On Complexity Issue of Online Learning Algorithms

Yuan Yao

Abstract— In this paper, some new probabilistic upper bounds are presented for the online learning algorithm proposed in [1], and more generally for linear stochastic approximations in Hilbert spaces. With these upper bounds not only does one recover almost sure convergence, but also relaxes the square summable condition on the step size appeared in our early work. We also give two probabilistic upper bounds for an averaging process, both of which achieve the same rate with respect to sample size as in "batch learning" algorithms.

Index Terms — Online learning, regularization, stochastic approximation, averaging process, reproducing kernel Hilbert Space.

I. INTRODUCTION

$\mathbf{S}$ UPERVISED learning, or learning from examples, is to find a function in a hypothesis space $\mathscr{H}$, which associates an input $x \in \mathscr{X}$ to an output $y \in \mathscr{Y}$, by drawing examples $(x_t, y_t)_{t \in \mathbb{N}}$ at random from a probability measure ρ on $\mathscr{X} \times \mathscr{Y}$. By "online learning", we mean a sequential decision process $(f_t)_{t \in \mathbb{N}}$ in the hypothesis space, where each f_{t+1} is decided by the current observation and f_t which only depends on previous examples, i.e. $f_{t+1} = T_{z_t}(f_t)$ where $z_t = (x_t, y_t)$ (see, e.g. [1], [2]). As a contrast, "batch learning" refers to a decision utilizing the whole set of examples (see, e.g., [3], [4]).

In the scheme of regularization, one wants to approximate a function f_λ^* as a solution of the following optimization problem (see, e.g., [4], [5]),

$$\min_{f \in \mathscr{H}} \int_{\mathscr{X} \times \mathscr{Y}} V(f(x), y) d\rho + \lambda \|f\|_{\mathscr{H}}^2, \quad \lambda > 0, \quad (1)$$

where the hypothesis space $\mathscr{H}$ is associated with a norm $\| \cdot \|_{\mathscr{H}}$ and $V : \mathscr{H} \times \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ is a loss function, which measures the prediction cost of f at x against y. Among a variety of choices on V and $\mathscr{H}$, it leads to a simple structure but deeper understanding by selecting the quadratic loss $V(f(x), y) = (f(x) - y)^2$ and the hypothesis space $\mathscr{H} = \mathscr{H}_K$, the reproducing kernel Hilbert space (RKHS) associated with a Mercer kernel K. In this setting there exists a unique minimizer f_λ^*, satisfying the following linear equation,

$$(L_K + \lambda I) f = L_K f_\rho, \quad (2)$$

where $f_\rho(x) = \int_{\mathscr{Y}} y d\rho_{\mathscr{Y}|x}$, the conditional expectation of y, is called the *regression function*; and the integral operator $L_K : \mathscr{L}_\rho^2(\mathscr{X}) \to \mathscr{L}_\rho^2(\mathscr{X})$ is defined by $L_K(f) =$

$\int_{\mathscr{X}} K(x, t) f(t) d\rho_{\mathscr{X}}$. Since $L_K + \lambda I$ ($\lambda > 0$) is invertible, we may write $f_\lambda^* = (L_K + \lambda I)^{-1} L_K f_\rho$. Moreover, such a choice avoids the estimation of covering numbers of $\mathscr{H}$, which is difficult in most cases [3], [6]; it provides a simple estimate of optimal upper bounds asymptotically meeting lower bounds [7], [8]; it bridges over the linear inverse problem toward other regularization schemes [9], [10]; and more interestingly in this paper, it takes an especially simple form in online learning algorithms [1].

Given an independent and identically distributed random sequence $(x_t, y_t)_{t \in \mathbb{N}}$, the algorithm in [1] returns a sequence $(f_t)_{t \in \mathbb{N}} \in \mathscr{H}_K$ to approximate f_λ^*,

$$f_{t+1} = f_t - \gamma_t((f_t(x_t) - y_t) K_{x_t} + \lambda f_t), \quad (3)$$

where $f_1 \in \mathscr{H}_K$, e.g. $f_1 = 0$, and in this paper the step size $\gamma_t > 0$ is chosen as $\gamma_t = O(t^{-\theta})$ for some $\theta \in [0, 1)$.

The algorithm can be regarded as either the stochastic approximation of the gradient descent method for (1), or the stochastic approximation of the linear equation (2), which was originally proposed in [11], [12]. Traditional analysis on stochastic approximations has been focusing on convergence and asymptotic rates. A convergence result often used in applications, known as the Robbins-Siegmund Theorem [13], imposes a condition on the step size that $\sum_t \gamma_t = \infty$ and $\sum_t \gamma_t^2 < \infty$, and leads to the almost sure convergence (with probability one). For the step size chosen in this paper, $\gamma_t = O(t^{-\theta})$, this requires $\theta \in (1/2, 1)$. In this setting, the asymptotic rate has been shown as $O(\gamma_t^{1/2}) = O(t^{-\theta/2})$. Note that the condition $\sum_t \gamma_t = \infty$, is used to "forget" the error caused by initial choices. However the square summable condition, $\sum_t \gamma_t^2 < \infty$, is not necessary for the almost sure convergence. For example in [14] (or see the remarks in [15]), to ensure the almost sure convergence it is enough that for all $c > 0$,

$$\sum_t e^{-c/\gamma_t} < \infty.$$

This even justifies the use of $\gamma_t = 1/\log^{1+\epsilon} t$ for some $\epsilon > 0$, which is however not pursued in this paper. For more background on stochastic approximations, see for example [16], [17], and references therein.

In learning theory a fundamental goal is to approximate the regression function f_ρ. For this purpose, it is not enough to apply traditional results on convergence and asymptotic rates; since to approximate f_ρ arbitrarily well, we need to tune the regularization parameter λ arbitrarily small as sample size goes large. The influence of λ to convergence is hidden in the constants and thus we seek upper bounds to disclose it.

In [1], we present a probabilistic upper bound based on Markov's Inequality, that the following holds with probability at least $1 - \delta$ ($\delta \in (0, 1)$)

$$\|f_t - f_\lambda^*\|_K \leq O(\lambda^{-\frac{\theta}{2(1-\theta)}} t^{-\theta/2} \delta^{-1/2}), \quad \theta \in (1/2, 1).$$

This upper bound is tight in the asymptotic rate of t; however, it only implies that f_t converges to f_λ^* in probability, weaker than the almost sure convergence.

In this paper, we present two new probabilistic upper bounds by using exponential probabilistic inequalities for martingales in Hilbert spaces [18], both of which lead to almost sure convergence and extend the rate of step size to $\theta \in [0, 1)$, at the sacrifice of rates on λ.

The first upper bound (as Theorem A) says that with probability at least $1 - \delta$ ($\delta \in (0, 1)$),

$$\|f_t - f_\lambda^*\|_K \leq O(\lambda^{-1-\frac{1}{2(1-\theta)}} t^{-\theta/2} \log^{1/2} 1/\delta), \quad \theta \in [0, 1).$$

This upper bound implies almost sure convergence for all $\theta \in (0, 1)$, by changing $1/\delta$ to $\log 1/\delta$. Note that when $\theta = 0$, algorithm (3) is often called the *Adaline* or *Widrow-Hoff algoirthm* ([19], or see Chapter 5 in [20]), which is not guaranteed to converge in this setting.

The second upper bound (as Theorem B) is given for the *averaging process* proposed in [21], [22],

$$\bar{f}_t = \frac{1}{t} \sum_{j=1}^{t} f_j = \bar{f}_{t-1} + \frac{1}{t}(f_t - \bar{f}_{t-1}), \quad \bar{f}_1 = f_1, \quad (4)$$

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

$$\|\bar{f}_t - f_\lambda^*\|_K \leq O(\lambda^{-2} t^{-1/2} \log^{1/2} 1/\delta), \quad \theta \in [0, 1).$$

In contrast to "batch learning" case with a rate $O(\lambda^{-1} t^{-1/2})$ [7], this upper bound achieves the same fixed rate in t for all $\theta \in [0, 1)$, while losing the rate in λ.

It is possible to improve the rate in λ by turning back to Markov's Inequality. In fact, the reason of loss in λ lies in the application of the Hoeffding-style inequalities which, compared to Markov's Inequality, replace the variance by its uniform upper bounds. Based on this observation, we obtain the following result (as Theorem B*) for the averaging process by using Markov's Inequality,

$$\|\bar{f}_t - f_\lambda^*\|_K \leq O(\lambda^{-1} t^{-1/2} \delta^{-1/2}), \quad \theta \in [0, 1),$$

which holds with probability at least $1 - \delta$ ($\delta \in (0, 1)$). We conjecture that this can be improved to be $O(\lambda^{-1} t^{-1/2} \log^{1/2} 1/\delta)$ by using other variance-based inequalities, such as Bennet's or Bernstein's Inequality.

