Online Learning as Stochastic Approximation of Regularization Paths: Optimality and Almost-Sure Convergence

Pierre Tarrès and Yuan Yao

Abstract—**In this paper, an online learning algorithm is proposed as sequential stochastic approximation of a regularization path converging to the regression function in reproducing kernel Hilbert spaces (RKHSs). We show that it is possible to produce the best known strong (RKHS norm) convergence rate of batch learning, through a careful choice of the gain or step size sequences, depending on regularity assumptions on the regression function. The corresponding weak (mean square distance) convergence rate is optimal in the sense that it reaches the minimax and individual lower rates in this paper. In both cases, we deduce almost sure convergence, using Bernstein-type inequalities for martingales in Hilbert spaces. To achieve this, we develop a bias-variance decomposition similar to the batch learning setting; the bias consists in the approximation and drift errors along the regularization path, which display the same rates of convergence, and the variance arises from the sample error analyzed as a (reverse) martingale difference sequence. The rates above are obtained by an optimal tradeoff between the bias and the variance.**

Index Terms—**Online learning, stochastic approximations, regularization path, reproducing kernel Hilbert space.**

I. INTRODUCTION

CONSIDER the following classical problem of learning from examples: given a sequence of i.i.d. 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 seeks to approximate from some hypothesis space $\mathscr{H}$ the *regression function*

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

i.e. the conditional expectation of y given x. Recall that f_ρ minimizes the following mean square error

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

In the latter half of the last century till now, we have seen a large volume of literature on exploring the hypothesis space $\mathscr{H}$ as a reproducing kernel Hilbert space (RKHS) $\mathscr{H}_K$ for some positive semi-definite kernel K [2], [5], [22], [26], [33]. RKHS provides us a unified framework for nonparametric regressions including several important settings, e.g.

(i) generalized smooth spline functions in Sobolev spaces [33],
(ii) real analytic functions with bounded bandwidth [10] and their generalizations [28],
(iii) Gaussian processes [20]; [22].

In fact, any Hilbert space of functions on $\mathscr{X}$ with a bounded evaluation functional is a RKHS [2], [33]. By choosing suitable kernels, $\mathscr{H}_K$ can be used to approximate any function in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$, the square integrable functions with respect to the marginal probability measure $\rho_{\mathscr{X}}$. With such a dense function space $\mathscr{H}$, regularization is necessary where the following Tikhonov regularization is widely adopted [9], [14]. Let, for all $\lambda > 0$, f_λ be the solution of the regularized least square problem

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

Depending on assumptions on the Hilbert space $\mathscr{H}$ and on the regularity of f_ρ, f_λ converges to f_ρ in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ or $\mathscr{H}$-norm when $\lambda \to 0$. The map

$$\begin{aligned} f_{\cdot} : \mathbb{R}_+ &\longrightarrow \mathscr{H} \\ \lambda &\longmapsto f_\lambda \end{aligned}$$

is called *regularization path* of f_ρ in $\mathscr{H}$.

Regularization paths gained rising attention from statistics recently, particularly because that the regularization paths of LASSO [12] are piecewise linear, which enables one to track the entire path by locating a finite number of change points. This property generalizes to the case where the loss and the penalty are respectively piecewise quadratic and linear [25]. However Tikhonov regularization does not own piecewise linear paths.

In machine learning, a *batch learning* algorithm refers to a mapping to $\mathscr{H}$ from a sample set given once and for all at some fixed size m, *i.e.* $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$. For instance, Tikhonov regularization yields (also called Ridge Regression in statistics)

$$f_{\mathbf{z},\lambda} := \arg\min_{f\in\mathscr{H}} \left\{\hat{\mathscr{E}}_{\mathbf{z}}(f) + \lambda\|f\|^2_{\mathscr{H}}\right\} \tag{3}$$

Manuscript received April 11, 2013; revised January 9, 2014; accepted May 24, 2014. Date of publication June 24, 2014; date of current version August 14, 2014. P. Tarrès was supported by the Leverhulme Prize. Y. Yao was supported in part by the National Basic Research Program of China (973 Program) under Grant 2011CB809105 and Grant 2012CB825501, in part by the National Science Foundation of China under Grants 6137004 and 61071157, in part by the National Science Foundation under Grant NSF0325113, and in part by the Hundred Talents Program, Peking University, Beijing, China.

P. Tarrès is with the Mathematical Institute, University of Oxford, Oxford OX1 3LB, U.K., and also with the Universite Paris-Dauphine, Paris Cedex 16 75775, France (e-mail: tarres@maths.ox.ac.uk).

Y. Yao is with the School of Mathematical Sciences, LMAM-LMEQF-LMPR, Peking University, Beijing 100871, China (e-mail: yuany@math.pku.edu.cn).

Communicated by N. Cesa-Bianchi, Associate Editor for Pattern Recognition, Statistical Learning Learning and Inference.

Digital Object Identifier 10.1109/TIT.2014.2332531

where the empirical error is defined by

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

For more background on regularization of inverse problems, see for instance [13]. In modern statistics, an L_1-type regularization called LASSO [32], is proposed in pursuit of sparsity of f_ρ with respect to certain basis.

The regularization parameter λ is chosen as a function of the sample size m, and of some prior knowledge on the regularity of the function f_ρ, such that as $m \to \infty$, $\lambda_m \to 0$ and $f_{\mathbf{z},\lambda_m} \to f_\rho$. In this setting, rigorous probabilistic upper bounds of $\|f_{\mathbf{z},\lambda_m} - f_\rho\|_{\mathscr{H}}$ were obtained for instance in [9] and [29].

In a contrast, an *online learning algorithm* aims at obtaining this approximation of the regression function recursively, using at each time step the new example $z_t = (x_t, y_t)$ to update the current hypothesis f_{t-1} (approximating f_ρ) to f_t. In other words, $f_t = T_t(f_{t-1}, z_t)$ for some map $T_t : \mathscr{H} \times \mathscr{X} \times \mathscr{Y} \to \mathscr{H}$. For example, [27] proposed online learning algorithms as stochastic gradient descent or Robbins-Monro procedure [24] to solve (2) with a fixed regularization $\lambda > 0$. In this setting, tight probabilistic upper bounds for the convergence $f_t \to f_\lambda$ are derived in [36], which further shows an averaging process can achieve the same convergence rates as batch learning [29], being minimax optimal for stochastic approximation under strongly convex objective functions [3]; see also [37] for further bounds in expected L^2-distance to f_ρ, when λ is chosen as a function of the total sample size up to time T as in batch learning. Note that for fixed regularization parameter, stochastic gradient descent can be generalized to incremental methods that are effective to solve the equivalent (3) with general convex loss and regularization schemes [6].

However, these results are only with a fixed regularization $\lambda_t = \lambda > 0$ with bounds on $\|f_t - f_\lambda\|$. In online learning, the sample size t increases as time goes on, whence the regularization parameter λ_t needs to be updated such that f_t follows the regularization path f_{λ_t} with $\|f_t - f_{\lambda_t}\| \to 0$ and $f_{\lambda_t} \to f_\rho$. Note that [37] obtain bounds in expected L^2-distance on a large regularity class for f_ρ when $\lambda = 0$, although no almost-sure convergence is obtained. Recently [3] also proposed some bounds on the expected loss or risk when it is non-strongly convex, which provides a weaker convergence than in expected L^2-distance, but without any regularity assumption on f_ρ.

Our purpose in this paper is to iteratively define an "online" sequence of functions $(f_t)_{t\in\mathbb{N}} \in \mathscr{H}$, which will provide a stochastic approximation of the Tikhonov regularization path $(f_{\lambda_t})_{t\in\mathbb{N}} \in \mathscr{H}$. The main theorems in this paper provide some probabilistic upper bounds to guarantee 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 that $f_\rho \in \mathscr{H}_K$ has additional regularity. With an adequate choice of the regularization parameters $\lambda_t \to 0$ based on a bias-variance trade-off, the convergence rate in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ is optimal in the sense that it reaches the minimax and individual lower rate, and the convergence rate in $\mathscr{H}_K$ meets the same best rates as in batch learning [29]. Critically, both upper bounds depend on a logarithmic power $\alpha > 0$ of the confidence threshold δ (i.e. $O(\log^\alpha 1/\delta)$). They imply by Borel-Cantelli Lemma the almost sure convergence of f_t to f_ρ in $\mathscr{H}_K$ and $\mathscr{L}^2_{\rho_{\mathscr{X}}}$. Such a theorem improves our early result (see [35]), where in mean square distance the upper bounds depended polynomially on the confidence (i.e. $O(\delta^{-\alpha})$), and hence solves the open problem raised therein.

Our analysis starts in the setting of a general Hilbert space in Section III, with the study of an iteratively defined sequence, which is a stochastic approximation of the solution of some linear equation. This study will be specialized in later sections to the cases of $\mathscr{H}_K$ or $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ in order to show the main results of the paper. Two structural decomposition theorems are introduced in that Section III, the *reversed martingale decomposition* and the *martingale decomposition*, and play an important role in the proof of the main results, the former being suitable for strong convergence in $\mathscr{H}_K$ and the latter for weak convergence in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$.

Both decompositions lead to the breakdown of the total error $f_t - f_\rho$ into four parts: the *initial error* caused by the initial guess f_0, the *sample error* as a reverse martingale difference sequence, the *approximation error* $f_{\lambda_t} - f_\rho$, and the *drift error* along the regularization path (f_{λ_t}) caused by time-varying λ_t. By a suitable choice of step sizes, the initial error won't affect the convergence rates. Now a key observation is that the drift error, which does not appear in previous fixed regularization settings with $\lambda_t = \lambda$, has the same order as the approximation error. Bernstein-type inequalities for martingales in Hilbert spaces are then used to bound the sample error. Therefore we have a similar bias-variance decomposition in online learning as in batch learning, with the bias being the approximation and the drift errors, and the variance being the sample error. It is then possible to optimize in order to yield the same optimal rates in online learning as in batch learning.

The paper is organized as follows. Section II collects the main results. Section III studies stochastic approximations of regularization paths for linear operator equations in general Hilbert spaces, where the key martingale and reverse martingale decompositions are presented. Section IV collects some estimates on the drift along the regularization path, $\|f_\lambda - f_\mu\|$ ($\lambda, \mu > 0$), which are needed for the study of the bias, i.e. the approximation and drift errors. Sections V and VI yield upper bounds for convergence in $\mathscr{H}_K$ and $\mathscr{L}^2_{\rho_{\mathscr{X}}}$, respectively. Appendix A derives a probabilistic inequality from the Pinelis-Bernstein inequality for martingales in Hilbert spaces, which is used to derive the probabilistic upper bounds in this paper. Appendix B collects some preliminary upper bounds used in the paper. Appendix C gives proofs of some results in Section III-B.

II. Main Results

A. 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 let $\rho_{\mathscr{Y}|x}$ 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 let $\mathbb{E}[\cdot]$ be the expectation with respect to ρ.

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 $\mathscr{L}^2_{\rho_{\mathscr{X}}}$, where

$$\|f\|_\rho = \|f\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} = \left\{ \int_X |f(x)|^2 d\rho_{\mathscr{X}} \right\}^{1/2}.$$

Let $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ be a *Mercer kernel*, i.e. a continuous[1] 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 inner product, defined as the 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 that

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

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

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

Define the linear map

$$L_K : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{H}_K$$

by the following integral transform

$$L_K(f)(t) := \int_X K(t, x) f(x) d\rho_X(x).$$

It is well-known that L_K is well-defined, and that composition with the inclusion $\mathscr{H}_K \hookrightarrow \mathscr{L}^2_{\rho_{\mathscr{X}}}$ yields a compact positive self-adjoint operator on $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ [e.g. [9], [16]]. The restriction $L_K|_{\mathscr{H}_K} : \mathscr{H}_K \to \mathscr{H}_K$ is the *covariance operator* of $\rho_{\mathscr{X}}$ in $\mathscr{H}_K$: by the reproducing property,

$$L_K|_{\mathscr{H}_K} = \mathbb{E}_{x \sim \rho_{\mathscr{X}}}[\langle \cdot, K_x \rangle K_x].$$

Abusing notation, we will denote the three operators by L_K in the sequel.

