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. These upper bounds have improved rates where the usual regularized least square algorithm fails and achieve the minimax optimal rate $O(m^{-1/2})$ in some cases. We also discuss the implication of these results in the context of classification. Some connections are addressed with Boosting, Landweber iterations, and the on-line 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 a RKHS: 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 reduce bias. Solving this bias-variance trade-off leads to an early stopping rule.

In 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 values of convex combination coefficients, or l_1 norm) of the paths and thus restricting that radius implements regularization. For example,

Date: September 21, 2005.

2000 *Mathematics Subject Classification.* 62G08, 62L99, 68Q32, 68T05.

Key words and phrases. Gradient Descent Method, Early Stopping, Regularization, Boosting, Landweber Iteration, Reproducing Kernel Hilbert Space.

fixing radius is used in [Lugosi and Vayatis 2004; Blanchard, Lugosi, and Vayatis 2003]. Recently, early stopping regularization was systematically studied, 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 in [Zhao and Yu 2004] some backward steps are introduced to reduce 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.

It should be noted that the bias-variance decomposition here is close, but different to traditional work on early stopping regularization in ill-posed problems [e.g. see Hanke 1995; or reference in Chapter 5, Ong 2005]. In these works, the linear operators are fixed without randomization and only output noise is considered. The sample iteration is the noise perturbed path which typically first converges and eventually diverges (called *semi-convergence*). The population iteration is the unperturbed path convergent to the generalized inverse. In these works the early stopping was used to avoid the semi-convergence. However in the setting of our paper, only Monte-Carlo approximations of the linear operators can be obtained, and the sample iteration typically converges to an over-fitting solution. This increases the technical difficulty (e.g. the order optimality is still an open problem in this setting), though we benefit a lot from the similarity between the two settings.

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 by requiring the minimal number of iterations before stopping. 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, see for example [Smale and Zhou 2005; De Vito, Rosasco, Caponnetto, Giovannini, and Odone 2004], which is also called *penalized L_2 regression* or *ridge regression* in statistical literature, or *Tikhonov regularization* in inverse problems. Early stopping does not incur the *saturation* phenomenon that the rate no longer improves when the regression function goes beyond a certain level of regularity. The saturation problem was proposed and studied intensively in inverse problems [e.g. Engl, Hanke, and Neubauer 2000; Mathé 2004]. Our algorithms here can be regarded as finite rank Monte Carlo approximations of 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, 3.1 gives a comparison between early stopping and Tikhonov regularization; 3.2 discusses the connection to boosting in the view of gradient descent method; 3.3 discusses the connection to the Landweber iteration in linear inverse problems; 3.4 discusses the connection to on-line learning algorithms based on stochastic gradient method. Section 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. Appendix A collects some lemmas used in this paper and Appendix B provides some background on reproducing kernel Hilbert spaces, random operators and concentration inequalities in Hilbert spaces.

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 $\mathbf{z} = \{(x_i, y_i) \in X \times Y : i = 1, \dots, m\} \in Z^m$, drawn independently at random from a probability measure ρ on Z, one wants to minimize over $f \in \mathscr{H}$ the following quadratic functional

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

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

Here we recall some basic definitions on RKHSs and refer to Appendix B for an exposition with more details. Let $\mathscr{H}_K$ be the RKHS associated to a Mercer kernel $K : X \times X \to \mathbb{R}$, i.e. a continuous positive semi-definite function. Denote by $\langle\ ,\ \rangle_K$ and $\|\cdot\|_K$ the inner product and norm of $\mathscr{H}_K$. Let $K_x : X \to \mathbb{R}$ be the function defined by $K_x(s) = K(x, s)$ for $x, s \in X$.

Besides $\mathscr{H}_K$, another important function space, $\mathscr{L}^2_{\rho_X}$, is used throughout the paper. Denote by ρ_X the marginal probability measure on X. Then let $\mathscr{L}^2_{\rho_X}$ be the space of square integrable functions with respect to ρ_X and by $\|\cdot\|_\rho$ ($\langle , \rangle_\rho$) the norm (the inner product) in $\mathscr{L}^2_{\rho_X}$. Denote by $\rho_{Y|x}$ the conditional measure on Y given x. Our target function 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. In fact, by the relation

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

it can be seen that f_ρ is the minimizer of $\mathscr{E}(f)$ over $\mathscr{L}^2_{\rho_X}$

Thus the minimization of (1) is to equivalent to finding approximations of f_ρ from $\mathscr{H}_K$, a subspace (closed or not) in $\mathscr{L}^2_{\rho_X}$.

Here we define an integral operator which plays a central role in the theory. Let $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{H}_K$ be an integral operator defined by $(L_K f)(x') = \int K(x', x) f(x) d\rho_X$. Its restriction $L_K|_{\mathscr{H}_K}$ induces an operator from $\mathscr{H}_K$ into $\mathscr{H}_K$, which when its domain is clear from the context, is also denoted by L_K (in Appendix B denoted by $\bar{L}_K$).

Finally throughout the paper we assume the following.

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

2.2. **Gradient Descent Algorithms.** First we define two iterations: sample iteration and population iteration, then we show they are gradient descent algorithms with respect to proper objective functions.

Given a i.i.d. sample of size m, $\mathbf{z} =\in Z^m$, define the *sample iteration* as a 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). In this paper we choose $\gamma_t = \dfrac{1}{\kappa^2(t+1)^\theta}$ $(t \in \mathbb{N})$ for some $\theta \in [0, 1)$. Now define the *population iteration* as an averaged version of (2)

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

Clearly f_t is deterministic and $f_t^{\mathbf{z}}$ is a $\mathscr{H}_K$-valued random variable depending on $\mathbf{z}$.

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

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

Proposition 2.1. *The gradients of (1) and (4) are the maps from $\mathscr{H}_K$ into $\mathscr{H}_K$ given by*

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

and

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

Proof. The proof follows from Proposition A.1 in Appendix A by taking expectations, grad $V(f) = \mathbb{E}[(f(x) - y)K_x] = \int_{X\times Y}(f(x) - y)K_x d\rho = L_K f - L_K f_\rho$ and grad $\hat{V}(f) = \hat{\mathbb{E}}[(f(x) - y)K_x] = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)K_{x_i}$, 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 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 $\gamma_t = \dfrac{1}{\kappa^2(t+1)^\theta}$ $(t \in \mathbb{N})$ 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

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

where $C_{\rho,K,\delta} = \frac{8M}{1-\theta}\log^{1/2}\frac{2}{\delta} + R\left(\frac{2r\kappa^2}{e}\right)^r$, under the stopping rule

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

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

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

where $D_{\rho,K,\delta} = \frac{8M}{\kappa(1-\theta)^{3/2}}\log^{1/2}\frac{2}{\delta} + R\left(\frac{2(r-1/2)\kappa^2}{e}\right)^{r-1/2}$, under the stopping rule

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

Above $\lceil x \rceil$ denotes the smallest integer greater or equal than $x \in \mathbb{R}$.