The organization of this paper is as follows. In Section II, we present our main results and discussions. In Section III, we study a more general problem, linear stochastic approximation in Hilbert spaces, from which we derive Theorem A and B in a special case. We propose in Section IV a martingale decomposition for remainders, which is crucial for later development. All the proofs for the theorems in Section III are collected in Section V. In section VI we prove Theorem B* via a reverse martingale decomposition for remainders. Conclusion and open problems are summarized in Section VII. The last

section is an appendix collecting some crucial estimates used in this paper.

II. MAIN RESULTS

Before presenting the main results, we need some definitions and remarks on notation.

In this paper, let $\mathscr{X} \subseteq \mathbb{R}^n$ be compact, $\mathscr{Y} = \mathbb{R}$, and $\mathscr{Z} = \mathscr{X} \times \mathscr{Y}$. Assume that there is a $M_\rho > 0$ such that $\text{supp}(\rho) \subseteq \mathscr{X} \times [-M_\rho, M_\rho]$. Define

$$C_K := \max_{x \in \mathscr{X}} \sqrt{K(x, x)} < \infty. \quad (5)$$

and a constant only depending on $\theta \in [0, 1)$,

$$D_\theta = 1 + 2^{\frac{\theta}{1-\theta}} \left(1 + \Gamma\left(\frac{1}{1-\theta}\right)\right) \geq 1. \quad (6)$$

Assume that the examples $(x_t, y_t)_{t \in \mathbb{N}}$ are independent and identically distributed (i.i.d.) according to ρ.

In this paragraph, we provide a short background on reproducing kernel Hilbert spaces (RKHS). A function $K : X \times X \to \mathbb{R}$ is called a *Mercer kernel*, if it is a continuous symmetric real function which is *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 $c_i \in \mathbb{R}$ ($i = 1, \ldots, l$). Let $\mathscr{H}_K$ be the Reproducing Kernel Hilbert Space associated with a Mercer kernel K. Recall the definition as follows. Consider the vector space V_K generated by $\{K_x : x \in X\}$, i.e. all the finite linear combinations of K_x, where for each $x \in X$, the function $K_x : X \to \mathbb{R}$ is defined by $K_x(x') = K(x, x')$. A semi-definite inner product $\langle \ , \ \rangle_K$ on this vector space can be defined as the unique linear extension of $\langle K_x, K_{x'} \rangle_K := K(x, x')$. The induced semi-norm is $\|f\|_K = \sqrt{\langle f, f \rangle_K}$ for each $f \in V_K$. Notice that the zero set $V_0 = \{f \in V_K : \|f\|_K = 0\}$ is a subspace. Then the semi-definite inner product induces an inner product on the quotient space V_K/V_0. Let $\mathscr{H}_K$ be the completion of this inner product space V_K/V_0 with respect to $\| \cdot \|_K$. The most important property of RKHS is the so called *reproducing property*: for any $f \in \mathscr{H}_K$, $f(x) = \langle f, K_x \rangle_K$ ($x \in X$). RKHS can be regarded as a generalization of real analytic functions (or band-limited functions), see for example [23], [24].

Recall the definition of the *incomplete gamma function* restricted on $[0, \infty) \times [0, \infty)$,

$$\Gamma(a, x) = \int_x^\infty s^{a-1} e^{-s} ds, \qquad \text{where } a, x \geq 0.$$

The *gamma function* is defined by $\Gamma(a) = \Gamma(a, 0)$. Finally we make a remark on notation. When $n < m$, the product and summation, $\prod_{i=m}^{n} x_i$ and $\sum_{i=m}^{n} x_i$, are understood to be 1 and 0, respectively. We use $\mathbb{E}_z$ and $\mathbb{E}_{z_1|z_2}$ to denote the expectation and conditional expectation, respectively. Shorthand notation $\mathbb{E}$ is also used when its meaning is clear from the context.

The following are the main results in this paper.

A. Probabilistic Upper Bound with Almost Sure Convergence

Theorem A. *Let* $\lambda \leq \lambda_0$, $\gamma_t = t^{-\theta}/(C_K^2 + \lambda)$ *for some* $\theta \in [0, 1)$, *and* $f_1 = 0$. *Then for all* $t \in \mathbb{N}$ *there holds*

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

where

$$\mathscr{E}_{init}(t) \leq e^{\frac{\alpha}{1-\theta}(1-t^{1-\theta})}\|f_\lambda^*\|_K,$$

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

$$\mathscr{E}_{samp}(t) \leq C_{\rho,\theta,K}\left(\frac{1}{\lambda}\right)^{1+\frac{1}{2(1-\theta)}}\left(\frac{1}{t}\right)^{\frac{\theta}{2}}\log^{1/2}\frac{2}{\delta}.$$

Here $C_{\rho,\theta,K} = 16\sqrt{D_\theta}C_K M_\rho(\lambda_0 + C_K^2)^{1/2(1-\theta)}$.

The proof of Theorem A will be given in Section III as a corollary of Theorem 3.1.

Remark 2.1: The second inequality is equivalent to

$$\mathbf{Prob}\{\mathscr{E}_{samp}(t) \geq \varepsilon\} \leq 2e^{-c\varepsilon^2 t^\theta}$$

where $c = \lambda^{2+\frac{1}{1-\theta}}/C_{\rho,\theta,K}^2$. For each $\varepsilon > 0$, denote by A_t the event $\{\mathscr{E}_{samp}(t) \geq \varepsilon\}$. Then

$$\sum_{t\in\mathbb{N}}\mathbf{Prob}(A_t) \leq 2\sum_{t\in\mathbb{N}}e^{-c\varepsilon t^\theta} < \infty.$$

By the Borel-Cantelli Lemma, we have $\mathbf{Prob}(A_t \text{ i.o.}) = 0$, i.e. it is of zero probability that A_t happens for infinitely many values $t \in \mathbb{N}$, whence $\mathscr{E}_{samp}(t) \to 0$ almost surely (with probability one).

Remark 2.2: Note that when $\theta = 0$, the *Widrow-Hoff* algorithm [19] can't ensure its convergence by this upper bound. However, it can be combined with the averaging process to achieve a convergence rate of $O(t^{-1/2})$, which will be discussed in the next subsection.

B. Averaging Process

It is natural to consider the average of the ensemble $\{f_1, \ldots, f_t\}$ up to time t, which might improve the convergence rate since by intuition averaging may reduce variance. In stochastic approximation, this acceleration by averaging was firstly observed independently by [22] and [21] (or see [25]) based on asymptotic analysis; recently this phenomenon has also been noticed in learning theory society (see, e.g., [2]). A recent result [26] studies this averaging process in a more general framework of two-time-scale linear stochastic approximations with asymptotic analysis. Below we show a probabilistic upper bound with a fixed rate $O(t^{-1/2})$ for all $\theta \in [0, 1)$.

Theorem B. *Let* $\lambda \leq \lambda_0$, $\gamma_t = t^{-\theta}/(C_K^2 + \lambda)$ *for some* $\theta \in [0, 1)$, *and* $f_1 = 0$. *Then for all* $t \in \mathbb{N}$ *there holds for (4)*

$$\|\bar{f}_t - f_\lambda^*\|_K \leq \mathscr{E}_{init}(t) + \mathscr{E}_{samp}(t).$$

where

$$\mathscr{E}_{init}(t) \leq C_1\left(\frac{1}{\lambda t}\right),$$

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

$$\mathscr{E}_{samp}(t) \leq C_2\left(\frac{1}{\lambda}\right)^2\sqrt{\frac{1}{t}}\log^{1/2}\frac{2}{\delta}.$$

Here $C_1 = D_\theta(\lambda_0 + C_K^2)\|f_\lambda^*\|_K$ *and* $C_2 = 2^{3+\theta}D_\theta C_K M_\rho(\lambda_0 + C_K^2)$.

The proof of Theorem B will be given in Section III as a corollary of Theorem 3.2.

Remark 2.3: Assume without loss of generality that $\lambda_0 = C_K^2$. When $\theta = 0$, $D_0 = 3$ and this gives the following bound for combined *Adaline-Averaging* algorithm

$$\mathscr{E}_{init}(t) \leq 6C_K^2\|f_\lambda^*\|_K\left(\frac{1}{\lambda t}\right),$$

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

$$\mathscr{E}_{samp}(t) \leq 48C_K^3 M_\rho\left(\frac{1}{\lambda}\right)^2\sqrt{\frac{1}{t}}\log^{1/2}\frac{2}{\delta}.$$

The rate in λ can be improved. Let $\sigma_\lambda^2 = \mathbb{E}[\|(y - f_\lambda^*(x))K_x - \lambda f_\lambda^*\|_K^2$ for some $\sigma_\lambda \geq 0$. By Markov's Inequality we obtain the following theorem, whose proof will be given in Section VI.

