Found. Comput. Math. 145–170 (2006)

DOI: 10.1007/s10208-004-0160-z

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Online Learning Algorithms

Steve Smale[1] and Yuan Yao[2]

[1]Toyota Technological Institute at Chicago
1427 East 60th Street
Chicago, IL 60637, USA
smale@math.berkeley.edu

[2]Department of Mathematics
University of California at Berkeley
Berkeley, CA 94720, USA
Current address
Toyota Technological Institute at Chicago
1427 East 60th Street
Chicago, IL 60637, USA
yao@math.berkeley.edu

Abstract. In this paper, we study an online learning algorithm in Reproducing Kernel Hilbert Spaces (RKHSs) and general Hilbert spaces. We present a general form of the stochastic gradient method to minimize a quadratic potential function by an independent identically distributed (i.i.d.) sample sequence, and show a probabilistic upper bound for its convergence.

1. Introduction

Consider learning from examples $(x_t, y_t) \in X \times \mathbb{R}$ $(t \in \mathbb{N})$, drawn at random from a probability measure ρ on $X \times \mathbb{R}$. For $\lambda > 0$, one wants to approximate the

Date received: October 26, 2004. Final version received: April 25, 2005. Date accepted: April 28, 2005. Communicated by Filipe Cucker. Online publication: September 23, 2005.
AMS classification: 62L20, 68Q32, 68T05.
Key words and phrases: Online learning, Stochastic approximation, Regularization, Reproducing Kernel Hilbert Spaces.

function f_λ^* minimizing over $f \in \mathscr{H}$ the quadratic functional

$$\int_{X\times Y} (f(x) - y)^2 \, d\rho + \lambda \| f \|_{\mathscr{H}}^2,$$

where $\mathscr{H}$ is some Hilbert space. In this paper a scheme for doing this is given by using one example at a time t to update to f_t the current hypothesis f_{t-1} which depends only on the previous examples. The scheme chosen here is based on the stochastic approximation of the gradient of the quadratic functional of f displayed above, and takes an especially simple form in the setting of a "Reproducing Kernel Hilbert Space." Such a stochastic approximation procedure was first proposed by Robbins and Monro [**18**], and also by Kiefer and Wolfowitz [**14**]. Its convergence and asymptotic rates have been widely studied (e.g., see [**19**], [**9**], [**1**], [**12**], [**21**], or [**13**] and references therein). For more background on stochastic algorithms see, for example, [**2**], [**10**], [**16**]. The main goal in our development of the algorithm is to give error estimates which characterize in probability the distance of our updated hypothesis to f_λ^* (and eventually the "regression function" of ρ). By choosing a quadratic functional to optimize, one is able to give a deeper understanding of this *online learning* phenomenon.

In contrast, in the more common setting for Learning Theory, the learner is presented with the whole set of examples in one batch. One may call this type of work *batch learning*.

The organization of this paper is as follows. Section 2 presents an online learning algorithm in Reproducing Kernel Hilbert Spaces (RKHSs), and states Theorem A for a probabilistic upper bound on initial error and sample error. Section 3 presents a general form of the stochastic gradient method in Hilbert spaces and Theorem B, together with a derivation of Theorem A from Theorem B. Section 4 gives the proof of Theorem B and the various bounds appearing in Section 3. Section 5 compares our results with the case of "batch learning." Section 6 discusses the Adaline or Widrow–Hoff algorithm and related works. Appendix A collects some estimates used throughout the paper, and Appendix B presents a generalized Bennett's inequality for independent sums in Hilbert spaces.

1.1 Notation. Let X be a closed subset of $\mathbb{R}^n$, $Y = \mathbb{R}$ and $Z = X \times Y$. Let ρ be a probability measure on Z, let ρ_X be the induced marginal probability measure on X, and let $\rho_{Y|x}$ be the conditional probability measure on Y with respect to $x \in X$. Define $f_\rho\colon\ X \to Y$ by

$$f_\rho(x) = \int_Y y \, d\rho_{Y|x},$$

the *regression function of* ρ. In other words, for each $x \in X$, $f_\rho(x)$ is the average of y with respect to $\rho_{Y|x}$. Let $\mathscr{L}^2_{\rho_X}(X)$ be the Hilbert space of square integrable functions with respect to ρ_X, and denoted by $\mathscr{L}^2_\rho(X)$ for simplicity. In the sequel, $\| \ \|_\rho$ denotes the norm in $\mathscr{L}^2_\rho(X)$ and $\| \ \|_\infty$ denotes the supreme norm with respect to ρ_X (i.e., $\| f \|_\infty = \operatorname{ess\,sup}_{\rho_X} |f(x)|$). We assume that $\| f_\rho \|_\infty < \infty$ and

$f_\rho \in \mathscr{L}^2_\rho(X)$. Finally, we use such a convention on the summation and product sign: as the index $m > n$, let $\sum_{i=m}^{n} x_i = 0$ and $\prod_{i=m}^{n} x_i = 1$, for any summand x_i.

2. An Online Learning Algorithm in RKHS

Let $K\colon X \times X \to \mathbb{R}$ be a *Mercer kernel*, i.e., a continuous symmetric real function which is *positive semidefinite* 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 $c_i \in \mathbb{R}$ $(i = 1, \ldots, l)$. Note $K(x, x) \geq 0$ for all x. In the following $\|\cdot\|$ and $\langle\ ,\ \rangle$ denote the Euclidean norm and the Euclidean inner product in $\mathbb{R}^n$, resp ectively. We give two typical examples of Mercer kernels. The first is the Gaussian kernel $K\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by $K(x, x') = \exp(-\|x - x'\|^2/c^2)$ $(c > 0)$. The second is the linear kernel $K\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by $K(x, x') = \langle x, x' \rangle + 1$. The restriction of these functions on $X \times X$ will induce the corresponding kernels on subsets of $\mathbb{R}^n$.

Let $\mathscr{H}_K$ be the Reproducing Kernel Hilbert Space (RKHS) associated with a Mercer kernel K. Recall the definition as follows. Consider the vector space V_K generated by $\{K_t\colon t \in X\}$, i.e., all the finite linear combinations of K_t, where, for each $t \in X$, the function $K_t\colon X \to \mathbb{R}$ is defined by $K_t(x) = K(x, t)$. An inner product $\langle\ ,\ \rangle_K$ on this vector space can be defined as the unique linear extension of $\langle K_x, K_{x'} \rangle_K := K(x, x')$ and its induced norm[1] is $\|f\|_K = \sqrt{\langle f, f \rangle_K}$ for each $f \in V_K$. Let $\mathscr{H}_K$ be the completion of this inner product space V_K / V_0. It follows that for any $f \in \mathscr{H}_K$, $f(x) = \langle f, K_x \rangle_K$ $(x \in X)$. This is often called the *reproducing property* in literature. Define a linear map $L_K\colon \mathscr{L}^2_\rho(X) \to \mathscr{H}_K$ by $L_K(f)(x) = \int_X K(x, t) f(t)\, d\rho_X$. The operator $L_K + \lambda I\colon \mathscr{H}_K \to \mathscr{H}_K$ is an isomorphism if $\lambda > 0$ (endomorphism if $\lambda \geq 0$), where $L_K\colon \mathscr{H}_K \to \mathscr{H}_K$ is the restriction of $L_K\colon \mathscr{L}^2_\rho(X) \to \mathscr{H}_K$.

Given a sequence of examples $z_t = (x_t, y_t) \in X \times Y$ $(t \in \mathbb{N})$, our online learning algorithm in RKHS is

$$f_{t+1} = f_t - \gamma_t((f_t(x_t) - y_t)K_{x_t} + \lambda f_t) \qquad \text{for some} \quad f_1 \in \mathscr{H}_K, \text{ e.g., } f_1 = 0, \tag{1}$$

where:

(1) for each $t \in \mathbb{N}$, (x_t, y_t) is drawn identically and independently according to ρ;
(2) the regularization parameter $\lambda \geq 0$; and
(3) the step size $\gamma_t > 0$.

[1] Note that the zero set $V_0 = \{f \in V_K\colon \|f\|_K = 0\} = \{0\}$, whence $\|\cdot\|_K$ is in fact a norm. To see this, by the reproducing property and Cauchy–Schwartz inequality,

$$\|f\|_K = 0 \quad \Rightarrow \quad |f(t)| = |\langle f, K_t \rangle_K| \leq \|f\|_K \|K_t\| = 0, \qquad \forall\, t \in X \quad \Rightarrow \quad f = 0,$$

which implies $V_0 = \{0\}$.

Note that for each f, the map $X \times Y \to \mathbb{R}$ given by $(x, y) \mapsto f(x) - y$ is a real-valued random variable and $K_x\colon X \to \mathscr{H}_K$ is a $\mathscr{H}_K$-valued random variable. Thus f_{t+1} is a random variable with values in $\mathscr{H}_K$ depending on $(z_i)_{i=1}^t$. Moreover, we see that $f_{t+1} \in \operatorname{span}\{f_1, K_{x_i}\colon 1 \le i \le t\}$, a finite-dimensional subspace of $\mathscr{H}_K$. The derivation of (1) is given in the next section from a stochastic gradient algorithm in general Hilbert spaces.

In the sequel we assume that

$$C_K := \sup_{x \in X} \sqrt{K(x, x)} < \infty. \tag{2}$$

For example, the following typical kernels have $C_K = 1$:

(1) Gaussian kernel: $K\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ such that $K(x, x') = e^{-\|x-x'\|^2/c^2}$.
(2) Homogeneous polynomial kernel: $K\colon R^n \times R^n \to \mathbb{R}$ such that $K(x, x') = \langle x, x' \rangle^d$. By the scaling property, we can restrict K to the sphere $S^{n-1} \times S^{n-1}$.
(3) Translation invariant kernels: Any $K\colon X \times X \to \mathbb{R}$ such that $K(x, x') = K(x - x')$ and $K(0) = 1$.