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

Remark 2.2. The first upper bound holds for all $r > 0$. In the second upper bound, $r > 1/2$ implies $f_\rho \in \mathscr{H}_K$ as $L_K^{1/2} : \mathscr{L}^2_{\rho_X} \to \mathscr{H}_K$ is a Hilbert space isometry. In particular, when $r \to \infty$, we approaches the asymptotic rate $\|f^{\mathbf{z}}_{t^*(m)} - f_\rho\|_\rho \leq O(m^{-1/2})$ [1] 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, as only a finite number of eigenvalues of L_K are nonzero whence arbitrarily large r is allowed. Such a result improves the upper bounds for the usual regularized least square algorithm [Minh 2005; or 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 $r > 1/2$). 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 2.3. Some minimax lower rate [DeVore, Kerkyacharian, Picard, and Temlyakov 2004, Temlyakov 2004] and individual lower rate [Caponnetto and De Vito 2005], suggest that for $r > 0$ the convergence rate $O(m^{-r/(2r+1)})$ is optimal in both senses, which has been achieved by the usual regularized least square algorithm when r varies over a suitable range. This implies that in the sense above we can not obtain a rate faster than $O(m^{-1/2})$ which is achieved at $r \to \infty$ by our algorithms. It is an open problem whether the rates in the Main Theorem can be improved to meet the lower rate.

Remark 2.4. The upper bounds show an interesting result, that shrinking the step size γ_t might only affect the early stopping time, but not the rate of convergence. 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.

[1] This implies a rate $O(1/m)$ for the *generalization error* $\mathscr{E}(f^{\mathbf{z}}_{t^*(m)}) - \mathscr{E}(f_\rho) = \|f^{\mathbf{z}}_{t^*(m)} - f_\rho\|^2_\rho$.

Now we discuss a direct application of the main theorem to the setting of classification. Notice that when $Y = \{\pm 1\}$, algorithm (2) may provide a classification rule $\mathrm{sign} f_{t^*}^{\mathbf{z}}$. Hence we may consider such a rule as an approximation of the Bayes rule, $\mathrm{sign} f_\rho$. The following result gives an upper bound on the distance $\|\mathrm{sign} f_t^{\mathbf{z}} - \mathrm{sign} f_\rho\|_\rho$.

Theorem 2.5. *Suppose $Y = \{\pm\}$ and Tsybakov's noise condition*

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

for some $q \in [0, \infty]$ and $B_q \ge 0$. Under the same condition of the Main Theorem, for all $r > 0$ and $t \in \mathbb{N}$, the following holds with probability at least $1 - \delta$ ($\delta \in (0, 1)$),

$$\|\mathrm{sign} f_{t^*(m)}^{\mathbf{z}} - \mathrm{sign} f_\rho\|_\rho \le C_{\rho,K,\delta} m^{-\frac{\alpha r}{2(r+1)(2-\alpha)}}.$$

where $\alpha = q/(q+1)$ and $C_{\rho,K,\delta} = \frac{32(B_q+1)M}{1-\theta} \log^{1/2} \frac{2}{\delta} + 4(B_q+1)R \left(\frac{2r\kappa^2}{e}\right)^r$, under the stopping rule

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

Again the proof together with a detailed introduction on the background, is given in Section 6. Some remarks follow.

Remark 2.6. This result implies that as $\alpha = 1$ and $r \to \infty$, the convergence rate may approach $O(1/\sqrt{m})$ arbitrarily. This happens when using a finite dimensional $\mathscr{H}_K$ with a hard margin on f_ρ (i.e. $\rho_X(|f_\rho| \le c) = 0$ for some $c > 0$). However as in the Main Theorem, the constants here blow up exponentially as $r \to \infty$.

Remark 2.7. Consider Bayes consistency. Define the risk of f by

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

and let $R(f_\rho)$ be the *Bayes risk*. Combined with Proposition 6.2-2, the Main Theorem leads to an upper bound,

$$R(f_{t^*(m)}^{\mathbf{z}}) - R(f_\rho) \le O(m^{-\frac{2r}{2(r+1)(2-\alpha)}}),$$

whose asymptotic rate approaches to $O(1/m)$ as $\alpha = 1$ and $r \to \infty$, which is shown to be optimal [e.g. see Bartlett, Jordan, and McAuliffe 2003, Tsybakov 2004 and reference therein].

Next we present upper bounds for the sample error and the approximation error, respectively, which are used to prove the Main Theorem.

Theorem 2.8 (Sample Error)**.** *With probability at least $1 - \delta$ ($\delta \in (0, 1)$) there holds for all $t \in \mathbb{N}$,*

$$\|f_t^{\mathbf{z}} - f_t\|_\rho \le C_1 \frac{t^{1-\theta}}{\sqrt{m}},$$

where $C_1 = \dfrac{4M}{1-\theta} \log^{1/2} \dfrac{2}{\delta}$; and

$$\|f_t^{\mathbf{z}} - f_t\|_K \le C_2 \sqrt{\frac{t^{3(1-\theta)}}{m}},$$

where $C_2 = \dfrac{4M}{\kappa(1-\theta)^{3/2}} \log^{1/2} \dfrac{2}{\delta}$.

Theorem 2.9 (Approximation Error)**.** *Suppose $f_\rho \in L_K^r(B_R)$ for some $R, r > 0$ and $f_0 = 0$. Then for all $t \in \mathbb{N}$,*

$$\|f_t - f_\rho\|_\rho \le C_3 t^{-r(1-\theta)},$$

where $C_3 = R\left(\frac{2r\kappa^2}{e}\right)^r$; and if moreover $r > 1/2$, then $f_\rho \in \mathscr{H}_K$ and

$$\|f_t - f_\rho\|_K \le C_4 t^{-(r-1/2)(1-\theta)},$$

where $C_4 = R\left(\frac{2(r-1/2)\kappa^2}{e}\right)^{r-1/2}$.

Their proofs are given in Section 5.

Remark 2.10. 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 Theorem 2.8 and 2.9, we have

$$\|f_t^{\mathbf{z}} - f_\rho\|_\rho \le C_1 \frac{t^{1-\theta}}{\sqrt{m}} + C_3 t^{-r(1-\theta)}.$$

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

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

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

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

$$\|f_{t^*}^{\mathbf{z}} - f_\rho\|_\rho \le (\beta^{1-\theta} C_1 + \beta^{-r(1-\theta)} C_3) m^{-r/(2r+2)} \le (2C_1 + C_3) m^{-r/(2r+2)}.$$

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 on-line learning algorithms as stochastic approximations of the gradient descent method.

3.1. **Early Stopping vs. Tikhonov Regularization.** In this subsection we give some comparisons of early stopping regularization vs. the usual regularized least square algorithm [e.g. Cucker and Smale 2002; Smale and Zhou 2005], where following the tradition of inverse problems the latter is roughly called Tikhonov regularization here to emphasize some motivations behind our studies.

Let

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

be the solution of problem (1) with Tikhonov regularization. It is known [e.g. Cucker and Smale 2002] that

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

Moreover for simplicity, let $\gamma_t = \gamma_0 = 1$ (whence $\kappa = 1$), then by induction the population iteration (3) becomes

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

Now let $(\mu_i, \phi_i)_{i\in\mathbb{N}}$ be an eigen-system of the compact operator $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}$. Then we have decompositions

$$f_\lambda = \sum_i \frac{\mu_i}{\mu_i + \lambda}\langle f_\rho, \phi_i\rangle\phi_i,$$

and

$$f_t = \sum_i (1 - (1 - \mu_i)^t)\langle f_\rho, \phi_i\rangle\phi_i.$$

Compactness of L_K implies that $\lim_{i\to\infty}\mu_i = 0$. Therefore for most μ_i which are sufficiently small, $\mu_i/(\mu_i + \lambda) \approx 0$ and $1 - (1 - \mu_i)^t$ converges to 0 with gap dropping at least exponentially with t. Therefore both early stopping and Tikhonov regularization can be regarded as low pass filters on f_ρ, which tends to project f_ρ to the eigenfunctions corresponding to large eigenvalues. Such an observation in statistical estimation can be traced back to [Wahba 1987], which further derives generalized cross-validation criteria for data-dependent early stopping rules [see also Wahba, Johnson, Gao, and Gong 1995 for numerical experiments]. 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), there is a one-to-one correspondence between the two regularization methods.

