theorem 2.6.

Proof of Theorem 2.6. (1) This follows from the Main Theorem(1) with Proposition 6.2(6).

(2) By Proposition 6.2(5),

$$\begin{aligned}\|\mathrm{sign}(f^{\mathbf{z}}_{t^*}) - \mathrm{sign}(f_\rho)\|_\rho &\leq 2T^{1/2}(\|f^{\mathbf{z}}_{t^*} - f_\rho\|_\infty) \leq 2B_q^{1/2}\|f^{\mathbf{z}}_{t^*} - f_\rho\|_\infty^{q/2} \\ &\leq 2\sqrt{\kappa B_q}\|f^{\mathbf{z}}_{t^*} - f_\rho\|_K^{q/2}.\end{aligned}$$

The result then follows from the Main Theorem(2).

(3) From Proposition 6.4,

$$\begin{aligned}\|\mathrm{sign}(f^{\mathbf{z}}_{t^*}) - \mathrm{sign}(f_\rho)\|_\rho &= 4\rho_X(X_{f^{\mathbf{z}}_{t^*}}) \leq \mathbf{Prob}\{\mathbf{z} \in Z^m : \|f_{\mathbf{z}} - f_\rho\|_\infty > \gamma\} \\ &\leq \mathbf{Prob}\{\mathbf{z} \in Z^m : \|f_{\mathbf{z}} - f_\rho\|_K > \gamma/\kappa\}.\end{aligned}$$

Now apply the Main Theorem(2), by setting

$$D_{\rho,K\delta} m^{-(r-1/2)/(2r+2)} \geq \frac{\gamma}{\kappa},$$

which gives

$$\log^{1/2}\frac{2}{\delta} \geq \frac{1}{\kappa C_9} m^{(r-1/2)/(2r+2)} - \frac{C_{10}}{C_9}, \tag{24}$$

where

$$C_9 = \frac{4(1+\sqrt{2})M}{\kappa(1-\theta)^{3/2}} \qquad \text{and} \qquad C_{10} = R\left(\frac{2(r-\frac{1}{2})\kappa^2}{e}\right)^{r-1/2}.$$

Replacing $\log^{1/2} 2/\delta$ by its upper bound $\log 2/\delta \cdot \log^{-1} 2$, inequality (24) leads to an upper bound on δ,

$$\delta \le C_{11} e^{-C_{12}\gamma m^{(r-1/2)/(2r+2)}},$$

where $C_{11} = \frac{1}{2}\exp(C_{10}\log 2/C_9)$ and $C_{12} = \log 2/(\kappa C_9)$. This gives the upper bound,

$$\begin{aligned}\|\operatorname{sign}(f_{t^*}^{\mathbf{z}}) - \operatorname{sign}(f_\rho)\|_\rho &\le \mathbf{Prob}\{\mathbf{z} \in Z^m : \|f_{\mathbf{z}} - f_\rho\|_K > \gamma/\kappa\} \\ &\le C_{11} e^{-C_{12}\gamma m^{(r-1/2)/(2r+2)}},\end{aligned}$$

which completes the proof. ■

7. Conclusion and Open Problems

In this paper we present some upper bounds for early stopping regularization to approximate the regression function f_ρ from a reproducing kernel Hilbert space $\mathscr{H}_K$ when f_ρ lies in the image of L_K^r $(r > 0)$. These upper bounds have asymptotic rates of $O(m^{-r/(2r+2)})$ for $\mathscr{L}^2_{\rho_X}$-convergence when $r > 0$, and $O(m^{-(r-1/2)/(2r+2)})$ for a stronger $\mathscr{H}_K$-convergence when $r > \frac{1}{2}$. A direct application of these upper bounds in classifications leads to some fast convergence rates for the plug-in classifiers. In particular, exponential convergence rates are achieved when $r > \frac{1}{2}$ and a hard margin condition holds for f_ρ.

On the other hand, these upper bounds are suboptimal and it is still an open problem as to *how to achieve the optimal* $\mathscr{L}^2_{\rho_X}$*-convergence rates* $O(m^{-r/(2r+1)})$ *for all* $r > 0$.

Acknowledgments. The authors would like to acknowledge Ernesto De Vito for sharing with us some results on the generalized Mercer Theorem; Ha Quang Minh for sharing with us his results on convergence rates of (Tikhonov) regularized least square learning algorithms; Grace Wahba for pointing out her early work on this topic; Peter Bickel and Bin Yu for many helpful discussions on Boosting; and, especially, Steve Smale for his support and encouragement. The authors also thank Bo Li, Gang Liang, Michele Piana, Rob Schapire, Pierre Tarres, D.-X. Zhou, and many reviewers for helpful discussions and comments.