Theorem B*. *Let* $\lambda \leq \lambda_0$, $\gamma_t = t^{-\theta}/(C_K^2 + \lambda)$ *for some* $\theta \in [0, 1)$, *and* $f_1 = 0$. *Then the following holds for all* $t \in \mathbb{N}$,

$$\|\bar{f}_t - f_\lambda^*\|_K \leq \mathscr{E}_{init}(t) + \mathscr{E}_{samp}(t),$$

where

$$\mathscr{E}_{init}(t) \leq D_\theta(\lambda_0 + C_K^2)\|f_\lambda^*\|_K\left(\frac{1}{\lambda t}\right),$$

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

$$\mathscr{E}_{samp}(t) \leq \frac{2^\theta D_\theta \sigma_\lambda}{\lambda\sqrt{\delta t}}.$$

Remark 2.4: Proposition 6.5-3 gives an estimate on σ_λ,

$$\sigma_\lambda \leq M_\rho\sqrt{5(\lambda_0 + C_K^2)}.$$

Remark 2.5: If $\sigma_\lambda^2 = 0$, we obtain the following upper bound in a deterministic setting,

$$\|\bar{f}_t - f_\lambda^*\|_K \leq D_\theta(\lambda_0 + C_K^2)\|f_\lambda^*\|_K\left(\frac{1}{\lambda t}\right).$$

C. Comparison with "Batch Learning" Results

Given a sample $\mathbf{z} = \{(x_i, y_i) : i = 1, \ldots, t\}$, "batch learning" means solving the following *regularized least square* problem (see, e.g., [4], [3])

$$\min_{f\in\mathscr{H}_K}\frac{1}{t}\sum_{i=1}^t(f(x_i) - y_i)^2 + \lambda\|f\|_K^2, \quad \lambda > 0.$$

There exists a unique minimizer $f_{\lambda,\mathbf{z}}$ satisfying

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

where $a = (a_1, \ldots, a_t)$ is the solution of the linear equation

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

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 $\mathbf{y} = (y_1, \ldots, y_t) \in \mathbb{R}^t$.

A probabilistic upper bound for $\|f_{\lambda,\mathbf{z}} - f_\lambda^*\|_K$ is given in [27], and this has been substantially improved by [28] using also some ideas from [29]. Moreover, [30] gives error bounds expressed in a different form. A recent result (Theorem 1 in [7]) shows that,

Theorem 2.6: With probability at least $1 - \delta$ ($\delta \in (0, 1)$) there holds

$$\|f_{\lambda,\mathbf{z}} - f_\lambda^*\|_K \leq \frac{6 C_K M_\rho \log(2/\delta)}{\lambda \sqrt{t}}.$$

Remark 2.7: A recent result [8] shows that the rate $O(\lambda^{-1} t^{-1/2})$ is optimal in the sense that it leads to a convergence rate asymptotically meeting the minimax lower bound. Theorem B tells us that the averaging process achieves $O(\lambda^{-2} t^{-1/2})$, which is optimal in t but suboptimal in λ. Theorem B* improves this to $O(\lambda^{-1} t^{-1/2})$, though it only leads to convergence in probability.

III. LINEAR STOCHASTIC APPROXIMATION IN HILBERT SPACES

In this section we study a more general problem, stochastic approximation of linear equations in Hilbert spaces. Some general upper bounds are given and they lead to Theorem A and B in a special case.

Let W be a Hilbert space, $A(z) : W \to W$ a random positive operator depending on $z \in \mathscr{Z}$ and $B(z) \in W$ a random vector. Define $\hat{A} = \mathbb{E}_z[A(z)]$ and $\hat{B} = \mathbb{E}_z[B(z)]$. Consider the following linear equation

$$\hat{A}w = \hat{B}, \tag{7}$$

whose unique solution is $w^* = \hat{A}^{-1}\hat{B}$.

In the sequel, we assume that almost surely,

Finiteness Condition. A. $\mu_{\min} I \leq A(z) \leq \mu_{\max} I$ ($0 < \mu_{\min} \leq \mu_{\max} < \infty$) and let $\alpha = \mu_{\min}/\mu_{\max} \in (0, 1]$;
B. $\|B(z)\| \leq \beta < \infty$;
C. $\mathbb{E}\|A(z)w^* - B(z)\|^2 = \sigma^2 < \infty$.

Given an i.i.d. sequence $(z_t)_{t \in \mathbb{N}}$, define a sequence $\{w_t\}_{t \in W}$ as successive stochastic approximations of w^*,

$$w_{t+1} = w_t - \gamma_t(A_t w_t - B_t), \qquad w_1 \in W \tag{8}$$

where $A_t = A(z_t)$, $B_t = B(z_t) \in W$ and $\gamma_t = 1/\mu_{\max} t^\theta$ for some $\theta \in [0, 1)$.

Define a *remainder* sequence $(r_t)_{t \in \mathbb{N}}$ by

$$r_t = w_t - w^*,$$

which measures the deviation between w_t and w^*. It can be seen that both w_t and r_t are W-valued random variables depending on $z_1, \ldots, z_{t-1}$. In this note we assume that $(z_t)_{t \in \mathbb{N}}$ is a i.i.d. sequence, but the method we used here can be extended to more general cases.

The main results in this section are in the following.

Theorem 3.1: Let $\gamma_t = t^{-\theta}/\mu_{\max}$ ($\theta \in [0, 1)$) and $w_1 = 0$. Then for all $t \in \mathbb{N}$, there holds

$$\|w_t - w^*\| \leq \mathscr{E}_{init}(t) + \mathscr{E}_{samp}(t),$$

where

$$\mathscr{E}_{init}(t) \leq e^{\frac{\alpha}{1-\theta}(1 - t^{1-\theta})}\|r_1\|,$$

and with probability at least $1 - \delta$,

$$\mathscr{E}_{samp}(t) \leq \frac{16\sqrt{D_\theta}\beta}{\mu_{\max}}\left(\frac{1}{\alpha}\right)^{1 + \frac{1}{2(1-\theta)}}\left(\frac{1}{t}\right)^{\theta/2}\log^{1/2}\frac{2}{\delta}.$$

For the averaged sequence $\bar{w}_t = \frac{1}{t}\sum_{j=1}^t w_j$, we have

Theorem 3.2: Let $\gamma_t = t^{-\theta}/\mu_{\max}$ ($\theta \in [0, 1)$) and $w_1 = 0$. Then for all $t \in \mathbb{N}$ there holds

$$\|\bar{w}_t - w^*\| \leq \mathscr{E}_{init}(t) + \mathscr{E}_{samp}(t),$$

where

$$\mathscr{E}_{init}(t) \leq D_\theta\left(\frac{1}{\alpha t}\right)\|r_1\|,$$

and with probability at least $1 - \delta$,

$$\mathscr{E}_{samp}(t) \leq \frac{2^{3+\theta}D_\theta\beta}{\mu_{\max}}\left(\frac{1}{\alpha}\right)^2\sqrt{\frac{1}{t}}\log^{1/2}\frac{2}{\delta}.$$

A. Proofs of Theorem A and B

Proof of Theorem A: We first show that the algorithm given by (3) can be derived from equation (8); then Theorem A follows from Theorem 3.1.

Let $E_x : \mathscr{H}_K \to \mathbb{R}$ be the evaluation operator such that $E_x(f) = f(x)$. Let $E_x^* : \mathbb{R} \to \mathscr{H}_K$ be the adjoint of E_x defined by $\langle y, E_x(f)\rangle_{\mathbb{R}} = \langle E_x^*(y), f\rangle_{\mathscr{H}_K}$, whence by reproducing property $f(x) = \langle f, K_x\rangle$, we have $E_x^*(y) = yK_x$. Now take $W = \mathscr{H}_K$, define $A(z) : \mathscr{H}_K \to \mathscr{H}_K$ by $f \mapsto E_x^* E_x(f) + \lambda$ and $B(z) = E_x^*(-y)$. Then $\hat{A} = L_K + \lambda I$ and $\hat{B} = -L_K f_\rho$. By this substitution, equation (8) becomes (3).

Notice that $\mu_{\max} = \lambda + C_K^2$, $\mu_{\min} = \lambda$, and $\beta = C_K M_\rho$. Theorem A thus follows from Theorem 3.1. ∎

Proof of Theorem B: In a similar way to the proof of Theorem A, Theorem B follows from Theorem 3.2. ∎

IV. MARTINGALE DECOMPOSITION OF REMAINDERS

In this section, we decompose the remainder r_t and its average $\bar{r}_t$ into the sum of two parts: one is deterministic reflecting the error caused by initial choice, and the other is a martingale reflecting the fluctuation caused by random sampling. Upper bounds for them will be given in the next section. Such a decomposition is somehow close to the treatment in Robbins-Siegmund Theorem [13], where $\|r_t\|^2$ is transformed into a supermartingale. But our problem benefits from the linear structure and get a direct decomposition on r_t. We note that such a martingale decomposition can be extended to $\|r_t\|^2$ in nonlinear stochastic approximations, which however is not pursued here.