In the sequel, we decompose $\|f_t - f_\rho\|_\rho$ into several parts and give upper bounds for each of them. Before that, we introduce several important quantities.

First consider the minimization of the regularized least squares problem in $\mathscr{H}_K$,

$$\min_{f \in \mathscr{H}_K} \int_Z (f(x) - y)^2 \, d\rho + \lambda \|f\|_K^2, \qquad \lambda > 0.$$

The existence and uniqueness of a minimizer is guaranteed by Proposition 7 in Chapter III of Cucker and Smale [**7**] which exhibits it as

$$f_\lambda^* = (L_K + \lambda I)^{-1} L_K f_\rho, \tag{3}$$

where $f_\rho \in \mathscr{L}_\rho^2(X)$ is the regression function. In fact, f_λ^* defined in this way is also the equilibrium of the averaged update equation of (1),

$$\mathbb{E}[f_{t+1}] = \mathbb{E}[f_t] - \gamma_t(\mathbb{E}[(f_t(x_t) - y_t)K_{x_t} + \lambda f_t]). \tag{4}$$

In other words, f_λ^* satisfies

$$\mathbb{E}[(f_\lambda^*(x) - y)K_x + \lambda f_\lambda^*] = 0. \tag{5}$$

To see this, it is enough to notice that, by $L_K(f)(x) = \int_X K(x, t) f(t) \, d\rho_X$, we have

$$L_K(f_\lambda^*) = \mathbb{E}_x[f_\lambda^*(x) K_x],$$

and

$$L_K(f_\rho) = \mathbb{E}_x[[\mathbb{E}_{y|x} y] K_x],$$

whence equation (5) turns out to be $L_K(f_\lambda^*) + \lambda f_\lambda^* = L_K(f_\rho)$, which leads to the definition of f_λ^* in (3).

Notice that the map $(x, y) \mapsto (f_\lambda^*(x) - y)K_x + \lambda f_\lambda^*$ is a $\mathscr{H}_K$-valued random variable, with zero mean. Thus the following variance

$$\sigma^2 = \mathbb{E}_z[\|(f_\lambda^*(x) - y)K_x + \lambda f_\lambda^*\|_K^2], \tag{6}$$

characterizes the fluctuation about the equilibrium caused by the random sample $z = (x, y)$. If $\sigma^2 = 0$, we have the deterministic gradient method (see Section 3). If $M_\rho > 0$ is a constant, such that $\text{supp}(\rho) \subseteq X \times [-M_\rho, M_\rho]$, then Proposition 3.4 in the next section implies

$$\sigma^2 \leq \left(\frac{2C_K M_\rho(\lambda + C_K^2)}{\lambda}\right)^2.$$

The main purpose in this paper is to obtain a probabilistic upper bound for

$$\|f_t - f_\rho\|_\rho.$$

By the triangle inequality we may write

$$\|f_t - f_\rho\|_\rho \leq \|f_t - f_\lambda^*\|_\rho + \|f_\lambda^* - f_\rho\|_\rho. \tag{7}$$

The second part of the right-hand side in (7), $\|f_\lambda^* - f_\rho\|_\rho$, is called the *approximation error*. An upper bound for the approximation error will be given at the end of this section. In the following we will give a probabilistic upper bound on $\|f_t - f_\lambda^*\|_K$, whence via $\|f_t - f_\lambda^*\|_\rho \leq C_K\|f_t - f_\lambda^*\|_K$ we obtain an upper bound on the sample error. Before the statement of the theorem, we define

$$\alpha = \frac{\lambda}{\lambda + C_K^2}, \tag{8}$$

whose meaning, as the inverse condition number, will be discussed in the next section.

Theorem A. *Let* $\gamma_t = 1/(\lambda + C_K^2)t^\theta$ $(t \in \mathbb{N})$ *for some* $\theta \in \left(\frac{1}{2}, 1\right)$. *Then, for each* $t \in \mathbb{N}$, *we may write*

$$\|f_t - f_\lambda^*\|_K \leq \mathscr{E}_{\text{init}}(t) + \mathscr{E}_{\text{samp}}(t), \tag{9}$$

where

$$\mathscr{E}_{\text{init}}(t) \leq e^{[\alpha/(1-\theta)](1-t^{1-\theta})}\|f_1 - f_\lambda^*\|_K;$$

and with probability at least $1 - \delta$ $(\delta \in (0, 1))$ *in the space* Z^{t-1},

$$\mathscr{E}_{\text{samp}}^2(t) \leq \frac{C_\theta \sigma^2}{\delta(\lambda + C_K^2)^2}\left(\frac{1}{\alpha}\right)^{\theta/1-\theta}\left(\frac{1}{t}\right)^\theta.$$

Here σ^2 is the variance in (6) *and the positive constant C_θ satisfies*

$$C_\theta = 8 + \frac{2}{2\theta - 1}\left(\frac{\theta}{e(2-2^\theta)}\right)^{\theta/1-\theta}.$$

The proof will be deferred to later sections.

Remark 2.1. Assume $\lambda \le 1$ and consider the upper bound

$$\sigma^2 \le (2C_K M_\rho(\lambda + C_K^2)/\lambda)^2.$$

Then the following holds with probability at least $1 - \delta$ ($\delta \in (0, 1)$),

$$\|f_t - f_\lambda^*\|_K \le e^{C_1\lambda(1-t^{1-\theta})}\|f_1 - f_\lambda^*\|_K + \frac{C_2}{\sqrt{\delta}}\left(\frac{1}{\lambda}\right)^{(2-\theta)/2(1-\theta)}\left(\frac{1}{t}\right)^{\theta/2}, \quad (10)$$

where

$$C_1 = \frac{1}{(1-\theta)(1+C_K^2)} \qquad \text{and} \qquad C_2 = 2C_K M_\rho\sqrt{C_\theta}\left(1 + C_K^2\right)^{\theta/2(1-\theta)}.$$

Remark 2.2. In decomposition (9) in Theorem A, $\mathscr{E}_{\text{init}}(t)$ has a deterministic bound and characterizes the accumulated effect from the initial choice, which is called the *initial error*. $\mathscr{E}_{\text{samp}}(t)$ depends on the random sample and thus has a probabilistic bound, which is called the *sample error*. We can also give upper bounds on the *approximation error*, $\|f_\lambda^* - f_\rho\|_\rho$.

The approximation error can be bounded if we put some regularity assumptions on the regression function f_ρ. For example, the following result appears in Theorem 4 of Smale and Zhou [**20**].

Theorem 2.3.

(1) *Suppose $L_K^{-r} f_\rho \in L_\rho^2(X)$ for some $r \in (0, 1]$. Then*

$$\|f_\lambda^* - f_\rho\|_\rho \le \lambda^r \|L_K^{-r} f_\rho\|_\rho.$$

(2) *Suppose $L_K^{-r} f_\rho \in L_\rho^2(X)$ for some $r \in (\frac{1}{2}, 1]$. Then*

$$\|f_\lambda^* - f_\rho\|_K \le \lambda^{r-1/2} \|L_K^{-r} f_\rho\|_\rho.$$

Notice that since $L_K^{-1/2}$ is an isomorphism, $\mathscr{H}_K \to L_\rho^2(X)$, the second condition implies $f_\rho \in \mathscr{H}_K$.

3. A Stochastic Gradient Algorithm in Hilbert Spaces

In this section, we extend the setting in the first section to general Hilbert spaces. Let W be a Hilbert space with inner product $\langle\ ,\ \rangle$. Consider the quadratic potential map $V\colon W \to \mathbb{R}$ given by

$$V(w) = \tfrac{1}{2}\langle Aw, w\rangle + \langle B, w\rangle + C, \tag{11}$$

where $A\colon W \to W$ is a positive definite bounded linear operator whose inverse is bounded, i.e., $\|A^{-1}\| < \infty$, $B \in W$, and $C \in \mathbb{R}$. Then the gradient $\operatorname{grad} V\colon W \to W$ is given by

$$\operatorname{grad} V(w) = Aw + B,$$

V has a unique minimal point $w^* \in W$ such that $\operatorname{grad} V(w^*) = Aw^* + B = 0$, i.e.,

$$w^* = -A^{-1}B.$$

Our concern is to find an approximation of this point, when A, B, and C are random variables on a space Z. We give a sample complexity analysis (i.e., the sample size sufficient to achieve an approximate minimizer with high probability) of the so-called *stochastic gradient method* given by the update formula

$$w_{t+1} = w_t - \gamma_t \operatorname{grad} V(w_t), \qquad \text{for} \quad t = 1, 2, 3, \ldots, \tag{12}$$

with γ_t a positive step size. For each example z, the stochastic gradient of V_z, $\operatorname{grad} V_z\colon W \to W$, is given by the affine map $\operatorname{grad} V_z(w) = A(z)w + B(z)$, with $A(z)$, $B(z)$ denoting the values of random variables A, B at $z \in Z$. Our analysis will benefit from this affine structure and independent sampling. Thus (12) becomes:

For $t = 1, 2, 3, \ldots$, let z_t be a sample sequence and define an update by

$$w_{t+1} = w_t - \gamma_t(A_t w_t + B_t) \qquad \text{for some} \quad w_1 \in W, \tag{13}$$

where:

(1) $z_t \in Z$ ($t \in \mathbb{N}$) are drawn independently and identically according to ρ;
(2) the step size $\gamma_t > 0$; and
(3) the map $A\colon Z \to SL(W)$ is a random variable depending on z with values in $SL(W)$, the vector space of symmetric bounded linear operators on W, and $B\colon Z \to W$ is a W-valued random variable depending on z. For each $t \in \mathbb{N}$, let $A_t = A(z_t)$ and $B_t = B(z_t)$.

