ONLINE LEARNING AS STOCHASTIC APPROXIMATIONS OF REGULARIZATION PATHS

PIERRE TARRÈS AND YUAN YAO

Abstract. In this paper, we study an online learning algorithm as stochastic approximations of a regularization path convergent to the regression function. Some probabilistic upper bounds are given for the convergence of the algorithm, under certain regularity assumption on the regression function. In the case of a strong convergence (convergence in reproducing kernel Hilbert spaces), the convergence rate obtained is the same as the best known rate in batch learning; and in the case of a weak convergence (convergence in mean square distance), the convergence rate is optimal in the sense that it reaches the minimax and the individual lower rates.

1. Introduction

Consider the following problem of learning from examples: given a sequence of random examples $(z_t = (x_t, y_t))_{t\in\mathbb{N}}$ drawn from a probability measure ρ on $\mathscr{X} \times \mathscr{Y}$, one wants to approximate the *regression function*, $f_\rho(x) := \int_{\mathscr{Y}} y d\rho_{\mathscr{Y}|x}$, *i.e.* the conditional expectation of y given x. In an optimization view, f_ρ minimizes the following quadratic functional

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

However, minimizing (1) without constraints may put f_ρ in a too large space to search. So one typically turns to some regularization methods to solve (1). In this paper, we focus on the following regularized least square problem

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

where $\mathscr{H}$ is some Hilbert space and $\lambda > 0$ is the regularization parameter. This scheme is well-known as ridge regression in statistics and is also called Tikhonov regularization in inverse problems [Engl, Hanke, and Neubauer 2000]. One may choose a suitable $\mathscr{H}$, such that when $\lambda \to 0$, $f_\lambda \to f_\rho$. The map $f_\lambda : \mathbb{R}_+ \to \mathscr{H}$ defined by $\lambda \mapsto f_\lambda$, characterizes a function path in $\mathscr{H}$, which may start from some f_{λ_0} and go toward f_ρ as $\lambda \to 0$. Therefore it is called here a *regularization path* of f_ρ in $\mathscr{H}$.

Our purpose is to seek a sequence of functions $(f_t)_{t\in\mathbb{N}} \in \mathscr{H}$ with certain dependence on the examples up to time t, and a sequence of regularization parameters (λ_t) tending to 0, such that f_t follows f_{λ_t} closely and converges to f_ρ. It is crucial to restrict the permissible form of the

Date: April 29, 2006.

2000 *Mathematics Subject Classification.* 62L20, 68Q32, 68T05.

Key words and phrases. Online Learning, Stochastic Approximations, Regularization Path, Reproducing Kernel Hilbert Space.

This work was supported by NSF grant 0325113.

dependence of f_t on historical examples. For example, f_t may depend explicitly on all the available examples at time t, $(z_i)_{i=1}^t$, which is often called as *batch learning*. In this setting existing results [e.g. Cucker and Smale 2002; Smale and Zhou 2005] already give such a sequence f_t, which is the minimizer of an empirical counterpart of (2) by replacing the integral with a sum over the sample set at time t.

In this paper, our departure from those batch learning results lies on the following recursive structure imposed on (f_t): for each t, f_t depends on the example z_t and f_{t-1} which only depends on previous examples $z_1, \ldots, z_{t-1}$, i.e. $f_t = T_t(f_{t-1}, z_t)$ for some map $T_t : \mathscr{H} \times \mathscr{X} \to \mathscr{H}$. Such a sequence (f_t) is called an *online learning* sequence, as a contrast to the *batch learning*. In particular, in online learning, the sample size t is changing over time, thus one can not choose a fixed regularization parameter based on *a priori* knowledge on a fixed sample size as in batch learning. This feature forces online learning to track the entire regularization path, which increases the technical difficulty in the treatment of online learning vs. batch learning. We note that recent works on support vector machines [Hastie, Rosset, Tibshirani, and Zhu 2004] enables one to construct batch learning algorithms to follow the entire regularization path with the same amount of computational cost as a single fixed regularization.

An attractive feature of this online learning scheme, lies in that its computational complexity to follow the entire regularization path may be as small as linear $O(t)$ (the algorithm in this paper, however, requires $O(t^2)$ in the worst case). In a contrast, the batch learning scheme typically involves inverting a matrix, which is $O(t^3)$.

The construction of such an online learning sequence (f_t), as we shall see soon, is a stochastic approximation of the gradient descent method to solve (2), for each fixed λ_t. In this way, our algorithms can be regarded as stochastic approximations of the regularization path f_λ. For fixed regularization parameters, $\lambda_t = \lambda$, this kind of online learning algorithms in a similar setting has been studied in [Smale and Yao 2005], with improved upper bounds in [Yao 2005]. References [Duflo 1996; Kushner and Yin 2003] provide more background on stochastic approximations.

As in previous results, in this paper we choose $\mathscr{H}$ a reproducing kernel Hilbert spaces (RKHS), $\mathscr{H}_K$. RKHS enables us to develop results in a coordinate-free way, where the gradiend descent method takes an especially simple form [e.g. Kivinen, Smola, and Williamson 2004]. Moreover, RKHS provides an unified framework to include several important settings, e.g. (1) generalized smooth spline functions in Sobelev spaces [Wahba 1990], (2) real analytic functions with bounded bandwidth [Daubechies 1992] and their generalizations [Smale and Zhou 2004], (3) an unified framework for stochastic processes [Loéve 1948; Parzen 1961]. By choosing suitable kernels, $\mathscr{H}_K$ can be used to approximate any functions in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$, the square integrable functions with respect to the marginal probability measure $\rho_{\mathscr{X}}$. For a wider background on RKHS, see for example [Berlinet and Thomas-Agnan 2004].

In this paper, we present some probabilistic upper bounds for the convergence of $(f_t)_{t\in\mathbb{N}}$ to f_ρ, in $\mathscr{H}_K$ or $\mathscr{L}^2_{\rho_{\mathscr{X}}}$, under the assumption of $f_\rho \in \mathscr{H}_K$ with additional regularity. The convergence rate in $\mathscr{H}_K$ is shown the same as the best known rate in batch learning [Smale and Zhou 2005]; and the convergence rate in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$, is optimal in the sense that it reaches the minimax and individual lower rate.

Our treatment starts from more general Hilbert spaces, with stochastic approximations of the solutions for a sequence of linear operator equations, which gives the main results in this paper when specialized to the setting in $\mathscr{H}_K$. Two structural decomposition theorems, namely the *reversed*

martingale decomposition and *the martingale decomposition*, play an important role in the proof of the main results, where the former is suitable for the strong convergence in $\mathscr{H}_K$ and the latter, by exploiting the spectral decomposition, is suitable for the weak convergence in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$. A crucial estimate is about the drift along the regularization path, $\|f_{\lambda_t} - f_{\lambda_{t-1}}\|$, which has the same order as the approximation error $\|f_{\lambda_t} - f_\rho\|$. This key estimates lead to the same rate in online learning as in batch learning.

An *open problem* is also left in this paper: exponential concentration inequalities can not be applied to bound the weak convergence in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$, without sacrificing the rates. To overcome this issue, one might need a tighter bound on the growth of f_t in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$, which is not available at this moment up to the authors' knowledge. We leave this problem to further investigations.

The organization of this paper is as follows. Section 2 collects the main results in this paper. Section 3 studies stochastic approximations of regularization paths for linear operator equations in general Hilbert spaces, where two crucial structural decompositions are presented. Section 4 collects some estimates on the drifts along the regularization path, $\|f_\lambda - f_\mu\|$ $(\lambda, \mu > 0)$, which is the basis for later developments. Section 5 studies the upper bounds for convergence in $\mathscr{H}_K$ and Section 6 studies the upper bounds for convergence in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$. Appendix A derives a probabilistic inequality from the Pinelis-Bernstein inequality, which is used to derive the probabilistic upper bounds in this paper. Appendix B collects some basic estimates. Appendix C gives some estimates on the gap between the online learning sequence and the regularization path. Appendix D gives the proof of Theorem 3.6.

2. Main Results

2.1. **Notations and Assumptions.** Let $\mathscr{X} \subseteq \mathbb{R}^n$ be closed, $\mathscr{Y} = \mathbb{R}$ and $\mathscr{Z} = \mathscr{X} \times \mathscr{Y}$. Let ρ be a probability measure on $\mathscr{Z}$, $\rho_{\mathscr{X}}$ be the induced marginal probability measure on $\mathscr{X}$, and $\rho_{\mathscr{Y}|x}$ be the conditional probability measure on $\mathscr{Y}$ with respect to $x \in X$. Define $f_\rho : \mathscr{X} \to \mathscr{Y}$ by $f_\rho(x) = \int_{\mathscr{Y}} y d\rho_{\mathscr{Y}|x}$, the *regression function of* ρ. In the sequel, we denote by $\mathbb{E}[\ ]$ the expectation.

Let $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ be the Hilbert space of square integrable functions with respect to $\rho_{\mathscr{X}}$. In the sequel $\|\ \|_\rho$ denotes the norm in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$.

Let $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ be a *Mercer kernel*, i.e. a continuous symmetric real function which is *positive semi-definite* in the sense that $\sum_{i,j=1}^m c_i c_j K(x_i, x_j) \geq 0$ for any $m \in \mathbb{N}$ and any choice of $x_i \in X$ and $c_i \in \mathbb{R}$ $(i = 1, \ldots, m)$. A Mercer kernel K induces a function $K_x : \mathscr{X} \to \mathbb{R}$ $(x \in \mathscr{X})$ defined by $K_x(x') = K(x, x')$. Let $\mathscr{H}_K$ be the *reproducing kernel Hilbert space* (RKHS) associated with a Mercer kernel K, i.e. the completion of the span$\{K_x : x \in \mathscr{X}\}$ with respect to the following inner product: the unique linear extension of the bilinear form $\langle K_x, K_{x'} \rangle_K = K(x, x')$ $(x, x' \in \mathscr{X})$. The norm of $\mathscr{H}_K$ is denoted by $\|\ \|_K$. The most important property of RKHS is the *reproducing property*: for all $f \in \mathscr{H}_K$ and $x \in X$, $f(x) = \langle f, K_x \rangle_K$.

Throughout this paper, assume that

Finiteness Condition. (A) There exists a constant $\kappa \geq 0$ such taht

$$\kappa := \sup_{x \in \mathscr{X}} \sqrt{K(x, x)} < \infty.$$

(B) There exists a constant $M_\rho \geq 0$ such that

$$\mathrm{supp}(\rho) \subseteq \mathscr{X} \times [-M_\rho, M_\rho].$$

Define a linear map $L_K : \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{H}_K$ by $L_K(f)(x) = \int_X K(x,t) f(t) d\rho_X$. Together with the inclusion $J : \mathscr{H}_K \to \mathscr{L}^2_{\rho_\mathscr{X}}$, $J \circ L_K : \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{L}^2_{\rho_\mathscr{X}}$ is a compact operator on $\mathscr{L}^2_{\rho_\mathscr{X}}$ [e.g. Halmos and Sunder 1978]. The restriction $L_K|_{\mathscr{H}_K} : \mathscr{H}_K \to \mathscr{H}_K$ is the covariance operator of $\rho_\mathscr{X}$ in $\mathscr{H}_K$: $L_K|_{\mathscr{H}_K} f = \mathbb{E}\langle f, K_x\rangle K_x$ by the reproducing property. All the three operators, by abusing the notation, are all denoted by L_K in the sequel. It can be shown [e.g. Cucker and Smale 2002] that for any $\lambda \in \mathbb{R}_+$,

$$f_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho \in \mathscr{H}_K. \tag{3}$$

The compactness of $L_K : \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{L}^2_{\rho_\mathscr{X}}$ implies the existence of an orthonormal eigensystem $(\mu_\alpha, \phi_\alpha)_{\alpha\in\mathbb{N}}$, such that $L_K \phi_\alpha = \mu_\alpha \phi_\alpha$. Hence also $\phi_\alpha \in \mathscr{H}_K$. Define $L_K^r : \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{L}^2_{\rho_\mathscr{X}}$ by

$$\begin{array}{rccc} L_K^r : & \mathscr{L}^2_{\rho_\mathscr{X}} & \to & \mathscr{L}^2_{\rho_\mathscr{X}} \\ & \sum_\alpha a_\alpha \phi_\alpha & \mapsto & \sum_\alpha a_\alpha \mu_\alpha^r \phi_\alpha \end{array} \tag{4}$$

Note that $L_K^{1/2} : \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{H}_K$ is an isometrical isomorphism between the quotient space $\mathscr{L}^2_{\rho_\mathscr{X}} / \ker(L_K)$ and $\mathscr{H}_K$. For simplicity in this paper we assume that $\ker(L_K) = \{0\}$, which happens when $\mathscr{H}_K$ is dense in $\mathscr{L}^2_{\rho_\mathscr{X}}$ (e.g. Gaussian kernel). With $L_K^{1/2}$, $\langle \phi_\alpha, \phi_{\alpha'}\rangle_K = \langle L_K^{-1/2}\phi_\alpha, L_K^{-1/2}\phi_{\alpha'}\rangle_\rho = \mu_\alpha^{-1}\langle \phi_\alpha, \phi_{\alpha'}\rangle_\rho$, whence (ϕ_α) is a bi-orthogonal system in $\mathscr{H}_K$ and $\mathscr{L}^2_{\rho_\mathscr{X}}$. Finally denote by $L_x(f) = \langle f, K_x\rangle_K K_x = f(x) K_x$, whence the expectation with ρ_X is $\mathbb{E}[L_x] = L_K$.