Note that, by Cauchy-Schwarz inequality, $\|L_K f\|_\infty \leq \kappa^2 \|f\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$, so that

$$\|L_K\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{L}^2_{\rho_{\mathscr{X}}}} \leq \kappa^2. \tag{4}$$

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 $(\mu_\alpha, \phi_\alpha)_{\alpha \in \mathbb{N}}$ in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$. Recall that (e.g. [9]) L_K is a trace-class operator as

$$\sum_{\alpha \in \mathbb{N}} \mu_\alpha = \int_{\mathscr{X}} K(x, x) d\rho_{\mathscr{X}}(x) \leq \kappa^2.$$

We assume in this paper that all eigenvalues μ_α are positive, which implies that $L_K : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{L}^2_{\rho_{\mathscr{X}}}$ is injective and $L_K^{1/2} : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{H}_K$ is an isometrical isomorphism of Hilbert spaces [9]. Hence the eigenfunctions $(\phi_\alpha)_{\alpha \in \mathbb{N}}$ are orthogonal both in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ and $\mathscr{H}_K$. We can define, for all $r > 0$,

$$\begin{aligned} L_K^r : \ \mathscr{L}^2_{\rho_{\mathscr{X}}} &\to \mathscr{L}^2_{\rho_{\mathscr{X}}} \\ \sum_{\alpha \in \mathbb{N}} a_\alpha \phi_\alpha &\mapsto \sum_{\alpha \in \mathbb{N}} a_\alpha \mu_\alpha^r \phi_\alpha; \end{aligned} \tag{5}$$

L_K^r can be regarded as a low-pass filter, and $\|L_K^r\| = \max_{\alpha \in \mathbb{N}} \mu_\alpha^r = \|L_K\|^r$.

For all $f \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ and $r > 0$, we write $L_K^{-r} f \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ when f lies in the image of the mapping $L_K^r : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{L}^2_{\rho_{\mathscr{X}}}$. Note that, if $r \geq 1/2$, then this implies $f \in \mathscr{H}_K$ because of the isometry $L_K^{1/2}$ between $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ and $\mathscr{H}_K$.

For all $\lambda > 0$ and $r \in \mathbb{R} \setminus \{0\}$, we can similarly define $(L_K + \lambda I)^r : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{L}^2_{\rho_{\mathscr{X}}}$, which is a bijection; indeed, $\mu_\alpha \to_{\alpha \to \infty} 0$ implies $\lambda + \mu_\alpha \in [\lambda, A]$ for some $A > 0$, hence $\sum_{\alpha \in \mathbb{N}} a_\alpha^2 < \infty \iff \sum_{\alpha \in \mathbb{N}} a_\alpha^2 (\lambda + \mu_\alpha)^{2r} < \infty$.

It can be shown [e.g. [9]] that for any $\lambda \in \mathbb{R}_+$, the solution of (2) is

$$f_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho \in \mathscr{H}_K. \tag{6}$$

In this paper, by $B_1, C_1, D_1, B_2, C_2, D_2, \ldots$, we denote various constants, which are defined "locally" in the sense that the same notations appeared in different sections has different meanings.

B. Stochastic Gradient Algorithms

Let $\mathcal{F} = (\mathcal{F}_t)_{t \in \mathbb{N}_0} \in \mathscr{X} \times \mathscr{Y}$ be the filtration $\mathcal{F}_t = \sigma\{(x_i, y_i) : 1 \leq i \leq t\}$. In the sequel denote by $\mathbb{E}_t = \mathbb{E}[\cdot | \mathcal{F}_t]$, the conditional expectation w.r.t. $\mathcal{F}_t$. Consider the following $\mathcal{F}_t$-adapted process $(f_t)_{t \in \mathbb{N}}$ taking values in $\mathscr{H}_K$,

$$f_t = f_{t-1} - \gamma_t [(f_{t-1}(x_t) - y_t) K_{x_t} + \lambda_t f_{t-1}], \tag{7}$$

for some fixed $f_0 \in \mathscr{H}_K$, e.g. $f_0 := 0$, where
(I) for each t, (x_t, y_t) is independent and identically distributed (i.i.d.) according to ρ;
(II) the *gain (step size) sequence* $(\gamma_t)_{t \in \mathbb{N}}$ and *regularization sequence* $(\lambda_t)_{t \in \mathbb{N}}$ are taking values in $\mathbb{R}_+ := (0, \infty)$, and converging to 0 as t goes to infinity.

Remark II.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 $O(t^2)$ 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 merely linear $O(t)$ at the requirement of large memory and fast memory access.

[1] In computer science literature, one often bears in mind some implicit feature map $\Phi : \mathscr{X} \to \mathscr{H}$ which takes an input vector x to a high (or infinite) dimensional feature vector, say an element of a Hilbert space $\mathscr{H}$, and then one considers explicitly the inner product $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$ as the kernel. In this construction, the continuity of K is equivalent to continuity of the feature map Φ.

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

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

is given by grad $V_z(f) = (f(x)-y)K_x + \lambda f$ [e.g. [27]], as a random variable depending on z. Since the expectation $\mathbb{E}[V_z(f)] = 2(\mathscr{E}(f) + \lambda\|f\|_K^2)$, algorithm (7) can thus be regarded as stochastic approximations of gradient descent method to solve (2), for each $\lambda = \lambda_t$.

C. Main Theorems

Theorem A provides sufficient conditions for the convergence of the *online learning* sequence $(f_t)_{t\in\mathbb{N}_0}$ in (7) to the regression function f_ρ. Theorem B and C explicit the corresponding convergence rates, respectively in $\mathscr{H}_K$ and $\mathscr{L}^2_{\rho_{\mathscr{X}}}$.

Theorem A (SufficientConditionsforConvergence): Assume $f_\rho \in \mathscr{H}_K$, and let (f_t) be defined by equation (7), with assumptions (I)-(II). Then

$$\limsup_{t\to\infty}\mathbb{E}[\|f_t - f_\rho\|_K^2] = 0,$$

if the following conditions are satisfied:

(A) $\displaystyle\sum_{t=0}^{\infty}\gamma_t\lambda_t = \infty.$

(B) $\displaystyle\limsup_{t\to\infty}\sum_{k=1}^{t}\gamma_k^2\prod_{i=k+1}^{t}(1-\gamma_i\lambda_i)^2 = 0,$

(C) $\displaystyle\limsup_{t\to\infty}\sum_{k=1}^{t}\|f_{\lambda_k} - f_{\lambda_{k-1}}\|_K\prod_{i=k+1}^{t}(1-\gamma_i\lambda_i) = 0.$

This theorem will be proved in Section III, as a consequence of Theorem III.5 in the setting of Hilbert spaces. Assumptions (B) and (C) can be replaced by the stronger (but less technical) assumptions (B') and (C') in Corollary III.7 that $\gamma_t/\lambda_t \to 0$ and $\|f_{\lambda_t} - f_{\lambda_{t-1}}\|_K/(\lambda_t\gamma_t) \to 0$.

Remark II.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 decrease fast enough for the algorithm to follow the path. The drifts depend on regularity of f_ρ, that the smoother f_ρ is, the faster drifts go down.

In the next two theorems (B) and (C) we choose the sequences $(\gamma_t)_{t\in\mathbb{N}}$ and $(\lambda_t)_{t\in\mathbb{N}}$ in order to optimize the rates of convergence in $\mathscr{H}_K$ and $\mathscr{L}^2_{\rho_{\mathscr{X}}}$. This optimization is twofold.

First, the study of convergence of approximations of ordinary differential equations generically yields a phase transition between a slower rate with "shadowing" of mean-field trajectories, and a faster one, normally distributed after renormalization. Even though the picture is more complicated in our case, in particular because the vector f_t is infinite-dimensional, this justifies here that we choose $\gamma_t\lambda_t$ reciprocally linear in t.

Second, optimization over (γ_t) at fixed $(\gamma_t\lambda_t)$ yields a bias-variance trade-off similar to the one observed in statistical "batch" learning, which relies on the regularity assumption on the regression function f_ρ.

More precisely, let us first recall the phase transition in classical finite-dimensional stochastic approximation, in the rate of convergence towards a stable equilibrium. Naturally, we study the projections of the algorithm on the base of eigenvectors of the linearization of the ordinary differential equation at the equilibrium. Let $(\eta_t)_{t\in\mathbb{N}}$ be one of these projections, and assume for instance that the corresponding eigenvalue is -1, so that the stochastic recursion is of the form

$$\eta_{t+1} = \eta_t + \gamma_t(-\eta_t + \epsilon_{t+1} + r_{t+1}),$$

where $\mathbb{E}_{t-1}[\epsilon_t] = 0$, (ϵ_t) is bounded, and (r_t) is small. For simplicity we will assume that $r_t = 0$ (which corresponds to the special case $\lambda_t = \lambda$ is a constant), but the heuristics holds on to the general case where r_t is less than quadratic in all coordinates. Let, for all $t\in\mathbb{N}$, $\beta_t := \prod_{k=1}^{t}(1-\gamma_k)$. Then it is easy to show by induction that

$$\eta_t = \beta_t\left[\eta_0 + \sum_{j=1}^{t}\frac{\gamma_j}{\beta_j}\epsilon_j\right].$$

Now suppose for instance that $\gamma_t \sim c/t$ $(c>0)$; then $\beta_n n^c \xrightarrow[n\to\infty]{} C > 0$. Depending on the choice of c, η_t exhibits the following phase transition at $c = 1/2$ in its asymptotic dynamics.

- If $c < 1/2$ then $\sum(\gamma_j/\beta_j)^2 < \infty$, therefore $\sum_{j=1}^{t}\gamma_j\epsilon_j/\beta_j$ converges a.s. by Doob's convergence theorem, which implies that $\eta_t t^c \xrightarrow[t\to\infty]{} C'$ (where C' is a positive random variable). In other words, (η_t) asymptotically "shadows" one particular solution of the ODE
$$\frac{dx}{dt} = -cx,$$
in the sense that the distance of (η_t) to that solution converges to 0 faster than (η_t) converges to 0. For a review on shadowing in stochastic approximation, see [4], Chapter 8, for instance.
- On the contrary, if $c > 1/2$, then $\sum(\gamma_j/\beta_j)^2 = \infty$, and by the martingale convergence theorem (see for instance [34]), assuming for instance $\mathbb{E}_{t-1}[\epsilon_t^2] = D^2 > 0$ constant, and $\eta_t\sqrt{t}$ converges towards a centered normally distributed random variable with variance $c^2D^2/(2c-1)$, and follows an associated Ornstein-Uhlenbeck process, see [11], Chapter 4, for instance.

Therefore it suffices to choose $c > 1/2$ to achieve fast convergence rates. In this paper we will set $c = 1$ and choose $\gamma_t\lambda_t \sim 1/t$ to meet the heuristics above.

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.

Let $t_0 > 0$ and, for all $t\in\mathbb{N}$,

$$\bar{t} := t + t_0,$$

where t_0 is large enough which won't affect the speed of convergence. We assume, in the statement of Theorems B and C, that, for all $t \in \mathbb{N}$,

$$\gamma_t = a\left(\frac{1}{\bar{t}}\right)^{\frac{2r}{2r+1}}, \ \lambda_t = \frac{1}{a}\left(\frac{1}{\bar{t}}\right)^{\frac{1}{2r+1}}.$$

Theorem B (Upper Bounds for $\mathscr{H}_K$-Convergence): Assume $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_\mathscr{X}}$ for some $r \in (1/2, 3/2]$, $a \geq 1$, and $t_0 \geq (a\kappa^2 + 1)^{(2r+1)/(2r)}$. Then, for all $t \in \mathbb{N}$, with probability at least $1-\delta$,

$$\|f_t - f_\rho\|_K \leq \frac{C_0}{\bar{t}} + \left(C_1 a^{1/2-r}\log\frac{2}{\delta} + C_2 a\right)\left(\frac{1}{\bar{t}}\right)^{\frac{2r-1}{4r+2}},$$

where

$$C_0 := 2t_0^{\frac{4r+3}{4r+2}} M_\rho, \ \ C_1 := \frac{20r-2}{(2r-1)(2r+3)}\|L_K^{-r} f_\rho\|_\rho,$$

$$C_2 := \frac{20(\kappa+1)^2 M_\rho}{\kappa}.$$

Its proof is given in Section V.

Remark II.3: Given $\delta > 0$, M_ρ and $\|L_K^{-r} f_\rho\|_\rho$, one can optimize a in order to minimize

$$h(a) := C_1 a^{1/2-r}\log\frac{2}{\delta} + C_2 a.$$

This yields the choice $a^* := [C_1(r-1/2)\log(2/\delta)/C_2]^{(r+1/2)^{-1}} \vee 1$, with

$$h(a^*) = (r+1/2)\left[C_1 C_2^{r-1/2}(r-1/2)\log\frac{2}{\delta}\right]^{(r+1/2)^{-1}}$$

when $a^* > 1$.