From the stochastic gradient method in equation (12), we derive equation (1) for our online algorithm in RKHSs. Consider the Hilbert space $W = \mathscr{H}_K$. For fixed $z = (x, y) \in Z$, take the following quadratic potential map $V\colon \mathscr{H}_K \to \mathbb{R}$ defined by

$$V_z(f) = \tfrac{1}{2}\{(f(x) - y)^2 + \lambda\|f\|_K^2\}.$$

Recall that the gradient of V_z is a map $\operatorname{grad} V_z$: $\mathscr{H}_K \to \mathscr{H}_K$ such that, for all $g \in \mathscr{H}_K$,

$$\langle \operatorname{grad} V_z(f), g\rangle_K = DV_z(f)(g),$$

where the Frèchet derivative at f, $DV_z(f)$: $\mathscr{H}_K \to \mathbb{R}$, is the linear functional such that, for $g \in \mathscr{H}_K$,

$$\lim_{\|g\|\to 0} \frac{|V_z(f+g) - V_z(f) - DV_z(f)(g)|}{\|g\|} = 0.$$

Hence

$$DV_z(f)(g) = (f(x) - y)g(x) + \lambda\langle f, g\rangle_K = \langle (f(x) - y)K_x + \lambda f, g\rangle_K,$$

where the last step is due to the reproducing property $g(x) = \langle g, K_x\rangle_K$. This gives the following proposition:

Proposition 3.1. $\operatorname{grad} V_z(f) = (f(x) - y)K_x + \lambda f$.

Taking $f = f_t$ and $(x, y) = (x_t, y_t)$, by $f_{t+1} = f_t - \gamma_t \operatorname{grad} V_{z_t}(f_t)$, we have

$$f_{t+1} = f_t - \gamma_t((f_t(x_t) - y_t)K_{x_t} + \lambda f_t),$$

which establishes equation (1).

In the sequel we assume that

Finiteness Condition.

(1) For almost all $z \in Z$, $\mu_{\min} I \le A(z) \le \mu_{\max} I$ $(0 < \mu_{\min} \le \mu_{\max} < \infty)$; and
(2) $\|B(z)\| \le \beta < \infty$ for almost all $z \in Z$.

Consider the following averaging of equation (13), by taking the expectation over the truncated history $(z_i)_{i=1}^t$,

$$\mathbb{E}_{z_1,\dots,z_t}[w_{t+1}] = \mathbb{E}_{z_1,\dots,z_{t-1}}[w_t] - \gamma_t(\mathbb{E}_{z_t}[A_t]w_t + \mathbb{E}_{z_t}[B_t]), \tag{14}$$

where w_t depends on the truncated sample up to time $t-1$, $(z_i)_{i=1}^{t-1}$. Then the equilibrium for this averaged equation (14) will satisfy

$$\mathbb{E}_{z_t}[A_t]w_t + \mathbb{E}_{z_t}[B_t] = 0 \quad \Longleftrightarrow \quad w_t = -\mathbb{E}_{z_t}[A_t]^{-1}\mathbb{E}_{z_t}[B_t]. \tag{15}$$

This motivates the following definitions:

Definition A.

(1) The equilibrium $w^* = -\hat{A}^{-1}\hat{B}$ where $\hat{A} = \mathbb{E}_z[A(z)]$ and $\hat{B} = \mathbb{E}_z[B(z)]$.

(2) The inverse condition number for the family $\{A(z)\colon z \in Z\}$, $\alpha = \mu_{\min}/\mu_{\max} \in (0, 1]$.

For each $w \in W$, the stochastic gradient at w as a map $\operatorname{grad} V_z(w)\colon Z \to W$ such that $z \mapsto A(z)w + B(z)$, is a W-valued random variable depending on z. In particular, $\operatorname{grad} V_z(w^*)$ has zero mean, with variance defined by

$$\sigma^2 = \mathbb{E}[\|\operatorname{grad} V_z(w^*)\|^2] = \mathbb{E}_z[\|A_z w^* + B_z\|^2],$$

which reflects the fluctuations of $\operatorname{grad} V_z(w^*)$ caused by the randomness of sample z. Observe that when $\sigma^2 = 0$, we have the following deterministic gradient algorithm to minimize V,

$$w_{t+1} = w_t - \gamma_t \operatorname{grad} V(w_t)$$

where $\operatorname{grad} V(w) = \hat{A}w + \hat{B}$.

Now we are ready to state the general version of the main theorem for Hilbert spaces. Here we consider the product probability measure on Z^{t-1}, which makes sense since z_i $(1 \le i \le t-1)$ are i.i.d. random variables. As in the first section, we will decompose and give a deterministic bound on $\mathscr{E}_{\text{init}}$ and a probabilistic bound on $\mathscr{E}_{\text{samp}}$, respectively.

Theorem B. *Assume* (13) *and the finiteness condition. Let* $\gamma_t = 1/\mu_{\max} t^\theta$ $(t \in \mathbb{N})$ *for some* $\theta \in (\frac{1}{2}, 1)$. *Then, for each* $t \in \mathbb{N}$, *we have*

$$\|w_t - w^*\| \le \mathscr{E}_{\text{init}}(t) + \mathscr{E}_{\text{samp}}(t) \tag{16}$$

where

$$\mathscr{E}_{\text{init}}(t) \le e^{[\alpha/(1-\theta)](1-t^{1-\theta})}\|w_1 - w^*\|,$$

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

$$\mathscr{E}^2_{\text{samp}}(t) \le \frac{\sigma^2}{{\mu_{\max}}^2 \delta}\psi_\theta(t, \alpha).$$

Here

$$\psi_\theta(t, \alpha) = \sum_{k=1}^{t-1} \frac{1}{k^{2\theta}} \prod_{i=k+1}^{t-1} \left(1 - \frac{\alpha}{i^\theta}\right)^2.$$

Remark 3.2. As in the first section, $\mathscr{E}_{\text{init}}(t)$ has a deterministic upper bound and characterizes the accumulated effect from the initial choice, which is called the *initial error*, and $\mathscr{E}_{\text{samp}}(t)$ depends on the random sample and thus has a probabilistic bound, which is called the *sample error*.

Remark 3.3. In summary, w_t in equation (13) satisfies that for arbitrary integer $t \in \mathbb{N}$, the following holds with probability at least $1-\delta$ in the space of all samples of length $t-1$, i.e., Z^{t-1},

$$\|w_t - w^*\| \;\le\; e^{[\alpha/(1-\theta)](1-t^{1-\theta})}\|w_1 - w^*\| + \frac{\sqrt{\sigma^2}}{\mu_{\max}\sqrt{\delta}}\,\psi_\theta^{1/2}(t,\alpha).$$

When $\sigma^2 = 0$, we have the following convergence rate for the deterministic gradient algorithm

$$\|w_t - w^*\| \le e^{[\alpha/(1-\theta)](1-t^{1-\theta})}\|w_1 - w^*\|,$$

which is faster than any polynomial rate.

Proposition 3.4. *Let $\alpha \in (0,1]$ and $\theta \in (\frac{1}{2},1)$. The following upper bounds hold for all $t \in \mathbb{N}$:*

(1) $\sigma^2 \le (2\beta/\alpha)^2$; *and*
(2) $\psi_\theta(t,\alpha) \le C_\theta\,(1/\alpha)^{\theta/(1-\theta)}\,(1/t)^\theta$, *where*

$$C_\theta = 8 + \frac{2}{2\theta-1}\left(\frac{\theta}{e(2-2^\theta)}\right)^{\theta/(1-\theta)}.$$

Remark 3.5. In the setting of equation (1) in RKHS, we have

$$\beta = C_K M_\rho \qquad \text{and} \qquad \alpha = \frac{\lambda}{\lambda + C_K^2},$$

whence

$$\sigma^2 \le \left(\frac{2C_K M_\rho(\lambda + C_K^2)}{\lambda}\right)^2.$$

Remark 3.6. Choose the initialization $w_1 = 0$ for simplicity. Notice that $\|w^*\| = \|\hat{A}^{-1}\hat{B}\| \le \beta/\mu_{\min}$. Then we have the following bound, with probability at least $1-\delta$,

$$\|w_t - w^*\| \le \frac{\beta}{\mu_{\min}}\left(\frac{1}{t}\right)^{\theta/2}\left(t^{\theta/2}e^{[\alpha/(1-\theta)](1-t^{1-\theta})} + 2\sqrt{\frac{C_\theta}{\delta}}\right).$$

Remark 3.7. Consider the case that $\theta = 1$ and $\alpha \in (0,\frac{1}{2})$. Then, by Lemma A.2(3), we obtain that

$$\mathscr{E}_{\text{init}}(t) \le t^{-\alpha}\|w_1 - w^*\|$$

and

$$\mathscr{E}_{\text{samp}}(t) \le \frac{\sqrt{\sigma^2}}{\mu_{\max}\sqrt{\delta}}\,\psi_1^{1/2}(t,\alpha) \le \frac{4\beta}{\mu_{\min}\sqrt{\delta(1-2\alpha)}}t^{-\alpha}.$$

Choosing $w_1 = 0$ and using $\|w^*\| \le \beta/\mu_{\min}$, we obtain that

$$\|w_t - w^*\| \le \frac{\beta}{\mu_{\min}} \left(\frac{1}{t}\right)^{\alpha} \left(1 + \frac{4}{\sqrt{\delta(1-2\alpha)}}\right).$$

The proof of Theorem B and Proposition 3.4 will be given in Section 4. Here is the proof of Theorem A from Theorem B.

Proof of Theorem A. In this case $W = \mathscr{H}_K$. Before applying Theorem B, we need to rewrite equation (1) by the notations used in Theorem B.