First of all we introduce some short-hand notations. Define a random positive operator on W,

$$\Pi_k^t = \begin{cases} \prod_{i=k}^t (I - \gamma_i A_i), & k \le t; \\ I, & k > t. \end{cases} \quad (9)$$

If we replace A_i by $\hat{A}$, we obtain a deterministic positive operator, say $\hat{\Pi}_k^t$. Define $Y_t = A_t w^* - B_t$, a W-valued random variable depending on z_t. Clearly $\mathbb{E}_{z_t} Y_t = 0$ and by Finiteness Condition-C, $\mathbb{E}\|Y_t\|^2 = \sigma^2$ for all t.

The following proposition gives a decomposition of r_t into the sum of a deterministic part and a martingale.

Proposition 4.1: For all $t \in \mathbb{N}$,

$$r_t = \hat{\Pi}_1^{t-1} r_1 - \sum_{k=1}^{t-1} \gamma_k \hat{\Pi}_{k+1}^{t-1} \chi_k, \quad (10)$$

where $\chi_k = (A_k - \hat{A}) w_k + \hat{B} - B_k$ ($1 \le k \le t$).

Proof: By equation (8)

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

where we can check that

$$\begin{aligned} \chi_t &= (A_t - \hat{A}) r_t + Y_t \\ &= (A_t - \hat{A}) w_t + \hat{B} - B_t. \end{aligned}$$

Then equation (10) follows from induction on t. ∎

Note that $\hat{\Pi}_{k+1}^t$ is deterministic, r_k depends on $z_1, \ldots, z_{k-1}$, $A_k - \hat{A}$ and Y_k are both of zero means depending only on z_k. Recall that given a sequence of random variables $(\xi_k)_{k \in \mathbb{N}}$ such that ξ_k depends on random variables $\{z_i : 1 \le i \le k\}$, (ξ_k) is called a *martingale difference sequence* if $\mathbb{E}_{z_k | z_1, \ldots, z_{k-1}}[\xi_k] = 0$. The sum of a martingale difference sequence is called a *martingale*. This motivates the following definition of a martingale difference sequence,

$$\xi_k = \begin{cases} \gamma_k \hat{\Pi}_{k+1}^{t-1} \chi_k, & 1 \le k \le t; \\ \xi_k = 0, & k > t. \end{cases}$$

With this we write,

$$r_t = \hat{\Pi}_1^{t-1} r_1 - \sum_{k=1}^{t-1} \xi_k. \quad (11)$$

Now consider the averaging process. Define

$$\bar{w}_t = \frac{1}{t} \sum_{i=1}^t w_i = \bar{w}_{t-1} + \frac{1}{t}(w_t - \bar{w}_{t-1}), \quad \bar{w}_1 = w_1,$$

and we study upper bounds for the *averaged remainder sequence*

$$\bar{r}_t = \bar{w}_t - w^* = \frac{1}{t} \sum_{i=1}^t (w_i - w^*) = \frac{1}{t} \sum_{i=1}^t r_i.$$

The following proposition gives a decomposition of $\bar{r}_t$.

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

$$\bar{r}_t = \frac{1}{t} \left(\sum_{j=0}^{t-1} \hat{\Pi}_1^j \right) r_1 - \frac{1}{t} \sum_{k=1}^{t-1} \gamma_k \left(\sum_{j=k}^{t-1} \hat{\Pi}_{k+1}^j \right) \chi_k, \quad (12)$$

Proof: By equation (10),

$$\begin{aligned} \bar{r}_t &= \frac{1}{t} \sum_{j=1}^t r_j \\ &= \frac{1}{t} \left(1 + \sum_{j=1}^{t-1} \hat{\Pi}_1^j \right) r_1 - \frac{1}{t} \sum_{j=1}^{t-1} \sum_{k=1}^j \gamma_k \hat{\Pi}_{k+1}^j \chi_k \\ &= \frac{1}{t} \left(\sum_{j=0}^{t-1} \hat{\Pi}_1^j \right) r_1 - \frac{1}{t} \sum_{k=1}^{t-1} \gamma_k \left(\sum_{j=k}^{t-1} \hat{\Pi}_{k+1}^j \right) \chi_k \end{aligned}$$

which ends the proof. ∎

Let

$$\eta_k = \begin{cases} \frac{\gamma_k}{t} (\sum_{j=k}^{t-1} \hat{\Pi}_{k+1}^j) \chi_k, & 1 \le k \le t; \\ 0, & k > t. \end{cases}$$

Then $(\eta_k)_{k \in \mathbb{N}}$ is a martingale difference sequence and its sum is a martingale. With this we have

$$\bar{r}_t = \frac{1}{t} \left(\sum_{j=0}^{t-1} \hat{\Pi}_1^j \right) r_1 - \sum_{k=1}^{t-1} \eta_k. \quad (13)$$

Now define an *initial error* by $\mathscr{E}_{init}(t) = \|\hat{\Pi}_1^{t-1} r_1\|$ (or, $\mathscr{E}_{init}(t) = \|\frac{1}{t} \left(\sum_{j=0}^{t-1} \hat{\Pi}_1^j \right) r_1\|$ in averaging process), which is deterministic and reflects the propagated effect of r_1; and a *sample error* by $\mathscr{E}_{samp}(t) = \|\sum_{k=1}^{t-1} \xi_k\|$ (or, $\mathscr{E}_{samp}(t) = \|\sum_{k=1}^{t-1} \eta_k\|$ in averaging process), which is random and reflects the stochastic error caused by samples. The initial error can be bounded deterministically. For the sample error, we can obtain probabilistic upper bounds by using the exponential inequalities for martingale difference sequences in Hilbert spaces [18]. We will show this in the next section.

V. Proofs of Theorem 3.1 and Theorem 3.2

For simplicity, in this section we choose Hoeffding's inequality for martingale difference sequences in Hilbert spaces [31]. We note here that by choosing Bennet-type inequalities [18], one can get tighter bounds depending on variances $\mathbb{E}\|A_t - \hat{A}\|^2$ and $\mathbb{E}\|B_t - \hat{B}\|^2$, in the sense that when these variances approach to zero, they lead to deterministic upper bounds.

Before presenting the proofs, we need some preliminary results. The first one is an extension of Hoeffding's Inequality from real numbers to Hilbert spaces, which is due to Iosif Pinelis [31] (see also Theorem 3.5 in [18]).

Lemma 5.1 (Pinelis-Hoeffding): Let $(\xi_i)_{i \in \mathbb{N}} \in \mathscr{H}$ be a martingale difference sequence in a Hilbert space $\mathscr{H}$ such that for all i almost surely $\|\xi_i\| \le c_i < \infty$. Then for all $t \in \mathbb{N}$,

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

The following proposition collects some useful estimates.

Proposition 5.2: Let $\alpha' = \alpha/(1-\theta)$. The following holds for all $t \in \mathbb{N}$,

1. $\|\Pi_k^t\| \le e^{\alpha'[k^{1-\theta}-(t+1)^{1-\theta}]}$ when $k \le t$, and the same holds for $\|\hat{\Pi}_k^t\|$;

2. For all integers $k \in [0, t]$,

$$\left\|\sum_{j=k}^{t} \Pi_{k+1}^j\right\| \le \begin{cases} 2^\theta D_\theta \alpha^{-1} k^\theta, & 1 \le k \le t; \\ D_\theta \alpha^{-1}, & k = 0. \end{cases}$$

The same also holds for $\|\sum_{j=k}^{t} \hat{\Pi}_{k+1}^t\|$;

3. $\|w^*\| \le \beta/\mu_{\min}$;

4. $\|Y_t\| \le 2\beta/\alpha$;

5. $\|w_t\| \le e^{\alpha'(1-t^{1-\theta})}\|w_1\| + 3\beta/\mu_{\min}$;

6. $\|r_t\| \le e^{\alpha'(1-t^{1-\theta})}\|w_1\| + 4\beta/\mu_{\min}$;

7. $\|\chi_t\| \le 2\mu_{\max} e^{\alpha'(1-t^{1-\theta})}\|w_1\| + 8\beta/\alpha$.

Proof: 1. By Lemma A.3-1 with $p = 1$,

$$\|\Pi_k^t\| \le \prod_{i=k}^{t}\left(1-\frac{\alpha}{i^\theta}\right) \le e^{\alpha'[k^{1-\theta}-(t+1)^{1-\theta}]}.$$

Similar to $\|\hat{\Pi}_k^t\|$.