On the other hand, there are also significant differences between the two regularization approaches. One major difference which motivates our study is that early stopping regularization seems to have better upper bounds than Tikhonov regularization. In fact, it can be shown [Minh 2005; or Appendix by Minh, in Smale and Zhou 2005] that if $f_\rho \in L^r_K(B_R)$ for some $r > 0$,

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

and for $r > 1/2$,

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

We can see for large r, the upper bound can not go faster than t^{-1} in Tikhonov regularization. On the other hand in early stopping regularization, taking $\theta = 0$ in Theorem 2.9 we have that for $r > 0$,

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

and for $r > 1/2$,

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

We may roughly regard such a relationship between regularization parameters, $\lambda \sim 1/t$, where early stopping has faster rates than Tikhonov regularization for large r. A consequence of this leads to a fast convergence rate $O(m^{-1/2})$ in the Main Theorem as $r \to \infty$, in a contrast for Tikhonov regularization the best known upper convergence rate in $\mathscr{L}^2_{\rho_X}$ (or in $\mathscr{H}_K$) [Minh 2005; or Remark 2 in Appendix in Smale and Zhou 2005] is $O(m^{-1/3})$ (or $O(m^{-1/4})$), achieved at all $r \geq 1$ (or $r \geq 3/2$).

In traditional studies of inverse problems, there is a closely related *saturation* phenomenon [e.g. Engl, Hanke, and Neubauer 2000]: when approximating functions in a Hilbert scale (here $L^r_K(B_R)$ for $r > 0$ or $r > 1/2$), Tikhonov regularization can not achieve the optimal error order for high enough regularity levels. The saturation phenomenon has been studied intensively in inverse problems. However due to the random design setting in learning, different to the setting in traditional inverse problems (we shall discuss this in Section 3.3), such results can not be applied to learning directly. A thorough study on saturation of regularization in learning, requires both tight upper and lower bounds, which are still open at this moment.

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. $t^* << m$), early stopping regularization is favored. For those very ill-conditioned kernel matrices, conjugate gradient descent method or more complicated 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 weather a "weak" learning algorithm which performs just slightly better than random guessing (with success probability slightly larger than 1/2) can be "boosted" into a "strong" learning algorithm of high accuracy [Valiant 1984; or see the review by Schapire 2002 or Dietterich 1997]. Roughly speaking, the weak learning algorithm generate an ensemble of base functions (weak learners) and then some aggregation methods are applied to construct a function of high accuracy (strong learner). 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 overfiting. 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 a road to game theory [e.g. Breiman 1999; Freund and Schapire 1999; Schapire 2001; Stoltz and Lugosi 2004], on the first have been seen in [Rudin, Daubechies, and Schapire 2004] and on the second lead 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 also lies in.

In the view of gradient descent with L_2 loss, our algorithms can be also 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 Shannon Sampling Theorem [Smale and Zhou 2004].

Sacrificing some mathematical rigor, our paper may be regarded as extending some early results in [Bühlmann and Yu 2002] from Sobolev spaces with fixed designs to general Reproducing Kernel Hilbert Spaces (RKHSs) with random designs (see Chapter I in [Györfi, Kohler, Krzyżak, and Walk 2002] for more discussions on random design vs. fixed design), yet with suboptimal convergence rates. However, a technical difficulty to justify this claim rigorously, lies in the question that if our "smoothness" assumption on regression function, $f_\rho \in L_K^r(B_R)$, is equivalent to Sobolev spaces even for suitable spline kernels. In fact, for each integer r, following [Wahba 1990] we can construct a corresponding spline kernel K, thus get a Sobolev space W^r as a RKHS $\mathscr{H}_K$, which is identical (isometric) to the image of $L_K^{1/2}$ [Cucker and Smale 2002]. Sobolev spaces with index less than r, W^s $(0 < s < r)$, can be regarded as the interpolation spaces of the Sobolev space with smooth index r, W^r. It is true that the image of L_K^s , $(s < 1/2)$, does lie in the interpolation spaces of the image of $L_K^{1/2}$, i.e. $\mathscr{H}_K = W^r$; however the converse, if every interpolation space of W^r can be represented as the image of L_K^s for suitable $s < 1/2$, is not clear yet up to the author's knowledge. On the other hand, such a connection seems not harmful and does help understanding, whence we note it down here with the hope that further work can clarify this connection.

The gradient descent view on boosting triggers a series of studies on consistency and regularization in boosting [e.g. Jiang 2004; Breiman 2004; Lugosi and Vayatis 2004; Zhang and Yu 2003], going beyond margin analysis in early studies [Schapire, Freund, Bartlett, and Lee 1998]. As we mentioned in the beginning, a common perspective adopts the bias-variance decomposition. But our paper differs to other works above in the following aspects.

A. In stead of convex combinations or linear combinations of functions, we choose in particular Reproducing Kernel Hilbert Spaces (RKHSs), which are simple but general enough to include all finite dimensional subspaces of continuous functions. In these specific spaces, we may obtain upper bounds with faster rates (optimal in finite dimensional subspaces), than [Zhang and Yu 2003] and [Blanchard, Lugosi, and Vayatis 2003] which study general scaled convex hulls of functions.

B. To benefit the linear structure in this paper, instead of using the VC-dimension or Rademacher complexity, we directly exploit the martingale inequalities for random operators and vectors in a Hilbert space. The idea that norm convergence of operators leading to uniform convergence of sequences, is in fact not new in literature, e.g. [Yosida and Kakutani 1941] or see the comments in [Peskir 2000].

Moreover, we also investigate the influence of restricting step sizes (or learning rate) on convergence rates, by imposing a polynomial decreasing rate on step sizes. It is interesting to notice that: when the step sizes decrease faster, on one hand, the gap between the paths (sample error or variance) grows slower, as shown in Theorem 2.8; on the other hand, the population iteration converges slower too, as shown in Theorem 2.9. The final bias-variance trade-off, as shown in the Main Theorem, turns out that under the given stopping rule, the decreasing rate of step sizes might not affect the convergence rate of $f_t^{\mathbf{z}}$, but just the stopping time.

3.3. **Perspectives on Landweber Iteration.** In this subsection we show that there are some close relationship between the algorithms in this paper and the Landweber iteration for linear inverse problems [Engl, Hanke, and Neubauer 2000]. Below one can see the population iteration (3) can be regarded as the Landweber iteration for a specific linear operator equation and the sample iteration (2) is a Monte Carlo approximation of that.

We start by rephrasing the learning problem as a suitable linear inverse problem, and then discuss the difference between the two fields. For a broader discuss on this perspect, see [De Vito, Rosasco, Caponnetto, Giovannini, and Odone 2004]. First note that the minimization of (1) over $\mathscr{H}_K$ can be written equivalently as

$$\inf_{f\in\mathscr{H}_K} \|f - f_\rho\|_\rho. \tag{6}$$

Let $P_K : \mathscr{L}^2_{\rho_X} \to \overline{\mathscr{H}}_K$ be the projection from $\mathscr{L}^2_{\rho_X}$ onto the closure of $\mathscr{H}_K$ in $\mathscr{L}^2_{\rho_X}$. Note that $\mathscr{H}_K$ is closed if it is of finite dimension. With the aid of P_K, we have

$$\|P_K f_\rho - f_\rho\|_\rho = \inf_{f\in\mathscr{H}_K} \|f - f_\rho\|_\rho.$$

The population iteration converges in $\mathscr{H}_K$ to $P_K f_\rho$, which however under the condition that $f_\rho \in L^r_K(B_R)$, coincides with f_ρ exactly.