This asymptotic rate in $O(t^{-(2r-1)/(4r+2)})$ is the same as the best known rates in batch learning algorithms; see [Theorem 2, [29]].

Remark II.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 does not depend on r and is faster than the remaining terms. The second 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 rates up to different constants. The third term at a rate $O(t^{-(2r-1)/(4r+2)})$, reflects the error caused by random fluctuations by the i.i.d. sampling. Later as we will see, the second term is a bound on the bias and the third term is a bound on the variance.

Theorem C (UpperBoundsfor$\mathscr{L}^2_{\rho_\mathscr{X}}$-Convergence): Assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_\mathscr{X}}$ for some $r \in [1/2, 1]$. Assume $a \geq 4$, and $t_0 \geq (2 + 8\kappa^2 a)^{(2r+1)/(2r)}$.

Then, for all $t \in \mathbb{N}$, with probability at least $1-\delta$ ($\delta \in (0,1)$),

$$\|f_t - f_\rho\|_\rho \leq \frac{D_0}{\bar{t}} + \left(D_1 a^{-r} + \sqrt{a} D_2 \log\frac{2}{\delta}\right)\left(\frac{1}{\bar{t}}\right)^{\frac{r}{2r+1}}$$

$$+ \ldots + \left(a^{3/2} D_3 \sqrt{\log \bar{t}} + a^{5/2} D_4\right)(\log(2/\delta))^2\left(\frac{1}{\bar{t}}\right)^{\frac{4r-1}{4r+2}}.$$

where

$$D_0 := 2M_\rho t_0, \ \ D_1 := \frac{5r+1}{r(1+r)}\|L_K^{-r} f_\rho\|_\rho,$$

$D_2 := 10\kappa M_\rho$, $D_3 = 63\kappa^2 M_\rho$, and $D_4 := 50\kappa^2 M_\rho t_0^{1/2-\theta}$.

Its proof will be given in Section VI.

Remark II.5: When $r \in (1/2, 1]$, the first term of $O(1/t)$ and the third term of $O(t^{-2r-1/2/2r+1}\log^{1/2} t)$ both drop faster than the second term of $O(t^{-\frac{r}{2r+1}})$, whence they can be ignored asymptotically. The second term as the dominant one, roughly speaking has contributions from two parts: the one with constant D_1 comes from the bias, *i.e.* the approximation and the drift errors, while the other with constant D_2 comes from the variance, *i.e.* the sample error.

Remark II.6: A special case is $r = 1/2$, which is equivalent to say $f_\rho \in \mathscr{H}_K$. In this case $\gamma_t = \lambda_t = \bar{t}^{-1/2}$, whence it does not satisfy the Path Following Condition (B) in Theorem A (recall $\prod_{i=k+1}^{t}(1-\gamma_i\lambda_i) = (k+t_0)(t+t_0)^{-1}$). But Theorem C suggests a weaker notion that f_t follows the regularization path, *i.e.* $f_t \to f_\rho$ in $\mathscr{L}^2_{\rho_\mathscr{X}}$ rather than $\mathscr{H}_K$, which in fact converges at a rate of $O(t^{-1/4}\log^{1/2} t)$ uniformly for all $f_\rho \in \mathscr{H}_K$.

Remark II.7: Overall, the convergence in $\mathscr{L}^2_{\rho_\mathscr{X}}$ has rates $O(t^{-r/(2r+1)}\log^{1/2} t \cdot \log^2 1/\delta)$, a logarithmic polynomial on δ, whence the Borel-Cantelli Lemma implies almost sure convergence $\|f_t - f_\rho\|_{\mathscr{L}^2_{\rho_\mathscr{X}}} \overset{as}{\to} 0$. Note that no almost-sure tight convergence bounds were obtained so far in online learning in RKHS [27], [35], [37].

In [37], some tight convergence rates are presented for a weaker convergence in mean square distance $\mathbb{E}\|f_t - f_\rho\|_\rho^2 \to 0$. In particular when $\lambda_t = 0$, mean square distance convergence rates are studied under two choices of step sizes, time varying γ_t and constant $\gamma_t = \gamma(T)$, depending on total sample size T as in batch-learning. For chosen time varying step sizes γ_t and $\lambda_t = 0$, convergence rates

$$\mathbb{E}[\|f_t - f_\rho\|_\rho^2] \leq O(t^{-2r/(2r+1)}\log t) \tag{8}$$

are established for $r \in (0, 1/2]$, which differs from the complexity class in this paper. Again for (chosen) varying γ_t but constant $\lambda_t = \lambda(T) > 0$, rates (8) at time $t = T$ hold for $r \in (0, 1]$. Finally for (chosen) constant step size $\gamma_t = \gamma(T)$ and $\lambda_t = 0$, the rates (8) are established at time $t = T$ for all $r > 0$. Note that those constant choices of $\lambda(T)$ and $\gamma(T)$ imply that the algorithms are not truly online, as they need to know *a priori* the total sample size. In summary it is therefore an open problem whether the same type of almost-sure convergence as above can be established for the whole regularity range $r > 0$.

Remark II.8: To see the asymptotic optimality, consider the generalization error $\mathscr{E}(f) - \mathscr{E}(f_\rho) = \|f - f_\rho\|_\rho^2$ [see [9]]. Since the rate $O(t^{-r/(2r+1)})$ dominates when $r > 1/2$, then under the same condition of Theorem C, there holds with probability at least $1-\delta$ ($\delta \in (0,1)$), for all $t \in \mathbb{N}$,

$$\mathscr{E}(f_t) - \mathscr{E}(f_\rho) \leq O(t^{-2r/(2r+1)}).$$

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 [7]. 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 $|y| \leq M_\rho$; (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 [7],

$$\liminf_{t\to\infty} \inf_{(z_i)_1^t \mapsto f_t} \sup_{\rho\in\mathscr{P}(b,r)} \mathbf{Prob}\Big\{(z_i)_1^t \in \mathscr{Z}^t : \cdots$$

$$\mathscr{E}(f_t) - \mathscr{E}(f_\rho) > Ct^{-\frac{2rb}{2rb+1}}\Big\} = 1$$

for some constant $C > 0$ independent on t, where the infimum in the middle is taken over all algorithms as a map $\mathscr{Z}^t \ni (z_i)_1^t \mapsto f_t \in \mathscr{H}_K$.

Note that in the minimax lower rate, the probability measure may change for different data size t, which violates the fundamental identical distribution assumption in learning. Therefore [15] suggests a kind of individual lower rates for learning problems. The following individual lower rate was obtained as Theorem 3 in [7]: for every $B > b$,

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

where the infimum is taken over arbitrary sequences of functions $f_t : \mathscr{Z}^t \to \mathscr{H}_K$. It can be seen that the key difference in the individual lower rate, lies in that by putting $\limsup_{t\to\infty}$ before $\sup_{\rho\in\mathscr{P}(b,r)}$, the probability measure ρ is applied to all sufficiently large t.

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 [9], 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. Therefore, the upper bound by Theorem C reaches both the minimax and the individual lower rates.

Remark II.9: The condition $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ for $r \geq 1/2$ implies that $f_\rho \in \mathscr{H}_K$ with some additional regularity, since $L_K^{1/2} : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{H}_K$ is a Hilbertian isometry. Assuming that $f_\rho = \sum_{\alpha\in\mathbb{N}} a_\alpha \phi_\alpha$, where $(\mu_\alpha, \phi_\alpha)$ is an orthonormal eigensystem of L_K, this regularity condition $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ translates into

$$\sum_{\alpha\in\mathbb{N}} \frac{a_\alpha^2}{\mu_\alpha^{2r}} < \infty$$

which requires that $a_\alpha \to 0$ rapidly enough, in particular faster than μ_α^r converges to 0. Hence the larger r is, the more regularity f_ρ has.

For example, let $X = S^d$ be the d-sphere and let ρ_X be the uniform measure on S^d. Then, following [33], one can take the Sobolev space $W_d(S^d)$ as a RKHS $\mathscr{H}_K$, such that the associated integral L_K has eigenvalues $\lambda_\alpha \sim \alpha^{-1}$ ($\alpha \in \mathbb{N}$). Then with $r = s/d$, $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ implies that $f_\rho \in W_s(S^d)$. Theorem B and C gives upper $W_d(S^d)$- and $\mathscr{L}^2_{\rho_{\mathscr{X}}}$-convergence rates at $O(t^{-\frac{2s-d}{4s+2d}})$ and $O(t^{-\frac{s}{2s+d}})$, respectively for $s \leq d$.

III. Sequential Stochastic Approximations of Regularization Paths in Hilbert Spaces

In this section, we study some stochastic approximation sequences in the more general setting of general Hilbert spaces.

Let $\mathscr{W}$ be a Hilbert space with inner product $\langle,\rangle$ and associated norm $\|u\| := \sqrt{\langle u, u\rangle}$, and let $SL(\mathscr{W})$ be the vector space of self-adjoint bounded linear operators on $\mathscr{W}$, endowed with the canonical norm

$$\|A\| := \sup_{\|x\|\leq 1} \|Ax\|.$$

Let $\mathscr{X}$ and $\mathscr{Y}$ be two topological spaces (on which we make no other assumption), let $\mathscr{Z} := \mathscr{X} \times \mathscr{Y}$ and let ρ be a probability measure on the Borel σ-algebra of $\mathscr{Z}$. Let $A : \mathscr{Z} \to SL(\mathscr{W})$ and $b : \mathscr{Z} \to \mathscr{W}$ be random variables on the sample space $\mathscr{Z}$ taking values respectively in $SL(\mathscr{W})$ and $\mathscr{W}$, and let

$$\bar{A} := \mathbb{E}[A], \ \bar{b} := \mathbb{E}[b]$$

be their expectations on $(\mathscr{Z}, \rho)$.

Now assume that $\bar{A}$ is a strictly positive operator with an *unbounded* inverse. Knowing A and b, but not ρ (and subsequently not $\bar{A}$ and $\bar{b}$), and assuming $\bar{b} \in \bar{A}(\mathscr{W})$, the aim is to devise a stochastic algorithm approximating the solution $\bar{w}$ of the following linear equation

$$\bar{A}w = \bar{b}, \tag{9}$$

using as data an i.i.d sequence $(z_t)_{t\in\mathbb{N}}$ in $\mathscr{Z}$ with probability law ρ. As in the standard setting of Robbins-Monro (see [17], [24]), it is natural to consider a stochastic gradient descent algorithm.

More precisely, the search for the solution $\bar{w}$ of (9) is equivalent to the minimization of the quadratic potential map $\hat{V} : \mathscr{W} \to \mathbb{R}$

$$\hat{V}(w) := \frac{1}{2}\langle \bar{A}(w - \bar{w}), w - \bar{w}\rangle,$$

whose gradient grad $\hat{V} : \mathscr{W} \to \mathscr{W}$ is given by

$$\text{grad } \hat{V}(w) = \bar{A}w - \bar{b} = \mathbb{E}[Aw - b].$$

In the context of online learning presented in the first two sections, $\mathscr{W} := \mathscr{H}_K$, $A((x, y))(f) := f(x)K_x$, $b((x, y)) := yK_x$ (see Section III-C), so that $\bar{A} = L_K$, $\bar{b} = L_K f_\rho$ and $\bar{w} = f_\rho$, and $\hat{V}(w) = \|f - f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} = \mathscr{E}(f) - \mathscr{E}(f_\rho)$ is the generalization error.

A natural Robbins-Monro gradient descent algorithm would be

$$w_t = w_{t-1} - \gamma_t(A(z_t)w_{t-1} - b(z_t)), \tag{10}$$

since $\mathbb{E}_{z_t\sim\rho}[A(z_t)w_{t-1} - b(z_t)] = \bar{A}w_{t-1} - \bar{b}$.

However, the sample complexity analysis on Hilbert spaces, in order to estimate the sample size sufficient to approximate the minimizer with high probability, requires boundedness of $\bar{A}^{-1}$ (see for instance [27]).