2. By Lemma A.3-2,

$$\left\|\sum_{j=k}^{t}\Pi_{k+1}^j\right\| \le 1 + \sum_{j=k+1}^{t}\prod_{i=k+1}^{j}\left(1-\frac{\alpha}{i}\right) \le 1 + \frac{D_\theta - 1}{\alpha}(k+1)^\theta,$$

where if $k = 0$, r.h.s. $\le D_\theta \alpha^{-1}$, and if $k \ge 1$, r.h.s. $\le 2^\theta D_\theta \alpha^{-1} k^\theta$.

3. $\|w^*\| \le \|\hat{A}^{-1}\|\|\hat{B}\| \le \beta/\mu_{\min}$.

4. $\|Y_t\| = \|A_t w^* - B_t\| \le \mu_{\max}\beta/\mu_{\min} + \beta \le 2\beta/\alpha$, since $\alpha = \mu_{\min}/\mu_{\max}$.

5. By equation (8)

$$\begin{aligned} w_{t+1} &= w_t - \gamma_t(A_t w_t - B_t) \\ &= (I - \gamma_t A_t)w_t + \gamma_t B_t \\ &= \Pi_1^t w_1 + \sum_{k=1}^{t}\gamma_k \Pi_{k+1}^t B_k, \end{aligned}$$

whence

$$\begin{aligned} \|w_{t+1}\| &\le \|\Pi_1^t\|\|w_1\| + \beta\sum_{k=1}^{t-1}\gamma_k\|\Pi_{k+1}^t\| \\ &\le e^{\alpha'(1-(t+1)^{1-\theta})}\|w_1\| + \frac{3\beta}{\mu_{\min}}, \end{aligned}$$

where the last step follows from part 1 and Lemma A.3-3.

6. Since $\|r_t\| \le \|w_t\| + \|w^*\|$, using part 3 and 5 gives the result.

7. Since $\|\chi_t\| = \|(A_t - \hat{A})w_t + \hat{B} - B_t\| \le 2\mu_{\max}\|w_t\| + 2\beta$, apply part 5 and notice that $6\beta/\alpha + 2\beta \le 8\beta/\alpha$, which gives the result. ∎

Now we are ready to give the formal proofs of Theorem 3.1 and 3.2.

A. Proof of Theorem 3.1

Proof of Theorem 3.1: By equation (11) we have

$$\begin{aligned} \|r_t\| &\le \|\hat{\Pi}_1^{t-1} r_1\| + \|\sum_{k=1}^{t-1}\xi_k\| \\ &= \mathcal{E}_{init}(t) + \mathcal{E}_{samp}(t). \end{aligned}$$

The upper bound on $\mathcal{E}_{init}(t)$ follows from Proposition 5.2-1. For the upper bound on $\mathcal{E}_{samp}(t)$, by Proposition 5.2-6 with $w_1 = 0$, $\|\chi_k\| \le 8\beta/\alpha$, whence ξ_k is bounded by

$$\begin{aligned} \|\xi_k\| &\le \gamma_k\|\Pi_{k+1,t-1}\|\|\chi_k\| \\ &\le \frac{8\beta}{\mu_{\min}}\left[\frac{1}{k^\theta}\prod_{i=k+1}^{t-1}\left(1-\frac{\alpha}{i^\theta}\right)\right] = c_k. \end{aligned}$$

Applying Pinelis-Hoeffding inequality (Lemma 5.1), we obtain

$$\mathbf{Prob}\left\{\left\|\sum_{k=1}^{t-1}\xi_k\right\| \ge \epsilon\right\} \le 2\exp\left\{-\frac{\epsilon^2}{2\sum_{k=1}^{t-1}c_k^2}\right\}.$$

Let the right hand side equal δ, then

$$\epsilon^2 = 2\left(\sum_{k=1}^{t-1}c_k^2\right)\log\frac{2}{\delta} \le \frac{128\beta^2}{\mu_{\min}{}^2}\psi_\theta^2(t,\alpha)\log\frac{2}{\delta},$$

where

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

We complete the proof by applying the upper bound for $\psi_\theta^2(t,\alpha)$ in Lemma A.3-4. ∎

B. Proof of Theorem 3.2

Proof of Theorem 3.2: By equation (13) we have

$$\begin{aligned} \|\bar{r}_t\| &\le \frac{1}{t}\left\|\left(\sum_{j=0}^{t-1}\hat{\Pi}_1^j\right)r_1\right\| + \|\sum_{k=1}^{t-1}\eta_k\| \\ &= \mathcal{E}_{init}(t) + \mathcal{E}_{samp}(t). \end{aligned}$$

The initial error bound follows from Propoisition 5.2-2 with $k = 0$. As to the sample error bound, by Proposition 5.2-2 and Proposition 5.2-7 with $w_1 = 0$, we obtain

$$\begin{aligned} \|\eta_k\| &\le \frac{\gamma_k}{t}\left\|\sum_{j=k}^{t-1}\hat{\Pi}_{k+1}^j\right\|\|\chi_k\| \\ &\le \frac{2^{\theta+3}\beta D_\theta \mu_{\max}}{t\mu_{\min}{}^2} = c_\eta. \end{aligned} \tag{14}$$

Applying Pinelis-Hoeffding inequality (Lemma 5.1),

$$\mathbf{Prob}\left\{\left\|\sum_{k=1}^{t-1}\eta_k\right\| \ge \epsilon\right\} \le 2\exp\left\{-\frac{\epsilon^2}{2\sum_{k=1}^{t-1}c_\eta^2}\right\},$$

and setting the right hand to be δ, we obtain

$$\epsilon \le \sqrt{2t}c_\eta\log^{1/2}\frac{2}{\delta}.$$

The second bound follows from (14). ∎

VI. REVERSED MARTINGALE DECOMPOSITION AND PROOF OF THEOREM B*

In this section we give a proof of Theorem B*. Note that in the martingale decompostion in Section IV, $\chi_t = (A_t - \hat{A})r_t + \hat{B} - B_t$ whose variance grows in proportion to $\|r_t\|^2$, whence there is no improvement replacing Hoeffding's Inequality by Markov's inequality. However, we may avoid this by turning to the remainder decomposition used in [1] where we directly deal with the variance, $\sigma^2 = \mathbb{E}\|Y_t\|^2$. Yet this approach leads to a reversed martingale decomposition for remainders, as we shall see soon.

The following lemma is taken from [1], whose proof is included here for completeness.

Lemma 6.1: For all $t \in \mathbb{N}$,

$$r_t = \Pi_1^{t-1} r_1 - \sum_{k=1}^{t-1} \gamma_k \Pi_{k+1}^{t-1} Y_k.$$

Proof: Note that

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

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

It leads to the following decomposition for the averaged remainder.

Lemma 6.2: For all $t \in \mathbb{N}$,

$$\bar{r}_t = \frac{1}{t}\left(\sum_{j=0}^{t-1} \Pi_1^j\right) r_1 - \sum_{k=1}^{t-1} \frac{\gamma_k}{t}\left(\sum_{j=k}^{t-1} \Pi_{k+1}^j\right) Y_k.$$

For $k \in \mathbb{Z}$, define

$$\eta_k = \begin{cases} \dfrac{\gamma_k}{t}\left(\displaystyle\sum_{j=k}^{t-1} \Pi_{k+1}^j\right) Y_k, & 1 \leq k \leq t; \\ 0, & \text{otherwise.} \end{cases}$$

Recall that a sequence of random variables (x_k) is called a *reversed martingale difference sequence* if (x_{-k}) is a martingale difference sequence. Then (η_k) is a reversed martingale difference sequence; since it depends on $\{z_k, \ldots, z_{t-1}\}$ and $\mathbb{E}_{z_k|z_{k+1},\ldots,z_{t-1}}[\eta_k] = 0$, which implies that (η_{-k}) is a martingale difference sequence.

We will use the following well-known Markov's Inequality.

Lemma 6.3 (Markov): Let X be a nonnegative random variable. Then for any real number $\epsilon > 0$, we have

$$\mathbf{Prob}\{X \geq \epsilon\} \leq \frac{\mathbb{E}[X]}{\epsilon}.$$

Theorem 6.4: Let $\gamma_t = t^{-\theta}/\mu_{\max}$ ($\theta \in [0,1)$) and $w_1 = 0$. Define $\alpha = \mu_{\min}/\mu_{\max} \in (0,1]$. Then the following holds for all $t \in \mathbb{N}$,

$$\|\bar{w}_t - w^*\| \leq \mathscr{E}_{init}(t) + \mathscr{E}_{samp}(t).$$

Here

$$\mathscr{E}_{init}(t) \leq D_\theta \left(\frac{1}{\alpha t}\right) \|r_1\|,$$

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

$$\mathscr{E}_{samp}(t) \leq \frac{2^\theta D_\theta \sigma}{\sqrt{\delta} \mu_{\max}} \left(\frac{1}{\alpha}\right) \sqrt{\frac{1}{t}}.$$

Proof: The initial error bound follows from Propoistion 5.2-2 with $k = 0$.