In this paper, by $C_1, C_2, \ldots$, we denote various constants, which are defined 'locally' in the sense that the same notation may appear in different sections for different constants.

2.2. **Stochastic Gradient Algorithms.** Let $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathscr{Z}_+^*} \in X \times \mathbb{R}$ be the filtration $\mathcal{F}_t = \sigma\{(x_k, y_k) \,:\, 1 \leq k \leq t\}$. Denote by $\mathbb{E}_t$ the expectation w.r.t. $\mathcal{F}_t$ and by $\mathbb{E}$ the expectation over all the random variables. Consider the following $\mathcal{F}_t$-adapted process $(f_t)_{t\in\mathbb{N}}$ with values in $\mathscr{H}_K$,

Define an online learning sequence $(f_t)_{t\in\mathbb{N}}$ as follows,

$$f_t = f_{t-1} - \gamma_t[(f_{t-1}(x_t) - y_t)K_{x_t} + \lambda_t f_{t-1}], \qquad \text{for some } f_0 \in \mathscr{H}_K, \text{ e.g. } f_0 = 0 \tag{5}$$

where
(A) for each t, (x_t, y_t) is independent and identically distributed (i.i.d.) according to ρ;
(B) the step size $\gamma_t > 0$;
(C) the regularization parameter $\lambda_t > 0$.

Remark 2.1. The computational cost of this algorithm typically is $O(t^2)$. As each step t, the main computational cost is due to the evaluation $f_{t-1}(x_t)$ which needs to access all K_{x_i} $(1 \leq i \leq t)$ in $O(t)$ steps. Thus the total cost is of $O(1 + 2 + \ldots + t) = O(t^2)$ at time t. In the cases that one can store and access the values $f_t(x)$ for all x, e.g. on a grid of $\mathscr{X}$, the computational cost is only linear $O(t)$ at the requirement of large memory and fast memory access.

By reproducing property, we can see that the gradient map of

$$V_z(f) = \frac{1}{2}[(f(x) - y)^2 + \lambda \|f\|_K^2], \qquad z = (x, y) \in \mathscr{Z}$$

is given by $\text{grad}V_z(f) = (f(x) - y)K_x + \lambda f$ [Smale and Yao 2005], as a random variable depending on z. Since the expectation $\mathbb{E}[V_z(f)] = 2(\mathscr{E}(f) + \lambda\|f\|_K^2)$, algorithm (5) can thus be regarded as stochastic approximations of gradient descent method to solve (2), for each $\lambda = \lambda_t$.

2.3. **Main Theorems.** In the sequel we consider the approximation of $f_\rho \in \mathscr{H}_K$ by online learning sequence (f_t). The following theorems are the main results of this paper. Theorem A gives a sufficient condition that (f_t) follows the regularization path f_λ. Theorem B and C provides probabilistic upper bounds for the convergence rate of $f_t \to f_\rho$ in $\mathscr{H}_K$ and $\mathscr{L}^2_{\rho_{\mathscr{X}}}$, respectively.

First we need a definition for that (f_t) follows the regularization path f_λ.

Definition. An online learning sequence (f_t) is said to *follow the regularization path* $f_\lambda : \mathbb{R}_+ \to \mathscr{H}_K$, if there exists a sequence $(\lambda_t) \downarrow 0$, such that

$$\lim_{t\to\infty} \mathbb{E}\|f_t - f_{\lambda_t}\|_K = 0.$$

The following theorem gives some sufficient conditions for (f_t) following the regularization path.

Theorem A (Path Following Condition)**.** *The online learning sequence (f_t) defined by equation (5) follows the regularization path f_λ, if the following conditions are satisfied:*

(A) $\sum_{t\to\infty} \gamma_t\lambda_t = \infty.$

(B) $\lim_{t\to\infty} \dfrac{\gamma_t}{\lambda_t} = 0,$

(C) $\lim_{t\to\infty} \dfrac{\|f_{\lambda_t} - f_{\lambda_{t-1}}\|_K}{\lambda_t\gamma_t} = 0,$

This theorem will be proved in Section 3, following from Theorem 3.6 in the setting of general Hilbert spaces.

Remark 2.2. Although $\lambda_t \to 0$, condition (A) puts a restriction that $\gamma_t\lambda_t$ can not drop too fast, in fact this is necessary to "forget" the error caused by the initial guess f_0. Condition (B) says that the step size $\gamma_t \to 0$, and it has to drop faster than the regularization parameter λ_t. Such a condition is to attenuate the random fluctuation caused by sampling. Condition (C) implies that the drifts of the regularization path (f_{λ_t}) converges to zero, at a speed faster than $\gamma_t\lambda_t$. This condition says that in the long run, the drifts along the regularization path should be slow enough for the algorithm to follow the path.

The next two theorems present some probabilistic upper bounds which characterize the convergence rates in $\mathscr{H}_K$ and $\mathscr{L}^2_{\rho_{\mathscr{X}}}$, under certain regularity assumptions on the regression function f_ρ. Below by $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ $(r > 0)$, we mean that f_ρ lies in the image of the mapping $L_K^r : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{L}^2_{\rho_{\mathscr{X}}}$. Due to the isometry $L_K^{1/2} : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{H}_K$, for $r \geq 1/2$ this implies $f_\rho \in \mathscr{H}_K$, with additional regularity if $r > 1/2$. Using the spectral decomposition of L_K, L_K^r can be regarded as a low-pass filter as given in (4), *i.e.*

$$L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}} \Leftrightarrow f_\rho = \sum_\alpha a_\alpha \mu_\alpha^r \phi_\alpha, \quad \text{for some } \textstyle\sum_\alpha a_\alpha^2 < \infty.$$

Theorem B (Convergence Rates in $\mathscr{H}_K$)**.** *Assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ for some $r \in (1/2, 3/2]$. Let $t_0 \geq (\kappa+1)^4$. Then there is a choice of (γ_t) and (λ_t) such that with probability at least $1-\delta$, the following holds for all $t \in \mathbb{N}$,*

$$\|f_t - f_\rho\|_K \leq \frac{C_1}{t} + (C_2 \log\frac{2}{\delta} + C_3 \|L_K^{-r} f_\rho\|_\rho) \left(\frac{1}{t}\right)^{\frac{2r-1}{4r+2}},$$

where

$$C_1 = 2t_0^{\frac{4r+3}{4r+2}} M_\rho, \quad C_2 = \frac{4M_\rho}{3}(7\kappa+1), \quad C_3 = \frac{20r-2}{(2r-1)(2r+3)}.$$

One such choice is $\gamma_t = (t+t_0)^{-2r/(2r+1)}$ and $\lambda_t = (t+t_0)^{-1/(2r+1)}$.

Its proof is given in Section 5.

Remark 2.3. The asymptotic rate $O(t^{-(2r-1)/(4r+2)})$ is the same as the batch learning algorithms [Theorem 2, in Smale and Zhou 2005].

Remark 2.4. Note that the upper bound consists of three parts. The first term at a rate $O(t^{-1})$, captures the influence of the initial choice $f_0 = 0$, which is much faster than the remaining terms. The second term at a rate $O(t^{-(2r-1)/(4r+2)})$, reflects the error caused by random fluctuations by the i.i.d. sampling. The third term at a rate $O(\|L_K^{-r} f_\rho\|_\rho t^{-(2r-1)/(4r+2)})$, collects contributions from both drifts along the regularization path $f_{\lambda_t} - f_{\lambda_{t-1}}$ and the approximation error $f_{\lambda_t} - f_\rho$, since they share the same rate upto different constants.

Theorem C (Convergence Rates in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$)**.** *Assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ for some $r \in [1/2, 1]$. Let $t_0 \geq (\kappa^2+1)^4$. Then there is a choice of (γ_t) and (λ_t) such that with probability at least $1-\delta$ ($\delta \in (0,1)$), the following holds for all $t \in \mathbb{N}$,*

$$\|f_t - f_\rho\|_\rho \leq D_1 t^{-1} + \left(D_2\sqrt{\frac{1}{\delta}} + \left(D_3 + D_4\sqrt{\frac{1}{\delta}}\right)\|L_K^{-r} f_\rho\|_\rho\right) t^{-r/(2r+1)},$$

where

$$D_1 = M_\rho(t_0+1), \quad D_2 = \sqrt{6}\kappa M_\rho(1+\kappa\sqrt{\kappa^2+1}), \quad D_3 = \frac{5r+1}{r(r+1)}, \quad D_4 = 2\sqrt{2}\kappa.$$

One such choice is $\gamma_t = (t+t_0)^{-2r/(2r+1)}$ and $\lambda_t = (t+t_0)^{-1/(2r+1)}$.

Its proof will be given in Section 6.

Remark 2.5. A special case is $r = 1/2$, which is equivalent to say $f_\rho \in \mathscr{H}_K$. In this case $\gamma_t = \lambda_t = (t+t_0)^{-1/2}$, whence it does not satisfy the Path Following Condition (B) in Theorem A. But Theorem C suggests a weaker notion that f_t follows the regularization path, *i.e.* $\lim_{t\to\infty} \mathbb{E}[\|f_t - f_{\lambda_t}\|_\rho] = 0$, which in fact converges at a rate of $O(t^{-1/4})$ uniformly for all $f_\rho \in \mathscr{H}_K$.

Remark 2.6. It is still open whether the upper bound above can be improved by replacing $1/\delta$ with $\log 1/\delta$. One way to achieve this is to prove that $\|f_t\|_\rho \leq O(1/\sqrt{\lambda_t})$, which is open yet. For details, see more discussions in Remark 6.6, on the problem of using Bernstein's type inequality here.

Note that for any $f \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$, the generalization error $\mathscr{E}(f) - \mathscr{E}(f_\rho) = \|f - f_\rho\|_\rho^2$ [e.g. see Cucker and Smale 2002], which is often used to evaluate the performance of learning algorithms in literature. We have the following corollary of Theorem C.

Corollary 2.7. *Under the same condition of Theorem C, there holds with probability at least* $1-\delta$ *(*$\delta \in (0,1)$*), for all* $t \in \mathbb{N}$*,*

$$\mathscr{E}(f_t) - \mathscr{E}(f_\rho) \le 2D_1 t^{-2} + 2(D_2\sqrt{\frac{1}{\delta}} + D_3\|L_K^{-r} f_\rho\|_\rho + + D_4\sqrt{\frac{1}{\delta}}\|L_K^{-r} f_\rho\|_\rho)^2 t^{-2r/(2r+1)},$$

where $D_1, \ldots, D_4$ *are the same constants in Theorem C.*

Remark 2.8. For $r \in (1/2, 1]$, the asymptotic rate $O(t^{-2r/(2r+1)})$ has been shown to be optimal in the sense that it reaches the minimax and individual lower rate [Caponnetto and De Vito 2005]. To be precise, let $\mathscr{P}(b, r)$ ($b > 1$ and $r \in (1/2, 1]$) be the set of probability measure ρ on $\mathscr{X} \times \mathscr{Y}$, such that: (A) almost surely $|y| \le M_\rho$; (B) $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_\mathscr{X}}$; (C) the eigenvalues $(\mu_n)_{n\in\mathbb{N}}$ of $L_K : \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{L}^2_{\rho_\mathscr{X}}$, arranged in a nonincreasing order, are subject to the decay $\mu_n = O(n^{-b})$. Then the following minimax lower rate was given as Theorem 2 in [Caponnetto and De Vito 2005],

$$\liminf_{t\to\infty} \inf_{(z_i)_1^t \mapsto f_t} \sup_{\rho\in\mathscr{P}(b,r)} \mathbf{Prob}\left\{(z_i)_1^t \in \mathscr{Z}^t : \mathscr{E}(f_t) - \mathscr{E}(f_\rho) > Ct^{-\frac{2rb}{2rb+1}}\right\} = 1$$

for some constant $C > 0$ independent on t, where the infimum in the middle is taken over all algorithms as a map $\mathscr{Z}^t \ni (z_i)_1^t \mapsto f_t \in \mathscr{H}_K$. Moreover, for every $B > b$, the following individual lower rate was given as Theorem 3 in [Caponnetto and De Vito 2005],

$$\inf_{((z_i)_1^t \mapsto f_t)_{t\in\mathbb{N}}} \sup_{\rho\in\mathscr{P}(b,r)} \limsup_{t\to\infty} \frac{\mathbb{E}[\mathscr{E}(f_t)] - \mathscr{E}(f_\rho)}{t^{-\frac{2rB}{2rB+1}}} > 0,$$

where the infimum is taken over arbitrary sequences of functions $f_t : \mathscr{Z}^t \to \mathscr{H}_K$. Note that in both lower rates the permissible f_t is beyond online learning sequences. In the minimax lower rate, the probability measure may change for each fixed t; while in the individual lower rate, the probability measure is fixed for all large enough t, which is more suitable to the nature of learning. For more background on these lower rates, see for example [Györfi, Kohler, Krzyżak, and Walk 2002] and references therein.