In the perspective of linear inverse problem, we may consider the following linear operator equation

$$I_K f = f_\rho \tag{7}$$

where the linear map $I_K : \mathscr{H}_K \hookrightarrow \mathscr{L}^2_{\rho_X}$ is an embedding, i.e. a continuous (bounded) inclusion. I_K is compact in the setting of this paper (i.e. $X \subseteq \mathbb{R}^n$ is closed and K is a bounded Mercer's kernel). A least square solution of (7) satisfies the following normal equation

$$I_K^* I_K f = I_K^* f_\rho \tag{8}$$

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 $I_K^* I_K = L_K|_{\mathscr{H}_K} : \mathscr{H}_K \to \mathscr{H}_K$, which, abusing the notation, is also denoted by L_K. Then the normal equation (8) is simply

$$L_K f = L_K f_\rho. \tag{9}$$

In this way on 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 (7).

Now we develop a discrete version of the equations above. Given a finite sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in X^m \times Y^m$, consider to find a function $f \in \mathscr{H}_K$ such that $f(x_i) = y_i$ $(i = 1, \dots, m)$. Let $S_\mathbf{x} : \mathscr{H}_K \to \mathbb{R}^m$ be the sampling operator such that $S_\mathbf{x}(f) = (f(x_i))_{i=1}^m$ (see Appendix B for detail). Consider the linear operator equation

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

whose normal equation becomes

$$S_\mathbf{x}^* S_\mathbf{x} f = S_\mathbf{x}^* \mathbf{y}, \tag{11}$$

where $S_\mathbf{x}^* : \mathbb{R}^m \to \mathscr{H}_K$ is the conjugate of $S_\mathbf{x}$, defined by $S_\mathbf{x}^*(\mathbf{y}) = \frac{1}{m}\sum_{i=1}^m y_i K_{x_i}$ (see Appendix B for detail). In this way the sample iteration (2) written via the sampling operator as,

$$f^\mathbf{z}_{t+1} = f^\mathbf{z}_t - \gamma_t (S_\mathbf{x}^* S_\mathbf{x} f^\mathbf{z}_t - S_\mathbf{x}^* \mathbf{y}), \tag{12}$$

is the Landweber iteration to solve (10). Clearly the sample iteration converges to an overfitting solution f such that $f(x_i) = y_i$ exactly for all $i = 1, \dots, m$.

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

A. Since (x_i, y_i) $(i = 1, \ldots, m)$ are randomly sampled from probability measure ρ on Z, we have Monte Carlo approximations $S_{\mathbf{x}}^* S_{\mathbf{x}} \sim I_K^* I_K$ and $S_{\mathbf{x}}^* \mathbf{y} \sim I_K^* f_\rho$. In classical inverse problems, the sample $\mathbf{x}$ is fixed (i.e. *fixed design*) and only the output $\mathbf{y}$ are noise-perturbed.

B. Moreover, there are some difference on convergence considered in the two settings. To this aim it is important to focus on the existence of a solution for equation (7). Note that equation (7) has a least square solution if and only if $P_K f_\rho \in \mathscr{H}_K$ (or equivalently $f_\rho \in \mathscr{H}_K \bigoplus \overline{\mathscr{H}_K}^\perp$), where we can define $f_\rho^\dagger = P_K f_\rho$ to be the generalized solution [Engl, Hanke, and Neubauer 2000], i.e. the unique minimal norm least square solution of (7) in $\mathscr{H}_K$. In this case inverse problem typically studies the convergence

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

under the assumption $f_\rho^\dagger = (I_K^* I_K)^r g$ for some $\|g\|_K \leq R$ [Engl, Hanke, and Neubauer 2000].

If $P_K f_\rho \notin \mathscr{H}_K$ (or $f_\rho \notin \mathscr{H}_K \bigoplus \overline{\mathscr{H}_K}^\perp$), then $f_\rho^\dagger$ does not exists, which is however often met in learning theory. A typical example is that K is the Gaussian kernel ($\mathscr{H}_K$ is thus dense in $\mathscr{L}^2_{\rho_X}$), but $f_\rho \notin \mathscr{H}_K$. In this case, if measured by $\|\cdot\|_K$ norm, f_t diverges eventually. But we still have

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

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

3.4. **Perspectives on On-line Learning.** The on-line learning algorithms in [Smale and Yao 2005] are stochastic approximations of the gradient descent method for the following least square problem with Tikhonov regularization,

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

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

$$f_t = f_{t-1} - \gamma_t [(f_{t-1}(x_t) - y_t) K_{x_t} + \lambda f_{t-1}], \qquad \text{for some } f_0 \in \mathscr{H}_K \tag{13}$$

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

It is interesting to compare the step sizes (or learning rate) in this paper and the on-line learning algorithm (13). For convergence of (13), one need shrinking step sizes $\gamma_t \to 0$, but the shrinkage can't go too fast: in fact $\sum_t \gamma_t = \infty$ is necessary to "forget" the initial error [Smale and Yao 2005] or [Yao 2005]. However, one can see from the upper bounds in the Main Theorem that, all the step sizes in the family can achieve a common convergence rate $O(t^{-r/(2r+2)})$ and the constant step size is even favored since it leeds to the minimal number of iterations before stopping.

Some closer connections can be seen from the decomposition in Proposition 4.3 in the next section.

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. Theorem 2.8 and 2.9. 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 estimate 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 $t-k+1$,

$$\pi_k^t(x) = \begin{cases} \prod_{i=k}^{t} (1-\gamma_i x), & k \le t; \\ 1, & k > t. \end{cases} \tag{14}$$

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

$$\sum_{k=\tau}^{t-1} \gamma_k x \pi_{k+1}^{t-1}(x) = \sum_{k=\tau}^{t-1} (1-(1-\gamma_k x))\pi_{k+1}^{t-1}(x) = \sum_{k=\tau}^{t-1} (\pi_{k+1}^{t-1}(x) - \pi_k^{t-1}(x)) = 1 - \pi_\tau^{t-1}(x). \tag{15}$$

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

$$g_t(x) = \sum_{k=0}^{t-1} \gamma_k \pi_{k+1}^{t-1}(x); \tag{16}$$

and the *residue polynomial*

$$r_t(x) = 1 - x g_t(x) = \pi_0^{t-1}(x). \tag{17}$$

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

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) $\|g_t(T)\| \le \sum_0^{t-1} \gamma_k$;

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) $|g_t(x)| \le \sum_0^{t-1} \gamma_k$;
(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 (12). □

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

Remark 4.4. This result is similar to the remainder decomposition in [Yao 2005], where χ_k is a martingale difference sequence; however here we lose this martingale property since both $f_k^{\mathbf{z}}$ and $S_{\mathbf{x}}$ are random variables dependent on $\mathbf{x}$.

Proof. We use a new representation of $f_t^{\mathbf{z}}$ other than Proposition 4.2-2,

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

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

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

Subtracting on both sides Proposition 4.2-1, we obtain the result. □

Some useful upper bounds are collected in the following proposition.