For any $f \in \mathscr{H}_K$, let the evaluation functional at $x \in X$ be $E_x\colon \mathscr{H}_K \to \mathbb{R}$ be such that $E_x(f) = f(x)\,(\forall\, x \in X)$. Denote by $E_x^*\colon \mathbb{R} \to \mathscr{H}_K$ the adjoint operator of E_x such that $\langle E_x(f), y\rangle_{\mathbb{R}} = \langle f, E_x^*(y)\rangle_K$ $(y \in \mathbb{R})$. From this definition, we see that $E_x^*(y) = yK_x$.

Define the linear operator $A_x\colon \mathscr{H}_K \to \mathscr{H}_K$ by $A_x = E_x^* E_x + \lambda I$, i.e., $A_x(f) = f(x)K_x + \lambda f$, whence A_x is a random variable depending on x. Taking the expectation of A_x, we have $\hat{A} = \mathbb{E}_x[A_x] = L_K + \lambda I$.

Moreover, define $B_z = E_x^*(-y) = -yK_x \in \mathscr{H}_K$, which is a random variable depending on $z = (x, y)$. Notice that the expectation of B_z, $\hat{B} = \mathbb{E}_z[B_z] = \mathbb{E}_x[\mathbb{E}_y[-y]K_x] = -L_K f_\rho$. For simplicity below, we denote $A_t = A_{x_t}$ and $B_t = B_{z_t}$.

With these notations, equation (1) can be rewritten as

$$f_{t+1} = f_t - \gamma_t(A_t f_t + B_t).$$

Clearly, $f_\lambda^* = (L_K + \lambda I)^{-1} L_K f_\rho$ satisfies $0 = \mathbb{E}_z[A(z) f_\lambda^* + B(z)] = \hat{A} f_\lambda^* + \hat{B}$. Thus f_λ^* is the equilibrium of the averaged equation (4).

Notice that the positive operator L_K satisfies $\|L_K\| = \sup_{x\in X} K(x, x) = C_K^2$. Therefore $\mu_{\max} = \lambda + C_K^2$, $\mu_{\min} = \lambda$, and $\beta = C_K M_\rho$.

Finally, by identifying $w_t = f_t$ and $w^* = f_\lambda^*$, the upper bound on the initial error $\mathscr{E}_{\text{init}}(t)$ follows from Theorem B, and the upper bound on the sample error $\mathscr{E}_{\text{samp}}(t)$ follows from Theorem B and Proposition 3.4(2). □

Remark 3.8. If $\theta = 1$ and $\lambda < C_K^2$ (whence $\alpha \in (0, \frac{1}{2})$), by Remark 3.7, we have

$$\|f_t - f_\lambda^*\|_K \le \left(\frac{1}{t}\right)^{\alpha} \left(\|f_\lambda^*\|_K + \frac{\sqrt{\sigma^2}}{\sqrt{\delta}(\lambda + C_K^2)} \psi_1^{1/2}(t, \alpha)\right).$$

By Lemma A.2(3), we have an upper bound for $\psi_1(t, \alpha)$,

$$\psi_1(t, \alpha) \le \frac{4}{1-2\alpha} t^{-2\alpha}.$$

With this upper bound and $\sigma^2 \le (2\beta/\alpha)^2 = 4C_K^2 M_\rho^2 (\lambda + C_K^2)^2/\lambda^2$, we obtain that

$$\|f_t - f_\lambda^*\|_K \le \left(\frac{1}{t}\right)^{\alpha} \left(\|f_\lambda^*\|_K + \frac{4C_K M_\rho}{\lambda\sqrt{\delta(1-2\alpha)}}\right),$$

which holds with probability at least $1-\delta$. Notice that this upper bound has a polynomial decay $O(t^{-\alpha})$.

4. Proof of Theorem B

In this section we shall use $\mathbb{E}_z[\cdot]$ to denote the expectation with respect to z. When the underlying random variable in expectation is clear from the context, we will simply write $\mathbb{E}[\cdot]$.

Define the remainder vector at time t, $r_t = w_t - w^*$, which is a random variable depending on $(z_i)_{i=1}^{t-1} \in Z^{t-1}$ when $t \geq 2$. The following lemma gives a formula to compute r_{t+1}.

Lemma 4.1.

$$r_{t+1} = \prod_{i=1}^{t}(I - \gamma_i A_i)r_1 - \sum_{k=1}^{t}\gamma_k \left(\prod_{i=k+1}^{t}(I-\gamma_i A_i)\right)(A_k w^* + B_k).$$

Proof. Since $w_{t+1} = w_t + \gamma_t(A_t w_t + B_t)$, then

$$\begin{aligned} r_{t+1} &= w_{t+1} - w^* \\ &= w_t - \gamma_t(A_t w_t + B_t) - (I - \gamma_t A_t)w^* - \gamma_t A_t w^* \\ &= (I - \gamma_t A_t)r_t - \gamma_t(A_t w^* + B_t). \end{aligned}$$

The result then follows from induction on $t \in \mathbb{N}$. □

For simplicity we introduce the following notations, a symmetric linear operator X_{k+1}^t: $W \to W$ which depends on $z_{k+1}, \ldots, z_t$,

$$X_{k+1}^t(z_{k+1}, \ldots, z_t) = \prod_{i=k+1}^{t}(I - \gamma_i A_i) \qquad (X_{k+1}^t = I \text{ if } k \geq t),$$

and a vector $Y_k \in W$ which depends on z_k only,

$$Y_k(z_k) = A_k w^* + B_k.$$

Clearly, $\mathbb{E}[Y_k] = 0$ and $\mathbb{E}[\|Y_k\|^2] = \sigma^2$ for every $1 \leq k \leq t$. With this notation Lemma 4.1 can be written as

$$r_{t+1} = X_1^t r_1 - \sum_{k=1}^{t}\gamma_k X_{k+1}^t Y_k, \tag{17}$$

where the first term $X_1^t r_1$ reflects the accumulated error caused by the initial choice; the second term $\sum_{k=1}^{t-1}\gamma_k X_{k+1}^t Y_k$ is of zero mean and reflects the fluctuation caused by the random sample. Based on this observation we define the *initial error*:

$$\mathscr{E}_{\text{init}}(t+1) = \|X_1^t r_1\|, \tag{18}$$

and the *sample error*:

$$\mathscr{E}_{\text{samp}}(t+1) = \left\| \sum_{k=1}^{t} \gamma_k X_{k+1}^t Y_k \right\|. \tag{19}$$

The main concern in this section is to obtain upper bounds on the initial error and the sample error. The following estimates are crucial in the proofs of Theorem B and Proposition 3.4.

Proposition 4.2. *Let* $\gamma_t = 1/\mu_{\max} t^\theta$ *for some* $\theta \in (\frac{1}{2}, 1]$. *For all* $\alpha = \mu_{\min}/\mu_{\max} \in (0, 1]$, *the following holds*:

(1) *Let* $\alpha' = \alpha/(1-\theta)$. *Then*

$$\|X_1^t r_1\| \le \begin{cases} e^{\alpha'(1-(t+1)^{1-\theta})}\|r_1\|, & \theta \in (\frac{1}{2}, 1); \\ (t+1)^\alpha \|r_1\|, & \theta = 1. \end{cases}$$

(2) $\|Y_k\| \le 2\beta/\alpha$,

(3) $\mathbb{E}\left[\left\| \sum_{k=1}^{t} \gamma_k X_{k+1}^t Y_k \right\|^2 \right] \le \dfrac{\sigma^2}{{\mu_{\max}}^2} \psi_\theta(t+1, \alpha)$.

From this proposition and the following Markov's inequality, we give the proof of Theorem B.

Lemma 4.3 (Markov's Inequality). *Let* X *be a nonnegative random variable. Then, for any real number* $\varepsilon > 0$, *we have*

$$\mathbf{Prob}\{X \ge \varepsilon\} \le \frac{\mathbb{E}[X]}{\varepsilon}.$$

Proof of Theorem B. By (18) and the estimation (1) in Proposition 4.2, where $\theta \in \left(\frac{1}{2}, 1\right)$, we have

$$\mathscr{E}_{\text{init}}(t) \le e^{[\alpha/(1-\theta)](1-t^{1-\theta})} \|w_1 - w^*\|.$$

By (19) and the estimation (3) in Proposition 4.2 and Markov's inequality with $X = \mathscr{E}^2_{\text{samp}}(t)$, we obtain, for $t \ge 2$,

$$\mathbf{Prob}\{\mathscr{E}^2_{\text{samp}}(t) \le \varepsilon^2\} \le \frac{\sigma^2}{\varepsilon^2 {\mu_{\max}}^2} \psi_\theta(t, \alpha).$$

Setting the right-hand side to be $\delta \in (0, 1)$, we get the probabilistic upper bound on the *sample error*. It remains to check that when $t = 1$, $\mathscr{E}_{\text{init}}(t) = \|w_1 - w^*\|$ and $\mathscr{E}_{\text{samp}}(t) = 0$, whence the bound still holds. □

Next we give the proof of Proposition 4.2.

Proof of Proposition 4.2. (1) By $\mu_{\min} I \le A \le \mu_{\max} I$ and $\gamma_t = 1/\mu_{\max} t^\theta$ ($\theta \in \left(\frac{1}{2}, 1\right]$), then

$$\begin{aligned} \|X_{k+1}^t r_1\| &\le \prod_{i=k+1}^{t} \|I - \gamma_i A_i\| \|r_1\| \\ &\le \prod_{i=k+1}^{t} \left(1 - \frac{\alpha}{i^\theta}\right) \|r_1\|, \qquad \alpha = \mu_{\min}/\mu_{\max}. \end{aligned} \tag{20}$$

Setting $k = 0$ and by (1) in Lemma A.2, we obtain the result.