Now we compare these lower rates to our upper bound. Since $L_K : \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{L}^2_{\rho_\mathscr{X}}$ is a trace-class operator, its eigenvalues are summable. Therefore by taking $b = B = 1$, one may obtain an eigenvalue-independent lower rate $O(t^{-2r/(2r+1)})$ for all possible L_K. In this way, the upper bound in Corollary 2.7 reaches both the minimax and the individual lower rates.

3. Sequential Stochastic Approximations in Hilbert Spaces

In this section, we consider a more general setting: stochastic approximations of solutions for a sequence of linear operator equations in general Hilbert spaces. The sequence of linear operator equations are constructed in a spirit of regularization. A theorem for the convergence is given, which leads to Theorem A when specialized to the setting in this paper.

Let $\mathscr{W}$ be a Hilbert space and $\hat{A} : \mathscr{W} \to \mathscr{W}$ be a positive operator and $\hat{b} \in \mathscr{W}$. Consider the following linear equation

$$\hat{A}w = \hat{b}. \tag{6}$$

where $\hat{A}$ has an *unbounded* inverse. As in the standard setting of Robbins-Monro procedure [Robbins and Monro 1951], we assume that $\hat{A}$ and $\hat{b}$ are the expectations of some random operators and vectors, respectively. However due to the unboundedness of $\hat{A}^{-1}$, the complexity analysis fails for the standard algorithm [Smale and Yao 2005] .

To solve this ill-posed problem with unbounded $\hat{A}^{-1}$, one may construct a sequence $\hat{A}_t \to \hat{A}$ and $\hat{b}_t \to \hat{b}$, where each $\hat{A}_t$ has bounded inverse. Then one has a sequence $\hat{w}_t = \hat{A}_t^{-1}\hat{b}_t$ converging to the solution of (6) and $\hat{w}_t$ is continuous with respect to $\hat{A}_t$ and $\hat{b}_t$. Such a sequence $(\hat{w}_t)$, will be called a *regularization path* of the solution of equation (6).

It remains to approximate $(\hat{w}_t)$, since $\hat{A}_t$ and $\hat{b}_t$ are the means with respect to some unknown probability measure, as $\hat{A}$ and $\hat{b}$. Here we define a sequence of stochastic approximations

$$w_t = w_{t-1} - \gamma_t(A_t w_{t-1} - b_t),$$

where $A_t = A_t(z_t)$ and $b_t = b_t(z_t)$ are random variables depending on the sample z_t, such that $\mathbb{E}[A_t(z_t)] = \hat{A}_t$ and $\mathbb{E}[b_t(z_t)] = \hat{b}_t$.

In this section our purpose is to study the evolutions of (w_t) and $(\hat{w}_t)$, and give conditions of their gap converging to zero.

First of all we summarize the assumptions taken in this section.

Generalized Finiteness Condition. For each $t \in \mathbb{N}$, almost surely there holds
(A) $\hat{A}$ is invertible with *unbounded* inverse;
(B) A_t, $\hat{A}_t$ and $\hat{A}$ have operator norms bounded by $\overline{\alpha} \in (0, \infty)$;
(C) A_t and $\hat{A}_t$ are invertible with inverse operator norms bounded by $1/\underline{\alpha}_t$, with $\underline{\alpha}_t \to 0$;
(D) b_t and $\hat{b}_t$ have norms bounded by $\beta \in (0, \infty)$;

3.1. **Two Structural Decomposition Theorems.** In this subsection we presents two structural decompositions of the *remainder*,

$$r_t := w_t - \hat{w}_t. \tag{7}$$

Both ways decomposes r_t into three parts: one depending on r_0; one depending on the following defined *drifts* along the regularization path $(\hat{w}_t)$,

$$\Delta_j := \hat{w}_t - \hat{w}_{t-1}; \tag{8}$$

and one random variable of zero mean, either as a reversed martingale or as a martingale. Both decompositions are found useful, where the reversed martingale decomposition is enough to study the convergence in $\mathscr{H}_K$ and the martingale decomposition is crucial to obtain sharp convergence rates in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$.

Theorem 3.1 (Reversed Martingale Decomposition)**.** *Define a random operator on $\mathscr{W}$,*

$$\Pi_j^t(z_j, \dots, z_t) = \begin{cases} \prod_{i=j}^{t} (I - \gamma_i A_i), & j \le t; \\ I, & j > t. \end{cases}$$

Then for all $t \in \mathbb{N}$ and $t > t_0$,

$$r_t = \Pi_{t_0+1}^t r_{t_0} - \sum_{j=t_0+1}^{t} \gamma_j \Pi_{j+1}^t (A_j \hat{w}_j - b_j) - \sum_{j=t_0+1}^{t} \Pi_j^t \Delta_j \tag{9}$$

Remark 3.2. Note that Π_{j+1}^t is a random operator depending on $z_{j+1}, \dots z_t$, and $A_j\hat{w}_j - b_j$ is a zero mean random variable depending on z_j. By independence of $(z_t)_{t\in\mathbb{N}}$, the conditional expectation $\mathbb{E}[\gamma_j \Pi_{j+1}^t (A_j\hat{w}_j - b_j) | z_{j+1}, \dots, z_t] = 0$, whence for each t, $\gamma_j \Pi_{j+1}^t (A_j\hat{w}_j - b_j)$ is a *reversed*

martingale difference sequence whose sum is a *reversed martingale sequence* with zero mean. For more background on reversed martingale, see for example [Neveu 1975]. This decomposition will be used to derive Theorem A and Theorem B.

Proof of Theorem 3.1. By definition,

$$\begin{aligned} r_t &= w_t - \hat{w}_t \\ &= w_{t-1} - \gamma_t(A_t w_{t-1} - b_t) - \hat{w}_t \\ &= (I - \gamma_t A_t)(w_{t-1} - \hat{w}_{t-1}) - \gamma_t(A_t\hat{w}_t - b_t) - (I - \gamma_t A_t)(\hat{w}_t - \hat{w}_{t-1}) \end{aligned}$$

which gives

$$r_t = (I - \gamma_t A_t) r_{t-1} - \gamma_t(A_t\hat{w}_t - b_t) - (I - \gamma_t A_t)\Delta_t. \tag{10}$$

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

Theorem 3.3 (Martingale Decomposition)**.** *Let* $\chi_t = (\hat{A}_t - A_t)w_{t-1} + (b_t - \hat{b}_t)$*, and*

$$\hat{\Pi}_j^t = \begin{cases} \prod_{i=j}^{t} \left(I - \gamma_i \hat{A}_i\right), & j \le t; \\ I, & j > t. \end{cases}$$

Then for all $t \in \mathbb{N}$ *and* $t > t_0$*,*

$$r_t = \hat{\Pi}_{t_0+1}^t r_{t_0} + \sum_{j=t_0+1}^{t} \gamma_j \hat{\Pi}_{j+1}^t \chi_j - \sum_{j=t_0+1}^{t} \hat{\Pi}_j^t \Delta_j \tag{11}$$

Remark 3.4. This decomposition was proposed in [Yao 2005]. Note that in this decomposition only the second term is random. The operator $\hat{\Pi}_{j+1}^t$ is deterministic and χ_j is a zero mean random variable depending on $z_1, \ldots, z_j$. Therefore the conditional expectation $\mathbb{E}[\gamma_j \hat{\Pi}_{j+1}^t \chi_j | z_1, \ldots, z_{j-1}] = 0$, whence for each t, $\gamma_j \hat{\Pi}_{j+1}^t \chi_j$ is a *martingale difference sequence* for all $t \in \mathbb{N}$, whose sum is a *martingale sequence* of zero mean. Note that this martingale property holds even for dependent sampling $z_t(z_1, \ldots, z_{t-1})$.

Remark 3.5. An importance feature of this decomposition used in this paper, lies in that the operator $\hat{\Pi}_j^t$ is deterministic and when taking $\hat{A}_i = L_K + \lambda_i$, it has a spectral decomposition by the eigenfunctions of $L_K : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{L}^2_{\rho_{\mathscr{X}}}$. This feature plays a key role in the proof of Theorem C. But a disadvantage is that the term χ_t depends on w_{t-1}, which increases the difficulty to bound χ_t. In fact, the open problem how to improve Theorem C by replacing $1/\delta$ to $\log 1/\delta$, depends on how to get a tighter bound on $\|\chi_t\|$, see Remark 6.6 for details.

Proof of Theorem 3.3. By definition,

$$\begin{aligned} r_t &= w_t - \hat{w}_t \\ &= w_{t-1} - \gamma_t(A_t w_{t-1} - b_t) - \hat{w}_t \\ &= (I - \gamma_t \hat{A}_t)(w_{t-1} - \hat{w}_{t-1}) + \gamma_t[(\hat{A}_t - A_t)w_{t-1} + (b_t - \hat{A}_t\hat{w}_t)] - (I - \gamma_t\hat{A}_t)[\hat{w}_t - \hat{w}_{t-1}]. \end{aligned}$$

Using $\hat{A}_t\hat{w}_t = \hat{b}_t$ for all $t \in \mathbb{N}$, we obtain

$$r_t = (I - \gamma_t\hat{A}_t) r_{t-1} + \gamma_t[(\hat{A}_t - A_t)w_{t-1} + (b_t - \hat{b}_t)] - (I - \gamma_t\hat{A}_t)(\hat{w}_t - \hat{w}_{t-1}). \tag{12}$$

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

3.2. **Convergence of Remainder and The Proof of Theorem A.** Here we give an application of the reversed martingale decomposition, Theorem 3.1, to derive Theorem A. A general theorem is first given to study $\mathbb{E}\|r_t\| \to 0$, which leads to Theorem A when specialized to $r_t = f_t - f_{\lambda_t}$.

Theorem 3.6. *Suppose that the variance $\mathbb{E}\|A_t\hat{w}_t - b_t\|^2$ is uniformly bounded for all $t \in \mathbb{N}$. Then*

$$\mathbb{E}\|r_t\| \to 0,$$

if the following conditions hold:

(A) $\sum_t \gamma_t \underline{\alpha}_t = \infty$.

(B) $\lim_{t\to\infty} \frac{\gamma_t}{\underline{\alpha}_t} = 0$,

(C) $\lim_{t\to\infty} \frac{\|\Delta_t\|}{\underline{\alpha}_t \gamma_t} = 0$,

Its proof was included in Appendix D. Equipped with Theorem 3.6, we are in a position to prove Theorem A.

Proof of Theorem A. Let $\mathscr{W} = \mathscr{H}_K$ and $w_t = f_t$. Define $A_t = L_t + \lambda_t I$ ($L_t := L_K^{x_t} = \langle\ , K_{x_t}\rangle_K K_{x_t}$), $\hat{A}_t = L_K + \lambda_t I$, $b_t = y_t K_{x_t}$, and $\hat{b}_t = L_K f_\rho$. Therefore $\underline{\alpha}_t = \lambda_t$ and $\hat{w}_t = f_{\lambda_t}$. With these replacement, it suffices to check the uniform boundedness of

$$\mathbb{E}\|A_t\hat{w}_t - b_t\|^2 = \mathbb{E}\|(f_{\lambda_t}(x_t) - y_t)K_{x_t} + \lambda_t f_{\lambda_t}\|_K^2.$$

But this can be seen by Lemma 5.5(B) which gives $\mathbb{E}\|(f_{\lambda_t}(x_t) - y_t)K_{x_t} + \lambda_t f_{\lambda_t}\|_K^2 \le 10\kappa^2 M_\rho^2$. Theorem A then follows from Theorem 3.6. □

4. Estimates for The Drifts

In this section, we provide some estimates on the drift $\|f_\lambda - f_\mu\|$ ($\lambda, \mu > 0$), in $\mathscr{H}_K$-norm or $\mathscr{L}^2_{\rho_{\mathscr{X}}}$-norm. When specialized to $\mu = 0$, these estimates give the approximation errors $\|f_\lambda - f_\rho\|$. Note that we extend the range of r to $(0, \infty)$ (or $(1/2, \infty)$), from $(0, 1]$ (or $(1/2, 1]$), usually considered in literature [e.g. see Smale and Zhou 2005]. Such an extension was firstly pointed to us by H.Q. Minh. Note that for large enough r ($r > 1$ in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ norm or $r > 3/2$ in $\mathscr{H}_K$ norm), the rates in the upper bounds can not be improved. This phenomenon is related to the *saturation* problem of regularizations [Engl, Hanke, and Neubauer 2000].