To solve this ill-posed problem with unbounded $\bar{A}^{-1}$, one may construct sequences of random variables $(A_t)_{t\in\mathbb{N}}$ and $(b_t)_{t\in\mathbb{N}}$ on the sample space $\mathscr{Z}$ taking values respectively in $SL(\mathscr{W})$ and $\mathscr{W}$, with the assumption that, if

$$\bar{A}_t := \mathbb{E}[A_t], \quad \bar{b}_t := \mathbb{E}[b_t]$$

are their expectations on $(\mathscr{Z}, \rho)$, $\bar{A}_t$ has bounded inverse and $\bar{A}_t \to \bar{A}$, $\bar{b}_t \to \bar{b}$. Then the aim is to find assumptions ensuring that the stochastic approximation sequence $(w_t)_{t\in\mathbb{N}}$ iteratively defined by $w_0 := W_0$ deterministic, and

$$w_t = w_{t-1} - \gamma_t(A_t(z_t)w_{t-1} - b_t(z_t)), \tag{11}$$

where $(\gamma_t)_{t\in\mathbb{N}}$ is a real positive sequence, converges to the solution $\bar{w}$ of (9) as t goes to infinity. We note that such a

This question can be divided into two subquestions: first the *deterministic* convergence of

$$\bar{w}_t := \bar{A}_t^{-1}\bar{b}_t. \tag{12}$$

to $\bar{w}$, the path $t \mapsto \bar{w}_t$ being then called a *regularization path* of the solution of equation (9), and second the *probabilistic* convergence of the quantity

$$r_t := w_t - \bar{w}_t, \tag{13}$$

which we call the *remainder* (note that $\bar{w}_t = \bar{A}_t^{-1}\bar{b}_t \neq \mathbb{E}[w_t]$ in general). In the online learning case (see Section III-C), we choose $A_t := A + \lambda_t I$, $(\lambda_t)_{t\in\mathbb{N}}$ positive sequence, $b_t := b$, so that $\bar{w}_t = f_{\lambda_t} \to f_\rho$ in $\mathscr{H}_K$.

We provide in Section III-A two structural decompositions of r_t, respectively a reversed martingale and a martingale one. Both expand r_t into three parts: one depending on the initial value of $r_{.}$ called the *initial error*, one depending on the *drift*

$$\Delta_j := \bar{w}_j - \bar{w}_{j-1} \tag{14}$$

along the regularization path $(\bar{w}_t)$ called the *drift error*, and finally one random variable of zero mean called the *sample error*, respectively written as a reversed martingale and as a martingale at time t.

The reversed martingale decomposition will, on one hand, enable us to prove Theorem III.5 below, whose corollary is Theorem A in the context of online learning, and which provides sufficient assumptions on the asymptotic behaviour of the norms of A_t, A_t^{-1}, $\bar{A}_t^{-1}$ and b_t for the convergence of the variance of the remainder r_t. On the other hand, this reversed decomposition will yield Theorem B giving upper bounds on $f_t - f_\rho$ in $\mathscr{H}_K$ with high probability, proved in Section V.

The martingale decomposition will imply Theorem C giving upper bounds of $f_t - f_\rho$ in $\mathscr{L}^2_{\rho_\mathscr{X}}$ with high probability, proved in Section VI.

A. Two Structural Decomposition Theorems

For all $j,\, t \in \mathbb{N}$, let Π_j^t be the *random* operator on $\mathscr{W}$, on the sample space $\mathscr{Z}^{\mathbb{N}}$, defined by

$$\Pi_j^t((z_i)_{i\in\mathbb{N}}) = \begin{cases} \displaystyle\prod_{i=j}^{t} (I - \gamma_i A_i(z_i)) & \text{if} j \le t; \\ I & \text{otherwise.} \end{cases}$$

By a slight abuse of notation, we let $A_t := A_t(z_t)$ and $b_t := b_t(z_t)$ in the sequel, when there is no ambiguity.

Theorem III.1 (Reversed Martingale Decomposition): For all $s,\, t \in \mathbb{N}$, $t \ge s$,

$$r_t = \Pi_{s+1}^t r_s - \sum_{j=s+1}^{t} \gamma_j \Pi_{j+1}^t (A_j\bar{w}_j - b_j) - \sum_{j=s+1}^{t} \Pi_j^t \Delta_j \tag{15}$$

Remark III.2: Note that Π_{j+1}^t is an operator whose randomness only depends on $z_{j+1}, \dots z_t$, whereas the randomness in $A_j\bar{w}_j - b_j$, with zero mean, only depends on z_j. By independence of z_t, $t \in \mathbb{N}$, the conditional expectation $\mathbb{E}[\gamma_j \Pi_{j+1}^t(A_j\bar{w}_j - b_j)|z_{j+1}, \dots, z_t]$ is 0, whence for each t, $\gamma_j \Pi_{j+1}^t(A_j\bar{w}_j - b_j)$ is a *reversed martingale difference sequence* whose sum is a *reversed martingale sequence* with zero mean. For more background on reversed martingales, see for example [21].

Proof of Theorem III.1: By definition,

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

which implies

$$r_t = (I - \gamma_t A_t) r_{t-1} - \gamma_t(A_t\bar{w}_t - b_t) - (I - \gamma_t A_t)\Delta_t. \tag{16}$$

The result follows by induction on $t \in \mathbb{N}$, $t \ge s$. ■

For all $j,\, t \in \mathbb{N}$, let

$$\chi_t = (\bar{A}_t - A_t)w_{t-1} + (b_t - \bar{b}_t),$$

and let $\bar{\Pi}_j^t$ be the *deterministic* operator on $\mathscr{W}$ defined by

$$\bar{\Pi}_j^t = \begin{cases} \displaystyle\prod_{i=j}^{t} \big(I - \gamma_i \bar{A}_i\big) & \text{if } j \le t; \\ I, & \text{otherwise.} \end{cases}$$

Theorem III.3 (Martingale Decomposition): For all $s,\, t \in \mathbb{N}$, $t \ge s$,

$$r_t = \bar{\Pi}_{s+1}^t r_s + \sum_{j=s+1}^{t} \gamma_j \bar{\Pi}_{j+1}^t \chi_j - \sum_{j=s+1}^{t} \bar{\Pi}_j^t \Delta_j \tag{17}$$

Remark III.4: The martingale decomposition was proposed in [36]. Contrary to the reversed martingale decomposition, only the *sample error* is random here, the operator $\bar{\Pi}_{j+1}^t$ being deterministic. The process $(\gamma_j \bar{\Pi}_{j+1}^t \chi_j)_{j\in\mathbb{N}}$ is a *martingale difference sequence* since, for all $j \in \mathbb{N}$ and $t \ge j$, $\mathbb{E}_{j-1}[\gamma_j \bar{\Pi}_{j+1}^t \chi_j] = 0$. Note that the martingale property continues to hold for dependent sampling $z_t(z_1, \dots, z_{t-1})$, as long as $\mathbb{E}_{t-1}[A_t(z_t)] = \bar{A}_t$ and $\mathbb{E}_{t-1}[b_t(z_t)] = \bar{b}_t$. As a consequence, the same almost-sure convergence result can be proved in this case, using the decomposition of the Markov sampling process in [30], which has exponentially decreasing drift errors.

The non-randomness of the operator $\bar{\Pi}_j^t$ will play a key role in the proof of Theorem C in the online learning context, since it will enable us to make explicit calculations involving the spectral decomposition of $L_K : \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{L}^2_{\rho_\mathscr{X}}$ (recall that $\bar{A}_i = L_K + \lambda_i$ then). However, the fact that χ_t, contrary to $A_t\bar{w}_t - b_t$ in the reversed expansion, does not depend only on z_t but rather on the whole past $(z_i)_{0\le i\le t}$, makes it necessary to obtain a preliminary upper bound of χ_t in Appendix C, which explains the factor $(\log 2/\delta)^2$ in Theorem C, rather than $\log 2/\delta$ in Theorem B.

Proof of Theorem III.3: By definition,

$$\begin{aligned} r_t &= w_t - \bar{w}_t \\ &= w_{t-1} - \bar{w}_t - \gamma_t(A_t w_{t-1} - b_t) \\ &= (I - \gamma_t \bar{A}_t)(w_{t-1} - \bar{w}_t) + \gamma_t \chi_t, \text{ using } \bar{b}_t = \bar{A}_t \bar{w}_t \\ &= (I - \gamma_t \bar{A}_t) r_{t-1} + \gamma_t \chi_t - (I - \gamma_t \bar{A}_t)[\bar{w}_t - \bar{w}_{t-1}]. \end{aligned}$$

The result follows by induction on $t \in \mathbb{N}$, $t \geq s$. ■

B. Sufficient Conditions for the Convergence of the Remainder

The following Theorem III.3, which implies Theorem A in the context of online learning (see Section III-C), states the convergence of $\|r_t\|^2 = \|w_t - \bar{w}_t\|^2$ to zero in expectation, under some assumptions on the asymptotic behaviour of the gain sequence γ_t and of the norms of b_t and operators A_t, A_t^{-1} and $\bar{A}_t^{-1}$.

The corresponding *Generalized Finiteness Condition* on the asymptotic behaviour of A_t and b_t is a generalization of the *Finiteness Condition* in [27].

Generalized Finiteness Condition. Let $(\underline{\alpha}_t)_{t\in\mathbb{N}}$ and $(\overline{\alpha}_t)_{t\in\mathbb{N}}$ be deterministic positive sequences. For all $t \in \mathbb{N}$, assume that almost surely, A_t is positive, and the operators A_t, $\bar{A}_t$ and $\bar{A}$ are invertible (although $\bar{A}$ has an *unbounded* inverse), and that

$$\|A_t\| \leq \overline{\alpha}_t, \quad \|A_t^{-1}\| \leq \underline{\alpha}_t^{-1}.$$

Theorem III.5: Consider the stochastic approximation sequence $(w_t)_{t\in\mathbb{N}_0}$ and remainder $(r_t)_{t\in\mathbb{N}_0}$ defined in (11)-(13).

Suppose that the Generalized Finiteness Condition holds, and that the variance $\mathbb{E}\|A_t \bar{w}_t - b_t\|^2$ is uniformly bounded in $t \in \mathbb{N}$. Then

$$\mathbb{E}\|r_t\|^2 \to 0,$$

if the following assumptions hold:

(A) $\gamma_t \to 0$ and $\sum_t \gamma_t \underline{\alpha}_t = \infty$,

(B) $\limsup_{t\to\infty} \sum_{k=1}^{t} \gamma_k^2 \prod_{i=k+1}^{t} (1 - \gamma_i \underline{\alpha}_i)^2 = 0$,

(C) $\limsup_{t\to\infty} \sum_{k=1}^{t} \|\Delta_k\| \prod_{i=k+1}^{t} (1 - \gamma_i \underline{\alpha}_i) = 0$.

The following Lemma III.6 enables us to provide simple sufficient conditions for (B) and (C) in Corollary III.7.

Lemma III.6: Let $(a_t)_{t\in\mathbb{N}}$ and $(b_t)_{t\in\mathbb{N}}$ be two real positive sequences converging to 0 when t goes to infinity. Then

$$\limsup_{t\to\infty} a_t/b_t = 0 \textit{ and } \sum_{t\in\mathbb{N}} b_t = \infty$$

$$\Longrightarrow \limsup_{t\to\infty} \sum_{k=1}^{t} a_k \prod_{i=k+1}^{t} (1 - b_i) = 0.$$

Corollary III.7: In the statement of Theorem III.5, assumptions (B) and (C) may respectively be replaced by

(B') $\limsup_{t\to\infty} \dfrac{\gamma_t}{\underline{\alpha}_t} = 0$,

(C') $\limsup_{t\to\infty} \dfrac{\|\Delta_t\|}{\underline{\alpha}_t \gamma_t} = 0$.

Theorem III.5 and Lemma III.6 are proved in Appendix C, and imply Corollary III.7: Lemma III.6 with $a_t := \gamma_t^2$ (resp. $a_t := \|\Delta_t\|$) and $b_t := \underline{\alpha}_t \gamma_t$ shows that (B') (resp. (C')) implies (B) (resp. (C)).

The proof of Theorem III.5 makes use of the following preliminary Lemma III.8 (shown in Appendix C), which implies some upper bounds of the norms of operators Π_j^t, $t \geq j$, also used in Sections VI and V.

Lemma III.8: Let $j_0 \in \mathbb{N}$, and let $(\gamma_t)_{t\in\mathbb{N}}$, $(\underline{\alpha}_t)_{t\in\mathbb{N}}$ and $(\overline{\alpha}_t)_{t\in\mathbb{N}}$be real positive sequences, and let $(A_t)_{t\in\mathbb{N}}$ be a sequence of positive compact self-adjoint operators on the Hilbert space $\mathscr{W}$. Assume that, for all $t \geq j_0$, $\|A_t\| \leq \overline{\alpha}_t$, $\|A_t^{-1}\| \leq \underline{\alpha}_t^{-1}$ and $\gamma_t \overline{\alpha}_t \leq 1$.