Proposition 4.5. *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 repeated use Corollary 4.1 and the isometry property $L_K^{1/2} : \mathscr{L}^2_{\rho_X} \to \mathscr{H}_K$, i.e. equation (B-1).

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

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

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 $G_{ij} = \frac{1}{m} K(x_i, x_j)$. 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 on remainder decompositions 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) $\|f_t^{\mathbf{z}} - f_t\|_K \le \left(\sum_{k=0}^{t-1} \gamma_k \|\pi_{k+1}^t(L_K)\| \right) \sup_{1 \le k \le t-1} \|\chi_k\|_K \le (\sum_{k=0}^{t-1} \gamma_k) \sup_{1 \le k \le t-1} \|\chi_k\|_K.$

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,

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

The result follows by taking the square root. □

5. Bounds for Sample Error and Approximation Error

In this section, we present the proofs of Theorem 2.8 and 2.9.

5.1. **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 $\chi_t = (L_K - S_{\mathbf{x}}^* S_{\mathbf{x}}) f_t^{\mathbf{z}} + S_{\mathbf{x}}^* \mathbf{y} - L_K f_\rho$ using the concentration results in Appendix B.

Proposition 5.1. *With probability at least $1-\delta$ ($\delta \in (0,1)$) there holds for all $t \in \mathbb{N}$,*

$$\sup_{1\le k\le t-1} \|\chi_k\|_K \le \frac{4\kappa M}{\sqrt{1-\theta}} \log^{1/2} \frac{2}{\delta} \sqrt{\frac{t^{1-\theta}}{m}}.$$

Proof. Note that

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

By the upper bound in Lemma A.3 and the concentration results in Appendix B, we have

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

and

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

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 2.8.

Proof of Theorem 2.8. Using Proposition 4.5-5 and Proposition 5.1,

$$\begin{aligned} \|f_t^{\mathbf{z}} - f_t\|_\rho &\le \sqrt{\sum_{k=0}^{t-1} \gamma_k} \sup_{1\le k\le t-1} \|\chi_k\|_K \le \frac{1}{\kappa\sqrt{1-\theta}} t^{\frac{1-\theta}{2}} \frac{4\kappa M}{\sqrt{1-\theta}} \log^{1/2} \frac{2}{\delta} \sqrt{\frac{t^{1-\theta}}{m}} \\ &\le \frac{4M}{1-\theta} \log^{1/2} \frac{2}{\delta} \cdot \frac{t^{1-\theta}}{\sqrt{m}} \end{aligned}$$

which gives the first bound.

Using Proposition 4.5-4, the upper bound in Lemma A.3 and Proposition 5.1,

$$\begin{aligned} \|f_t^{\mathbf{z}} - f_t\|_K &\le \sum_{k=0}^{t-1} \gamma_k \sup_{1\le k\le t-1} \|\chi_k\|_K \le \frac{1}{\kappa^2(1-\theta)} t^{1-\theta} \frac{4\kappa M}{\sqrt{1-\theta}} \log^{1/2} \frac{2}{\delta} \sqrt{\frac{t^{1-\theta}}{m}} \\ &\le \frac{4M}{\kappa(1-\theta)^{3/2}} \log^{1/2} \frac{2}{\delta} \cdot \frac{t^{\frac{3}{2}(1-\theta)}}{\sqrt{m}} \end{aligned}$$

which gives the second bound. □

5.2. **A Deterministic Upper Bound for Approximation Error.** The following is the proof of Theorem 2.9 using similar technique in [Engl, Hanke, and Neubauer 2000].

Proof of Theorem 2.9. Let $f_\rho = L_K^r g$ with $\|g\|_\rho \le R$. 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

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

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

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

But the function

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

is maximized at $\lambda^* = r/(\sum_i\gamma_i)$ with $f(\lambda^*) = -r + r\log r - r\log\sum_i\gamma_i$. Taking $\gamma_t = (t+1)^{-\theta}/\kappa^2$, by the lower bound in Lemma A.3 we obtain

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

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

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

Replacing r by $r-1/2$ 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 2.5. The formal proof is presented in 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 $Y=\{\pm1\}$. Define the *misclassification set*

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

For classification problems, the following error measure is proposed in [Smale and Zhou 2005]

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

which is equivalent to the probability of misclassification by f,

$$\|\text{sign}f - \text{sign}f_\rho\|_\rho^2 = 4\rho_X(X_f). \tag{18}$$

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

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

which is minimized at the so called *Bayes rule*, $\text{sign}f_\rho$. It is easy to check that

$$R(f) - R(f_\rho) = \int_{X_f} |f_\rho(x)|d\rho_X(x). \tag{19}$$

6.1. **Tsybakov's Noise Condition.** We adopts the following approach to assess the regularity of the marginal probability measure ρ_X in a classification problem.

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

$$T_\rho(s) = \rho_X(\{x \in X \ : \ f_\rho(x) \in [-s, s]\}), \tag{20}$$

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

$$T_\rho(s) \le B_q s^q, \ \forall s \in [0,1], \tag{21}$$

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 $s \le \epsilon$ for some $\epsilon > 0$), indicating a nonzero hard margin, we have $q = \infty$.

The following equivalent condition is useful (see Tsybakov 2004 or Bousquet, Boucheron, and Lugosi 2004].

Lemma 6.1. *Tsybakov's condition (21) is equivalent*[2] *to that for all* $f \in \mathscr{L}^2_{\rho_X}$,

$$\rho_X(X_f) \le c_\alpha (R(f) - R(f_\rho))^\alpha, \tag{22}$$

where

$$\alpha = \frac{q}{q+1} \in [0,1] \tag{23}$$

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

Proof. (21) $\Rightarrow$ (22). Recalling (19) we have the following chains of inequalities

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

The proof follows taking

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

and setting α as in (23).

(22) $\Rightarrow$ (21). Define for $s > 0$,

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

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

$$\rho_X(X_f) = \rho_X(X_s) \le c_\alpha (R(f) - R(f_\rho))^\alpha \le c_\alpha (\int_{X_s} |f_\rho(x)| d\rho_X)^\alpha \le c_\alpha t^\alpha \rho_X(X_s)^\alpha = c_\alpha t^\alpha \rho_X(X_f)^\alpha$$

whence $\rho_X(X_f) \le c_\alpha^{1/(1-\alpha)} t^{\alpha/(1-\alpha)}$ which recovers (21) with $q = \alpha/(1-\alpha)$ and $B_q = c_\alpha^{1/(1-\alpha)}$. $\square$

[2]The uniform condition, for all $f \in \mathscr{L}^2_{\rho_X}$, is crucial for the direction (22)$\Rightarrow$(21) 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 2.5, or Proposition 6.2-5, only requires the direction (21)$\Rightarrow$(22).