As to the sample error, note that

$$
\begin{aligned}
\mathbb{E}\Big\|\sum_{k=1}^{t-1} \eta_k\Big\|^2 &\leq \sum_{k=1}^{t-1} \frac{\gamma_k^2}{t^2} \mathbb{E}\Big\|\sum_{j=k}^{t-1} \Pi_{k=k+1}^j\Big\|^2 \mathbb{E}\|Y_k\|^2 \\
&\leq \frac{2^{2\theta} D_\theta^2 \sigma^2}{\mu_{\max}{}^2 \alpha^2} t^{-1}.
\end{aligned}
$$

where the last is due to Proposition 5.2-2 and Finiteness Condition-C. The sample error bound then follows from Markov inequality by taking $X = \|\sum_{k=1}^{t-1} \eta_k\|^2$. ∎

Proof of Theorem B:* Setting $\mu_{\max} = \lambda + C_K^2$, $\mu_{\min} = \lambda$, $\alpha = \lambda/(\lambda + C_K^2)$, and $\sigma = \sigma_\lambda$, the result follows from Theorem 6.4. ∎

The following proposition gives an estimate of σ_λ.

Proposition 6.5: 1. $\|f_\lambda^*\|_K \leq M_\rho/\sqrt{\lambda}$;
2. $\|f_\lambda^*\|_{\mathscr{L}_\rho^2} \leq 2M_\rho$;
3. $\sigma_\lambda \leq M_\rho \sqrt{5(\lambda + C_K^2)}$.

Proof: 1. Note that

$$f_\lambda^* = \arg\min_{f \in \mathscr{H}_K} \|f - f_\rho\|_{\mathscr{L}_\rho^2}^2 + \lambda\|f\|_K^2.$$

Taking $f = 0$, we have

$$\|f_\lambda^* - f_\rho\|_{\mathscr{L}_\rho^2}^2 + \lambda\|f_\lambda^*\|_K^2 \leq \|f_\rho\|_{\mathscr{L}_\rho^2}^2 \leq M_\rho^2, \quad (15)$$

which leads to the result.

2. From (15), we obtain $\|f_\lambda^* - f_\rho\|_{\mathscr{L}_\rho^2} \leq M_\rho$. The result then follows from

$$\|f_\lambda^*\|_{\mathscr{L}_\rho^2} \leq \|f_\lambda^* - f_\rho\|_{\mathscr{L}_\rho^2} + \|f_\rho\|_{\mathscr{L}_\rho^2} \leq 2M_\rho.$$

3. Note that $\mathbb{E}[yK_x] = L_K f_\rho$ and $\mathbb{E}[f_\lambda^*(x)K_x] = L_K f_\lambda^*$. Then

$$
\begin{aligned}
\sigma_\lambda^2 &= \mathbb{E}\|(y - f_\lambda^*(x))K_x - \lambda f_\lambda^*\|_K^2 \\
&= \mathbb{E}[(f_\lambda^*(x) - y)^2 K(x,x)] \\
&\quad + 2\lambda\langle f_\lambda^*, L_K(f_\lambda^* - f_\rho)\rangle_K + \lambda^2\|f_\lambda^*\|_K^2
\end{aligned}
$$

where

$$
\begin{aligned}
&\mathbb{E}[(f_\lambda^*(x) - y)^2 K(x,x)] \\
&\leq C_K^2(\|f_\lambda^* - f_\rho\|_{\mathscr{L}_\rho^2}^2 + \mathbb{E}(f_\rho(x) - y)^2) \\
&\leq 5C_K^2 M_\rho^2,
\end{aligned}
$$

$$
\begin{aligned}
&\langle f_\lambda^*, L_K(f_\lambda^* - f_\rho)\rangle_K \\
&\leq \|L_K^{1/2} f_\lambda^*\|_K \cdot \|L_K^{1/2}(f_\lambda^* - f_\rho)\|_K \\
&= \|f_\lambda^*\|_{\mathscr{L}_\rho^2} \cdot \|f_\lambda^* - f_\rho\|_{\mathscr{L}_\rho^2} \\
&\leq 2M_\rho^2,
\end{aligned}
$$

and $\lambda^2\|f_\lambda^*\|_K^2 \leq \lambda M_\rho^2$. Thus we end the proof. ∎

VII. CONCLUSION AND OPEN PROBLEMS

In this paper, we have shown by probabilistic upper bounds that a two-stage online learning algorithm, the stochastic approximation of the gradient descent method followed by an averaging process, can achieve the almost sure convergence with an optimal asymptotic rate with respect to the sample size, as good as "batch learning". Moreover considering the regularization parameter and confidence, the best results obtained so far are, $O(\lambda^{-2}t^{-1/2}\log^{1/2}1/\delta)$ (Theorem B) or $O(\lambda^{-1}t^{-1/2}\delta^{-1/2})$ (Theorem B*).

Thus it is still an open problem, *if we can achieve* $O(\lambda^{-1}t^{-1/2}\log^{1/2}1/\delta)$, *the optimal rate known in "batch learning".*

APPENDIX A: SOME ESTIMATES BASED ON GAMMA FUNCTION

Lemma A.1: Let $\theta \in [0,1)$, $a > 0$ and $t \geq 2$. Then for any $\tau \in \mathbb{R}$,

$$e^{-at^{1-\theta}}\int_1^t x^{-(\theta+\tau)}e^{ax^{1-\theta}}dx = O(t^{-\tau}).$$

In fact, if $\tau \geq 0$,

$$A_{\theta,a}t^{-\tau} \leq e^{-at^{1-\theta}}\int_1^t x^{-(\theta+\tau)}e^{ax^{1-\theta}}dx \leq A'_{\theta,\tau,a}t^{-\tau},$$

and if $\tau < 0$,

$$B_{\theta,\tau,a}t^{-\tau} \leq e^{-at^{1-\theta}}\int_1^t x^{-(\theta+\tau)}e^{ax^{1-\theta}}dx \leq B'_{\theta,a}t^{-\tau}.$$

Here

$$A_{\theta,a} = \frac{1-e^{-a(2^{1-\theta}-1)}}{a(1-\theta)},$$

$$A'_{\theta,\tau,a} = \frac{2^{\tau/(1-\theta)}}{a(1-\theta)}\left(1+a^{-\tau/(1-\theta)}\Gamma\left(\frac{1+\tau-\theta}{1-\theta}\right)\right),$$

$$B_{\theta,\tau,a} = \frac{2^{\tau/(1-\theta)}(1-e^{-a})}{a(1-\theta)}, \quad \text{and} \quad B'_{\theta,a} = \frac{1}{a(1-\theta)}.$$

Proof: Let $y = t^{1-\theta} - x^{1-\theta}$. Then

$$e^{-at^{1-\theta}}\int_1^t x^{-(\theta+\tau)}e^{ax^{1-\theta}}dx \qquad (A-1)$$

$$= \frac{1}{1-\theta}\int_0^{t^{1-\theta}-1}(t^{1-\theta}-y)^{-\tau/(1-\theta)}e^{-ay}dy$$

$$= \frac{t^{-\tau}}{1-\theta}\int_0^{t^{1-\theta}-1}\left(1-\frac{y}{t^{1-\theta}}\right)^{-\tau/(1-\theta)}e^{-ay}dy.$$

1. (*For $A_{\theta,a}$*) For $0 \leq y \leq t^{1-\theta}-1$, $1-\frac{y}{t^{1-\theta}} \leq 1$. Thus if $\tau \geq 0$, equation (A-1) has

$$r.h.s. \geq \frac{t^{-\tau}}{1-\theta}\int_0^{t^{1-\theta}-1}e^{-ay}dy$$

$$= \frac{t^{-\tau}}{a(1-\theta)}(1-e^{-a(t^{1-\theta}-1)})$$

$$\geq \frac{t^{-\tau}}{a(1-\theta)}(1-e^{-a(2^{1-\theta}-1)}), \quad t \geq 2.$$