Theorem 4.1. *Let $\lambda > \mu \ge 0$. If $\mu = 0$ we define $f_\mu = f_\rho$. Assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ for some $r > 0$.*

(A) If $r \in (0, 1]$, then

$$\|f_\lambda - f_\mu\|_\rho \le |\lambda^r - \mu^r| \frac{\|L_K^{-r} f_\rho\|_\rho}{r};$$

(B) If $r \ge 1$, then for any $1 \le s \le r$,

$$\|f_\lambda - f_\mu\|_\rho \le \kappa^{2(s-1)} |\lambda - \mu| \|L_K^{-s} f_\rho\|_\rho;$$

(C) If $r \geq 1/2$, then

$$\|f_\lambda - f_\mu\|_K \leq \frac{|\lambda - \mu|}{\lambda}\|f_\rho\|_K;$$

(D) If $r \in (1/2, 3/2]$, then

$$\|f_\lambda - f_\mu\|_K \leq |\lambda^{r-1/2} - \mu^{r-1/2}|\frac{\|L_K^{-r} f_\rho\|_\rho}{r - \frac{1}{2}};$$

(E) If $r \geq 3/2$, then for any $3/2 \leq s \leq r$,

$$\|f_\lambda - f_\mu\|_K \leq \kappa^{2(s-3/2)}|\lambda - \mu|\|L_K^{-s} f_\rho\|_\rho.$$

Remark 4.2. From (A) and (B) (or (D) and (E)) we can see that $\|f_\lambda - f_\mu\|_\rho \leq O(|\lambda^{\min(r,1)} - \mu^{\min(r,1)}|)$ (or $\|f_\lambda - f_\mu\|_K \leq O(|\lambda^{\min(r-1/2,1)} - \mu^{\min(r-1/2,1)}|)$). In this way the upper bounds 'saturate' in the rates when f_ρ has large enough regularity indexed by $r > 1$ (or $r > 3/2$).

Proof. Assume that $\lambda \geq \mu$ for simplicity. By definition,

$$(L_K + \lambda I)f_\lambda = L_K f_\rho,\ (L_K + \mu I)f_\mu = L_K f_\rho,$$

which yields

$$f_\lambda - f_\mu = (\mu - \lambda)(L_K + \lambda I)^{-1}(L_K + \mu I)^{-1} L_K f_\rho. \tag{13}$$

(A) If $r \in (0, 1]$,

$$\begin{aligned}\|f_\lambda - f_\mu\|_\rho &\leq |\mu - \lambda|\|(L_K + \lambda I)^{-1} f_\rho\|_\rho \leq |\mu - \lambda|\|(L_K + \lambda I)^{-1} L_K^r\|\|L_K^{-r} f_\rho\|_\rho \\ &\leq |\mu - \lambda|\|(L_K + \lambda I)^{r-1}\|\|L_K^{-r} f_\rho\|_\rho = \Lambda(\mu)|\mu^r - \lambda^r|\|J\|\|L_K^{-r} f_\rho\|_\rho\end{aligned}$$

where

$$\Lambda(\mu) = \frac{1 - \frac{\mu}{\lambda}}{1 - \left(\frac{\mu}{\lambda}\right)^r}, \qquad \text{and} \qquad J = \lambda^{r-1}(L_K + \lambda I)^{r-1}.$$

Now $\|J\| \leq 1$ and

$$\Lambda(\mu) \leq \frac{1}{r},$$

where we use, for $u := 1 - \mu/\lambda$, that $u \leq (1 - (1 - u)^r)/r$, since $u \mapsto (1 - (1 - u)^r)/r$ (defined on $(-\infty, 1]$) is convex and remains above the tangent line at 0. In particular, $\Lambda(0) = 1$. This completes the proof of (A).

(B) For any $s \leq r$, $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ implies $L_K^{-s} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$. If $s \geq 1$, by equation (13),

$$\begin{aligned}\|f_\lambda - f_\mu\|_\rho &\leq |\mu - \lambda|\|(L_K + \lambda I)^{-1} f_\rho\|_\rho \leq |\mu - \lambda|\|(L_K + \lambda I)^{-1} L_K^s\|\|L_K^{-s} f_\rho\|_\rho \\ &\leq |\mu - \lambda|\|L_K^{s-1}\|\|L_K^{-s} f_\rho\|_\rho \leq \kappa^{2(s-1)}|\mu - \lambda|\|L_K^{-s} f_\rho\|_\rho\end{aligned}$$

(C) In particular if $r \geq 1/2$, this implies $f_\rho \in \mathscr{H}_K$, whence by (13)

$$\|f_\lambda - f_\mu\|_K \leq |\mu - \lambda|\|(L_K + \lambda I)^{-1}\|\|(L_K + \mu I)^{-1} L_K\|\|f_\rho\|_K \leq \frac{|\mu - \lambda|}{\lambda}\|f_\rho\|_K.$$

(D) If $r \in (1/2, 3/2]$, then similar to (A),

$$\|f_\lambda - f_\mu\|_K = \|L_K^{-1/2}(f_\lambda^* - f_\mu^*)\|_\rho \quad \leq \quad \Lambda(\mu)|\mu^{r-1/2} - \lambda^{r-1/2}|\|J\|\|L_K^{-r} f_\rho\|_\rho$$

where

$$\Lambda(\mu) = \frac{1-\frac{\mu}{\lambda}}{1-\left(\frac{\mu}{\lambda}\right)^{r-1/2}}, \qquad \text{and} \qquad J = \lambda^{3/2-r}(L_K+\lambda I)^{r-3/2}.$$

We complete the proof by replacing r with $r-1/2$ in (A).

(E) It follows from (B) by replacing s with $s-1/2$. □

5. Upper Bounds for Convergence in $\mathscr{H}_K$

In this section we are going to give a probabilistic upper bound for

$$\|f_t - f_\rho\|_K$$

as a proof for Theorem B. Throughout this section, we assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ for some $r \in (1/2, 3/2]$, which implies $f_\rho \in \mathscr{H}_K$ with additional regularity.

The idea in the proof starts from the triangle inequality

$$\|f_t - f_\rho\|_K \le \|f_t - f_{\lambda_t}\|_K + \|f_{\lambda_t} - f_\rho\|_K.$$

The first term can be further decomposed into the reversed martingale decomposition as Theorem 3.1, which can be rewritten as

$$r_t = \Pi_1^t r_0 - \sum_{j=1}^t \gamma_j \Pi_{j+1}^t (A_j \hat{w}_j - b_j) - \sum_{j=1}^t \Pi_j^t \Delta_j \tag{14}$$

where $r_t = f_t - f_{\lambda_t}$, $A_t = L_t + \lambda_t I$ ($L_t := L_K^{x_t} = \langle\, , K_{x_t}\rangle_K K_{x_t}$), $b_t = y_t K_{x_t}$, $\hat{w}_t = f_{\lambda_t}$, $\Delta_t = f_{\lambda_t} - f_{\lambda_{t-1}}$, and

$$\Pi_j^t(x_j, \dots, x_t) = \begin{cases} \prod_{i=j}^t \left(I - \gamma_i (L_i + \lambda_i I)\right), & j \le t; \\ I, & j > t, \end{cases}$$

with the choice that

$$\gamma_t = \frac{1}{(t+t_0)^\theta}, \qquad \lambda_t = \frac{1}{(t+t_0)^{1-\theta}}, \quad \text{for some } \theta \in [0,1]$$

For convenience, we make the following definitions.

[Definitions of Errors]

(A) *Initial Error*: $\mathscr{E}_{init}(t) = \|\Pi_1^t r_0\|$, which reflects the propagation error by the initial choice f_0;
(B) *Sample Error*: $\mathscr{E}_{samp}(t) = \|\sum_{j=1}^t \gamma_j \Pi_{j+1}^t (A_j \hat{w}_j - b_j)\|$, where $\xi_j = \gamma_j \Pi_{j+1}^t (A_j \hat{w}_j - b_j)$ is a reversed martingale difference sequence, reflecting the random fluctuation caused by sampling;
(C) *Drift Error*: $\mathscr{E}_{drift}(t) = \|\sum_{j=1}^t \Pi_j^t \Delta_j\|$, which measures the error caused by drifts from $f_{\lambda_{t-1}}$ to f_{λ_t} along the regularization path;
(D) *Approximation Error*: $\mathscr{E}_{approx}(t) = \|f_{\lambda_t} - f_\rho\|_K$, which measures the distance between the regression function and the regularization path at time t.

In this way we may bound

$$\|f_t - f_\rho\|_K \le \mathscr{E}_{init}(t) + \mathscr{E}_{samp}(t) + \mathscr{E}_{drift}(t) + \mathscr{E}_{approx}(t).$$

In the remaining of this section, we are going to provide upper bounds for each of the four errors, which, roughly speaking, are

$$\begin{aligned}\mathscr{E}_{approx}(t) &\leq O(t^{-(r-1/2)(1-\theta)})\\ \mathscr{E}_{drift}(t) &\leq O(t^{-(r-1/2)(1-\theta)})\\ \mathscr{E}_{init}(t) &\leq O(t^{-1})\\ \mathscr{E}_{samp}(t) &\leq O(t^{\frac{1}{2}-\theta})\end{aligned}$$

It is not surprising that the approximation error and drift error have the same rate, as both of them come from the estimates on drifts in Theorem 4.1. Theorem B then follows from these bounds by setting $\theta = 2r/(2r+1)$.

5.1. **Approximation Error.** The approximation error is derived from Theorem 4.1(D) by setting $\lambda = \lambda_t$ and $\mu = 0$.

Theorem 5.1 (Approximation Error)**.** *For $r \in (1/2, 3/2]$,*

$$\|f_{\lambda_t} - f_\rho\|_K \leq C_1 (t+t_0)^{-(r-1/2)(1-\theta)},$$

where $C_1 = (r-1/2)^{-1}\|L_K^{-r} f_\rho\|_\rho$.

5.2. **Drift Error.**

Theorem 5.2 (Drift Error)**.** *Assume that $a = 1$ and $t_0^\theta \geq \kappa^2 + 1$.*

$$\mathscr{E}_{drift}(t) \leq C_2 (t+t_0)^{-(r-1/2)(1-\theta)}$$

where $C_2 = \dfrac{4(1-\theta)}{1-(r-1/2)(1-\theta)}\|L_K^{-r} f_\rho\|_\rho$.

Proof. By Theorem 4.1(D), it follows that

$$\|\Delta_t\| = \|f_{\lambda_t} - f_{\lambda_{t-1}}\|_K \leq 4(1-\theta)(t+t_0)^{-(r-1/2)(1-\theta)-1}\|L_K^{-r} f_\rho\|_\rho,$$

where we use

$$\begin{aligned}|\lambda_t^{r-1/2} - \lambda_{t-1}^{r-1/2}| &= |(t+t_0)^{-(r-1/2)(1-\theta)} - (t+t_0-1)^{-(r-1/2)(1-\theta)}|\\ &\leq (r-1/2)(1-\theta)(t+t_0-1)^{-(r-1/2)(1-\theta)-1}\\ &\leq 4(r-1/2)(1-\theta)(t+t_0)^{-(r-1/2)(1-\theta)-1}, \quad \frac{a+1}{b+1} \geq \frac{a}{b} \text{ if } b > a > 0, \qquad (15)\end{aligned}$$

where the second last step is due to the Mean Value Theorem with $h(x) = x^{-r(1-\theta)}$ and $h'(x) = -r(1-\theta)x^{-r(1-\theta)-1}$, such that

$$|h(t+t_0) - h(t+t_0-1)| = |h'(\eta)| \leq |h'(t+t_0-1)|, \qquad \text{for some } \eta \in (t+t_0-1, t+t_0).$$

By Lemma B.2(B),

$$\|\Pi_j^t\| \leq \frac{j+t_0}{t+t_0+1},$$

whence

$$\begin{aligned}\mathscr{E}_{drift}(t) &= \|\sum_{j=1}^{t}\Pi_j^t\Delta_j\|_K \le \frac{4(1-\theta)\|L_K^{-r}f_\rho\|_\rho}{t+t_0+1}\sum_{j=1}^{t}(j+t_0)^{-(r-1/2)(1-\theta)} \\ &\le \frac{4(1-\theta)\|L_K^{-r}f_\rho\|_\rho}{1-(r-1/2)(1-\theta)}(t+t_0)^{-(r-1/2)(1-\theta)}\end{aligned}$$

since

$$\sum_{j=1}^{t}(j+t_0)^{-(r-1/2)(1-\theta)} \le \int_0^t (x+t_0)^{-(r-1/2)(1-\theta)}dx \le \frac{(t+t_0)^{1-(r-1/2)(1-\theta)}}{1-(r-1/2)(1-\theta)}$$

□

5.3. **Initial Error.**

Theorem 5.3 (Initial Error)**.** *Let $t_0 \ge (\kappa^2+1)^{1/\theta}$. Then for all $t \in \mathbb{N}$,*

$$\mathscr{E}_{init}(t) \le C_3(t+t_0)^{-1},$$

where $C_3 = (t_0+1)\|r_0\|$.

Proof. By Lemma B.2(B) with $j=1$,

$$\|\Pi_1^t\| \le \frac{t_0+1}{t+t_0+1} \le \frac{t_0+1}{t+t_0}.$$

This gives $\mathscr{E}_{init}(t) \le (t_0+1)\|r_0\|(t+t_0)^{-1}$. □

5.4. **Sample Error.**

Theorem 5.4 (Sample Error)**.** *Let $a=1$, $t_0^\theta = (\kappa+1)^2)$ and $\gamma_0 = t_0^{-\theta}$. The following holds with probability at least $1-\delta$ ($\delta \in (0,1)$) in the space Z^t,*

$$\mathscr{E}_{samp} \le C_4(t+t_0)^{1/2-\theta}$$

where $C_4 = \frac{4M_\rho}{3}(7\kappa+1)\log\frac{2}{\delta}$.

Before the formal presentation of the proof, we need some auxilary estimates.