Then, for all $t \geq j_0$ and $j_0 \leq j \leq t$,

(A) $\|I - \gamma_t A_t\| \leq 1 - \gamma_t \underline{\alpha}_t$;

(B) $\|\Pi_j^t\| \leq \prod_{i=j}^{t} (1 - \gamma_j \underline{\alpha}_j)$.

In particular, if the two sequences $(\gamma_t)_{t\in\mathbb{N}}$ and $(\underline{\alpha}_t)_{t\in\mathbb{N}}$ are such that, for all $t \geq j_0$, $\gamma_t \underline{\alpha}_t := c\bar{t}^{-1}$ (recall $\bar{t} = t + t_0$) for some c, $t_0 > 0$, then (B) yields

$$\|\Pi_j^t\| \leq \left(\frac{j + t_0}{\bar{t} + 1}\right)^c.$$

C. Application to Online Learning and Proof of Theorem A

The online learning sequence $(f_t)_{t\in\mathbb{N}_0}$ defined in (7), with assumptions (I)-(II), can be interpreted as a sequential stochastic approximation algorithm $(w_t)_{t\in\mathbb{N}_0}$ in (10), taking values in the Hilbert space $\mathscr{W} := \mathscr{H}_K$: letting $z_t = (x_t, y_t)$,

$$\begin{aligned} A((x, y)) &: = \langle ., K_x \rangle_K K_x, b((x, y)) := y K_x \\ A_t &: = A(z_t) + \lambda_t I, b_t := b, \end{aligned}$$

so that

$$\begin{aligned} \bar{A} &= L_K, \quad \bar{b} = L_K f_\rho, \quad \bar{w} = f_\rho, \\ \bar{A}_t &= L_K + \lambda_t I, \quad \bar{w}_t = f_{\lambda_t}, \\ r_t &= f_t - f_{\lambda_t}, \quad \Delta_t = f_{\lambda_t} - f_{\lambda_{t-1}}. \end{aligned}$$

Let us emphasize that the operator A is only defined from $\mathscr{H}_K$ to $\mathscr{H}_K$ here (we would not be able to define $f(x)$ for $f \in \mathscr{L}^2_{\rho_\mathscr{X}}$). The properties mentioned below will only hold on $\mathscr{H}_K$ in general, and in particular the norms of operators $\|.\|$ are assumed to be $\|.\|_{\mathscr{H}_K \to \mathscr{H}_K}$, although operators defined on $\mathscr{L}^2_{\rho_\mathscr{X}}$ and commuting with $L_K^{1/2}$ (which is an isometry between $\mathscr{L}^2_{\rho_\mathscr{X}}$ and H_K) have the same norm in either spaces.

Note that $A(z)$ is positive for all $z = (x, y) \in \mathscr{Z}$ (which implies $\bar{A} = L_K$ positive as well), since

$$\langle A((x, y))(f), f \rangle = \langle f(x) K_x, f \rangle = f(x)^2 \geq 0$$

for all $f \in \mathscr{H}_K$.

Also, for all $f \in \mathscr{H}_K$, $\|Af\| = |\langle K_x, f \rangle| \|K_x\| \leq \|K_x\|^2 \|f\|$, so that

$$\|A\| \leq \kappa^2, \quad \|\bar{A}\| \leq \mathbb{E}(\|A\|) \leq \kappa^2.$$

Hence

$$\|A_t\| \leq \overline{\alpha}_t := \lambda_t + \kappa^2, \quad \|A_t^{-1}\|^{-1} \geq \underline{\alpha}_t := \lambda_t. \tag{18}$$

With these definitions, we are now ready to prove Theorem A.

Proof of Theorem A: Under the assumptions of Theorem A, the *Generalized Finiteness Condition* of Section III-B is satisfied. Now $f_\rho \in \mathscr{H}_K$ implies $\|f_\lambda - f_\rho\|_K \to 0$ when $\lambda \to 0$, using for example Theorem IV.1 (C) with $r = 1/2$. Therefore the conclusion follows from the convergence of $\mathbb{E}[\|\omega_t - \bar{\omega}_t\|^2] = \mathbb{E}[\|f_t - f_{\lambda_t}\|^2]$ to 0 in Theorem III.5, while the condition of uniform boundedness of $\mathbb{E}\|A_t\bar{\omega}_t - b_t\|^2$ is shown in Lemma V.5 (B). ■

For convenience, we will use, in Sections V, VI and in the Appendix, the notation

$$L_t := A(z_t) = \langle ., K_{x_t}\rangle_K K_{x_t} \tag{19}$$

We will assume that

$$\gamma_t = \frac{a}{\bar{t}^{\theta}}, \qquad \lambda_t = \frac{b}{\bar{t}^{1-\theta}}, \quad \text{for some } \theta \in [0,1],\ t_0 > 0, \tag{20}$$

and then study the $\mathscr{H}_K$ or $\mathscr{L}^2_{\rho_{\mathscr{X}}}$- norm of the error $f_t - f_\rho$, based using a reverse martingale (resp. martingale) decomposition in Section V (resp. Section VI). We will then optimize the upper bounds in θ, a and b by using some prior information on the regularity of f_ρ.

Finally observe that Lemma III.8 implies, using (18), that for all $j, t \in \mathbb{N}$, $t \geq j$,

$$\|I - \gamma_t A_t\| \leq 1 - \gamma_t\lambda_t = 1 - \frac{ab}{\bar{t}}, \qquad \|\Pi_j^t\| \leq \left(\frac{j + t_0}{\bar{t} + 1}\right)^{ab} \tag{21}$$

if $t_0^\theta \geq a(\kappa^2 + b)$ (and, therefore, $\gamma_t\overline{\alpha}_t = \gamma_t\lambda_t + \gamma_t\kappa^2 \leq abt_0^{-1} + a\kappa^2 t_0^{-\theta} \leq 1$).

Similarly, for all $j, t \in \mathbb{N}$, $t \geq j$,

$$\|I - \gamma_t \bar{A}_t\| \leq 1 - \frac{ab}{\bar{t}}, \qquad \|\bar{\Pi}_j^t\| \leq \left(\frac{j + t_0}{\bar{t} + 1}\right)^{ab} \tag{22}$$

if $t_0^\theta \geq a(\kappa^2 + b)$; the norm in (22) can be $\|.\|_{\mathscr{H}_K \to \mathscr{H}_K}$ as well as $\|.\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{L}^2_{\rho_{\mathscr{X}}}}$.

IV. Estimates of Drift on the Regularization Path

This section is devoted to estimates on the drift $\|f_\lambda - f_\mu\|$ $(\lambda, \mu \geq 0)$, along the regularization path $\lambda \to f_\lambda$, in $\mathscr{H}_K$-norm or $\mathscr{L}^2_{\rho_{\mathscr{X}}}$-norm, assuming that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ for some $r > 0$. These estimates enable us to upper bound on the one hand the approximation error $\|f_\lambda - f_\rho\|$ (when specialized to $\mu = 0$), and on the other hand the *drift error* in the martingale and reversed martingale decompositions.

Note that the estimate $\|f_\lambda - f_\mu\|_K = O(|\lambda - \mu|)$ in the case $r = 1$ is not improved by increasing r. This is related to a phenomenon usually refered to as the *saturation* problem in regularizations [13].

Theorem IV.1: Let $\lambda > \mu \geq 0$. Assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ for some $r \geq -1$.

(A) If $r \in [-1, 1] \setminus \{0\}$, then

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

(B) If $r \geq 1$, then for any $1 \leq s \leq r$,

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

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

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

If $r = 1/2$, then $\|f_\lambda - f_\rho\|_K \to 0$ as $\lambda \to 0$.

(D) If $r \in [-1/2, 3/2] \setminus \{1/2\}$, 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 $3/2 \leq s \leq r$,

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

Proof: Fix $\lambda > \mu$, assume $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ for some $r \in [-1, 1]$ and let $\|.\| := \|.\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{L}^2_{\rho_{\mathscr{X}}}}$. We first prove that, for all $u \geq -1$, if we let

$$J_{u,\lambda,\mu} := (\mu - \lambda)(L_K + \lambda I)^{-1}(L_K + \mu I)^{-1} L_K^{1+u}$$

then, for all $t \in [-1, 1] \setminus \{0\}$, $u \geq t$,

$$\|J_{u,\lambda,\mu}\| \leq \kappa^{2(u-t)}|\lambda^t - \mu^t|/|t|. \tag{23}$$

This will be useful, since

$$\begin{aligned} f_\lambda - f_\mu &= (\mu - \lambda)(L_K + \lambda I)^{-1}(L_K + \mu I)^{-1} L_K f_\rho \\ &= J_{r,\lambda,\mu} L_K^{-r} f_\rho, \end{aligned} \tag{24}$$

using that

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

Let us prove (23): using $\|L_K^{u-t}\| = \|L_K\|^{u-t} \leq \kappa^{2(u-t)}$ by (4), and $\max(t, 0) + \min(0, t) = t$,

$$\begin{aligned} \|J_{u,\lambda,\mu}\| &\leq |\lambda - \mu| \|(L_K + \lambda I)^{\max(t,0)-1} \dots \\ &\qquad \cdot (L_K + \mu I)^{\min(t,0)} L_K^{-(t+1)} L_K^{1+u}\| \\ &\leq |\lambda - \mu|\lambda^{-1}\lambda^{\max(t,0)}\mu^{\min(t,0)}\|L_K^{u-t}\| \\ &\leq \kappa^{2(u-t)}|\lambda - \mu|\lambda^{-1}\max(\lambda^t, \mu^t) \\ &= \kappa^{2(u-t)}\Lambda(\mu)|\lambda^t - \mu^t|, \end{aligned}$$

where

$$\Lambda(\mu) := \begin{cases} \frac{1 - \mu/\lambda}{1-(\mu/\lambda)^t} & \text{if } t > 0 \\ \frac{1 - \mu/\lambda}{1-(\lambda/\mu)^t} & \text{if } t < 0 \end{cases}$$

Now

$$\Lambda(\mu) \leq \frac{1}{|t|}.$$

Indeed, if $t > 0$, then this is a consequence of $x \leq (1 - (1 - x)^t)/t$ applied to $x := 1 - \mu/\lambda$, using that $x \mapsto (1 - (1 - x)^t)/t$ (defined on $(-\infty, 1]$) is convex and thus remains above the tangent line at 0. Similarly, we use $x \leq (1 - (1 - x)^{-t})/(-t)$ if $t < 0$.

Now (23)-(24) implies (A) with $u := r$ and $t := r$, and (B) with $u := s$ and $t := 1$, since $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ implies $L_K^{-s} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ for any $s \leq r$. Similarly, (D) (resp. (E)) follows from

$$L_K^{-1/2}(f_\lambda - f_\mu) = J_{r-1/2,\lambda,\mu} L_K^{-r} f_\rho,$$

and (23) applied to $u := r - 1/2$ and $t := u$ (resp. $u := s - 1/2$ and $t := 1$).

Let us now prove (C): if $r \ge 1/2$, then $f_\rho \in \mathscr{H}_K$, and the first part of equality (24) implies

$$\begin{aligned}&\|f_\lambda - f_\mu\|_K\\ &\le |\mu-\lambda|\|(L_K+\lambda I)^{-1}\|\|(L_K+\mu I)^{-1}L_K\|\|L_K^{-1/2}f_\rho\|_\rho\\ &\le \frac{|\mu-\lambda|}{\lambda}\|f_\rho\|_K.\end{aligned}$$

It remains to show the second part of (C): let us use again the notation $(\mu_\alpha, \phi_\alpha)_{\alpha\in\mathbb{N}}$ for the orthonormal eigensystem in $\mathscr{L}^2_{\rho_\mathscr{X}}$. Now, $f_\lambda - f_\rho = -\lambda(L_K+\lambda I)^{-1}f_\rho$, so that

$$\begin{aligned}\|f_\lambda - f_\rho\|_K^2 &= \sum_{\alpha\in\mathbb{N}} \frac{a_\alpha^2}{\mu_\alpha}\left(\frac{\lambda}{\mu_\alpha+\lambda}\right)^2\\ &\le \sum_{\alpha\ge N_0}\frac{a_\alpha^2}{\mu_\alpha} + \lambda^2\sum_{\alpha<N_0}\frac{a_\alpha^2}{\mu_\alpha^3};\end{aligned} \qquad (25)$$

we choose N_0 so that for any given $\epsilon > 0$, $\sum_{\alpha\ge N_0} a_\alpha^2/\mu_\alpha \le \epsilon/2$, using $\|f_\rho\|_K^2 = \sum_{\alpha\in\mathbb{N}} a_\alpha^2/\mu_\alpha < \infty$. For such fixed N_0, fix λ_ϵ such that $\lambda_\epsilon^2\sum_{\alpha<N_0}\frac{a_\alpha^2}{\mu_\alpha^3} < \epsilon/2$. Then, for any $\lambda \le \lambda_\epsilon$,

$$\|f_\lambda - f_\rho\|_K^2 < \epsilon$$

which establishes the convergence. Note that this convergence can also be derived, using an RKHS density argument, see [38]. ■

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

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, and assume that the sequences $(\gamma_t)_{t\in\mathbb{N}}$ and $(\lambda_t)_{t\in\mathbb{N}}$ are chosen in (20).