This work was done while the authors visited the Toyota Technical Institute at Chicago and the Computer Science Department at the University of Chicago, in the Program in Learning Theory and Related Areas during March–May, 2005. The authors are supported by NSF grant 0325113. Lorenzo Rosasco and Andrea Caponneto are also partially funded by the FIRB Project ASTAA and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

J. Y. AUDIBERT, A. B. TSYBOKOV (2005): *Fast learning rates for plug-in classifiers under the margin condition*. Ann. Statist. Accepted.

P. L. BARTLETT, M. J. JORDAN, J. D. MCAULIFFE (2003): *Convexity, classification, and risk bounds*. J. Amer. Statist. Assoc. Preprint.

F. BAUER, S. PEREVERZEV, L. ROSASCO (2006): *On regularization algorithms in learning theory*. Technical Report. DISI, Universitá di Genova.

P. J. BICKEL, Y. RITOV, A. ZAKAI (2005): *Some theory for generalized boosting algorithms*. Preprint.

N. BISSANTZ, T. HOHAGE, A. MUNK, F. RUYMGAART (2006): *Convergence rates of general regularization methods for statistical inverse problems and applications*. Preprint Nr. 2006-02. Institute for Mathematical Stochastics, Georgia Augusta University Goettingen.

G. BLANCHARD, G. LUGOSI, N. VAYATIS (2003): *On the rate of convergence of regularized boosting classifiers*. J. Mach. Learn. Res. (4), 861–894.

O. BOUSQUET, S. BOUCHERON, G. LUGOSI (2004): *Theory of classification: A survey of recent advances*. ESAIM Probab. Statist. To appear.

L. BREIMAN (1999): *Prediction games and arcing algorithms*. Neural Comput., **11**:1493–1517.

L. BREIMAN (2004): *Population theory for boosting ensembles*. Ann. Statist., **32**:1–11.

P. BÜHLMANN, B. YU (2002): *Boosting with the l_2-loss: Regression and classification*. J. Amer. Statist. Assoc., **98**:324–340.

A. CAPONNETTO (2006): *Optimal rates for regularization operators in learning theory*. Techinical Report. TTI-Chicago.

A. CAPONNETTO, E. DE VITO (2005): *Optimal rates for regularized least squares algorithm*. Preprint.

F. CUCKER, S. SMALE (2002): *On the mathematical foundations of learning*. Bull. Amer. Math. Soc., **29**(1):1–49.

E. DE VITO, L. ROSASCO, A. CAPONNETTO, U. D. GIOVANNINI, F. ODONE (2004): *Learning from examples as an inverse problem*. J. Mach. Learn. Res. Preprint.

R. DEVORE, G. KERKYACHARIAN, D. PICARD, V. TEMLYAKOV (2004): *Mathematical methods for supervised learning*. IMI Research Reports 04:22. Department of Mathematics, University of South Carolina.

T. G. DIETTERICH (1997): *Machine learning research: Four current directions*. AI Mag., **18**(4):97–136.

H. W. ENGL, M. HANKE, A. NEUBAUER (2000): Regularization of Inverse Problems. Amsterdam: Kluwer Academic.

H. FLEMING (1990): *Equivalence of regularization and truncated iteration in the solution of ill-posed image reconstruction problems*. Linear Algebra Appl., **130**:133–150.

Y. FREUND, R. E. SCHAPIRE (1997): *A decision-theoretic generalization of online learning and an application to boosting*. J. Comput. System Sci., **55**(1):119–139.

Y. FREUND, R. E. SCHAPIRE (1999): *Adaptive game playing using multiplicative weights*. Games Econom. Behav., **29**:79–103.

J. FRIEDMAN, T. HASTIE, R. TIBSHIRANI (2000): *Additive logistic regression: A statistical view of boosting*. Ann. Statist., **38**(2):337–374.

J. H. FRIEDMAN (2001): *Greedy function approximation: A gradient boosting machine*. Ann. Statist., **29**:1189–1232.

L. GYÖRFI, M. KOHLER, A. KRZYŻAK, H. WALK (2002): A Distribution-Free Theory of Nonparametric Regression. New York: Springer-Verlag.

M. HANKE (1995): Conjugate Gradient-Type Methods for Ill-Posed Problems. Boston: Pitman Research Notes in Mathematics Series. Longman Scientific & Technical.

W. JIANG (2004): *Process consistency for adaboost*. Ann. Statist., **32**:13–29.

G. LUGOSI, N. VAYATIS (2004): *On the Bayes-risk consistency of regularized boosting methods*. Ann. Statist., **32**:30–55.

L. MASON, J. BAXTER, P. BARTLETT, M. FREAN (2000): *Functional gradient techniques for combining hypotheses*. In: A. J. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (Eds.), Advances in Large Margin Classifiers. Cambridge, MA: MIT Press.