Lemma 5.5. *Let $A_t\hat{w}_t - b_t = (f_{\lambda_t}(x_t)-y_t)K_{x_t} + \lambda_t f_{\lambda_t}$.*

(A) $\|A_t\hat{w}_t - b_t\|_K \le (\kappa+1)^2 M_\rho/\sqrt{\lambda_t}$;

(B) $\mathbb{E}[\|A_t\hat{w}_t - b_t\|_K^2] \le 4\kappa^2 M_\rho^2$.

Proof. Recall that

$$A_t\hat{w}_t - b_t = (f_{\lambda_t}(x_t)-y_t)K_{x_t} + \lambda_t f_{\lambda_t}.$$

Then

(A) Using $\|f_\lambda\|_K \le M_\rho/\sqrt{\lambda}$ in Lemma B.3(A),

$$\|A_t\hat{w}_t - b_t\| \le \|f_{\lambda_t}(x_t)K_{x_t}\|_K + |y_t|\|K_{x_t}\|_K + \lambda_t\|f_{\lambda_t}\|_K \le M_\rho\kappa^2/\sqrt{\lambda_t} + M_\rho\kappa + M_\rho\sqrt{\lambda_t}$$

since $\|f_{\lambda_t}(x_t)K_{x_t}\|_K = |\langle f_{\lambda_t}, K_{x_t}\rangle|\|K_{x_t}\|_K \le \|f_{\lambda_t}\|_K\|K_{x_t}\|_K^2 \le M_\rho\kappa^2/\sqrt{\lambda_t}$. It remains to see

$$M_\rho\kappa^2/\sqrt{\lambda_t} + M_\rho\kappa + M_\rho\sqrt{\lambda_t} \;\le\; (\kappa^2+\kappa+1)M_\rho/\sqrt{\lambda_t} \le (\kappa+1)^2 M_\rho/\sqrt{\lambda_t}$$

(B) Using $\lambda_t f_\lambda = L_K f_\rho - L_K f_\lambda$ we obtain

$$(f_{\lambda_t}(x_t) - y_t)K_{x_t} + \lambda_t f_{\lambda_t} = (L_t - L_K)f_{\lambda_t} + L_K f_\rho - y_t K_{x_t}.$$

$$\begin{aligned}
\mathbb{E}[\|A_t\hat{w}_t - b_t\|^2] &= \mathbb{E}\|(L_t - L_K)f_{\lambda_t} + L_K f_\rho - y_t K_{x_t}\|_K^2 \\
&\le 2\mathbb{E}[\|(L_t - L_K)f_{\lambda_t}\|_K^2 + \|L_K f_\rho - y_t K_{x_t}\|_K^2] \\
&\le 2\mathbb{E}[\|L_t f_{\lambda_t}\|_K^2 + \|y_t K_{x_t}\|_K^2] \le 2\kappa^2(\|f_{\lambda_t}\|_\rho^2 + M_\rho^2) = 4\kappa^2 M_\rho^2
\end{aligned}$$

since $\mathbb{E}[L_t] = L_K$, $\mathbb{E}[y_t K_{x_t}] = L_K f_\rho$ and $\|f_\lambda\|_\rho \le M_\rho$ by Lemma B.3(B). □

Now we are ready to give the proof of the sample error bounds, Theorem 5.4.

Proof of Theorem 5.4. Denote $X_j = \gamma_j \Pi_{j+1}^t (A_j\hat{w}_j - b_j) = \gamma_j \Pi_{j+1}^t \xi_j$, which is a reversed martingale difference sequence. It suffices to prove that

$$\|X_j\| \le 2\gamma_t \lambda_t^{-1/2}(\kappa+1)^2 M_\rho \tag{16}$$

$$\mathbb{E}_{j-1}\|X_j\|^2 \le 16\kappa^2 M_\rho^2 \gamma_t^2 \tag{17}$$

Then by Proposition A.3, a varied form of Pinelis-Bernstein inequality,

$$\begin{aligned}
\mathscr{E}_{samp}(t) &= \|\sum_{j=1}^t X_j\| \le 2\left(\frac{2}{3}\gamma_t\lambda_t^{-1/2}(\kappa+1)^2 M_\rho + 4\kappa M_\rho\gamma_t\sqrt{t}\right)\log\frac{2}{\delta} \\
&\le M_\rho\left(\frac{4}{3}(\kappa+1)^2(t+t_0)^{-\theta/2} + 8\kappa\right)(t+t_0)^{1/2-\theta}\log\frac{2}{\delta} \\
&\le M_\rho\left(\frac{4}{3}(\kappa+1) + 8\kappa\right)(t+t_0)^{1/2-\theta}\log\frac{2}{\delta}
\end{aligned}$$

since $t_0^\theta \ge (\kappa+1)^2$.

To see (16) and (17), note that for $t_0^\theta \ge (\kappa+1)^2 \ge \kappa^2+1$, $\|\Pi_j^t\| \le \prod_{i=j}^t(1-\gamma_i\lambda_i)$, whence

$$\begin{aligned}
\|\gamma_j\Pi_{j+1}^t\| &\le \frac{1}{\lambda_j}\gamma_j\lambda_j\prod_{i=j+1}^t(1-\gamma_i\lambda_i) = \frac{1}{\lambda_j}\left[\frac{1}{j+t_0}\prod_{i=j+1}^t(1-\frac{1}{i+t_0})\right], \quad \gamma_t\lambda_t = \frac{1}{t+t_0} \\
&\le \frac{1}{\lambda_j}\left[\frac{1}{j+t_0}\cdot\frac{j+t_0+1}{t+t_0+1}\right], \qquad \text{by Lemma B.1} \\
&\le \frac{2}{\lambda_j(t+t_0+1)}, \qquad\qquad \frac{a+1}{b+1} \ge \frac{a}{b} \text{ for } b \ge a > 0.
\end{aligned}$$

Then it follows from Lemma 5.5,

$$\|X_j\| \le \|\gamma_j\Pi_{j+1}^t\|\|\xi_j\| \le \frac{2(\kappa+1)^2 M_\rho}{\lambda_j^{3/2}(t+t_0+1)} \le \frac{2(\kappa+1)^2 M_\rho}{\lambda_t^{3/2}(t+t_0)} = \frac{2\gamma_t(\kappa+1)^2 M_\rho}{\sqrt{\lambda_t}}$$

and

$$\mathbb{E}\|X_j\|^2 \;\leq\; \frac{4}{\lambda_j^2(t+t_0+1)^2}\mathbb{E}\|\xi_j\|_K^2 \leq \frac{16\kappa^2 M_\rho^2}{\lambda_j^2(t+t_0+1)^2} \leq \frac{16\kappa^2 M_\rho^2}{\lambda_t^2(t+t_0)^2} = 16\gamma_t^2\kappa^2 M_\rho^2,$$

as desired. □

5.5. **Total Error and Proof of Theorem B.** Finally we derive Theorem B by combining the four error bounds obtained above.

Proof of Theorem B. By

$$\|f_t - f_\rho\|_K \leq \mathscr{E}_{approx}(t) + \mathscr{E}_{drift}(t) + \mathscr{E}_{init}(t) + \mathscr{E}_{samp}(t)$$

with Theorem 5.1, 5.2, 5.3, and 5.4, we obtain

$$\|f_t - f_\rho\|_K \leq (C_1 + C_2)(t+t_0)^{-(r-1/2)(1-\theta)} + C_3(t+t_0)^{-1} + C_4(t+t_0)^{1/2-\theta}.$$

Setting $\theta = 2r/(2r+1)$, we obtain

$$\|f_t - f_\rho\|_K \leq (C_1 + C_2 + C_4)t^{-(2r-1)/(4r+2)} + C_3 t^{-1}. \tag{18}$$

Finally,

$$C_1 + C_2 = \left(\frac{2}{2r-1} + \frac{8}{2r+3}\right)\|L_K^{-r} f_\rho\|_\rho = \frac{20r-2}{(2r-1)(2r+3)}\|L_K^{-r} f_\rho\|_\rho.$$

By Lemma 5.5(A) with $f_0 = 0$

$$\|r_0\| = \|f_{\lambda_0}\| \leq \frac{M_\rho}{\sqrt{\lambda_0}} = t_0^{\frac{1}{4r+2}} M_\rho$$

whence $C_3 = (t_0+1)\|f_{\lambda_0}\| \leq 2t_0^{\frac{4r+3}{4r+2}} M_\rho$. We end the proof by plugging these constants into (18). □

6. Upper Bounds for Convergence in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$

In this section we are going to give a probabilistic upper bound for

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

as a proof for Theorem C. Throughout this section, we assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ for some $r \in [1/2, 3/2]$, implying $f_\rho \in \mathscr{H}_K$ with additional regularity. Note that the case $r = 1/2$ is not included in the study of convergence rates for $\|f_t - f_\rho\|_K$.

Similar to Section 5, the idea in the proof starts from the triangle inequality

$$\|f_t - f_\rho\|_\rho \leq \|f_t - f_{\lambda_t}\|_\rho + \|f_{\lambda_t} - f_\rho\|_\rho.$$

But for the first term, instead of using the reversed martingale decomposition, here we need the martingale decomposition in Theorem 3.3, which can be rewritten as

$$r_t = \hat{\Pi}_1^t r_0 + \sum_{j=1}^t \gamma_j \hat{\Pi}_{j+1}^t \chi_j - \sum_{j=1}^t \hat{\Pi}_j^t \Delta_j$$

where $\chi_t = (L_K - L_t)f_{t-1} + (y_t K_{x_t} - L_K f_\rho)$ ($L_t := L_K^{x_t} = \langle\ , K_{x_t}\rangle_K K_{x_t}$), $\Delta_t = f_{\lambda_t} - f_{\lambda_{t-1}}$, and

$$\hat{\Pi}_j^t = \begin{cases} \prod_{i=j}^{t} (I - \gamma_i(L_K + \lambda_i I)), & j \le t; \\ I, & j > t. \end{cases} \tag{19}$$

with the choice that

$$\gamma_t = \frac{1}{(t+t_0)^\theta}, \qquad \lambda_t = \frac{1}{(t+t_0)^{1-\theta}}.$$

The reason of using such a decomposition, is that due to the isometry $L_K^{1/2} : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{H}_K$ such that $\|r_t\|_\rho = \|L_K^{1/2} r_t\|_K$, one can benefit from the spectral decomposition of $L_K^{1/2}\hat{\Pi}_j^t$ to get a tighter estimate. However such a nice feature is lost in the reversed martingale decomposition in that $L_K^{1/2}\Pi_j^t$ does not have an obvious spectral decomposition. On the other hand, due to χ_t depends on f_{t-1}, which increases the difficulty to estimate $\|\chi_t\|_\rho$. For this reason, we can not directly apply Pinelis-Bernstein's inequality to improve Theorem C by replacing $1/\delta$ with $\log 1/\delta$.

As in Section 5, we make the following definitions for convenience.

[Definitions of Errors]

(A) *Initial Error*: $\mathscr{E}_{init}(t) = \|\hat{\Pi}_1^t r_0\|_\rho$, which reflects the propagation error by the initial choice f_0;
(B) *Sample Error*: $\mathscr{E}_{samp}(t) = \|\sum_{j=1}^t \gamma_j \hat{\Pi}_{j+1}^t \chi_j\|_\rho$, where χ_j is a martingale difference sequence, reflecting the random fluctuation caused by sampling;
(C) *Drift Error*: $\mathscr{E}_{drift}(t) = \|\sum_{j=1}^t \hat{\Pi}_j^t \Delta_j\|_\rho$, which measures the error caused by drifts from $f_{\lambda_{j-1}}$ to f_{λ_j} along the regularization path;
(D) *Approximation Error*: $\mathscr{E}_{approx}(t) = \|f_{\lambda_t} - f_\rho\|_\rho$, which measures the distance between the regression function and the regularization path at time t.

In this way we may bound

$$\|f_t - f_\rho\|_\rho \le \mathscr{E}_{init}(t) + \mathscr{E}_{samp}(t) + \mathscr{E}_{drift}(t) + \mathscr{E}_{approx}(t).$$

In the remaining of this section, we are going to provide upper bounds for each of the four errors, which, roughly speaking, are

$$\begin{aligned} \mathscr{E}_{approx}(t) &\le O(t^{-r(1-\theta)}) \\ \mathscr{E}_{drift}(t) &\le O(t^{-r(1-\theta)}) \\ \mathscr{E}_{init}(t) &\le O(t^{-1}) \\ \mathscr{E}_{samp}(t) &\le O(t^{-\theta/2}) \end{aligned}$$

Theorem C then follows from these bounds by setting $\theta = 2r/(2r+1)$.

6.1. **Approximation Error.**

Theorem 6.1 (Approximation Error)**.** *For $r \in (0, 1]$,*

$$\mathscr{E}_{approx}(t) \le C_5 (t+t_0)^{-r(1-\theta)},$$

where $C_5 = r^{-1}\|L_K^{-r} f_\rho\|_\rho$.