Our goal is to provide a probabilistic upper bound for

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

in order to prove Theorem B. We start with the triangle inequality

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

and apply the reversed martingale decomposition of $(f_t)_{t\in\mathbb{N}}$ developed in Section III, Theorem III.1:

$$r_t = \Pi_1^t r_0 - \sum_{j=1}^t \gamma_j \Pi_{j+1}^t (A_j\bar\omega_j - b_j) - \sum_{j=1}^t \Pi_j^t \Delta_j. \qquad (26)$$

We make use of the corresponding notation of Section III, in particular Section III-C, so that

$$A_j\bar\omega_j - b_j = (L_t + \lambda_t I)f_{\lambda_t} - y_t K_{x_t},$$

and

$$\Pi_j^t(x_j,\ldots,x_t) = \begin{cases}\displaystyle\prod_{i=j}^t (I - \gamma_i(L_i+\lambda_i I)) & \text{if } j \le t;\\ I & \text{otherwise.}\end{cases}$$

Now

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

where we define the errors as follows:
(A) *Initial Error*: $\mathscr{E}_{init}(t) = \|\Pi_1^t r_0\|_K$ comes from the initial choice f_0;
(B) *Approximation Error*: $\mathscr{E}_{approx}(t) = \|f_{\lambda_t} - f_\rho\|_K$, measures the distance between the regression function and the regularization path at time t;
(C) *Drift Error*: $\mathscr{E}_{drift}(t) = \|\sum_{j=1}^t \Pi_j^t\Delta_j\|_K$ comes from the drift along the regularization path $t \mapsto f_{\lambda_t}$;
(D) *Sample Error*: $\mathscr{E}_{samp}(t) = \|\sum_{j=1}^t \gamma_j\Pi_{j+1}^t(A_j\bar\omega_j - b_j)\|_K$, where $\xi_j = \gamma_j\Pi_{j+1}^t(A_j\bar\omega_j - b_j)$ is a reversed martingale difference sequence, reflecting the random fluctuation caused by sampling.

In the remainder of this section, we are going to provide upper bounds for each of the four errors, which, roughly speaking when $ab = 1$, are

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

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 IV.1. This suggests our explanation that the bias $= \mathscr{E}_{approx}(t) + \mathscr{E}_{drift}(t)$ and the variance $= \mathscr{E}_{samp}(t)$. Theorem B then follows from these bounds by setting $\theta = 2r/(2r+1)$.

A. Initial Error

Theorem V.1 (Initial Error): Let $t_0^\theta \ge a(\kappa^2 + b)$. Then for all $t \in \mathbb{N}$,

$$\mathscr{E}_{init}(t) \le B_3\bar{t}^{-ab},$$

where $B_3 = (t_0+1)^{ab}\|r_0\|_K$.

Proof:

$$\begin{aligned}\mathscr{E}_{init}(t) \le \|\Pi_1^t\|\|r_0\|_K &\le \left(\frac{t_0+1}{\bar t+1}\right)^{ab}\|r_0\|_K\\ &\le \left(\frac{t_0+1}{\bar t}\right)^{ab}\|r_0\|_K\end{aligned}$$

where the second last step uses Lemma III.8 (B) with $j = 1$. ■

B. Approximation Error

The approximation error is derived from Theorem IV.1(D) by setting $\lambda = \lambda_t$ and $\mu = 0$.

Theorem V.2 (Approximation Error): For $r \in (1/2, 3/2]$ and $L_K^{-r}f_\rho \in \mathscr{L}^2_{\rho_\mathscr{X}}$,

$$\|f_{\lambda_t} - f_\rho\|_K \le B_1 b^{r-1/2}\bar{t}^{-(r-1/2)(1-\theta)},$$

where $B_1 = (r-1/2)^{-1}\|L_K^{-r}f_\rho\|_\rho$.

C. Drift Error

Theorem V.3 (Drift Error): Let $t_0^\theta \geq [a(\kappa^2+b) \vee 1]$. *Then for* $r \in (1/2, 3/2]$ *and* $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_\mathscr{X}}$,

$$\mathscr{E}_{drift}(t) \leq \begin{cases} B_2 b^{r-1/2} \bar{t}^{-(r-1/2)(1-\theta)}, & \cdots \\ \quad \text{if } ab > (r-1/2)(1-\theta), \\ B_2 b^{r-1/2} \bar{t}^{-ab}, & \cdots \\ \quad \text{if } ab < (r-1/2)(1-\theta), \end{cases}$$

where $B_2 = \dfrac{4(1-\theta)}{|ab-(r-1/2)(1-\theta)|} \|L_K^{-r} f_\rho\|_\rho$.

Proof: We are going to provide an upper bound of

$$\mathscr{E}_{drift}(t) = \| \sum_{j=1}^{t} \Pi_j^t \Delta_j \|_K.$$

First, Lemma III.8 implies, using (18), that for all $j, t \in \mathbb{N}$, $t \geq j$,

$$\|\Pi_j^t\| \leq \left(\frac{j+t_0}{\bar{t}+1}\right)^{ab}, \tag{27}$$

if $t_0^\theta \geq a(\kappa^2+b)$ (and, therefore, $\gamma_t \overline{\alpha}_t = \gamma_t \lambda_t + \gamma_t \kappa^2 \leq abt_0^{-1} + a\kappa^2 t_0^{-\theta} \leq 1$).

Second, by Theorem IV.1(D),

$$\begin{aligned} &\|\Delta_t\|_K \qquad\qquad (28)\\ &= \|f_{\lambda_t} - f_{\lambda_{t-1}}\|_K \leq \left|\lambda_t^{r-1/2} - \lambda_{t-1}^{r-1/2}\right| \frac{\|L_K^{-r} f_\rho\|_\rho}{r-\frac{1}{2}} \\ &\leq b^{r-1/2}(1-\theta)(\bar{t}-1)^{-(r-1/2)(1-\theta)-1} \|L_K^{-r} f_\rho\|_\rho, \end{aligned}$$

where we use

$$\begin{aligned} &|\lambda_t^{r-1/2} - \lambda_{t-1}^{r-1/2}| \qquad\qquad (29)\\ &= b^{r-1/2} \left|\bar{t}^{-(r-1/2)(1-\theta)} - (\bar{t}-1)^{-(r-1/2)(1-\theta)}\right| \\ &\leq b^{r-1/2}(r-1/2)(1-\theta)(\bar{t}-1)^{-(r-1/2)(1-\theta)-1}, \end{aligned}$$

due to the Mean Value Theorem with $h(x) = x^{-(r-1/2)(1-\theta)}$ and $h'(x) = -(r-1/2)(1-\theta)x^{-(r-1/2)(1-\theta)-1}$, such that for some $\eta \in (\bar{t}-1, \bar{t})$,

$$|h(\bar{t}) - h(\bar{t}-1)| = |h'(\eta)| \leq |h'(\bar{t}-1)|.$$

Now combining (27) and (28) gives

$$\begin{aligned} \mathscr{E}_{drift}(t) &= \| \sum_{j=1}^{t} \Pi_j^t \Delta_j \|_K \\ &\leq b^{r-1/2}(1-\theta)\|L_K^{-r} f_\rho\|_\rho \cdots \\ &\quad \cdot \left(\sum_{j=1}^{t} \left(\frac{j+t_0}{\bar{t}+1}\right)^{ab} (j+t_0-1)^{-(r-1/2)(1-\theta)-1} \right) \\ &\leq \frac{4b^{r-1/2}(1-\theta)\|L_K^{-r} f_\rho\|_\rho}{(\bar{t}+1)^{ab}} \cdots \\ &\quad \cdot \sum_{j=1}^{t} (j+t_0)^{ab-1-(r-1/2)(1-\theta)}. \end{aligned}$$

It suffices to bound

$$\begin{aligned} &\sum_{j=1}^{t} (j+t_0)^{ab-1-(r-1/2)(1-\theta)} \\ &\qquad \leq \int_0^{t+1} (x+t_0)^{ab-1-(r-1/2)(1-\theta)} dx =: I_t \end{aligned}$$

Now, if $ab > (r-1/2)(1-\theta)$, then

$$I_t \leq \frac{(\bar{t}+1)^{ab-(r-1/2)(1-\theta)}}{ab-(r-1/2)(1-\theta)};$$

whereas $ab < (r-1/2)(1-\theta)$ implies

$$I_t \leq \frac{t_0^{ab-(r-1/2)(1-\theta)}}{(r-1/2)(1-\theta)-ab} \leq \frac{1}{|ab-(r-1/2)(1-\theta)|},$$

with $t_0 \geq 1$. ■

D. Sample Error

Theorem V.4 (Sample Error): Assume that $t_0^\theta \geq [a(\kappa^2+b) \vee b \vee 1]$, $t_0^{1-\theta} \geq b$ *and* $ab \neq \theta - 1/2$ *or* $(3\theta-1)/2$. *Then, with probability at least* $1-\delta$ ($\delta \in (0,1)$),

$$\mathscr{E}_{samp}(t) \leq B_4 ab^{-1/2} \bar{t}^{-\left[ab \wedge \frac{3\theta-1}{2}\right]} + B_5 a \bar{t}^{-[ab \wedge (\theta-1/2)]}$$

where $B_4 = 2(\kappa+1)^2 M_\rho / 3 \log 2/\delta$ *and* $B_5 = 8\kappa M_\rho / \sqrt{|ab-(\theta-1/2)|} \log 2/\delta$.

The proof of Theorem requires some auxilary estimates. Recall that we assume here that

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

Lemma V.5: We have

(A) $\|A_t \bar{w}_t - b_t\|_K \leq (\kappa+1)^2 M_\rho / \sqrt{\lambda_t}$, if $t_0^{1-\theta} \geq b$;

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

Proof: (A) Using $\|f_\lambda\|_K \leq M_\rho/\sqrt{\lambda}$ in Lemma B.1(A),

$$\begin{aligned} &\|A_t \bar{w}_t - b_t\| \\ &\quad \leq \|f_{\lambda_t}(x_t) K_{x_t}\|_K + |y_t| \|K_{x_t}\|_K + \lambda_t \|f_{\lambda_t}\|_K \\ &\quad \leq M_\rho \kappa^2/\sqrt{\lambda_t} + M_\rho \kappa + M_\rho \sqrt{\lambda_t} \end{aligned}$$

since $\|f_{\lambda_t}(x_t) K_{x_t}\|_K = |\langle f_{\lambda_t}, K_{x_t}\rangle| \|K_{x_t}\|_K \leq \|f_{\lambda_t}\|_K \|K_{x_t}\|_K^2 \leq M_\rho \kappa^2/\sqrt{\lambda_t}$. Now,

$$\begin{aligned} M_\rho \kappa^2/\sqrt{\lambda_t} + M_\rho \kappa + M_\rho \sqrt{\lambda_t} &\leq (\kappa^2+\kappa+1) M_\rho/\sqrt{\lambda_t} \\ &\leq (\kappa+1)^2 M_\rho/\sqrt{\lambda_t} \end{aligned}$$

where the second last inequality is due to $t_0^{1-\theta} \geq b \Rightarrow \lambda_t \leq 1$.

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

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

since $\mathbb{E}[L_t] = L_K$, $\mathbb{E}[y_t K_{x_t}] = L_K f_\rho$ and $\|f_\lambda\|_\rho \leq M_\rho$ by Lemma B.1(B). ■

Now we are ready to give the proof of the sample error bounds, Theorem V-D.

Proof of Theorem V-D: We are going to bound

$$\mathscr{E}_{samp}(t) = \left\| \sum_{j=1}^{t} \xi_j \right\|_K$$

where $\xi_j = \gamma_j \Pi_{j+1}^t (A_j \bar{w}_j - b_j)$ is a reversed martingale difference sequence. To apply the Pinelis-Bernstein inequality in Proposition A.3, we need bounds on $\|\xi_j\|_K$ and $\mathbb{E}_{j+1}\|\xi_j\|_K^2$ where $\mathbb{E}_{j+1}[\cdot]$ is the expectation conditional on examples after time j.