P. MATHÉ (2004): *Saturation of regularization methods for linear ill-posed problems in Hilbert spaces*. SIAM J. Numer. Anal., **42**(3):968–973.

P. MATHÉ, S. PEREVERZEV (2002): *Moduli of continuity for operator monotone functions*. Numer. Funct. Anal. Optim., **23**:623–631.

H. Q. MINH (2005): Personal communications.

C. S. ONG (2005): *Kernels*: *Regularization and optimization*. PhD Thesis. Australian National University.

G. PESKIR (2000): From Uniform Laws of Large Numbers to Uniform Ergodic Theorems. Lecture Notes Series, Vol. 66. Department of Mathematics, University of Aarhus, Denmark.

I. PINELIS (1992): *An approach to inequalities for the distributions of infinite-dimensional martingales*. In: R. M. Dudley, M. G. Hahn, J. Kuelbs (Eds.), Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference, pp. 128–134.

C. RUDIN, I. DAUBECHIES, R. E. SCHAPIRE (2004): *The dynamics of adaboost*: *Cyclic behavior and convergence of margins*. J. Mach. Learn. Res. To appear.

R. SCHAPIRE (2001): *Drifting games*. Mach. Learn. **43**(3):265–291.

R. E. SCHAPIRE (2002): *The boosting approach to machine learning*: *An overview*. In: MSRI Workshop on Nonlinear Estimation and Classification. To appear.

R. E. SCHAPIRE, Y. FREUND, P. BARTLETT, W. S. LEE (1998): *Boosting the margin*: *A new explanation for the effectiveness of voting methods*. Ann. Statist., **26**(5):1651–1686.

S. SMALE, Y. YAO (2005): *Online learning algorithms*. Found. Comput. Math. Accepted.

S. SMALE, D.-X. ZHOU (2004): *Shannon sampling and function reconstruction from point values*. Bull. Amer. Math. Soc., **41**(3):279–305.

S. SMALE, D.-X. ZHOU (2005): *Learning theory estimates via integral operators and their approximations*. Preprint.

I. STEINWART, C. SCOVEL (2005): *Fast rates for support vector machines using gaussian kernels*. Ann. Statist. Submitted.

G. STOLTZ, G. LUGOSI (2004): *Learning correlated equilibria in games with compact sets of strategies*. Preprint.

V. N. TEMLYAKOV (2004): *Optimal estimators in learning theory*. Technical report. IMI Research Reports 04:23. Department of Mathematics, University of South Carolina.

A. B. TSYBAKOV (2004): *Optimal aggregation of classifiers in statistical learning*. Ann. Statist., **32**:135–166.

L. G. VALIANT (1984): *A theory of the learnable*. In: Proc. 16th Annual ACM Symposium on Theory of Computing, pp. 135–166. New York: ACM Press.

G. WAHBA (1983): *Bayesian "confidence intervals" for the cross-validated smoothing spline*. J. Roy. Statist. Soc. Ser. B, pp. 133–150.

G. WAHBA (1987): *Three topics in ill-posed problems*. In: H. Engl, C. Groetsch (Eds.), Proceedings of the Alpine–U.S. Seminar on Inverse and Ill-Posed Problems, pp. 385–408. New York: Academic Press.

G. WAHBA (1990): Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59. Philadelphia, PA: SIAM.

Y. YAO (2005): *On the complexity issue of online learning algorithms*. IEEE Trans. Inform. Theory. Submitted.

K. YOSIDA, S. KAKUTANI (1941): *Operator-theoretical treatment of Markoff's process and mean ergodic theorem*. Ann. of Math., **42**(1):188–288.

T. ZHANG, B. YU (2003): *Boosting with early stopping*: *Convergence and consistency*. Technical Report 635. Department of Statistics, University of California at Berkeley.

P. ZHAO, B. YU (2004): *Boosted lasso*. Technical Report 678. Department of Statistics, University of California at Berkeley (December, 2004; revised and submitted to J. Roy. Statist. Soc. Ser. B in April, 2005).

Y. Yao
Department of Mathematics
University of California
Berkeley, CA 94720
USA
yao@math.berkeley.edu

L. Rosasco
C.B.C.L.
Massachusetts Institute of Technology
Bldg. E25-201
45 Carleton St.
Cambridge, MA 02142
USA
and
DISI
Università di Genova
Via Dodecaneso 35
16146 Genova
Italy
rosasco@disi.unige.it

A. Caponnetto
C.B.C.L.
Massachusetts Institute of Technology
Bldg. E25-201
45 Carleton St.
Cambridge, MA 02142
USA
and
DISI
Università di Genova
Via Dodecaneso 35
16146 Genova
Italy
caponnet@mit.edu