6.2. **Comparison Results and Proof of Theorem 2.5.** 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 (22) hold then $R(f) - R(f_\rho) \leq 4c_\alpha \|f - f_\rho\|_\rho^{2/(2-\alpha)}$

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

4) $\|\mathrm{sign} f - \mathrm{sign} f_\rho\|_\rho^2 \leq T(\|f - f_\rho\|_\infty)$

5) If (22) hold then $\|\mathrm{sign} f - \mathrm{sign} f_\rho\|_\rho \leq 4c_\alpha \|f - f_\rho\|_\rho^{\frac{\alpha}{2-\alpha}}$

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 $1/2$ 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

$$|f_\rho(x)| \leq |f(x) - f_\rho(x)| \tag{24}$$

when $x \in X_f$. In fact from (19)

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

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 (19) we get

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

where for $|f_\rho(x)| > t$, $|f_\rho(x)| = |f_\rho(x)|^2/|f_\rho(x) < \frac{1}{t}|f_\rho(x)|^2$. Then by conditions (22) and (24) we have

$$R(f) - R(f_\rho) \leq t\rho_X(X_f) + \frac{1}{t}\int_{X_f} |f_\rho(x)|_\rho^2 d\rho_X(x) \leq t c_\alpha (R(f) - R(f_\rho))^\alpha + \frac{1}{t}\|f - f_\rho\|_\rho^2.$$

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

3) From (19), simply using Schwartz Inequality we have

$$R(f) - R(f_\rho) = \frac{1}{2}\int_X f_\rho(x)(\mathrm{sign} f(x) - \mathrm{sign} f_\rho(x)) d\rho_X(x) \leq \frac{1}{2}\|f_\rho\|_\rho \|\mathrm{sign} f - \mathrm{sign} f_\rho\|_\rho$$

4) See Proposition 2 in [Smale and Zhou 2005].

5) The proof follows from (18) by plugging in (22) and part 2). □

Now we are ready to give the proof of Theorem 2.5.

Proof of Theorem 2.5. It's a direct application of the Main Theorem with Proposition 6.2-5. □

Acknowledgments

The authors would like to acknowledge Ernesto De Vito for sharing us some results on generalized Mercer's Theorem; Ha Quang Minh for sharing 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 has been done while the authors visited Toyota Technical Institute at Chicago. 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.

Appendix A

Proposition A.1. *The gradient of* $V = \frac{1}{2}(f(x)-y)^2$ *is the map*

$$\begin{aligned} \mathrm{grad}V : \mathscr{H}_K &\to \mathscr{H}_K \\ f &\mapsto (f(x)-y)K_x. \end{aligned}$$

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

$$\lim_{\|g\|_K \to 0} \frac{|V(f+g) - V(f) - DV(f)(g)|}{\|g\|_K} = 0.$$

Note that

$$\begin{aligned} V(f+g) - V(f) &= \frac{1}{2}\left\{(f(x)+g(x)-y)^2 - (f(x)-y)^2\right\} \\ &= g(x)(f(x)-y) + \frac{1}{2}g(x)^2 \\ &= \langle g, (f(x)-y)K_x\rangle_K + \frac{1}{2}\langle g, K_x\rangle_K^2 \\ &\le \langle g, (f(x)-y)K_x\rangle_K + \frac{1}{2}\kappa^2\|g\|_K^2, \end{aligned}$$

whence

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

Recall the definition of $\mathrm{grad}V : \mathscr{H}_K \to \mathscr{H}_K$ as for all $g \in \mathscr{H}_K$,

$$\langle \mathrm{grad}V(f), g\rangle_K = DV(f)(g).$$

Thus we obtain the result. □

Lemma A.2. *For* $x \in [0,1]$ *and* $a, b > 0$,

$$x^a(1-x)^b \le \left(\frac{a}{a+b}\right)^a \left(\frac{b}{a+b}\right)^b.$$

Lemma A.3. *For all $t \in \mathbb{N}$, $\gamma_t = \kappa^{-2}(t+1)^{-\theta}$. Then for all $t \in \mathbb{N}$*

$$\frac{1}{2\kappa^2} t^{1-\theta} \leq \sum_{k=0}^{t-1} \gamma_k \leq \frac{1}{\kappa^2(1-\theta)} t^{1-\theta}.$$

Proof. 1) For the lower bound,

$$\begin{aligned} \sum_{k=0}^{t-1} \gamma_k &\geq \kappa^{-2} \int_1^t x^{-\theta} dx = \frac{1}{\kappa^2(1-\theta)} (1 - t^{\theta-1}) t^{1-\theta} \\ &\geq \frac{1}{\kappa^2(1-\theta)} (1 - 2^{\theta-1}) t^{1-\theta}, \qquad t \geq 2 \\ &\geq \frac{1}{2\kappa^2} t^{1-\theta}, \end{aligned}$$

where the last step is due to

$$\frac{1 - 2^{\theta-1}}{1-\theta} = \frac{2^{1-\theta} - 1}{2^{1-\theta}(1-\theta)} \geq \frac{(1-\theta)2^{-\theta}}{2^{1-\theta}(1-\theta)} = \frac{1}{2}.$$

It remains to check that for $t = 1$, $\gamma_0 = \kappa^{-2} \geq \kappa^{-2}/2$, i.e. the lower bound holds.

2) As to the upper bound, note that for all $t \in \mathbb{N}$

$$\begin{aligned} \sum_{k=0}^{t-1} \gamma_k &\leq \kappa^{-2} \left(1 + \int_1^t x^{-\theta} dx \right) \\ &= \kappa^{-2} \left(1 + \frac{1}{1-\theta} (t^{1-\theta} - 1) \right) \leq \frac{1}{\kappa^2(1-\theta)} t^{1-\theta}. \end{aligned}$$

□

Appendix B: RKHS, Random Operators and Measure-Concentration

In this appendix, we collect some facts on reproducing kernel Hilbert spaces, random Hilbert-schmidt operators and concentration inequalities in Hilbert space.

A function $K : X \times X \to \mathbb{R}$ is called a *Mercer kernel*, if it is a continuous, symmetric and *positive semi-definite* in the sense that $\sum_{i,j=1}^{l} c_i c_j K(x_i, x_j) \geq 0$ for any $l \in \mathbb{N}$ and any choice of $x_i \in X$ and $c_i \in \mathbb{R}$ $(i = 1, \ldots, l)$.

Given a Mercer kernel K, the associated *reproducing kernel Hilbert space* $\mathscr{H}_K$ can be constructed as follows. Let $K_t : X \to \mathbb{R}$ be a function defined by $K_t(x) = K(x, t)$. Define V_K as a vector space generated by $\{K_t : t \in X\}$, i.e. all the finite linear combinations of K_t. An inner product $\langle\ ,\ \rangle_K$ on V_K is defined as the unique linear extension of $\langle K_x, K_{x'} \rangle_K := K(x, x')$. With this inner product we have the *reproducing property*: for any $f \in V_K$, $f(x) = \langle f, K_x \rangle_K$ $(x \in X)$. The induced norm is defined by $\|f\|_K = \sqrt{\langle f, f \rangle_K}$ for each $f \in V_K$. Now define $\mathscr{H}_K$ to be the completion of this inner product space V_K. Examples of RKHS include Sobolev spaces [Wahba 1990], real analytic functions (band-limited functions) [Daubechies 1992] and their generalizations [Smale and Zhou 2004].

Let $\mathscr{C}(X)$ be the Banach space of real continuous function on X. Define a linear map $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{C}(X)$ by $(L_K f)(x') = \int K(x', x) f(x) d\rho_X$. Composition with the inclusion $\mathscr{C}(X) \hookrightarrow \mathscr{L}^2_{\rho_X}$ yields a linear operator $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}$, which abusing the notation, will be also denoted by L_K.