Notice that for $t_0^\theta \geq a(\kappa^2 + b)$ and $j \geq 1$, using $1 + x \leq e^x$ for all $x \in \mathbb{R}$,

$$\begin{aligned}
\|\gamma_j \Pi_{j+1}^t\| &\leq \frac{a}{(j+t_0)^\theta} \left(\frac{j + t_0 + 1}{\bar{t} + 1} \right)^{ab} \\
&\leq \frac{a(j+t_0)^{ab-\theta}}{(\bar{t}+1)^{ab}} (1 + t_0^{-1})^{ab} \\
&\leq \frac{ea(j+t_0)^{ab-\theta}}{(\bar{t}+1)^{ab}},
\end{aligned}$$

where e is the Euler constant.

Now Lemma V.5 (B) implies

$$\mathbb{E}[\|A_t \bar{w}_t - b_t\|_K^2] \leq 4\kappa^2 M_\rho^2.$$

Hence

$$\mathbb{E}_{j+1}\|\xi_j\|^2 \leq \frac{4e^2(a\kappa M_\rho)^2 (j+t_0)^{2ab-2\theta}}{(\bar{t}+1)^{2ab}},$$

so that, if $t_0 \geq 2$,

$$\begin{aligned}
&\sum_{j=1}^{t} \mathbb{E}_{j+1}\|\xi_j\|^2 \qquad (30) \\
&\leq \begin{cases} \dfrac{2e^2(a\kappa M_\rho)^2}{ab - (\theta - \frac{1}{2})} (\bar{t}+1)^{-2\theta-1}, \text{ if } ab > \theta - \frac{1}{2}; \\ \dfrac{2e^2(a\kappa M_\rho)^2}{(\theta - \frac{1}{2}) - ab} (\bar{t}+1)^{-2ab}, \quad \text{if } ab < \theta - \frac{1}{2}. \end{cases}
\end{aligned}$$

On the other hand, if $t_0^{1-\theta} \geq b$, Lemma V.5 (A) implies

$$\|A_j \bar{w}_j - b_j\|_K \leq (\kappa+1)^2 M_\rho / \sqrt{\lambda_j},$$

whence

$$\begin{aligned}
\|\xi_j\|_K &\leq \frac{ea(\kappa+1)^2 M_\rho}{\sqrt{b}} \cdot \frac{(j+t_0)^{ab-(3\theta-1)/2}}{(\bar{t}+1)^{ab}} \\
&\leq \begin{cases} \dfrac{ea(\kappa+1)^2 M_\rho}{\sqrt{b}} \bar{t}^{-(3\theta-1)/2}, \\ \quad \text{if } ab \geq (3\theta-1)/2; \\ \dfrac{ea(\kappa+1)^2 M_\rho}{\sqrt{b}} \bar{t}^{-ab}, \\ \quad \text{if } ab \leq (3\theta-1)/2. \end{cases} \qquad (31)
\end{aligned}$$

The final bound is obtained by Pinelis-Bernstein inequality in Proposition A.3 with (30) and (31). ■

E. Proof of Theorem B

We choose $\theta = 2r/(2r+1)$, $a \geq 1$, $b \leq 1$ such that $ab = 1$, and assume $t_0^\theta \geq a\kappa^2 + 1$; hence the assumptions of Theorems V.1–V.2 are satisfied. Let us check that the conditions of Theorem V-D also hold: $t_0 \geq 1$, $b \leq 1$ and $\theta < 1$ readily imply $t_0^{1-\theta} \geq b$, and $ab = 1 \neq \theta - 1/2$ or $(3\theta - 1)/2$. Using theorems V.1–V-D, we deduce

$$\begin{aligned}
&\|f_t - f_\rho\|_K \\
&\quad \leq \mathscr{E}_{init}(t) + \mathscr{E}_{approx}(t) + \mathscr{E}_{drift}(t) + \mathscr{E}_{samp}(t) \\
&\quad \leq B_3 \bar{t}^{-ab} + [(B_1 + B_2)a^{1/2-r} \cdots \\
&\qquad + (B_4 t_0^{-\theta} \sqrt{a} + B_5)a] \bar{t}^{-(2r-1)/(4r+2)}.
\end{aligned}$$

Note that, by Lemma V.5(A) with $f_0 = 0$,

$$\begin{aligned}
B_3 &= (t_0 + 1)\|r_0\| = (t_0+1)\|f_{\lambda_0}\| \\
&\leq C_0 := 2t_0 \frac{M_\rho}{\sqrt{\lambda_0}} = 2t_0^{\frac{4r+3}{4r+2}} M_\rho
\end{aligned}$$

On the other hand,

$$\begin{aligned}
C_1 := B_1 + B_2 &= \left(\frac{2}{2r-1} + \frac{8}{2r+3} \right) \|L_K^{-r} f_\rho\|_\rho \\
&= \frac{20r - 2}{(2r-1)(2r+3)} \|L_K^{-r} f_\rho\|_\rho
\end{aligned}$$

and, using Â $\sqrt{a} t_0^{-\theta} \leq \kappa^{-1}$ and $t_0 \geq 1$,

$$\begin{aligned}
B_4 t_0^{-\theta} \sqrt{a} + B_5 &\leq \frac{2(\kappa+1)^2 M_\rho}{3\kappa} + \frac{8\kappa M_\rho}{\sqrt{3/4}} \\
&\leq C_2 := \frac{20(\kappa+1)^2 M_\rho}{\kappa},
\end{aligned}$$

which concludes the proof of Theorem B.

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

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, and assume the sequences $(\gamma_t)_{t \in \mathbb{N}}$ and $(\lambda_t)_{t \in \mathbb{N}}$ are chosen in (20). Note that the case $r = 1/2$ is included here, whereas it was not in Section V and Theorem B.

Our goal is to provide a probabilistic upper bound of

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

in order to prove Theorem C. As in Section V, we start with the triangle inequality

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

but apply here the martingale decomposition of $(f_t)_{t \in \mathbb{N}_0}$ in $\mathscr{L}^2_{\rho_\mathscr{X}}$ developed in Theorem III.3 instead:

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

We make use of the corresponding notation of Section III, in particular III-C, so that

$$\chi_t = (L_K - L_t) f_{t-1} + (y_t K_{x_t} - L_K f_\rho),$$

and

$$\bar{\Pi}_j^t = \begin{cases} \prod_{i=j}^{t} (I - \gamma_i (L_K + \lambda_i I)) & \text{if } j \le t; \\ I & \text{otherwise.} \end{cases} \tag{32}$$

The martingale decomposition enables us to make use of the isometry $L_K^{1/2} : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{H}_K$, in the sense that one can benefit from the spectral decomposition of $L_K^{1/2}\bar{\Pi}_j^t$ to get a tighter estimate. This was not possible with the reversed martingale decomposition, since $L_K^{1/2}\Pi_j^t$ does not have an obvious spectral decomposition.

Note however that χ_t depends on f_{t-1}, so that we need preliminary estimates of $\|\chi_t\|_\rho$, provided in Appendix B.

As in Section V, we introduce the following definitions for convenience.

[Definitions of Errors]

(A) *Initial Error*: $\mathscr{E}_{init}(t) = \|\bar{\Pi}_1^t r_0\|_\rho$, which reflects the propagation error by the initial choice f_0;

(B) *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;

(C) *Drift Error*: $\mathscr{E}_{drift}(t) = \|\sum_{j=1}^t \bar{\Pi}_j^t \Delta_j\|_\rho$, which measures the error caused by drifts from $f_{\lambda_{j-1}}$ to f_{λ_j} along the regularization path;

(D) *Sample Error*: $\mathscr{E}_{samp}(t) = \|\sum_{j=1}^t \gamma_j \bar{\Pi}_{j+1}^t \chi_j\|_\rho$, where χ_j is a martingale difference sequence, reflecting the random fluctuation caused by sampling.

Our aim is to 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 remainder of this section, we are going to provide upper bounds for each of the four errors, which, roughly speaking when $ab = 1$, are

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

This suggests our explanation that the bias $= \mathscr{E}_{approx}(t) + \mathscr{E}_{drift}(t) = O(t^{-r(1-\theta)})$ and the variance $= \mathscr{E}_{samp}(t) = O(t^{-\theta/2})$ similar to the batch learning setting. Theorem C then follows from these bounds by setting $\theta = 2r/(2r+1)$.

A. Initial Error

Theorem VI.1 (Initial Error): Let $t_0^\theta \ge a(\kappa^2 + b)$. Then for all $t \in \mathbb{N}$,

$$\mathscr{E}_{init}(t) \le B_6 \bar{t}^{-ab},$$

where $B_6 = M_\rho (t_0 + 1)^{ab}$.

Proof: Lemma III.8(B) with $j = 1$ and (18) imply that, if $t_0^\theta \ge a(\kappa^2 + b)$,

$$\mathscr{E}_{init}(t) \le \|\bar{\Pi}_1^t\| \|r_0\| \le \left(\frac{t_0 + 1}{\bar{t} + 1}\right)^{ab} \|r_0\|.$$

For $f_0 = 0$, using Lemma B.1(B), $\|r_0\|_\rho = \|f_{\lambda_0}\|_\rho \le M_\rho$. ■

B. Approximation Error

Theorem VI.2 (Approximation Error): For $r \in (0, 1]$ and $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$,

$$\mathscr{E}_{approx}(t) \le B_7 b^r \bar{t}^{-r(1-\theta)},$$

where $B_7 = r^{-1}\|L_K^{-r} f_\rho\|_\rho$.

Proof: Follows from Theorem IV.1(A) with $\lambda = \lambda_t$ and $\mu = 0$. ■

C. Drift Error

Theorem VI.3 (Drift Error): Assume $t_0^\theta \ge [a(\kappa^2 + b) \vee 1]$. Then, if $r \in (0, 1]$ and $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$,

$$\mathscr{E}_{drift}(t) \le \begin{cases} B_8 b^r \bar{t}^{-r(1-\theta)}, & \text{if } ab > r(1-\theta); \\ B_8 b^r \bar{t}^{-ab}, & \text{if } ab < r(1-\theta), \end{cases}$$

where $B_8 = \dfrac{4(1-\theta)}{|ab - r(1-\theta)|}\|L_K^{-r} f_\rho\|_\rho$.

Proof: Similar to the proof of Theorem V.3, replacing $r - 1/2$ with r. ■

D. Sample Error

In this section we assume $b = a^{-1}$ for simplicity; this is necessary for the bounds in Appendix B, in particular Corollary B.7, and is enough to provide the optimal bounds we need (see discussion after statement of Theorem A).

Theorem VI.4 (Sample Error): Assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ for some $r \in [1/2, 1]$, $\theta \in [1/2, 2/3]$, $ab = 1$, $a \ge 4$ and $t_0^\theta \ge 2 + 8\kappa^2 a$. Then, for all $t \in \mathbb{N}$, with probability at least $1 - \delta$ ($\delta \in (0, 1)$)

$$\begin{aligned} \mathscr{E}_{samp}(t) \le \sqrt{a} B_9 \frac{\log(2/\delta)}{\bar{t}^{\theta/2}} \cdots \\ + \left(a^{3/2} B_{10} \sqrt{\log \bar{t}} + a^{5/2} B_{11}\right) \frac{(\log(2/\delta))^2}{\bar{t}^{(3\theta-1)/2}}, \end{aligned}$$

where

$$B_9 := 10\kappa M_\rho, \quad B_{10} := 63\kappa^2 M_\rho, \quad B_{11} := 50\kappa^2 M_\rho t_0^{1/2-\theta}.$$

Fix $t \in \mathbb{N}$, $\delta \in [0, 1]$, and let

$$A_{t,\delta} := \kappa M_\rho a \left[12 a t_0^{1/2-\theta} + 15\sqrt{\log \bar{t}}\right] \log \frac{2}{\delta}. \tag{33}$$

For all $j \in \mathbb{N}$, let us define the martingale difference sequence

$$X_j := \gamma_j \bar{\Pi}_{j+1}^t \chi_j \mathbf{1}_{\{\|h_{j-1}\|_K (j+t_0)^{\theta-1/2} \le A_{t,\delta}\}},$$

where we make use of the notation of Appendix B. Recall that Corollary B.7 implies, with probability at least $1 - \delta$, all the indicator function events for $1 \le j < t$ hold, which will be assumed in the computation below.