(2) Note that $\|w^*\| \le \beta/\mu_{\min}$. Thus we have

$$\begin{aligned} \|Y_k\| &= \|A_k w^* + B_k\| \le \|A_k\| \|w^*\| + \|B_k\| \\ &\le \mu_{\max}\beta/\mu_{\min} + \beta = \beta(\alpha^{-1} + 1) \le 2\beta/\alpha, \end{aligned}$$

since $\alpha \in (0, 1]$. This gives part (2).

(3) Note that

$$\begin{aligned} \mathbb{E}\left[\left\|\sum_{k=1}^{t} \gamma_k X_{k+1}^t Y_k\right\|^2\right] &= \mathbb{E}\left\langle \sum_{k=1}^{t} \gamma_k X_{k+1}^t Y_k, \sum_{k=1}^{t} \gamma_k X_{k+1}^t Y_k \right\rangle, \\ &= \sum_{k,l=1}^{t} \gamma_k \gamma_l \mathbb{E}\langle X_{k+1}^t Y_k, X_{l+1}^t Y_l\rangle, \end{aligned}$$

where, if $k \ne l$, say $k < l$,

$$\gamma_k \gamma_l \mathbb{E}_{z_k,\ldots,z_t} \langle X_{k+1}^t Y_k, X_{l+1}^t Y_l\rangle = \gamma_k \gamma_l \mathbb{E}_{z_{k+1},\ldots,z_t} [\mathbb{E}_{z_k|z_{k+1},\ldots,z_t} [Y_k]^T X_{k+1}^t X_{l+1}^t Y_l] = 0,$$

by $\mathbb{E}[Y_k] = 0$. Thus we have

$$\begin{aligned} \sum_{k,l=1}^{t} \gamma_k \gamma_l \mathbb{E}\langle X_{k+1}^t Y_k, X_{l+1}^t Y_l\rangle &= \sum_{k=1}^{t} \gamma_k^2 \mathbb{E}\langle X_{k+1}^t Y_k, X_{k+1}^t Y_k\rangle \\ &\le \sum_{k=1}^{t} \gamma_k^2 \mathbb{E}[\|X_{k+1}^t\|^2 \|Y_k\|^2] \\ &\le \frac{\sigma^2}{{\mu_{\max}}^2} \psi_\theta(t+1, \alpha), \end{aligned}$$

where the last inequality is due to $\mathbb{E}\|Y_k\|^2 = \sigma^2$ for all k and

$$\sum_{k=1}^{t} \gamma_k^2 \|X_{k+1}^t\|^2 \le \sum_{k=1}^{t} \frac{1}{{\mu_{\max}}^2 k^{2\theta}} \prod_{i=k+1}^{t} \left(1 - \frac{\alpha}{i^\theta}\right)^2 = \frac{1}{{\mu_{\max}}^2} \psi_\theta(t+1, \alpha). \quad \square$$

Finally, we derive the upper bounds for σ^2 and $\psi(t, \alpha)$ as in Proposition 3.4.

Proof of Proposition 3.4. The first upper bound follows from estimation (2) in Proposition 4.2,

$$\sigma^2 \leq (\|Y_k\|)^2 \leq \left(\frac{2\beta}{\alpha}\right)^2$$

for all $1 \leq k \leq t$.

For $t \geq 2$, the second upper bound is an immediate result from Lemma A.1; for $t = 1$, note that $\psi_\theta(t, \alpha) = 0$ whence the upper bound still holds.

5. Comparison with "Batch Learning" Results

The name "batch learning" is coined for the purpose of emphasizing the case when the sample of size $t \in \mathbb{N}$ is exposed to the learner in one batch, instead of one-by-one as in "online learning" in this paper. In the context of RKHS, given a sample $\mathbf{z} = \{z_i\colon\ i = 1, \ldots, t\}$, "batch learning" means solving the *regularized least squares* problem [**11**], [**7**]

$$f_{\lambda,\mathbf{z}} = \arg\min_{f \in \mathscr{H}_K} \frac{1}{t} \sum_{i=1}^{t} (f(x_i) - y_i)^2 + \lambda \langle f, f \rangle_K, \qquad \lambda > 0.$$

The existence and uniqueness of $f_{\lambda,\mathbf{z}}$ given as in Section 6 of [**7**] says

$$f_{\lambda,\mathbf{z}}(x) = \sum_{i=1}^{t} a_i K(x, x_i)$$

where $a = (a_1, \ldots, a_t)$ is the unique solution of the well-posed linear system in $\mathbb{R}^t$,

$$(\lambda t I + K_{\mathbf{z}})a = y,$$

with $(t \times t)$-identity matrix I, $(t \times t)$-matrix $K_{\mathbf{z}}$ whose (i, j) entry is $K(x_i, x_j)$ and $y = (y_1, \ldots, y_t) \in \mathbb{R}^t$.

A probabilistic upper bound for $\|f_{\lambda,\mathbf{z}} - f_\lambda^*\|_\rho$ is given by Cucker and Smale [**6**], and this has been substantially improved by De Vito, Caponnetto, and Rosasco [**8**], using also some ideas from Bousquet and Elisseeff [**3**]. Moreover, error bounds expressed in a different form were given by Zhang [**23**]. A recent result, shown in Smale and Zhou [**20**], is

Theorem 5.1.

$$\|f_{\lambda,\mathbf{z}} - f_\lambda^*\|_K \leq \frac{C_{\rho,K}}{\sqrt{\delta}} \left(\frac{1}{\lambda\sqrt{t}}\right),$$

where $C_{\rho,K} = C_K^2\sqrt{\sigma_\rho^2} + 3C_K^2\|f_\rho\|_\rho$ *and*

$$\sigma_\rho^2 = \int_{X\times Y}(y - f_\rho(x))^2\,d\rho.$$

Remark 5.2. Notice that if $\lambda \le 1$ without loss of generality, equation (10) in Remark 2.1 shows the following convergence rate:

$$\|f_t - f_\lambda^*\|_K \le O\left(\left(\frac{1}{\lambda}\right)^{(2-\theta)/2(1-\theta)}\left(\frac{1}{t}\right)^{\theta/2}\right),$$

where $\theta \in \left(\frac{1}{2}, 1\right)$. Since the function $\tau(\theta) = (2-\theta)/2(1-\theta) = 1/2(1-\theta) + \frac{1}{2}$, is an increasing function of θ, then $\tau(\theta) \in (\frac{3}{4}, \infty)$ as $\theta \in (\frac{1}{2}, 1)$. For small λ, when θ is close to $\frac{1}{2}$, the upper bound is close to $O(\lambda^{-3/4}t^{-1/4})$ which is tighter in λ but looser in t in comparison with the theorem above; on the other hand, when θ increases, the upper bound becomes tighter in t but much looser in λ.

6. Adaline

Example 6.1 (Adaline or Widrow–Hoff Algorithm). The Adaline or Widrow–Hoff algorithm [**5**, p. 23] is a special case of the online learning algorithm (1) where the step size γ_t is a constant η, the regularization parameter $\lambda = 0$, and the reproducing kernel is the linear kernel such that $K(x, x') = \langle x, x'\rangle + 1$ for $x, x' \in X = \mathbb{R}^n$. To see that, define two kernels by $K_0(x, x') = \langle x, x'\rangle$ and $K_1(x, x') = 1$. Then $\mathscr{H}_K = \mathscr{H}_{K_0} \oplus \mathscr{H}_{K_1}$. Notice that $\mathscr{H}_{K_0} \backsimeq \mathbb{R}^n$ and $\mathscr{H}_{K_1} \backsimeq \mathbb{R}$, whence $\mathscr{H}_K \backsimeq \mathbb{R}^{n+1}$. In fact, for $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$, a function in $\mathscr{H}_K$ can be written as $f(x) = \sum_{i=1}^n w^i x^i + b$ for $x \in X$. By the use of the Euclidean inner product in $\mathbb{R}^{n+1}$, we can write $f(x) = \langle (w, b), (x, 1)\rangle$. Therefore, the Adaline update formula

$$(w_{t+1}, b_{t+1}) = (w_t, b_t) + \eta(\langle w, x_t\rangle + b - y_t)(x_t, 1), \qquad t \in \mathbb{N},$$

can be written as the following formula, by taking the Euclidean inner product of both sides with the vector $(x, 1) \in \mathbb{R}^{n+1}$,

$$f_{t+1} = f_t + \eta(f_t(x_t) - y_t)K_{x_t}.$$

This is equivalent to setting $\gamma_t = \eta$ and $\lambda = 0$ in the online learning algorithm (1).

The case for fixed step size and zero regularization parameter is not included in Theorems A or B. In the case of nonstochastic samples, Cesa-Bianchi et al. [**4**] have some worst-case analysis on the upper bounds for the following quantity:

$$\sum_{t=1}^T(\langle w_t, x_t\rangle - y_t)^2 - \min_{\|w\|\le W}\sum_{t=1}^T(\langle w, x_t\rangle - y_t)^2.$$

Adam Kalai has shown us how one might convert these results of Cesa-Bianchi et al. to a form comparable to Theorem A. Beyond the square loss function above, some related works include [**15**] which presents a general gradient descent method in RKHS for bounded differentiable functions, and [**24**] which studies the gradient method with arbitrary differentiable convex loss functions. These works suggest different schemes on choosing the step size parameter and how these choices might affect the convergence rate under various conditions.