For closed $X \subseteq \mathbb{R}^n$ and the bounded Mercer kernel K ($\sup_{x \in X} K(x, x) < \infty$), $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}$ is a compact operator [Carmeli, DeVito, and Toigo 2005] and there exists an orthonormal eigensystem of $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}$, $(\lambda_i, \phi_i)_{i \in \mathbb{N}}$, such that

$$L_K f = \sum_i \lambda_i a_i \phi_i, \qquad \text{where } f = \textstyle\sum_i a_i \phi_i.$$

Moreover, given $r > 0$, define L^r_K by

$$L^r_K f = \sum_i \lambda^r_i a_i \phi_i, \qquad \text{where } f = \textstyle\sum_i a_i \phi_i.$$

For all $r > 0$, the image of L^r_K gives a scale of subspaces compactly embedded in $\mathscr{L}^2_{\rho_X}$. When $r = 1/2$, $L^{1/2}_K : \mathscr{L}^2_{\rho_X} \to \mathscr{H}_K$ is an isometry, i.e.

$$\langle L^{1/2}_K f, L^{1/2}_K g \rangle_K = \langle f, g \rangle_\rho, \qquad \text{for all } f, g \in \mathscr{L}^2_{\rho_X}; \tag{B-1}$$

and when $r \neq 1/2$, the image of L^r_K depends on ρ_X which is unknown.

Since the image of L_K lies in $\mathscr{H}_K$, then its restriction $L_K|_{\mathscr{H}_K}$ induces an operator $\bar{L}_K : \mathscr{H}_K \to \mathscr{H}_K$ such that $\bar{L}_K f = L_K f$ for $f \in \mathscr{H}_K$. Moreover in operator norms, $\|\bar{L}_K\| = \|L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}\|$. To see this by definition $\|\bar{L}_K\| := \sup_{\|f\|_K = 1} \|\bar{L}_K f\|_K / \|f\|_K$ where

$$\|\bar{L}_K f\|_K / \|f\|_K = \|L_K f\|_K / \|f\|_K = \|L^{1/2}_K f\|_\rho / \|L^{-1/2}_K f\|_\rho = \|L_K g\|_\rho / \|g\|_\rho, \quad g = L^{-1/2}_K f.$$

As $L^{-1/2}_K : \mathscr{H}_K \to \mathscr{L}^2_{\rho_X}$ is an isometry, so

$$\|\bar{L}_K\| = \sup_{\|f\|_K = 1} \|\bar{L}_K f\|_K / \|f\|_K = \sup_{\|g\|_\rho = 1} \|L_K g\|_\rho / \|g\|_\rho = \|L_K\| \leq \kappa^2.$$

Let $E_x : \mathscr{H}_K \to \mathbb{R}$ be the *evaluation functional* defined by $E_x(f) = f(x) = \langle f, K_x \rangle_K$, by the reproducing property. Let $E^*_x : \mathbb{R} \to \mathscr{H}_K$ be its adjoint such that $\langle E_x(f), y \rangle_\mathbb{R} = \langle f, E^*_x(y) \rangle_K$, whence $E^*_x(y) = y K_x$. They are bounded rank-one operators, $\|E_x\| = \|E^*_x\| \leq \kappa$. $E^*_x E_x : \mathscr{H}_K \to \mathscr{H}_K$ is a self-adjoint operator, with bound $\|E^*_x E_x\| \leq \kappa^2$.

With the aid of the reproducing property we can generalize the evaluation functional to the *sampling operators* on $\mathscr{H}_K$. Let $\mathbf{z} = \{(x_i, y_i) : i = 1, \ldots, m\}$ be a set of i.i.d. examples drawn from ρ. Define $\mathbf{x} = (x_i) \in X^m$ and $\mathbf{y} = (y_i) \in \mathbb{R}^m$. Let $(\mathbb{R}^m, \langle , \rangle_m)$ be an inner product space with $\langle u, v \rangle_m = \frac{1}{m} \sum_{i=1}^m u_i v_i$ for $u, v \in \mathbb{R}^m$. Define a *sampling operator* $S_\mathbf{x} : \mathscr{H}_K \to (\mathbb{R}^m, \langle , \rangle_m)$ by $S_\mathbf{x}(f) = (E_{x_i} f)_{i=1,\ldots,m} \in \mathbb{R}^m$. Let $S^*_\mathbf{x} : (\mathbb{R}^m, \langle , \rangle_m) \to \mathscr{H}_K$ be the adjoint of $S_\mathbf{x}$ such that $\langle S_\mathbf{x}(f), \mathbf{y} \rangle_m = \langle f, S^*_\mathbf{x} \mathbf{y} \rangle_K$. Thus $S^*_\mathbf{x}(\mathbf{y}) = \frac{1}{m} \sum_{i=1}^m y_i K_{x_i} = \frac{1}{m} \sum_{i=1}^m E^*_{x_i}(y_i)$. Both $S_\mathbf{x}$ and $S^*_\mathbf{x}$ are bounded random operators depending on $\mathbf{x}$, with bounds $\|S_\mathbf{x}\| = \|S^*_\mathbf{x}\| \leq \kappa$. Such sampling operators are used in a generalization of Shannon Sampling Theorem, [Smale and Zhou 2004].

Define a random operator $T_\mathbf{x} : \mathscr{H}_K \to \mathscr{H}_K$

$$T_\mathbf{x} = S^*_\mathbf{x} S_\mathbf{x} = \frac{1}{m} \sum_{i=1}^m E^*_{x_i} E_{x_i}. \tag{B-2}$$

Its expectation $\mathbb{E}[T_\mathbf{x}] = \bar{L}_K$.

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 contain 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. Note that $T_{\mathbf{x}}$ and $\bar{L}_K$ are Hilbert-Schmidt operators, thus we are going to bound the deviation $T_{\mathbf{x}} - \bar{L}_K$.

The following result is due to Iosif Pinelis [Pinelis 1992].

Lemma B.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 $t \in \mathbb{N}$,*

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

Proposition B.2. *L_K, $E_x^*E_x$ and $T_{\mathbf{x}}$ are Hilbert-Schmidt operators. Moreover,*

1) $\mathrm{tr}(L_K^2) \le \kappa^4$*;*

2) $\mathrm{tr}(E_x^*E_x) \le \kappa^2$*;*

3) $\mathrm{tr}(E_x^*E_xE_t^*E_t) \le \kappa^4$*;*

4) $\mathrm{tr}(T_{\mathbf{x}}) \le \kappa^2$*;*

5) $\mathrm{tr}(T_{\mathbf{x}}^2) \le \kappa^4$

Proof. 1) See Corollary 3 in Section 2, Chapter III, [Cucker and Smale 2002];

2) Since $E_x^*E_x$ is a rank one operator, then $\mathrm{tr}(E_x^TE_x) \le \|E_x^TE_x\| \le \kappa^2$;

3) Noting that $E_xE_t^* = K(x,t) \le \kappa^2$, whence $\mathrm{tr}(E_x^*E_xE_t^*E_t) = k(x,t)\mathrm{tr}(E_x^TE_x) \le \kappa^4$;

4) By $\mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B)$, $\mathrm{tr}(T_{\mathbf{x}}) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{tr}(E_{x_i}^TE_{x_i}) \le \kappa^2$;

5) Similar to 4, noting that

$$\mathrm{tr}(T_{\mathbf{x}}^2) = \frac{1}{m^2}\sum_{i,j=1}^{m}\mathrm{tr}(E_{x_i}^*E_{x_i}E_{x_j}^*E_{x_j}).$$

The result follows from part 3. □

Let $\xi_i = E_{x_i}^*E_{x_i} - \bar{L}_K$. Note that $\mathrm{tr}(L_K^2|_{\mathscr{H}_K}) \le \mathrm{tr}(L_K^2) \le \kappa^4$. Thus setting $c_i^2 = 2\kappa^4$, and $\epsilon = n\epsilon$, we obtain