Proof. It follows from Theorem 4.1(A) with $\lambda = \lambda_t$ and $\mu = 0$. □

6.2. **Drift Error.**

Theorem 6.2 (Drift Error)**.** *Assume $t_0^\theta \geq \kappa^2 + 1$. For $r \in (0,1]$ and $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_X}$,*

$$\mathscr{E}_{drift}(t) \leq C_6 (t+t_0)^{-r(1-\theta)}$$

where $C_6 = \dfrac{4(1-\theta)}{1-r(1-\theta)} \|L_K^{-r} f_\rho\|_\rho$.

Proof. By Theorem 4.1(A), it follows that

$$\|\Delta_t\| = \|f_{\lambda_t} - f_{\lambda_{t-1}}\|_\rho \leq 4(1-\theta)(t+t_0)^{-r(1-\theta)-1} \|L_K^{-r} f_\rho\|_\rho,$$

use inequality (15) by replacing $r - 1/2$ with r. By Lemma B.1,

$$\|\Pi_j^t\| \leq \frac{j+t_0}{t+t_0+1},$$

whence

$$\begin{aligned} \mathscr{E}_{drift}(t) &= \|\sum_{j=1}^t \Pi_j^t \Delta_j\|_K \leq \frac{4(1-\theta)\|L_K^{-r} f_\rho\|_\rho}{t+t_0+1} \sum_{j=1}^t (j+t_0)^{-r(1-\theta)} \\ &\leq \frac{4(1-\theta)\|L_K^{-r} f_\rho\|_\rho}{1-r(1-\theta)} (t+t_0)^{-r(1-\theta)} \end{aligned}$$

since

$$\sum_{j=1}^t (j+t_0)^{-r(1-\theta)} \leq \int_0^t (x+t_0)^{-r(1-\theta)} dx \leq \frac{1}{1-r(1-\theta)} (t+t_0)^{1-r(1-\theta)}$$

□

6.3. **Initial Error.**

Theorem 6.3 (Initial Error)**.** *Let $a = 1$ and $t_0 \geq (\kappa^2+1)^{1/\theta}$. Then for all $t \in \mathbb{N}$,*

$$\mathscr{E}_{init}(t) \leq C_7 (t+t_0)^{-1},$$

where $C_7 = M_\rho (t_0 + 1)$.

Proof. Similar to Theorem 5.3, by Lemma B.2(D) with $j = 1$ we obtain

$$\|\hat{\Pi}_1^t\| \leq \frac{t_0+1}{t+t_0},$$

which gives

$$\mathscr{E}_{init}(t) \leq (t_0+1)\|r_0\|(t+t_0)^{-1}.$$

For $f_0 = 0$, using Lemma B.3(B), $\|r_0\|_\rho = \|f_{\lambda_0}\|_\rho \leq M_\rho$. □

6.4. **Sample Error.**

Theorem 6.4 (Sample Error)**.** *Assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ for some $r \in [1/2, 1]$ and $t_0^\theta \geq \kappa^2 + 1$. Then with probability at least $1 - \delta$ ($\delta \in (0,1)$), there holds for all $t \in \mathbb{N}$,*

$$\mathscr{E}_{samp}(t) \leq C_8 t^{-\theta/2}$$

where

$$C_8 = \sqrt{\frac{2}{\delta}}\kappa(\sqrt{3}M_\rho + 2\|L_K^{-r} f_\rho\|_\rho + \sqrt{3(\kappa^2+1)}\kappa M_\rho)$$

Before presenting the formal proof, we need an auxillary estimate.

Lemma 6.5. *Assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ for some $r \in [1/2, 1]$ and $t_0^\theta \geq \kappa^2 + 1$. Then for all $t \in \mathbb{N}$, there holds*

$$\mathbb{E}\|\chi_t\|_K^2 \leq C_9.$$

where

$$C_9 = 2\kappa^2(3M_\rho^2 + 4\|L_K^{-r} f_\rho\|_\rho^2 + 3(\kappa^2+1)\kappa^2 M_\rho^2)$$

Proof. By definition

$$\chi_t = (\hat{A}_t - A_t)w_{t-1} + b_t - \hat{b}_t = (L_K - L_t)f_{t-1} + y_t K_{x_t} - L_K f_\rho$$

where $L_t := \langle\ , K_{x_t}\rangle K_{x_t}$. Then

$$\mathbb{E}\|\chi_t\|_K^2 \leq 2\mathbb{E}\|(L_K - L_t)f_{t-1}\|_K^2 + 2\mathbb{E}\|y_t K_{x_t} - L_K f_\rho\|_K^2 \leq 2\mathbb{E}\|L_t f_{t-1}\|_K^2 + 2\mathbb{E}\|y_t K_{x_t}\|_K^2$$

using for $\mathbb{E}[X] = \mu$, $\mathbb{E}\langle X - \mu, X - \mu\rangle = \mathbb{E}\|X\|^2 - \|\mu\|^2 \leq \mathbb{E}\|X\|^2$, with the replacement that $X = L_t$ and $\mu = L_K$, or that $X = y_t K_{x_t}$ and $\mu = L_K f_\rho$, respectively.

Note that the second term $\mathbb{E}\|y_t K_{x_t}\|_K^2 \leq \kappa^2 M_\rho^2$. It remains to bound the first term,

$$\mathbb{E}\|L_t f_{t-1}\|_K^2 = \mathbb{E}\|f_{t-1}(x_t)K_{x_t}\|_K^2 \leq \kappa^2 \mathbb{E}\|f_{t-1}\|_\rho^2 \leq 2\kappa^2(\mathbb{E}\|f_{t-1} - f_{\lambda_{t-1}}\|_\rho^2 + \|f_{\lambda_{t-1}}\|_\rho^2),$$

where using Lemma B.3, $\|f_\lambda\|_\rho \leq M_\rho$, and Corollary C.3 for the bound on $\mathbb{E}\|f_{t-1} - f_{\lambda_{t-1}}\|_\rho^2$,

$$r.h.s. \leq 2\kappa^2(2M_\rho^2 + 4\|L_K^{-r} f_\rho\|_\rho^2 + 3(\kappa^2+1)\kappa^2 M_\rho^2).$$

Putting two terms together gives the result. □

Now we are ready to prove the bound for the sample error.

Proof of Theorem 6.4. Denote $X_j = \gamma_j \hat{\Pi}_{j+1}^t \chi_j$, which is a martingale difference sequence. It suffices to show

$$\mathbb{E}[\|\sum_{j=1}^t X_j\|_\rho^2] \leq C_9(t + t_0)^{-\theta}. \tag{20}$$

Then it follows from the Markov inequality

$$\mathbf{Prob}\left\{(z_i)_1^t \in \mathscr{Z}^t : \|\sum_{j=1}^t X_j\|_\rho \geq \epsilon\right\} \leq \frac{\mathbb{E}[\|\sum_{j=1}^t X_j\|_\rho^2]}{\epsilon^2} \leq \frac{C_9}{\epsilon^2} t^{-\theta}.$$

Setting the right hand side to be δ, and noticing that

$$\sqrt{\frac{C_9}{\delta}} = \sqrt{\frac{2}{\delta}}\kappa(3M_\rho^2+4\|L_K^{-r}f_\rho\|_\rho^2+3(\kappa^2+1)\kappa^2M_\rho^2)^{1/2} \le \sqrt{\frac{2}{\delta}}\kappa(\sqrt{3}M_\rho+2\|L_K^{-r}f_\rho\|_\rho+\sqrt{3(\kappa^2+1)}\kappa M_\rho)$$

using $(a^2+b^2+c^2)^{1/2} \le a+b+c$ for $a,b,c>0$, we obtain the result.

It remains to prove (20). By isometry $L_K^{1/2}: \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{H}_K$,

$$\begin{aligned}\mathbb{E}[\|\sum_{j=1}^t X_j\|_\rho^2] &= \mathbb{E}\|L_K^{1/2}\sum_{j=1}^t X_j\|_K^2 = \sum_{j=1}^t \gamma_j^2\mathbb{E}\|L_K^{1/2}\hat{\Pi}_{j+1}^t\chi_j\|_K^2 \\ &\le \sum_{j=1}^t \gamma_j^2\|\hat{\Pi}_{j+1}^t L_K\hat{\Pi}_{j+1}^t\|\cdot\mathbb{E}\|\chi_j\|_K^2\end{aligned}$$

Using Lemma 6.5, we have $\mathbb{E}\|\chi_j\|_K^2 \le C_9$. To estimate $\sum_{j=1}^t \gamma_j^2\|\hat{\Pi}_{j+1}^t L_K\hat{\Pi}_{j+1}^t\|$, we use the spectral decomposition of $L_K: \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{L}^2_{\rho_\mathscr{X}}$. Let $(\mu_\alpha,\phi_\alpha)_{\alpha\in\mathbb{N}}$ be an orthonormal eigen-system of L_K, by Mercer's Theorem. For simplicity, denote $a_i = \gamma_i\lambda_i + \gamma_i\mu_\alpha$, then

$$\begin{aligned}&\sum_{j=1}^t \gamma_j^2\|\hat{\Pi}_{j+1}^t L_K\hat{\Pi}_{j+1}^t\| \le \sup_{\mu_\alpha}\sum_{j=1}^t \gamma_j^2\mu_\alpha\prod_{i=j+1}^t(1-a_i)^2 \\ =\ &\sup_{\mu_\alpha}\sum_{j=1}^t\left[\gamma_j\prod_{i=j+1}^t(1-a_i)\right]\cdot\left[\gamma_j\mu_\alpha\prod_{i=j+1}^t(1-a_i)\right] \\ \le\ &\sup_{\mu_\alpha}\left\{\left[\sup_j\gamma_j\prod_{i=j+1}^t(1-a_i)\right]\cdot\left[\sum_{j=1}^t\gamma_j\mu_\alpha\prod_{i=j+1}^t(1-a_i)\right]\right\}\end{aligned}$$

where for large enough t_0,

$$\sup_j\gamma_j\prod_{i=j+1}^t(1-a_i) \le \sup_j\gamma_j\prod_{i=j+1}^t(1-\gamma_i\lambda_i) \le \sup_j\frac{1}{(j+t_0)^\theta}\cdot\frac{j+t_0}{t+t_0+1} \le (t+t_0)^{-\theta}, \tag{21}$$

and

$$\sum_{j=1}^t\gamma_j\mu_\alpha\prod_{i=j+1}^t(1-a_i) \le \sum_{j=1}^t(1-(1-\gamma_j\mu_\alpha))\prod_{i=j+1}^t(1-\gamma_i\mu_\alpha) = 1-\prod_{i=1}^t(1-\gamma_i\mu_\alpha) \le 1,$$

which gives (20). □

Remark 6.6. It is still an open problem, if we can improve this bound to replace $1/\delta$ with $\log 1/\delta$, by using the Pinelis-Bernstein inequality for the martingale difference sequence. The difficulty seems that, the Pinelis-Bernstein inequality needs a uniform bound on $\|\gamma_j\hat{\Pi}_{j+1}^t\chi_j\|_\rho$, which so far is only $O(t^{-(1-2\theta)})$, by Lemma B.4 such that $\|\chi_t\|_K \le O(\|f_{t-1}\|_K) \le O(1/\lambda_t)$ and $\|\gamma_j L_K^{1/2}\hat{\Pi}_{j+1}^t\| \le O(t^{-\theta})$ as in the proof above. Then using the Pinelis-Bernstein inequality in Proposition A.3, we have

$$\mathscr{E}_{samp}(t) \le O(t^{-(1-2\theta)}) + O(t^{-\theta/2}),$$

where the first term has a decreasing rate slower than $O(t^{-\theta/2})$ when $\theta = 2r/(2r+1)$ for $r \in [1/2,1]$. The successful application of Pinelis-Bernstein, may rely on an improved estimate $\|f_t\|_\rho \le O(1/\sqrt{\lambda_t})$, which is still open at this moment.

6.5. **Total Error and the Proof of Theorem C.** Choosing $\theta \in [0,1]$ properly, we obtain the Theorem C.

Proof of Theorem C. By triangle inequality,

$$\|f_t - f_\rho\|_\rho \leq \mathscr{E}_{approx}(t) + \mathscr{E}_{drift}(t) + \mathscr{E}_{init}(t) + \mathscr{E}_{samp}(t)$$

Combining Theorem 6.1, 6.2, 6.3, and 6.4, and setting $\theta = 2r/(2r+1)$, we obtain

$$\|f_t - f_\rho\|_\rho \leq (C_5 + C_6 + C_8)(t+t_0)^{-r/(2r+1)} + C_7(t+t_0)^{-1}$$

where

$$C_5 + C_6 = \left(\frac{1}{r} + \frac{4}{r+1}\right)\|L_K^{-r} f_\rho\|_\rho = \frac{5r+1}{r(r+1)}\|L_K^{-r} f_\rho\|_\rho$$

whence

$$C_5 + C_6 + C_8 \quad = \quad \sqrt{\frac{6}{\delta}}\kappa M_\rho(1 + \kappa\sqrt{\kappa^2+1}) + \left(2\sqrt{2}\kappa\sqrt{\frac{1}{\delta}} + \frac{5r+1}{r(r+1)}\right)\|L_K^{-r} f_\rho\|_\rho,$$

which ends the proof. □

Appendix A: A Probabilistic Inequality

The following result is quoted from [Theorem 3.4 in Pinelis 1994].