Appendix A: Some Estimates

The following lemma gives an upper bound for

$$\psi_\theta(t,\alpha) = \sum_{k=1}^{t-1} \frac{1}{k^{2\theta}} \prod_{i=k+1}^{t-1} \left(1 - \frac{\alpha}{i^\theta}\right)^2.$$

Lemma A.1 (Main Analytic Estimate). *Let* $\alpha \in (0,1]$ *and* $\theta \in \left(\frac{1}{2}, 1\right)$. *Then for* $t \in \mathbb{N}$,

$$\psi_\theta(t+1,\alpha) \;\le\; C_\theta \left(\frac{1}{\alpha}\right)^{\theta/(1-\theta)} \left(\frac{1}{t+1}\right)^\theta,$$

where

$$C_\theta = 8 + \frac{2}{2\theta - 1} \left(\frac{\theta}{e(2-2^\theta)}\right)^{\theta/(1-\theta)}.$$

Proof. The following fact will be used repeatedly in this section,

$$\ln(1+x) \le x, \qquad \text{for all} \quad x > -1. \tag{A.1}$$

Thus we have

$$\sum_{i=k+1}^{t} \ln\left(1 - \frac{\alpha}{i^\theta}\right)^2 \;\le\; -2\alpha \sum_{i=k+1}^{t} \frac{1}{i^\theta} \le -2\alpha \int_{k+1}^{t+1} \frac{1}{x^\theta}\, dx,$$

which equals

$$\frac{2\alpha}{1-\theta}((k+1)^{1-\theta} - (t+1)^{1-\theta})$$

if $\theta \in \left(\frac{1}{2}, 1\right)$.

From this estimate it follows that

$$\psi_\theta(t+1,\alpha) \le e^{-2\alpha'(t+1)^{1-\theta}} \sum_{k=1}^{t} \frac{1}{k^{2\theta}} e^{2\alpha'(k+1)^{1-\theta}} = S_1 + S_2,$$

where $\alpha' = \alpha/(1-\theta)$ and

$$S_1 = e^{-2\alpha'(t+1)^{1-\theta}} \sum_{k=1}^{\lfloor (t-1)/2 \rfloor} \frac{1}{k^{2\theta}} e^{2\alpha'(k+1)^{1-\theta}},$$

$$S_2 = e^{-2\alpha'(t+1)^{1-\theta}} \sum_{k=\lfloor (t+1)/2 \rfloor}^{t} \frac{1}{k^{2\theta}} e^{2\alpha'(k+1)^{1-\theta}},$$

where $\lfloor x \rfloor$ denotes the largest integer no larger than x.

Next we give upper bounds on S_1 and S_2. First,

$$\begin{aligned} S_1 &\le e^{-2\alpha'(1-2^{\theta-1})(t+1)^{1-\theta}} \sum_{k=1}^{\lfloor (t-1)/2 \rfloor} \frac{1}{k^{2\theta}} \le e^{-2\alpha'(1-2^{\theta-1})(t+1)^{1-\theta}} \int_{1/2}^{t/2} \frac{1}{x^{2\theta}}\, dx \\ &= e^{-2\alpha'(1-2^{\theta-1})(t+1)^{1-\theta}} \frac{1}{1-2\theta} \left(\left(\frac{t}{2}\right)^{1-2\theta} - \left(\frac{1}{2}\right)^{1-2\theta} \right) \\ &\le \frac{2}{2\theta-1}\, e^{-2\alpha'(1-2^{\theta-1})(t+1)^{1-\theta}} \end{aligned}$$

as $\theta \in \left(\frac{1}{2}, 1\right)$. To give a polynomial upper bound for $\exp\{-2\alpha'(1-2^{\theta-1})(t+1)^{1-\theta}\}$, we use the fact that for any $c > 0$, $a > 0$, and $x \in (0, \infty)$,

$$e^{-cx} \le \left(\frac{a}{ec}\right)^a x^{-a}.$$

To see this, it is enough to observe that the function $f(x) = x^a/e^{cx}$ is maximized at $x = a/c$. Let $a = (1/\theta - 1)^{-1}$, $c = 2\alpha'(1-2^{\theta-1})$, and $x = (t+1)^{1-\theta} = (t+1)^{\theta(1/\theta-1)}$, then

$$e^{-2\alpha'(1-2^{\theta-1})(t+1)^{1-\theta}} \le \left(\frac{\theta}{e\alpha(2-2^{\theta})}\right)^{\theta/(1-\theta)} (t+1)^{-\theta},$$

Thus, for $\theta \in \left(\frac{1}{2}, 1\right)$, $\alpha \in (0,1)$, and $t \in \mathbb{N}$,

$$S_1 \le \frac{2}{2\theta-1} \left(\frac{\theta}{e\alpha(2-2^{\theta})}\right)^{\theta/(1-\theta)} (t+1)^{-\theta}.$$

Second, notice that $1/\lfloor (t+1)/2 \rfloor \le 2/t \le 4/(t+1)$ (for $t \in \mathbb{N}$), then let $p(t) = e^{2\alpha'(t+1)^{1-\theta}}/t^{\theta}$ and we have

$$\begin{aligned} S_2 &\le e^{-2\alpha'(t+1)^{1-\theta}} \frac{4^{\theta}}{(t+1)^{\theta}} \left(p(t) + \sum_{k=\lfloor \frac{t+1}{2} \rfloor}^{t-1} \frac{1}{k^{\theta}} e^{2\alpha'(k+1)^{1-\theta}} \right) \\ &\le 2^{2\theta} e^{-2\alpha'(t+1)^{1-\theta}} (t+1)^{-\theta} \left(p(t) + \int_{t/2-1}^{t} \frac{1}{x^{\theta}} e^{2\alpha'(x+1)^{1-\theta}} dx \right) \quad \text{for} \quad t \ge 4, \end{aligned}$$

where

$$\begin{aligned}
&\int_{t/2-1}^{t} \frac{1}{x^\theta} e^{2\alpha'(x+1)^{1-\theta}} dx \\
&\quad \le \int_{t/2-1}^{t} \frac{2^\theta}{(x+1)^\theta} e^{2\alpha'(x+1)^{1-\theta}}\, dx \quad \text{by } \frac{1}{x} \le \frac{2}{x+1} \text{ for } t \ge 4 \\
&\quad = \frac{2^\theta}{1-\theta} \int_{(t/2)^{1-\theta}}^{(t+1)^{1-\theta}} e^{2\alpha' y}\, dy \qquad \text{by } y = (x+1)^{1-\theta} \\
&\quad = \frac{2^{\theta-1}}{\alpha'(1-\theta)} e^{2\alpha'(t+1)^{1-\theta}} (1 - e^{2\alpha'((t/2)^{1-\theta} - (t+1)^{1-\theta})}) \le \frac{1}{\alpha} e^{2\alpha'(t+1)^{1-\theta}},
\end{aligned}$$

whence

$$S_2 \le 2^{2\theta}(t+1)^{-\theta} \left(t^{-\theta} + \frac{1}{\alpha} \right) \le \frac{8}{\alpha}(t+1)^{-\theta} \qquad \text{for} \quad t \ge 4.$$

It is easy to check that $\psi(t+1, \alpha) \le 2 \le (8/\alpha)(t+1)^{-\theta}$ for $1 \le t \le 3$.

Therefore, for $t \in \mathbb{N}$,

$$\begin{aligned}
\psi_\theta(t+1, \alpha) &\le S_1 + S_2 \le \left(\frac{2}{2\theta - 1} \left(\frac{\theta}{e\alpha(2 - 2^\theta)} \right)^{\theta/(1-\theta)} + \frac{8}{\alpha} \right) (t+1)^{-\theta} \\
&= \left(\frac{2}{2\theta - 1} \left(\frac{\theta}{e(2 - 2^\theta)} \right)^{\theta/(1-\theta)} + 8\alpha^{(2\theta-1)/(1-\theta)} \right) \\
&\quad \times \left(\frac{1}{\alpha} \right)^{\theta/(1-\theta)} (t+1)^{-\theta} \\
&\le \left(\frac{2}{2\theta - 1} \left(\frac{\theta}{e(2 - 2^\theta)} \right)^{\theta/(1-\theta)} + 8 \right) \left(\frac{1}{\alpha} \right)^{\theta/(1-\theta)} (t+1)^{-\theta},
\end{aligned}$$

where the last step is due to $\alpha^{(2\theta-1)/(1-\theta)} < 1$ as $\alpha \in (0, 1)$. □

The following lemma is also useful in the various upper bound estimations in Proposition 4.2.

Lemma A.2. (1) *For* $\alpha \in (0, 1]$ *and* $\theta \in [0, 1]$,

$$\prod_{i=k+1}^{t} \left(1 - \frac{\alpha}{i^\theta} \right) \le \begin{cases} \exp\left(\dfrac{2\alpha}{1-\theta} ((k+1)^{1-\theta} - (t+1)^{1-\theta}) \right), & \theta \in [0, 1), \\ \left(\dfrac{k+1}{t+1} \right)^\alpha, & \theta = 1. \end{cases}$$

(2) *For* $\alpha \in (0, 1]$ *and* $\theta \in [0, 1]$,

$$\sum_{k=1}^{t} \frac{1}{k^\theta} \prod_{i=k+1}^{t} \left(1 - \frac{\alpha}{i^\theta}\right) \le \frac{3}{\alpha}.$$

(3) *If* $\theta = 1$, *and for* $\alpha \in (0, 1]$,

$$\begin{aligned} \psi_1(t+1, \alpha) &= \sum_{k=1}^{t} \frac{1}{k^2} \prod_{i=k+1}^{t} \left(1 - \frac{\alpha}{i}\right)^2 \\ &\le \begin{cases} \dfrac{4}{1-2\alpha}(t+1)^{-2\alpha}, & \alpha \in \left(0, \frac{1}{2}\right); \\ 4(t+1)^{-1}\ln(t+1), & \alpha = \frac{1}{2}; \\ \dfrac{6}{2\alpha-1}(t+1)^{-1}, & \alpha \in \left(\frac{1}{2}, 1\right); \\ 6(t+1)^{-1}, & \alpha = 1. \end{cases} \end{aligned}$$

Proof. (1) By inequality (A.1), we have, for $\theta \in [0, 1]$,

$$\ln\left(1 - \frac{\alpha}{i^\theta}\right) \le \frac{-\alpha}{i^\theta}.$$

Thus

$$\sum_{i=k+1}^{t} \ln\left(1 - \frac{\alpha}{i^\theta}\right) \le -\alpha \sum_{i=k+1}^{t} \frac{1}{i^\theta} \le -\alpha \int_{k+1}^{t+1} \frac{1}{x^\theta}\, dx \tag{A.2}$$

which equals

$$\frac{\alpha}{1-\theta}((k+1)^{1-\theta} - (t+1)^{1-\theta}),$$

if $\theta \in [0, 1)$, and

$$\ln\left(\frac{k+1}{t+1}\right)^{\alpha},$$

if $\theta = 1$. Taking the exponential gives the inequality.