2. (*For $B'_{\theta,a}$*) Similarly if $\tau < 0$, equation (A-1) has

$$r.h.s. \leq \frac{t^{-\tau}}{1-\theta}\int_0^{t^{1-\theta}-1}e^{-ay}dy$$

$$= \frac{t^{-\tau}}{a(1-\theta)}(1-e^{-a(t^{1-\theta}-1)})$$

$$\leq \frac{t^{-\tau}}{a(1-\theta)}.$$

3. (*For $A'_{\theta,\tau,a}$*) Note that for $0 \leq y \leq t^{1-\theta}-1$, $s = y/t^{1-\theta} \in (0,1)$, whence

$$\frac{1}{1-s} = 1+\sum_{n=1}^{\infty}s^n = 1+\frac{s}{1-s}$$

$$\leq 1+\frac{y/t^{1-\theta}}{1-(t^{1-\theta}-1)/t^{1-\theta}}$$

$$= 1+y. \qquad (A-2)$$

Thus for $\tau > 0$ the right hand side of (A-1) is bounded by

$$r.h.s. \leq \frac{t^{-\tau}}{1-\theta}\int_0^{t^{1-\theta}-1}(1+y)^{\tau/(1-\theta)}e^{-ay}dy$$

$$\leq \frac{2^{\tau/(1-\theta)}t^{-\tau}}{1-\theta}\int_0^1 e^{-ay}dy$$

$$+ \frac{2^{\tau/(1-\theta)}t^{-\tau}}{1-\theta}\int_1^{t^{1-\theta}-1}y^{\tau/(1-\theta)}e^{-ay}dy$$

$$\leq \frac{2^{\tau/(1-\theta)}}{1-\theta}t^{-\tau}\left\{\int_0^{\infty}e^{-ay}dy\right.$$

$$\left.+ \int_0^{\infty}y^{(1+\tau-\theta)/(1-\theta)-1}e^{-ay}dy\right\}$$

$$\leq \frac{2^{\tau/(1-\theta)}}{a(1-\theta)}(1$$

$$+ a^{-\tau/(1-\theta)}\Gamma\left(\frac{1+\tau-\theta}{1-\theta}\right))t^{-\tau}.$$

4. (*For $B_{\theta,\tau,a}$*) By equation (A-2), for $\tau < 0$ the right hand side of (A-1) is bounded by

$$r.h.s. \geq \frac{t^{-\tau}}{1-\theta}\int_0^{t^{1-\theta}-1}(1+y)^{\tau/(1-\theta)}e^{-ay}dy$$

$$\geq \frac{2^{\tau/(1-\theta)}t^{-\tau}}{1-\theta}\int_0^1 e^{-ay}dy$$

$$+ \frac{2^{\tau/(1-\theta)}t^{-\tau}}{1-\theta}\int_0^{t^{1-\theta}-1}y^{\tau/(1-\theta)}e^{-ay}dy$$

$$\geq \frac{2^{\tau/(1-\theta)}t^{-\tau}}{1-\theta}\int_0^1 e^{-ay}dy$$

$$= \frac{2^{\tau/(1-\theta)}(1-e^{-a})}{a(1-\theta)}t^{-\tau}.$$

This completes the proof. ∎

Lemma A.2: Let $\theta \in (0,1)$. Then

$$C_\theta t^\theta \leq e^{t^{1-\theta}}\int_t^{\infty}e^{-x^{1-\theta}}dx \leq C'_\theta t^\theta,$$

where $C_\theta = 1/(1-\theta)$ and $C'_\theta = 2^{\theta/(1-\theta)}(1+\Gamma(1/(1-\theta)))$.

Proof: 1. *Lower bound.* Consider the continuous function

$$f(x) = x^{1-\theta}.$$

By the mean value theorem, when $x \geq t > 0$, there exists a $\zeta \in (t, x)$ such that

$$
\begin{aligned}
f(t) - f(x) &= f'(\zeta)(t-x) = (1-\theta)\zeta^{-\theta}(t-x) \\
&\geq -(1-\theta)x^{-\theta}(x-t),
\end{aligned}
$$

whence

$$
\begin{aligned}
& e^{t^{1-\theta}} \int_t^\infty e^{-x^{1-\theta}} dx \\
&\geq \int_t^\infty e^{-(1-\theta)t^{-\theta}(x-t)} dx \\
&= e^{(1-\theta)t^{1-\theta}} \int_t^\infty e^{-(1-\theta)t^{-\theta}x} dx \\
&= \frac{t^\theta}{1-\theta}.
\end{aligned}
$$

2. *Upper bound.* It is enough to show that for $x \geq 1$ and $a \geq 1$,

$$\Gamma(a, x) \leq G_a e^{-x} x^{a-1}, \qquad G_a = 2^{a-1}(1 + \Gamma(a)). \quad \text{(A-3)}$$

If this is true, the result follows from setting $a = 1/(1-\theta) \geq 1$, $C'_\theta = G_{1/(1-\theta)}$, and replacing x by $t^{1-\theta}$.

To show (A-3), by setting $s = x + \tau$,

$$
\begin{aligned}
\Gamma(a, x) &= \int_x^\infty s^{a-1} e^{-s} ds \\
&= x^{a-1} e^{-x} \int_0^\infty e^{-\tau}(1 + \tau/x)^{a-1} d\tau, \\
&\leq x^{a-1} e^{-x} \int_0^\infty e^{-\tau}(1 + \tau)^{a-1} d\tau, \\
&\quad \text{(by } x \geq 1 \text{ and } a \geq 1\text{)}, \\
&\leq x^{a-1} e^{-x} \left\{ 2^{a-1} \int_0^1 e^{-\tau} d\tau \right. \\
&\quad \left. + 2^{a-1} \int_1^\infty e^{-\tau} \tau^{a-1} d\tau \right\} \\
&\leq 2^{a-1}(1 + \Gamma(a)) x^{a-1} e^{-x}.
\end{aligned}
$$

This completes the proof. ∎

Lemma A.3: 1. For $\alpha \in (0, 1]$, $p > 0$, and $\theta \in [0, 1]$,

$$
\prod_{i=k}^{t-1} \left(1 - \frac{\alpha}{i^\theta}\right)^p
$$

$$
\leq \begin{cases}
\exp\left\{ \dfrac{\alpha p}{1-\theta}(k^{1-\theta} - t^{1-\theta}) \right\}, & \theta \in [0, 1) \\
\left(\dfrac{k}{t}\right)^{\alpha p}, & \theta = 1
\end{cases}
$$

2. For $\alpha \in (0, 1]$, $\theta \in [0, 1)$, and all $t \in \mathbb{N}$,

$$\psi_\theta^0(t, k, \alpha) := \sum_{j=k}^{t-1} \prod_{i=k}^j \left(1 - \frac{\alpha}{i^\theta}\right) \leq \frac{D_\theta - 1}{\alpha} k^\theta;$$

3. For $\alpha \in (0, 1]$, $\theta \in [0, 1]$, and all $t \in \mathbb{N}$,

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

4. For $\alpha \in (0, 1]$, $\theta \in [0, 1)$, and all $t \in \mathbb{N}$,

$$
\begin{aligned}
\psi_\theta^2(t, \alpha) &:= \sum_{k=1}^{t-1} \frac{1}{k^{2\theta}} \prod_{i=k+1}^{t-1} \left(1 - \frac{\alpha}{i^\theta}\right)^2 \\
&\leq 2D_\theta \left(\frac{1}{\alpha}\right)^{\frac{1}{1-\theta}} \left(\frac{1}{t}\right)^\theta.
\end{aligned}
$$

Proof: The following fact will be used repeatedly in the proof,

$$\ln(1 + x) \leq x, \qquad \text{for all } x > -1. \quad \text{(A-4)}$$

1. By the inequality (A-4), we have for $\theta \in [0, 1]$,

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

Thus

$$
\begin{aligned}
\sum_{i=k}^{t-1} \ln\left(1 - \frac{\alpha}{i^\theta}\right)^p &\leq -\alpha p \sum_{i=k}^{t-1} \frac{1}{i^\theta} \\
&\leq -\alpha p \int_k^t \frac{1}{x^\theta} dx
\end{aligned}
$$

which equals

$$\frac{\alpha p}{1-\theta}\left(k^{1-\theta} - t^{1-\theta}\right),$$

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

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

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

2. Notice that Let $\alpha' = \alpha/(1-\theta)$. Using part 1 with $p = 1$, we obtain

$$\prod_{i=k}^j \left(1 - \frac{\alpha}{i^\theta}\right) \leq e^{\alpha'[k^{1-\theta} - (j+1)^{1-\theta}]},$$

whence

$$
\begin{aligned}
& \sum_{j=k}^{t-1} \prod_{i=k}^j \left(1 - \frac{\alpha}{i^\theta}\right) \\
&\leq \sum_{j=k}^{t-1} e^{\alpha'[k^{1-\theta} - (j+1)^{1-\theta}]} \\
&\leq e^{\alpha' k^{1-\theta}} \int_k^\infty e^{-\alpha' x^{1-\theta}} dx \\
&\leq (\alpha')^{-1/(1-\theta)} C'_\theta [(\alpha')^{1/(1-\theta)} k]^\theta, \\
&\quad \text{(by Lemma A.2)} \\
&\leq 2^{\theta/(1-\theta)} \left\{1 + \Gamma\left(\frac{1}{1-\theta}\right)\right\} \left(\frac{1}{\alpha}\right) k^\theta \\
&= \frac{D_\theta - 1}{\alpha} k^\theta.
\end{aligned}
$$

3. Notice that

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

The first term is bounded by $1/\alpha$ for $t > 1$. It is sufficient to show the second term is bounded by $2/\alpha$. To see this, we consider seperately two cases $\theta \in [0, 1)$ and $\theta = 1$.