Lemma A.1 (Pinelis-Bennett)**.** *Let ξ_i be a martingale difference sequence in a Hilbert space. Suppose that almost surely $\|\xi_i\| \leq M$ and $\sum_{i=1}^t \mathbb{E}_{i-1}\|\xi_i\|^2 \leq \sigma_t^2$. Then*

$$\mathbf{Prob}\left\{\left\|\sum_{i=1}^t \xi_i\right\| \geq \epsilon\right\} \leq 2\exp\left\{-\frac{\sigma_t^2}{M^2} g\left(\frac{M\epsilon}{\sigma_t^2}\right)\right\},$$

where $g(x) = (1+x)\log(1+x) - x$ for $x > 0$.

Using the lower bound $g(x) \geq \frac{x^2}{2(1+x/3)}$, one may obtain the following generalized Bernstein's inequality.

Corollary A.2 (Pinelis-Bernstein)**.** *Let ξ_i be a martingale difference sequence in a Hilbert space. Suppose that almost surely $\|\xi_i\| \leq M$ and $\sum_{i=1}^t \mathbb{E}_{i-1}\|\xi_i\|^2 \leq \sigma_t^2$. Then*

$$\mathbf{Prob}\left\{\left\|\sum_{i=1}^t \xi_i\right\| \geq \epsilon\right\} \leq 2\exp\left\{-\frac{\epsilon^2}{2(\sigma_t^2 + M\epsilon/3)}\right\}. \tag{A-1}$$

The following result will be used as a basic probabilistic inequality to derive various bounds.

Proposition A.3. *Let ξ_i be a martingale difference sequence in a Hilbert space. Suppose that almost surely $\|\xi_i\| \leq M$ and $\sum_{i=1}^t \mathbb{E}_{i-1}\|\xi_i\|^2 \leq \sigma_t^2$. Then the following holds with probability at least $1-\delta$ ($\delta \in (0,1)$),*

$$\left\|\sum_{i=1}^t \xi_i\right\| \leq 2\left(\frac{M}{3} + \sigma_t\right)\log\frac{2}{\delta}.$$

Proof. Taking the right hand side of (A-1) to be δ, then we arrive at the following quadratic equation for ϵ,

$$\epsilon^2 - \frac{2M}{3}\epsilon \log\frac{2}{\delta} - 2\sigma_t^2 \log\frac{2}{\delta} = 0.$$

Note that $\epsilon > 0$, then

$$\begin{aligned}
\epsilon &= \frac{1}{2}\left\{\frac{2M}{3}\log\frac{2}{\delta} + \sqrt{\frac{4M^2}{9}\log^2\frac{2}{\delta} + 8\sigma_t^2\log\frac{2}{\delta}}\right\} \\
&= \frac{M}{3}\log\frac{2}{\delta} + \sqrt{\left(\frac{M}{3}\right)^2\log^2\frac{2}{\delta} + 2\sigma_t^2\log\frac{2}{\delta}} \\
&\leq \frac{2M}{3}\log\frac{2}{\delta} + \sqrt{2\sigma_t^2\log\frac{2}{\delta}},
\end{aligned}$$

where the second last step is due to $\sqrt{a^2+b^2} \leq a+b$ $(a,b>0)$ with

$$a = \frac{M}{3}\log\frac{2}{\delta}, \quad \text{and} \quad b = \sqrt{2\sigma_t^2\log\frac{2}{\delta}}.$$

We complete the proof by relaxing $\sqrt{2\sigma_t^2\log 2/\delta} \leq 2\sigma_t \log 2/\delta$ since $2\log 2/\delta > 1$ for $\delta \in (0,1)$. □

Appendix B: Basic Estimates

Lemma B.1. *For all $t_0 > 0$ and $a > 0$,*

$$\prod_{i=j}^{t}\left(1 - \frac{a}{i+t_0}\right) \leq \left(\frac{j+t_0}{t+t_0+1}\right)^a$$

Proof.

$$\begin{aligned}
\prod_{i=j}^{t}\left(1 - \frac{a}{i+t_0}\right) &\leq \exp\{-a\sum_{i=j}^{t}(i+t_0)^{-1}\} \\
&\leq \exp\{-a\int_{j+t_0}^{t+t_0+1} x^{-1}dx\} \\
&= \exp\left\{\ln\left(\frac{j+t_0}{t+t_0+1}\right)^a\right\} = \left(\frac{j+t_0}{t+t_0+1}\right)^a
\end{aligned}$$

□

Lemma B.2. *If $t_0^\theta \geq \kappa^2 + 1$, then the following holds for all $t \in \mathbb{N}$,*

(A) $\|I - \gamma_t A_t\| \leq 1 - \dfrac{1}{t+t_0}$;

(B) $\|\Pi_j^t\| \leq \dfrac{j+t_0}{t+t_0+1}$;

(C) $\|I - \gamma_t \hat{A}_t\| \leq 1 - \dfrac{1}{t+t_0}$;

(D) $\|\hat{\Pi}_j^t\| \le \dfrac{j+t_0}{t+t_0+1}$.

Proof. (A) First we show that $\gamma_t(\kappa^2+\lambda_t)<1$. In fact, for $t\in\mathbb{N}$,

$$\gamma_t\lambda_t+\gamma_t\kappa^2=\frac{1}{t+t_0}+\frac{\kappa^2}{(t+t_0)^\theta}\le\frac{1+\kappa^2}{t_0^\theta}\le 1.$$

Therefore

$$\|I-\gamma_tA_t\|=\|I-\gamma_tL_t-\gamma_t\lambda_t\|\le 1-\gamma_t\lambda_t=1-\frac{1}{t+t_0},$$

since $\|L_t\|\le\kappa^2$.

(B) To show this,

$$\|\Pi_j^t\|=\|\prod_{i=j}^t(I-\gamma_iA_i)\|\le\prod_{i=j}^t(1-\gamma_i\lambda_i)=\prod_{i=j}^t(1-\frac{1}{i+t_0})=\frac{j+t_0-1}{t+t_0}\le\frac{j+t_0}{t+t_0+1}$$

where the last step is due to Lemma B.1.

(C) and (D) are similar to (A) and (B), respectively. □

Lemma B.3. *For any* $\lambda>0$,

(A) $\|f_\lambda\|_K\le M_\rho/\sqrt{\lambda}$;

(B) $\|f_\lambda\|_\rho\le M_\rho$.

Proof. (A) Note that

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

Taking $f=0$, we have

$$\|f_\lambda-f_\rho\|_\rho^2+\lambda\|f_\lambda\|_K^2\le\|f_\rho\|_\rho^2\le M_\rho^2, \tag{B-1}$$

which leads to the result.

(B) By definition we obtain

$$\|f_\lambda\|_\rho=\|(L_K+\lambda I)^{-1}L_Kf_\rho\|_\rho\le\|(L_K+\lambda I)^{-1}L_K\|\cdot\|f_\rho\|_\rho\le\|f_\rho\|_\rho\le M_\rho.$$

□

Lemma B.4. *If* $f_0=0$, *then for all* $t\in\mathbb{N}$,

$$\|f_t\|_K\le\frac{\kappa M_\rho}{\lambda_t}$$

Proof. Since

$$f_t=f_{t-1}-\gamma_t((f_{t-1}(x_t)-y_t)K_{x_t}+\lambda_tf_{t-1})=(1-\gamma_t\lambda_t-\gamma_tL_K^{x_t})f_{t-1}+\gamma_ty_tK_{x_t}$$

then for $t_0\ge[(1+\kappa^2)]^{(2r+1)/2r}$, $\gamma_t\kappa^2+\gamma_t\lambda_t\le 1$, whence

$$\|f_t\|_K\le\|1-\gamma_t\lambda_t-\gamma_tL_K^{x_t}\|\|f_{t-1}\|_K+\gamma_t\|y_tK_{x_t}\|_K\le(1-\gamma_t\lambda_t)\|f_{t-1}\|_K+\gamma_t\kappa M_\rho.$$

By induction on t, we have

$$\|f_t\|_K \leq \prod_{i=1}^{t}(1-\gamma_i\lambda_i)\|f_0\|_K + \kappa M_\rho \sum_{j=1}^{t}\gamma_j \prod_{i=j+1}^{t}(1-\gamma_i\lambda_i).$$

The first term is 0 since $f_0 = 0$. In the second term

$$\sum_{j=1}^{t}\gamma_j \prod_{i=j+1}^{t}(1-\gamma_i\lambda_i) \leq \max_{1\leq j\leq t}(\frac{1}{\lambda_j})\sum_{j=1}^{t}\gamma_j\lambda_j \prod_{i=j+1}^{t}(1-\gamma_i\lambda_i) \leq \frac{1}{\lambda_t}$$

since

$$\sum_{j=1}^{t}\gamma_j\lambda_j \prod_{i=j+1}^{t}(1-\gamma_i\lambda_i) = 1 - \prod_{i=1}^{t}(1-\gamma_i\lambda_i).$$

This gives the bound. □

Appendix C: Estimates for the Path Gap

In this section we derive some direct estimates for the remainder variance $\mathbb{E}\|f_t - f_{\lambda_t}\|^2$, in $\mathscr{H}_K$ norm or $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ norm. The result in Lemma C.2 will be used to derive a constant upper bound for $\mathbb{E}\|f_t - f_{\lambda_t}\|^2_\rho$ when $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ with $r \in [1/2, 1]$, i.e. Corollary C.3. Another application of Lemma C.2, which is not pursued in this paper, is that it can be used to derive Theorem A directly.

Lemma C.1. *Let $f \in \mathscr{H}_K$, and $\gamma, \lambda \in \mathbb{R}_+$. For all $z = (x, y) \in X \times \mathbb{R}$, let*

$$f^z := f - \gamma((f(x) - y)K_x + \lambda f) = f - \gamma((L_K^x + \lambda f) - g^z), \quad \text{where } g^z := yK_x.$$

Then

$$\mathbb{E}\|f^z - f_\lambda\|^2_\star \leq (1-\gamma\lambda)^2\|f - f_\lambda\|^2_\star - 2\gamma(1 - \gamma\lambda - 2\gamma\kappa^2)\|L_K^{1/2}(f - f_\lambda)\|^2_\star + \gamma^2 C_\star.$$

where $\star$ stands for either ρ or K, and

$$C_\star = \begin{cases} 6\kappa^2 M_\rho^2, & \star = K \\ 3(\kappa^2+1)\kappa^2 M_\rho^2, & \star = \rho \end{cases}$$

Proof. Using the expression $\lambda f_\lambda = L_K f_\rho - L_K f_\lambda$ from $(L_K + \lambda I)f_\lambda = L_K f_\rho$,

$$\begin{aligned} f^z - f_\lambda &= f - f_\lambda - \gamma((L_K^x + \lambda I)f - g^z) \\ &= [I - \gamma(L_K + \lambda I)](f - f_\lambda) + \gamma(L_K + \lambda I)(f - f_\lambda) - \gamma((L_K^x + \lambda I)f - g^z) \\ &= [I - \gamma(L_K + \lambda I)](f - f_\lambda) + \gamma(L_K - L_K^x)f - \gamma(L_K f_\rho - g^z) \end{aligned}$$

Therefore

$$\mathbb{E}\|f^z - f_\lambda\|^2_\star = \|[I - \gamma(L_K + \lambda I)](f - f_\lambda)\|^2_\star + \gamma^2\zeta(f) \tag{C-1}$$

where

$$\zeta(f) := \mathbb{E}[\|(L_K - L_K^x)f - (L_K f_\rho - g^z)\|^2_\star]$$

since

$$\mathbb{E}[(L_K - L_K^x)f - (L_K f_\rho - g^z)] = 0.$$

Let us now study the two terms in equation (C-1). First,

$$\begin{aligned}
&\|[I-\gamma(L_K+\lambda I)](f-f_\lambda)\|_\star^2\\
&=(1-\gamma\lambda)^2\|f-f_\lambda\|_\star^2-2\gamma(1-\gamma\lambda)\langle L_K(f-f_\lambda),f-f_\lambda\rangle_\star+\gamma^2\|L_K(f-f_\lambda)\|_\star^2\\
&\le(1-\gamma\lambda)^2\|f-f_\lambda\|_\star^2-2\gamma(1-\gamma\lambda)\|L_K^{1/2}(f-f_\lambda)\|_\star^2+\gamma^2\kappa^2\|L_K^{1/2}(f-f_\lambda)\|_\star^2
\end{aligned}$$