(2) If $\theta \in [0, 1)$, from (1) we have

$$\frac{1}{k^\theta} \prod_{i=k+1}^{t} \left(1 - \frac{\alpha}{i^\theta}\right) \le e^{-[2\alpha/(1-\theta)](t+1)^{1-\theta}} \frac{1}{k^\theta} e^{[2\alpha/(1-\theta)](k+1)^{1-\theta}},$$

whence

$$\begin{aligned} &\sum_{k=1}^{t} \frac{1}{k^\theta} \prod_{i=k+1}^{t} \left(1 - \frac{\alpha}{i^\theta}\right) \\ &\quad\le e^{-[2\alpha/(1-\theta)](t+1)^{1-\theta}} \left(\sum_{k=1}^{t-1} \frac{1}{k^\theta} e^{[2\alpha/(1-\theta)](k+1)^{1-\theta}} + \frac{1}{t^\theta} e^{[2\alpha/(1-\theta)](t+1)^{1-\theta}}\right) \end{aligned}$$

where

$$\sum_{k=1}^{t-1} \frac{1}{k^\theta} e^{[2\alpha/(1-\theta)](k+1)^{1-\theta}} \le 2^\theta \sum_{k=1}^{t-1} \left(\frac{1}{k+1}\right)^\theta e^{[2\alpha/(1-\theta)](k+1)^{1-\theta}}$$
$$\le 2\int_2^{t+1} e^{[2\alpha/(1-\theta)]x^{1-\theta}} x^{-\theta}\, dx \le \frac{1}{\alpha} e^{[2\alpha/(1-\theta)](t+1)^{1-\theta}}.$$

Therefore

$$e^{-[2\alpha/(1-\theta)](t+1)^{1-\theta}} \left(\sum_{k=1}^{t-1} \frac{1}{k^\theta} e^{[2\alpha/(1-\theta)](k+1)^{1-\theta}} + \frac{1}{t^\theta} e^{[2\alpha/(1-\theta)](t+1)^{1-\theta}}\right) \le \frac{1}{\alpha} + t^{-\theta}$$
$$< \frac{3}{\alpha}.$$

If $\theta = 1$, from inequality (A.2),

$$\sum_{k=1}^{t-1} \frac{1}{k} \prod_{i=k+1}^{t} \left(1 - \frac{\alpha}{i}\right) \le \sum_{k=1}^{t-1} \frac{1}{k} \left(\frac{k+1}{t+1}\right)^\alpha$$
$$\le \frac{2}{t^\alpha} \sum_{k=1}^{t-1} \frac{(k+1)^\alpha}{k+1} = \frac{2}{t^\alpha} \sum_{k=1}^{t-1} (k+1)^{\alpha-1}$$
$$\le \frac{2}{t^\alpha} \int_1^t x^{\alpha-1}\, dx,$$

where, if $\alpha = 1$,

$$\frac{2}{t^\alpha} \int_1^t x^{\alpha-1}\, dx = 2;$$

and, if $0 < \alpha < 1$,

$$\frac{2}{t^\alpha} \int_1^t x^{\alpha-1}\, dx = \frac{2}{\alpha} \left(\frac{t^\alpha - 1}{t^\alpha}\right) \le \frac{2}{\alpha}.$$

Therefore

$$\sum_{k=1}^{t} \frac{1}{k} \prod_{i=k+1}^{t} \left(1 - \frac{\alpha}{i}\right) \le t^{-\theta} + \frac{2}{\alpha} \le \frac{3}{\alpha},$$

which completes the proof of part (2).

(3) If $\theta = 1$, using inequality (A.1), we have

$$\sum_{i=k+1}^{t} \ln\left(1 - \frac{\alpha}{i}\right)^2 \le -2\alpha \sum_{i=k+1}^{t} \frac{1}{i} \le -2\alpha \int_{k+1}^{t+1} \frac{1}{x}\, dx = \ln\left(\frac{k+1}{t+1}\right)^{2\alpha}.$$

Thus

$$\psi_1(t+1,\alpha) \le t^{-2} + \sum_{k=1}^{t-1} \frac{1}{k^2}\left(\frac{k+1}{t+1}\right)^{2\alpha} \le \frac{4}{(t+1)^2} + \frac{2^{2\alpha}}{(t+1)^{2\alpha}} \sum_{k=1}^{t-1} k^{2\alpha-2}$$
$$\le \frac{4}{(t+1)^2} + \frac{2^{2\alpha}}{(t+1)^{2\alpha}} \int_{1/2}^{t-1/2} x^{2\alpha-2}\, dx,$$

where, if $\alpha \in (0, \frac{1}{2})$,

$$\text{r.h.s.} = \frac{4}{(t+1)^2} + \frac{2^{2\alpha}}{1-2\alpha}(t+1)^{-2\alpha}(2^{1-2\alpha} - (t-\tfrac{1}{2})^{2\alpha-1})$$
$$\le \left(2 + \frac{2}{1-2\alpha}\right)(t+1)^{-2\alpha} \le \frac{4}{1-2\alpha}(t+1)^{-2\alpha};$$

if $\alpha = \frac{1}{2}$,

$$\text{r.h.s.} = \frac{4}{(t+1)^2} + \frac{2}{t+1}(\ln(t-\tfrac{1}{2}) - \ln\tfrac{1}{2})$$
$$\le \left(\frac{2}{t+1} + \ln 2\right)\frac{2}{t+1}\ln(t+1) \le \frac{4}{t+1}\ln(t+1);$$

if $\alpha \in (\frac{1}{2}, 1)$,

$$\text{r.h.s.} = \frac{4}{(t+1)^2} + \frac{2^{2\alpha}}{2\alpha-1}(t+1)^{-2\alpha}((t-\tfrac{1}{2})^{2\alpha-1} - (\tfrac{1}{2})^{2\alpha-1})$$
$$\le \left(\frac{4}{t+1} + \frac{4}{2\alpha-1}\right)(t+1)^{-1} \le \frac{6}{2\alpha-1}(t+1)^{-1};$$

and, if $\alpha = 1$,

$$\text{r.h.s.} = \frac{4}{(t+1)^2} + \frac{4}{(t+1)^2}(t-1) \le 6(t+1)^{-1}.$$

This finishes the proof of the fourth part. □

Appendix B: Generalized Bennett's Inequality

In the direction of proving an exponential version of the main theorems with $1/\delta$ replaced by $\log 1/\delta$, it has seemed useful for us to consider Bennett's inequality for random variables in a Hilbert space. In the meantime, such a theorem was found useful in other work yet to appear. Thus we include Appendix B.

The following theorem might be considered as a generalization of Bennett's inequality for independent sums in Hilbert spaces, whose counterpart in real random variables is given in Theorem 3 of Smale and Zhou [**20**].

Theorem B.1 (Generalized Bennett). *Let $\mathscr{H}$ be a Hilbert space, let $\xi_i \in \mathscr{H}$ $(i = 1, \ldots, n)$ be independent random variables and let T_i: $\mathscr{H} \to \mathscr{H}$ be deterministic linear operators. Define $\tau_i = \|T_i\|$ and $\tau_\infty = \sup_i \tau_i$. Suppose that for all i almost surely $\|\xi_i\| \le M < \infty$. Define $\sigma_i^2 = \mathbb{E}\|\xi_i\|^2$ and $\sigma_\tau^2 = \sum_{i=1}^n \tau_i \sigma_i^2$. Then*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^n T_i(\xi_i - \mathbb{E}\xi_i)\right\| \ge \varepsilon\right\} \le 2\exp\left\{-\frac{\sigma_\tau^2}{\tau_\infty M^2} g\left(\frac{M\varepsilon}{\sigma_\tau^2}\right)\right\}$$

where $g(t) = (1+t)\log(1+t) - t$ for all $t \ge 0$. Considering that $g(t) \ge (t/2)\log(1+t)$, then

$$\mathbb{P}\left\{\left\|\sum_{i=1}^n T_i(\xi_i - \mathbb{E}\xi_i)\right\| \ge \varepsilon\right\} \le 2\exp\left\{-\frac{\varepsilon}{2\tau_\infty M}\log\left(1 + \frac{M\varepsilon}{\sigma_\tau^2}\right)\right\}.$$

The proof needs the following lemma due to Pinelis and Sakhanenko [**17**]. Its current form is taken from Theorem 3.3.4(a) in Yurinsky [**22**].