If $\theta \in [0, 1)$, from part 1 with $p = 1$, we have

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

$$\leq \quad e^{-\frac{\alpha}{1-\theta} t^{1-\theta}} \sum_{k=1}^{t-2} \frac{1}{k^\theta} e^{\frac{\alpha}{1-\theta}(k+1)^{1-\theta}}$$

where

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

$$\leq \quad 2^\theta \sum_{k=1}^{t-2} \left(\frac{1}{k+1}\right)^\theta e^{\frac{\alpha}{1-\theta}(k+1)^{1-\theta}}$$

$$\leq \quad 2 \int_1^t e^{\frac{\alpha}{1-\theta} x^{1-\theta}} x^{-\theta} dx$$

$$\leq \quad \frac{2}{\alpha} e^{\frac{\alpha}{1-\theta} t^{1-\theta}},$$

as desired.

If $\theta = 1$, from part 1 ($p = 1$),

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

$$\leq \quad \sum_{k=1}^{t-2} \frac{1}{k} \left(\frac{k+1}{t}\right)^\alpha$$

$$\leq \quad \frac{2}{t^\alpha} \sum_{k=1}^{t-2} \frac{(k+1)^\alpha}{k+1}$$

$$\leq \quad \frac{2}{t^\alpha} \int_1^t x^{\alpha-1} dx,$$

where if $\alpha = 1$,

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

and if $0 < \alpha < 1$,

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

4. Notice that

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

The first term is bounded by

$$\frac{1}{(t-1)^{2\theta}} \leq \frac{2^{2\theta}}{t^{2\theta}}.$$

Below we are going to give an upper bound on the second term. Let $\alpha' = \alpha/(1-\theta)$. By part 1 with $p = 2$,

$$\prod_{i=k+1}^{t-1} \left(1 - \frac{\alpha}{i^\theta}\right)^2 \leq \exp\{2\alpha'[(k+1)^{1-\theta} - t^{1-\theta}]\}.$$

Then

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

$$\leq \quad 2^{2\theta} \sum_{k=1}^{t-2} \frac{1}{(k+1)^{2\theta}} e^{2\alpha'[(k+1)^{1-\theta} - t^{1-\theta}]}$$

$$\leq \quad 2^{2\theta} e^{-2\alpha' t^{1-\theta}} \int_1^t x^{-2\theta} e^{2\alpha' x^{1-\theta}} dx$$

$$\leq \quad \frac{2^{\theta/(1-\theta)+2\theta-1}}{\alpha'(1-\theta)} \Big\{ 1 + (2\alpha')^{-\theta/(1-\theta)}$$

$$\cdot \Gamma\left(\frac{1}{1-\theta}\right) \Big\} t^{-\theta}$$

(by Lemma A.1 with $\tau = \theta$).

where by

$$(2\alpha')^{-\theta/(1-\theta)} \leq \left(\frac{1}{\alpha}\right)^{\theta/(1-\theta)},$$

we obtain

$$r.h.s. \leq 2^{\theta/(1-\theta)+2\theta-1} \left(1 + \Gamma\left(\frac{1}{1-\theta}\right)\right) \left(\frac{1}{\alpha}\right)^{1/(1-\theta)}.$$

Combining two terms together, we obtain

$$\psi_\theta^2(t, \alpha)$$

$$\leq \quad 2^{2\theta} t^{-2\theta} + 2^{\theta/(1-\theta)+2\theta-1} \left\{ 1 + \Gamma\left(\frac{1}{1-\theta}\right) \right\}$$

$$\cdot \left(\frac{1}{\alpha}\right)^{1/(1-\theta)} t^{-\theta}$$

$$\leq \quad 2^\theta \left\{ \alpha^{1/(1-\theta)} (2/t)^\theta + 2^{\theta/(1-\theta)} \left[1 + \Gamma\left(\frac{1}{1-\theta}\right) \right] \right\}$$

$$\cdot \left(\frac{1}{\alpha}\right)^{1/(1-\theta)} t^{-\theta}$$

$$\leq \quad 2 D_\theta \left(\frac{1}{\alpha}\right)^{1/(1-\theta)} t^{-\theta},$$

for $t \geq 2$, as desired. For $t = 1$, we complete the proof by noting that $\psi_\theta^2(1, \alpha) = 0$. ∎

REFERENCES

[1] S. Smale and Y. Yao, "Online learning algorithms," *Foundation of Computational Mathematics*, 2004, submitted.

[2] N. Cesa-Bianchi, A. Conconi, and C. Gentile, "On the generalization ability of on-line learning algorithms," *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 2050–2057, 2004.

[3] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bull. of the Amer. Math. Soc.*, vol. 29, no. 1, pp. 1–49, 2002.

[4] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances of Computational Mathematics*, vol. 13, no. 1, pp. 1–50, 1999.

[5] T. Poggio and S. Smale, "The mathematics of learning: Dealing with data," *Notices of the AMS*, vol. 50, no. 5, pp. 537–544, 2003.

[6] D.-X. Zhou, "Capacity of reproducing kernel spaces in learning theory," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1743–1752, 2003.

[7] S. Smale and D.-X. Zhou, "Learning theory estimates via integral operators and their approximations," *to appear*, 2005.

[8] A. Caponnetto and E. D. Vito, "Fast rates for regularized least squares algorithm," *CBCL Paper/AI Memo*, 2005, preprint.

[9] E. De Vito, L. Rosasco, A. Caponnetto, U. D. Giovannini, and F. Odone, "Learning from examples as an inverse problem," *Journal of Machine Learning Research*, 2004, to appear.

[10] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*. Kluwer Academic Publishers, 2000.

[11] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[12] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, vol. 23, pp. 462–466, 1952.

[13] H. Robbins and D. Siegmund, "A convergence theorem for nonnegative almost supermartingales and some applications," in *Optimizing Methods in Statistics*, J. S. Rustagi, Ed. Academic Press, New York, 1971, pp. 233–257.

[14] M. Duflo, "Cibles atteignables avec une probabilité positive d'aprés m. benaim," *Unpublished manuscript*, 1997.

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

[16] M. Duflo, *Algorithmes Stochastiques*. Berlin, Heidelberg: Springer-Verlag, 1996.

[17] H. J. Kushner and G. G. Yin, *Stochastic Approximations and Recursive Algorithms and Applications*. Berlin, Heidelberg: Springer-Verlag, 2003.

[18] I. Pinelis, "Optimum bounds for the distributions of martingales in banach spaces," *The Annals of Probability*, vol. 22, no. 4, pp. 1679–1706, 1994.

[19] B. Widrow and M. Hoff, "Adaptive switching circuits," *IRE WESCON Convention Record*, no. 4, pp. 96–104, 1960.

[20] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge Unversity Press, 2000.

[21] B. T. Polyak, "New method of stochastic approximation type," *Automation and Remote Control*, vol. 51, pp. 937–946, 1990.

[22] D. Ruppert, "Efficient estimators from a slowly convergent robbins-monro procedure," Technical Report 781, School of Operations Research and Industrial Engineering, Cornell University, Tech. Rep., 1988.

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

[24] S. Smale and D.-X. Zhou, "Shannon sampling and function reconstruction from point values," *Bull. of the Amer. Math. Soc.*, vol. 41, no. 3, pp. 279–305, 2004.

[25] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM Journal of Control and Optimization*, vol. 30, no. 4, pp. 835–855, 1992.

[26] V. R. Konda and J. N. Tsitsiklis, "Convergence rate of linear two-time-scale stochastic approximation," *The Annals of Applied Probability*, vol. 14, no. 2, pp. 796–819, 2004.

[27] F. Cucker and S. Smale, "Best choices for regularization parameters in learning theory," *Foundations Comput. Math.*, vol. 2, no. 4, pp. 413–428, 2002.

[28] E. De Vito, A. Caponnetto, and L. Rosasco, "Model selection for regularized least-squares algorithm in learning theory," *Foundations of Computational Mathematics*, 2004, to appear.

[29] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, no. 2, pp. 499–526, 2002.

[30] T. Zhang, "Leave-one-out bounds for kernel methods," *Neural Computation*, vol. 15, pp. 1397–1437, 2003.

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