It remains to estimate $\zeta(f)$:

$$\begin{aligned}
\zeta(f) &= \mathbb{E}[\|(L_K-L_K^x)(f-f_\lambda)+(L_K-L_K^x)f_\lambda+(L_Kf_\rho-g^z)\|_\star^2]\\
&\le 3\mathbb{E}[\|(L_K-L_K^x)(f-f_\lambda)\|_\star^2+\|(L_K-L_K^x)f_\lambda\|_\star^2+\|(L_Kf_\rho-g^z)\|_\star^2]\\
&\le 3\mathbb{E}[\|L_K^x(f-f_\lambda)\|_\star^2+\|L_K^xf_\lambda\|_\star^2+\|g^z\|_\star^2]\\
&\le 3\kappa^2[\|L_K^{1/2}(f-f_\lambda)\|_\star^2+\|L_K^{1/2}f_\lambda\|_\star^2+M_\rho^2]
\end{aligned}$$

where if $\star=K$,

$$r.h.s.=3\kappa^2[\|L_K^{1/2}(f-f_\lambda)\|_K^2+\|L_K^{1/2}f_\lambda\|_K^2+M_\rho^2]\le 3\kappa^2\|f-f_\lambda\|_\rho^2+6\kappa^2M_\rho^2$$

using $\|L_K^{1/2}f_\lambda\|_K=\|f_\lambda\|_\rho\le M_\rho$, and if $\star=\rho$,

$$r.h.s.=3\kappa^2[\|L_K^{1/2}(f-f_\lambda)\|_\rho^2+\|L_K^{1/2}f_\lambda\|_\rho^2+M_\rho^2]\le 3\kappa^2\|L_K^{1/2}(f-f_\lambda)\|_\rho^2+3(\kappa^2+1)\kappa^2M_\rho^2$$

using $\|L_K^{1/2}f_\lambda\|_\rho=\kappa\|f_\lambda\|_\rho\le\kappa M_\rho$. Combining the two estimates gives the result. □

Lemma C.2. *Let*

$$\pi_k^t=\begin{cases}\displaystyle\prod_{i=k}^t(1-\gamma_i\lambda_i), & k\le t;\\ I, & k>t.\end{cases}\tag{C-2}$$

If for all $t\in\mathbb{N}$, $\gamma_t(\lambda_t+2\kappa^2)\le 1$, then

$$\mathbb{E}\|f_t-f_{\lambda_t}\|_\star^2\le\pi_1^t\|f_0-f_{\lambda_0}\|_\star^2+\sum_{k=1}^t\pi_{k+1}^t\frac{\|f_{\lambda_k}-f_{\lambda_{k-1}}\|_\star^2}{\gamma_k\lambda_k}+C_\star\sum_{k=1}^t\gamma_k^2\pi_{k+1}^t.$$

where $\star$ stands for either ρ or K, and

$$C_\star=\begin{cases}6\kappa^2M_\rho^2, & \star=K\\ 3(\kappa^2+1)\kappa^2M_\rho^2, & \star=\rho\end{cases}$$

Proof. Using Lemma C.1, for $\gamma_t(\lambda_t+2\kappa^2)\le 1$,

$$\mathbb{E}\|f_t-f_{\lambda_t}\|_\star^2\le(1-\gamma_t\lambda_t)^2\|f_{t-1}-f_{\lambda_t}\|_\star^2+C_\star.$$

Note that

$$\begin{aligned}
\|f_{t-1}-f_{\lambda_t}\|_\star &\le \|f_{t-1}-f_{\lambda_{t-1}}\|_\star+\|f_{\lambda_t}-f_{\lambda_{t-1}}\|_\star\\
&= \|f_{t-1}-f_{\lambda_{t-1}}\|_\star+\delta_t, \qquad \text{define } \delta_t:=\|f_{\lambda_t}-f_{\lambda_{t-1}}\|_\star\\
&\le \|f_{t-1}-f_{\lambda_{t-1}}\|_\star^2(1+\gamma_t\lambda_t)+\delta_t^2(1+1/(\gamma_t\lambda_t))
\end{aligned}$$

using that, for all $a,b,c\in\mathbb{R}_+$,

$$(a+b)^2\le a^2(1+c)+b^2(1+1/c)$$

with $x:=\|f_{t-1}-f_{\lambda_{t-1}}\|_\star$, $a:=\delta_t$ and $b:=\gamma_t\lambda_t$.

This gives the iteration formula,

$$\mathbb{E}\|f_t - f_{\lambda_t}\|_\star^2 \le (1-\gamma_t\lambda_t)\|f_{t-1} - f_{\lambda_{t-1}}\|_\star^2 + (1-\gamma_t\lambda_t)\frac{\delta_t^2}{\gamma_t\lambda_t} + \gamma_t^2 C_\star,$$

which, by induction, leads to

$$\mathbb{E}\|f_t - f_{\lambda_t}\|_\star^2 \le \pi_1^t\|f_0 - f_{\lambda_0}\|_\star^2 + \sum_{k=1}^t \pi_{k+1}^t \frac{\delta_k^2}{\gamma_k\lambda_k} + C_\star \sum_{k=1}^t \gamma_k^2 \pi_{k+1}^t$$

which ends the proof. □

Corollary C.3. *Assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ with $r \in [1/2, 1]$ and $t_0^\theta \ge \kappa^2 + 1$. Then*

$$\mathbb{E}\|f_t - f_{\lambda_t}\|_\rho^2 \le M_\rho^2 + 4\|L_K^{-r} f_\rho\|_\rho^2 + 3(\kappa^2+1)\kappa^2 M_\rho^2$$

Proof. For $t_0^\theta \ge \kappa^2 + 1$, we have $\gamma_t\kappa^2 + \gamma_t\lambda_t \le 1$, whence by Lemma C.2 with $f_0 = 0$,

$$\mathbb{E}\|f_t - f_{\lambda_t}\|_\rho^2 \le \pi_1^t\|f_{\lambda_0}\|_\rho^2 + \sum_{j=1}^t \gamma_j\lambda_j\pi_{j+1}^t \frac{\|f_{\lambda_j} - f_{\lambda_{j-1}}\|_\rho^2}{\gamma_j\lambda_j} + C_\rho \sum_{j=1}^t \frac{\gamma_j}{\lambda_j}\gamma_j\lambda_j\pi_{j+1}^t \tag{C-3}$$

The first term is not larger than M_ρ^2, using $\pi_1^t \le 1$ and $\|f_\lambda\|_\rho \le M_\rho$ in Lemma B.3(B).

Now consider the second term. By Theorem 4.1(A) with $r \in [1/2, 1]$,

$$\|f_{\lambda_t} - f_{\lambda_{t-1}}\|_\rho \le |\lambda_t - \lambda_{t-1}| \frac{\|L_K^{-r} f_\rho\|_\rho}{r} \le 4(1-\theta)(t+t_0)^{-r(1-\theta)-1}\|L_K^{-r} f_\rho\|_\rho$$

where we use

$$\begin{aligned} |\lambda_t^r - \lambda_{t-1}^r| &= |(t+t_0)^{-r(1-\theta)} - (t+t_0-1)^{-r(1-\theta)}| \\ &\le r(1-\theta)(t+t_0-1)^{-r(1-\theta)-1} \le 4r(1-\theta)(t+t_0)^{-r(1-\theta)-1}, \end{aligned}$$

using the Mean Value Theorem and $(a+1)/(b+1) \ge a/b$ for $b > a > 0$. This gives for all $t \in \mathbb{N}$, $t_0 \ge 1$ and $\theta \in [1/2, 1]$,

$$\frac{\|f_{\lambda_t} - f_{\lambda_{t-1}}\|_\rho^2}{\gamma_t\lambda_t} \le 16(1-\theta)^2\|L_K^{-r} f_\rho\|_\rho^2 (t+t_0)^{-2r(1-\theta)-1} \le 4\|L_K^{-r} f_\rho\|_\rho^2$$

Using the telescope sum

$$\sum_{j=1}^t \gamma_j\lambda_j\pi_{j+1}^t = 1 - \pi_1^t \le 1, \tag{C-4}$$

we have a bound for the second term

$$\sum_{j=1}^t \gamma_j\lambda_j\pi_{j+1}^t \frac{\|f_{\lambda_j} - f_{\lambda_{j-1}}\|_\rho^2}{\gamma_j\lambda_j} \le 4\|L_K^{-r} f_\rho\|_\rho^2$$

It remains to bound the third term. Note that for $\theta \in [1/2, 1]$,

$$\frac{\gamma_t}{\lambda_t} = (t+t_0)^{-(2\theta-1)} \le 1.$$

Together with the telescope sum (C-4), the third term is not larger than C_ρ. This completes the proof. □

Appendix D: Proof of Theorem 3.6

Proof of Theorem 3.6. For convenience suppose that

$$\mathbb{E}\|(A_j\hat{w}_j - b_j)\|^2 \le C < \infty. \tag{D-1}$$

By Generalized Finiteness Condition (C) $\underline{\alpha}_t \to 0$ and condition (B) above $\gamma_t/\underline{\alpha}_t \to 0$, we have $\gamma_t \to 0$. Thus for a sufficiently large t_0, we have for all $t > t_0$, $\gamma_t\overline{\alpha} + \gamma_t\underline{\alpha}_t \le 1$ and thus

$$\|\Pi_j^t\| \le \prod_{i=j}^{t}(1-a_i), \qquad \text{where } a_i := \gamma_i\underline{\alpha}_i \text{ and } t \ge j > t_0. \tag{D-2}$$

Note that by condition (A), $\sum_i a_i = \sum_i \gamma_i\underline{\alpha}_i = \infty$.

The first term in (9) converges, by $\sum_i a_i = \infty$ and thus

$$\|\Pi_{t_0+1}^t r_{t_0}\| \le \prod_{i=t_0+1}^{t}(1-a_i)\|r_{t_0}\| \le e^{-\sum_{i=t_0+1}^{t} a_i}\|r_{t_0}\| \to 0.$$

Now consider the second term in (9). By independence of (z_t),

$$\mathbb{E}\left\|\sum_{j=t_0+1}^{t}\gamma_j\Pi_{j+1}^t(A_j\hat{w}_j - b_j)\right\|^2 = \sum_{j=t_0+1}^{t}\gamma_j^2\mathbb{E}\|\Pi_{j+1}^t(A_j\hat{w}_j - b_j)\|^2 \le C\sum_{j=t_0+1}^{t}\gamma_j^2\prod_{i=j+1}^{t}(1-a_i)^2 =: I.$$

using (D-1) and (D-2). I can be further split into two parts: for each $\epsilon > 0$, choose a large enough $J_1 \in \mathbb{N}$ such that for all $j \ge J_1$, $\gamma_j/\underline{\alpha}_j \le \sqrt{\epsilon/2C}$ (by condition (B) $\gamma_j/\underline{\alpha}_j \to 0$), whence

$$I \le C\sum_{j=t_0+1}^{J_1}\gamma_j^2\prod_{i=j+1}^{t}(1-a_i)^2 + \frac{\epsilon}{2}\sum_{j=J_1+1}^{t}a_j^2\prod_{i=j+1}^{t}(1-a_i)^2 =: I_1 + I_2.$$

Clearly $I_2 \le \epsilon/2$, by using telescope sum

$$\sum_{j=J}^{t}a_j\prod_{i=j+1}^{t}(1-a_i) = \sum_{j=J}^{t}[1-(1-a_j)]\prod_{i=j+1}^{t}(1-a_i) = 1 - \prod_{i=t_0+1}^{t}(1-a_i) \le 1. \tag{D-3}$$

Also $I_1 \le \epsilon/2$, since there is a large enough $T_2 \in \mathbb{N}$ such that for all $t \ge T_2$, $\displaystyle\sum_{i=J_1}^{t}a_i \ge \frac{1}{2}\log\frac{2CJ_1}{\epsilon\overline{\alpha}^2}$ (as condition (A) $\sum_i a_i = \infty$). Therefore $I \le \epsilon$ which shows the convergence of the second term.

As to the third term in (9), denoting $\delta_t = \|\Delta_t\|$, by (D-2) we obtain

$$\left\|\sum_{j=t_0+1}^{t}\Pi_j^t\Delta_j\right\| \le \sum_{j=t_0+1}^{t}\delta_j\prod_{i=j+1}^{t}(1-a_i).$$

By condition (C) $\delta_t/a_t \to 0$, using a similar splitting treatment in the second term, we can obtain the convergence of the third term. □

References

Berlinet, A. and C. Thomas-Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers.

Caponnetto, A. and E. De Vito (2005). Optimal rates for regularized least squares algorithm. *preprint*.

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

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

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

Duflo, M. (1996). *Algorithmes Stochastiques*. Berlin, Heidelberg: Springer-Verlag.

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

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

Halmos, R. P. and V. S. Sunder (1978). *Bounded Integral Operators in L^2 Spaces*. Vol. 96 of Ergebnisse der Mathematik und ihrer Grenzgebiete (Results in Mathematics and Related Areas). Berlin: Springer-Verlag.

Hastie, T., S. Rosset, R. Tibshirani, and J. Zhu (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research 5*, 1391–1415.

Kivinen, J., A. J. Smola, and R. C. Williamson (2004). Online learning with kernels. *IEEE Transactions on Signal Processing 52*(8), 2165–2176.

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

Loéve, M. (1948). Fonctions alèatoires du second ordre. In P. Lèvy (Ed.), *Processus Stochastiques et Mouvement Brownien*, Paris. Gauthier-Villars.

Neveu, J. (1975). *Discrete-Parameter Martingales*. North-Holland Publishing Company.

Parzen, E. (1961). An approach to time series analysis. *The Annals of Mathematical Statistics 32*(4), 951–989.

Pinelis, I. (1994). Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability 22*(4), 1679–1706.

Robbins, H. and S. Monro (1951). A stochastic approximation method. *The Annals of Mathematical Statistics 22*(3), 400–407.

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

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

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

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

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

Pierre Tarrès, Mathematical Institute, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, U.K.

E-mail address: tarres@maths.ox.ac.uk

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

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

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