Lemma B.2 (Pinelis and Sakhanenko, 1985). *Let $\xi_i \in \mathscr{H}$ $(i = 1, \ldots, n)$ be a sequence of independent random variables with values in a Hilbert space $\mathscr{H}$ and $\mathbb{E}[\xi_i] = 0$. Then, for any $t > 0$,*

$$\mathbb{E}\left[\cosh\left(t\left\|\sum_{i=1}^n \xi_i\right\|\right)\right] \le \prod_{j=1}^n \mathbb{E}(e^{t\|\xi_j\|} - t\|\xi_j\|).$$

Proof of Theorem B.1. Without loss of generality we assume $\mathbb{E}[\xi_i] = 0$. For arbitrary $s > 0$, by Markov's inequality,

$$\begin{aligned}
\mathbb{P}\left\{\left\|\sum_{i=1}^n T_i\xi_i\right\| \ge \varepsilon\right\} &= \mathbb{P}\left\{\exp\left(s\left\|\sum_{i=1}^n T_i\xi_i\right\|\right) \ge e^{s\varepsilon}\right\} \\
&\le e^{-s\varepsilon}\mathbb{E}\exp\left(s\left\|\sum_{i=1}^n T_i\xi_i\right\|\right) \\
&\le 2e^{-s\varepsilon}\mathbb{E}\cosh\left(s\left\|\sum_{i=1}^n T_i\xi_i\right\|\right),
\end{aligned}$$

where the last inequality is due to $e^x \le e^x + e^{-x} = 2\cosh(x)$. Then, by Lemma B.1,

$$\mathbb{P}\left\{\left\|\sum_{i=1}^n T_i\xi_i\right\| \ge \varepsilon\right\} \le 2e^{-s\varepsilon}\prod_{j=1}^n \mathbb{E}(e^{s\|T_j\xi_j\|} - s\|T_j\xi_j\|).$$

Denote

$$I = 2e^{-s\varepsilon}\prod_{j=1}^n \mathbb{E}(e^{s\|T_j\xi_j\|} - s\|T_j\xi_j\|).$$

For each $1 \le j \le n$, considering $\mathbb{E}\|\xi_j\|^2 = \sigma_j^2$ and $\|\xi_j\| \le M$ almost surely,

$$\begin{aligned}\mathbb{E}(e^{s\|T_j\xi_j\|} - s\|T_j\xi_j\|) &= 1 + \sum_{k=2}^{n} \frac{s^k \mathbb{E}\|T_j\xi_j\|^k}{k} \\ &\le 1 + \sum_{k=2}^{n} \frac{s^k \tau_\infty^{k-1} M^{k-2}}{k} \tau_j \sigma_j^2 \\ &\le \exp\left(\sum_{k=2}^{n} \frac{s^k \tau_\infty^{k-1} M^{k-2}}{k} \tau_j \sigma_j^2 \right) \\ &= \exp\left(\frac{e^{s\tau_\infty M} - 1 - s\tau_\infty M}{\tau_\infty M^2} \tau_j \sigma_j^2 \right),\end{aligned}$$

where the second last inequality is due to $1 + x \le e^x$ for all x. Therefore

$$I \le \exp\left\{ -s\varepsilon + \frac{e^{s\tau_\infty M} - 1 - s\tau_\infty M}{\tau_\infty M^2} \sum_{j=1}^{n} \tau_j \sigma_j^2 \right\},$$

where the right-hand side is minimized at

$$s_0 = \frac{1}{\tau_\infty M} \log\left(1 + \frac{M\varepsilon}{\sum_{j=1}^{n} \tau_j \sigma_j^2} \right).$$

Notice that $\sigma_\tau^2 = \sum_{j=1}^{n} \tau_j \sigma_j^2$, then with this choice we arrive at

$$I \le \exp\left\{ -\frac{\sigma_\tau^2}{\tau_\infty M^2} g\left(\frac{M\varepsilon}{\sigma_\tau^2} \right) \right\},$$

where the function $g(t) = (1 + t)\log(1 + t) - t$ for all $t \ge 0$. This is the first inequality.

Moreover, we can check the lower bound of g,

$$g(t) \ge \frac{t}{2} \log(1 + t),$$

which leads to the second inequality. □

By taking $T_i = (1/n)I$, the following corollary gives a form of Bennett's inequality for random variables in Hilbert spaces.

Corollary B.3 (Bennett). *Let $\mathscr{H}$ be a Hilbert space and let $\xi_i \in \mathscr{H}$ $(i = 1, \ldots, n)$ be independent random variables such that $\|\xi_i\| \le M$ and $\mathbb{E}\|\xi_i\|^2 \le \sigma^2$ for all i. Then*

$$\mathbb{P}\left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} [\xi_i - \mathbb{E}\xi_i] \right\| \ge \varepsilon \right\} \le 2 \exp\left\{ -\frac{n\sigma^2}{M^2} g\left(\frac{M\varepsilon}{\sigma^2} \right) \right\}.$$

Noticing that $g(t) \geq t^2/2(1 + t/3)$, the corollary leads to the following Bernstein inequality for independent sums in Hilbert spaces.

Corollary B.4 (Bernstein). *Let $\mathscr{H}$ be a Hilbert space and let $\xi_i \in \mathscr{H}$ $(i = 1, \ldots, n)$ be independent random variables such that $\|\xi_i\| \leq M$ and $\mathbb{E}\|\xi_i\|^2 \leq \sigma^2$ for all i. Then*

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}[\xi_i - \mathbb{E}\xi_i]\right\| \geq \varepsilon\right\} \leq 2\exp\left\{-\frac{n\varepsilon^2}{2(\sigma^2 + M\varepsilon/3)}\right\}.$$

Yurinsky [**22**] also gives Bernstein's inequalities for independent sums in Hilbert spaces and Banach spaces. The following result is a varied form of Theorem 3.3.4(b) in [**22**]. Note that it is weaker than the form above in that the constant $\frac{1}{3}$ changes to 1.

Theorem B.5. *Let ξ_i be independent random variables with values in a Hilbert space $\mathscr{H}$. Suppose that for all i almost surely $\|\xi_i\| \leq M < \infty$ and $\mathbb{E}\|\xi_i\|^2 \leq \sigma^2 < \infty$. Then, for $n \geq 0$,*

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \mathbb{E}[\xi_i])\right\| \geq \varepsilon\right\} \leq 2\exp\left\{-\frac{n\varepsilon^2}{2(\sigma^2 + M\varepsilon)}\right\}.$$

Acknowledgment

The authors were supported by NSF grant 0325113.

The authors would like to acknowledge Peter Bartlett and Pierre Tarrès for their suggestions on stepsize rate; Yifeng Yu and Krishnaswami Alladi for their helpful discussions on proving the Main Analytic Estimate (Lemma A.1) and Lemma A.2; Iosif Pinelis and Yiming Ying for pointing out Lemma B.1 and their helpful comments. We thank the reviewers for many suggestions. We also thank Ding-Xuan Zhou, David McAllester, Adam Kalai, Gang Liang, Leon Bottou and, especially, Tommy Poggio, for many helpful discussions.

References

[1] M. Benaïm, Dynamics of stochastic approximations, in *Le Seminaire de Probabilites*, Lectures Notes in Mathematics, Vol. 1709, Springer-Verlag, New York, 1999, pp. 1–68.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.

[3] O. Bousquet and A. Elisseeff, Stability and generalization. *J. Mach. Learn. Res.* **2** (2002), 499–526.

[4] N. Cesa-Bianchi, P. M. Long, and M. K. Warmuth, Worst-case quadratic loss bounds for prediction using linear functions and gradient descent, *IEEE Trans. Neural Networks* **7**(3) (1996), 604–619.

[5] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, Cambridge, 2000.
[6] F. Cucker and S. Smale, Best choices for regularization parameters in learning theory, *Found. Comput. Math.* **2**(4) (2002), 413–428.
[7] F. Cucker and S. Smale, On the mathematical foundations of learning, *Bull. Amer. Math. Soc.* **29**(1) (2002), 1–49.
[8] E. De Vito, A. Caponnetto, and L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, *Found. Comput. Math.* **5**(1), 59–85.
[9] M. Duflo, Cibles atteignables avec une probabilité positive d'aprés M. Benaïm, Unpublished manuscript, 1997.
[10] M. Duflo, *Algorithmes Stochastiques*, Springer-Verlag, Berlin, 1996.
[11] T. Evgeniou, M. Pontil, and T. Poggio, Regularization networks and support vector machines, *Adv. Comput. Math.* **13**(1) (1999), 1–50.
[12] L. Györfi, Stochastic approximation from ergodic sample for linear regression, *Z. Wahrsch. Verw. Gebiete* **54** (1980), 47–55.
[13] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, New York, 2002.
[14] J. Kiefer and J. Wolfowitz, Stochastic estimation of the maximum of a regression function, *Ann. Math. Statist.* **23** (1952), 462–466.
[15] J. Kivinen, A. J. Smola, and R. C. Williamson, Online learning with kernels, *IEEE Trans. Signal Process.* **52**(8) (2004), 2165–2176.
[16] H. J. Kushner and G. G. Yin, *Stochastic Approximations and Recursive Algorithms and Applications*, Springer-Verlag, Berlin, 2003.
[17] I. F. Pinelis and A. I. Sakhanenko, Remarks on inequalities for probabilities of large deviations, *Theory Probab. Appl.* **30**(1) (1985), 143–148.
[18] H. Robbins and S. Monro, A stochastic approximation method, *Ann. Math. Statist.* **22**(3) (1951), 400–407.
[19] H. Robbins and D. Siegmund, A convergence theorem for nonnegative almost supermartingales and some applications, in (J. S. Rustagi, editor), *Optimizing Methods in Statistics*, Academic Press, New York, 1971, pp. 233–257.
[20] S. Smale and D.-X. Zhou, Shannon sampling ii. connections to learning theory, *Appl. Comput. Harmonic Anal.* (2005), submitted.
[21] V. B. Tadic, On the almost sure rate of convergence of linear stochastic approximation algorithms, *IEEE Trans. Inform. Theory* **50** (2004), 401–409.
[22] Y Yurinsky, *Sums and Gaussian Vectors*, Lecture Notes in Mathematics, Vol. 1617, Springer-Verlag, Berlin, 1995.
[23] T. Zhang, Leave-one-out bounds for kernel methods, *Neural Comput.* **15** (2003), 1397–1437.
[24] M. Zinkevich, Online convex programming and generalized infinitesimal gradient ascent, Technical report, CMU-CS-03-110, School of Computer Science, Carnegie Mellon University, 2003.