Proposition B.3.

$$\mathbf{Prob}\left\{\left\|T_{\mathbf{x}} - \bar{L}_K\right\|_{HS} \ge \epsilon\right\} \le 2\exp\left\{-\frac{m\epsilon^2}{4\kappa^4}\right\}.$$

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

$$\|T_{\mathbf{x}} - \bar{L}_K\| \le \left\|T_{\mathbf{x}} - \bar{L}_K\right\|_{HS} \le \frac{2\kappa^2}{\sqrt{m}}\log^{1/2}\frac{2}{\delta}.$$

Note that $S^*_{\mathbf{x}}\mathbf{y} = \frac{1}{m}\sum_{i=1}^m y_i K_{x_i}$ is a random vector in $\mathscr{H}_K$ with expectation $\mathbb{E}[S^*_{\mathbf{x}}\mathbf{y}] = L_K f_\rho$. Moreover $\|S^*_{\mathbf{x}}\mathbf{y}\| \le \|S^*_{\mathbf{x}}\|\|\mathbf{y}\| \le \kappa M$ and $\|L_K f_\rho\| \le \kappa M$. Thus

Proposition B.4.

$$\mathbf{Prob}\left\{\|S^*_{\mathbf{x}}\mathbf{y} - L_K f_\rho\| \ge \epsilon\right\} \le 2\exp\left\{-\frac{m\epsilon^2}{4\kappa^2 M_\rho^2}\right\}.$$

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

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

This kind of concentration results was obtained by [De Vito, Rosasco, Caponnetto, Giovannini, and Odone 2004] in the context of inverse problem for learning.

References

Bartlett, P. L., M. J. Jordan, and J. D. McAuliffe (2003). Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, U.C. Berkeley.

Bickel, P. J., Y. Ritov, and A. Zakai (2005). Some Theory for Generalized Boosting Algorithms. preprint.

Blanchard, G., G. Lugosi, and N. Vayatis (2003). On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research* (4), 861–894.

Bousquet, O., S. Boucheron, and G. Lugosi (2004). Theory of classification: A survey of recent advances. *ESAIM Probability and Statistics*. to appear.

Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation 11*, 1493–1517.

Breiman, L. (2004). Population theory for boosting ensembles. *Annals of Statistics 32*, 1–11.

Bühlmann, P. and B. Yu (2002). Boosting with the l_2-loss: Regression and classification. *Journal of American Statistical Association 98*, 324–340.

Caponnetto, A. and E. D. Vito (2005). Fast rates for regularized least squares algorithm. *CBCL Paper #258/AI Memo #2005-13*.

Carmeli, C., E. DeVito, and A. Toigo (2005). Reproducing kernel hilbert spaces and mercer theorem. *preprint*.

Cucker, F. and S. Smale (2002). On the mathematical foundations of learning. *Bull. of the Amer. Math. Soc. 29*(1), 1–49.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

De Vito, E., L. Rosasco, A. Caponnetto, U. D. Giovannini, and F. Odone (2004). Learning from examples as an inverse problem. *Journal of Machine Learning Research*. preprint.

DeVore, R., G. Kerkyacharian, D. Picard, and V. Temlyakov (2004). Mathematical methods for supervised learning. IMI research reports 04:22, Department of Mathematics, University of South Carolina.

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

Engl, H. W., M. Hanke, and A. Neubauer (2000). *Regularization of Inverse Problems*. Kluwer Academic Publishers.

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

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

Freund, Y. and R. E. Schapire (1999). Adaptive game playing using multiplicative weights. *Games and Economic Behaviour 29*, 79–103.

Friedman, J., T. Hastie, and R. Tibshirani (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics 38*(2), 337–374.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*, 1189–1232.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York.

Hanke, M. (1995). *Conjugate Gradient Type Methods for Ill-posed Problems*. Pitman Research Notes in Mathematics Series. Longman Scientific & Technical.

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

Lugosi, G. and N. Vayatis (2004). On the bayes-risk consistency of regularized boosting methods. *Annals of Statistics 32*, 30–55.

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

Mathé, P. (2004). Saturation of regularization methods for linear ill-posed problems in Hilbert spaces. *SIAM Journal on Numerical Analysis 42*(3), 968–973.

Minh, H. Q. (2005). personal communications.

Ong, C. S. (2005). Kernels: Regularization and Optimization. PhD Thesis, Australian National University.

Peskir, G. (2000). *From Uniform Laws of Large Numbers to Uniform Ergodic Theorems*. Lecture Notes Series, no. 66, University of Aarhus, Dept. of Mathematics, Denmark.

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

Rudin, C., I. Daubechies, and R. E. Schapire (2004). The dynamics of adaboost: Cyclic behavior and convergence of margins. *Journal of Machine Learning Research*.

Schapire, R. (2001). Drifting games. *Machine Learning 43*(3), 265–291.

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

Schapire, R. E., Y. Freund, P. Bartlett, and W. S. Lee (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics 26*(5), 1651–1686.

Smale, S. and Y. Yao (2005). Online learning algorithms. *Foundation of Computational Mathematics*. preprint.

Smale, S. and D.-X. Zhou (2004). Shannon sampling and function reconstruction from point values. *Bull. of the Amer. Math. Soc. 41*(3), 279–305.

Smale, S. and D.-X. Zhou (2005). Learning theory estimates via integral operators and their approximations. *to appear*.

Steinwart, I. and C. Scovel (2005). Fast rates for support vector machines using gaussian kernels. *Annals of Statistics*. submitted.

Stoltz, G. and G. Lugosi (2004). Learning correlated equilibria in games with compact sets of strategies. preprint.

Temlyakov, V. N. (2004). Optimal estimators in learning theory. IMI research reports 04:23, Department of Mathematics, University of South Carolina.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics 32*, 135–166.

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

Wahba, G. (1987). Three topics in ill posed problems. In H. Engl and C. Groetsch (Eds.), *Proceedings of the Alpine-U.S. Seminar on Inverse and Ill Posed Problems*, pp. 385–408. A. Deepak Publishing.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, 59.

Wahba, G., D. R. Johnson, F. Gao, and J. Gong (1995). Adaptive tuning of numerical weather prediction models: Randomized gcv in three- and four-dimensional data assimilation. *Monthly Weather Review 123*, 3358–3369.

Yao, Y. (2005). On complexity issue of online learning algorithms. *IEEE Transactions on Information Theory*. submitted.

Yosida, K. and S. Kakutani (1941). Operator-theoretical treatment of markoff's process and mean ergodic theorem. *The Annals of Mathematics 42*(1), 188–288.

Zhang, T. and B. Yu (2003). Boosting with early stopping: convergence and consistency. Technical Report 635, Department of Statistics, University of California at Berkeley.

Zhao, P. and B. Yu (2004). Boosted lasso. Tech. Report #678, Statistics, UC Berkeley (December, 2004; revised and submitted to JRSS(B) in April, 2005).

Yuan Yao, Department of Mathematics, University of California, Berkeley, CA 94720.

Current address: Toyota Technological Institute at Chicago, 1427 East 60th Street, Chicago, IL 60637.

E-mail address: yao@math.berkeley.edu

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

Current address: Toyota Technological Institute at Chicago, 1427 East 60th Street, Chicago, IL 60637.

E-mail address: rosasco@disi.unige.it

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

Current address: Toyota Technological Institute at Chicago, 1427 East 60th Street, Chicago, IL 60637.

E-mail address: caponnet@mit.edu