Recall that

$$\begin{aligned} \chi_j &= (\bar{A}_j - A_j) w_{j-1} + b_j - \bar{b}_j \\ &= (L_K - L_j) f_{j-1} + y_j K_{x_j} - L_K f_\rho \end{aligned}$$

where $L_j := \langle\ , K_{x_j}\rangle K_{x_j}$.

Using Lemma B.3 and the decomposition $f_j = f_\rho + g_j + h_j$ in Appendix B, we deduce that, for all $1 \le j < t$, if $\|h_{j-1}\|_K (j+t_0)^{\theta-1/2} \le A_{t,\delta}$,

$$\begin{aligned}
&\mathbb{E}_{j-1}\|\chi_j\|_K^2 \\
&\quad = \mathbb{E}_{j-1}\|y_j K_{x_j} - L_j f_{j-1} - (L_K f_\rho - L_K f_{j-1})\|_K^2 \\
&\quad \le \mathbb{E}_{j-1}\|y_j K_{x_j} - L_j f_{j-1}\|_K^2 \\
&\quad \le \mathbb{E}_{j-1}\|(y_j K_{x_j} - L_j f_\rho) - L_j g_{j-1} - L_j h_{j-1}\|_K^2 \\
&\quad \le 3\kappa^2[\mathbb{E}_{j-1}|y_j - f_\rho(x_j)|^2 + \mathbb{E}_{j-1}|g_{j-1}(x_j)|^2 \cdots \\
&\qquad \cdots + \mathbb{E}_{j-1}|h_{j-1}(x_j)|^2] \\
&\quad \le 3\kappa^2[4M_\rho^2 + \|g_{j-1}\|_\rho^2 + \|h_{j-1}\|_\rho^2] \\
&\quad \le 3\kappa^2[5M_\rho^2 + \kappa^2 (j+t_0)^{1-2\theta} A_{t,\delta}^2] =: A'_{j,t,\delta}
\end{aligned}$$

Now, using the isometry $L_K^{1/2} : \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{H}_K$,

$$\begin{aligned}
&\sum_{j=1}^t \mathbb{E}_{j-1}\|X_j\|_\rho^2 \\
&\quad = \sum_{j=1}^t \mathbb{E}_{j-1}\|L_K^{1/2} X_j\|_K^2 = \sum_{j=1}^t \gamma_j^2 \mathbb{E}_{j-1}\|L_K^{1/2}\bar{\Pi}_{j+1}^t \chi_j\|_K^2 \\
&\quad \le \sum_{j=1}^t \left(\gamma_j^2 \|\bar{\Pi}_{j+1}^t L_K \bar{\Pi}_{j+1}^t\|\right) \mathbb{E}_{j-1}\|\chi_j\|_K^2 \\
&\quad \le \sum_{j=1}^t \gamma_j^2 A'_{j,t,\delta} \|\bar{\Pi}_{j+1}^t L_K \bar{\Pi}_{j+1}^t\|
\end{aligned}$$

In order to estimate $\sum_{j=1}^t \gamma_j^2 A'_{j,t,\delta}\|\bar{\Pi}_{j+1}^t L_K \bar{\Pi}_{j+1}^t\|$, recall that $(\mu_\alpha, \phi_\alpha)_{\alpha\in\mathbb{N}}$ is an orthonormal eigen-system of $L_K : \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{L}^2_{\rho_\mathscr{X}}$. Let $a_i = \gamma_i \lambda_i + \gamma_i \mu_\alpha$ for simplicity; then

$$\begin{aligned}
&\sum_{j=1}^t \gamma_j^2 A'_{j,t,\delta}\|\bar{\Pi}_{j+1}^t L_K \bar{\Pi}_{j+1}^t\| \\
&\quad \le \sup_{\mu_\alpha} \sum_{j=1}^t \gamma_j^2 A'_{j,t,\delta} \mu_\alpha \prod_{i=j+1}^t (1-a_i)^2 \\
&\quad = \sup_{\mu_\alpha} \sum_{j=1}^t \left[\gamma_j A'_{j,t,\delta} \prod_{i=j+1}^t (1-a_i)\right] \cdot \left[\gamma_j \mu_\alpha \prod_{i=j+1}^t (1-a_i)\right] \\
&\quad \le \sup_{\mu_\alpha} \left\{\left[\sup_j \gamma_j A'_{j,t,\delta} \prod_{i=j+1}^t (1-a_i)\right] \cdots \right. \\
&\qquad \left. \cdot \left[\sum_{j=1}^t \gamma_j \mu_\alpha \prod_{i=j+1}^t (1-a_i)\right]\right\}
\end{aligned}$$

where for large enough t_0,

$$\begin{aligned}
&\sup_j \gamma_j A'_{j,t,\delta} \prod_{i=j+1}^t (1-a_i) \\
&\quad \le \sup_j \gamma_j A'_{j,t,\delta} \prod_{i=j+1}^t (1-\gamma_i\lambda_i) \\
&\quad \le 3a\kappa^2 \sup_j \frac{j+t_0}{\overline{t}} \cdot \left(\frac{5M_\rho^2}{(j+t_0)^\theta} + \frac{\kappa^2 A_{t,\delta}^2}{(j+t_0)^{3\theta-1}}\right) \\
&\quad \le 3a\kappa^2 \left(\frac{5M_\rho^2}{\overline{t}^\theta} + \frac{\kappa^2 A_{t,\delta}^2}{\overline{t}^{3\theta-1}}\right),
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{j=1}^t \gamma_j \mu_\alpha \prod_{i=j+1}^t (1-a_i) \\
&\quad \le \sum_{j=1}^t (1-(1-\gamma_j\mu_\alpha)) \prod_{i=j+1}^t (1-\gamma_i\mu_\alpha) \\
&\quad = 1 - \prod_{i=1}^t (1-\gamma_i\mu_\alpha) \le 1.
\end{aligned}$$

These two upper bounds give

$$\sum_{j=1}^t \mathbb{E}_{j-1}\|X_j\|_\rho^2 \le \frac{3a\kappa^2}{\overline{t}^\theta}\left(5M_\rho^2 + \frac{\kappa^2 A_{t,\delta}^2}{\overline{t}^{2\theta-1}}\right) =: \sigma_t^2. \quad (34)$$

Moreover, again if $\|h_{j-1}\|_K (j+t_0)^{\theta-1/2} \le A_{t,\delta}$, then, using Lemma B.3 (B) and Corollary B.7, we deduce

$$\begin{aligned}
&\|y_j K_{x_j} - L_j f_{j-1}\|_K \\
&\quad = \|y_j K_{x_j} - L_j(f_\rho + g_{j-1} + h_{j-1})\|_K \\
&\quad \le \kappa M_\rho + \frac{\kappa^2 M_\rho}{\sqrt{\lambda_j}} + \kappa^2 A_{t,\delta}(j+t_0)^{1/2-\theta} =: C_{j,t,\delta},
\end{aligned}$$

which implies

$$\begin{aligned}
\|\chi_j\|_K &= \|y_j K_{x_j} - L_j f_{j-1} - \mathbb{E}_j[y_j K_{x_j} - L_j f_{j-1}]\|_K \\
&\le 2C_{j,t,\delta}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\|X_j\|_\rho &\le \gamma_j \|L_K^{1/2}\bar{\Pi}_{j+1}^t \chi_j\|_K \le 2\gamma_j C_{j,t,\delta}\|\bar{\Pi}_{j+1}^t L_K \bar{\Pi}_{j+1}^t\|_K^{1/2} \\
&\le 2\kappa \sup_j \gamma_j C_{j,t,\delta} \prod_{i=j+1}^t (1-\gamma_i\lambda_i), \quad \|L_K\| \le \kappa^2 \\
&\le 2a\kappa^2 \sup_j \frac{j+t_0}{\overline{t}} \cdot \left(\frac{M_\rho}{(j+t_0)^\theta} \cdots \right. \\
&\qquad \left. \cdots + \frac{\kappa M_\rho \sqrt{a}}{(j+t_0)^{(3\theta-1)/2}} + \frac{\kappa A_{t,\delta}}{(j+t_0)^{2\theta-1/2}}\right) \\
&\le 2a\kappa^2 \left(\frac{M_\rho}{\overline{t}^\theta} + \frac{\kappa M_\rho\sqrt{a}}{\overline{t}^{(3\theta-1)/2}} + \frac{\kappa A_{t,\delta}}{\overline{t}^{2\theta-1/2}}\right) \\
&\le \frac{2\sqrt{a}\kappa}{\overline{t}^{\theta/2}}\left(M_\rho + \kappa\frac{\kappa M_\rho a + A_{t,\delta}}{\overline{t}^{\theta-1/2}}\right) =: M,
\end{aligned}$$

where we use $t_0 \ge \kappa^2$ twice in the last inequality.

Combining M and σ_t from (34), we obtain

$$\begin{aligned}
&2\left(\frac{M}{3} + \sigma_t\right) \\
&\quad = \frac{2\sqrt{a}\kappa}{\overline{t}^{\theta/2}}\left[\left(\sqrt{15}+\frac{2}{3}\right) M_\rho + \kappa\frac{(\sqrt{3}+1/3)A_{t,\delta} + \kappa M_\rho a/3}{\overline{t}^{\theta-1/2}}\right] \\
&\quad \le \frac{\sqrt{a}B_9}{\overline{t}^{\theta/2}} + \left(a^{3/2}B_{10}\sqrt{\log\overline{t}} + a^{5/2}B_{11}\right)\frac{\log(2/\delta)}{\overline{t}^{(3\theta-1)/2}},
\end{aligned}$$

where we use that

$$\begin{aligned} B_9 &= 10\kappa M_\rho \geq 2\kappa(\sqrt{15} + 2/3)M_\rho, \\ B_{10} &= 63\kappa^2 M_\rho \geq \kappa^2 M_\rho[30(\sqrt{3} + 1/3) + 2/(3\log 2)], \\ B_{11} &= 50\kappa^2 M_\rho t_0^{1/2-\theta} \geq 24\kappa^2 M_\rho t_0^{1/2-\theta}(\sqrt{3} + 1/3). \qquad \blacksquare \end{aligned}$$

E. Proof of Theorem C

We choose $\theta = 2r/(2r+1)$, $a \geq 1$, $b \leq 1$ such that $ab = 1$, and assume $t_0^\theta \geq 8a\kappa^2 + 2$; hence the assumptions of Theorems VI.1–VI.4 are satisfied, and $ab = 1 > r(1-\theta)$ in Theorem VI.3. Using Theorems VI.1–VI.4, we deduce

$$\begin{aligned} &\|f_t - f_\rho\|_\rho \\ &\quad \leq \mathscr{E}_{init}(t) + \mathscr{E}_{approx}(t) + \mathscr{E}_{drift}(t) + \mathscr{E}_{samp}(t) \\ &\quad \leq \frac{B_6}{\bar{t}} + \left((B_7 + B_8)a^{-r} + \sqrt{a}B_9 \log\frac{2}{\delta}\right)\left(\frac{1}{\bar{t}}\right)^{\frac{r}{2r+1}} \dots \\ &\qquad \dots + \left(a^{3/2}B_{10}\sqrt{\log\bar{t}} + a^{5/2}B_{11}\right)\frac{(\log(2/\delta)^2}{\bar{t}^{\frac{6r-1}{4r+2}}}, \end{aligned}$$

This enables us to conclude, with $D_0 := 2M_\rho t_0 \geq B_6 = M_\rho(t_0 + 1)$,

$$D_1 := B_7 + B_8 = \frac{5r+1}{r(1+r)}\|L_K^{-r} f_\rho\|_\rho,$$

$D_2 := B_9$, $D_3 := B_{10}$, and $D_4 := B_{11}$.

Appendix A
A Probabilistic Inequality

The following result is quoted from [Theorem 3.4 in [23]].

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

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

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

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

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 $1-\delta$ ($\delta \in (0,1)$),

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

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 $\delta \in (0,1)$. $\blacksquare$

Appendix B
Preliminary Upper Bounds

Appendix B is devoted to the proof of preliminary upper bounds on the online learning sequence $(f_t)_{t\in\mathbb{N}}$ defined in (7), and on the regularization path $\lambda \mapsto f_\lambda$. We make use of the notation of Section III, in particular Section III-C. For simplicity we assume $f_0 := 0$; note that another choice would correspond to adding $\Pi_1^t f_0$ to f_t at time t. We assume that the sequences $(\gamma_t)_{t\in\mathbb{N}}$ and $(\lambda_t)_{t\in\mathbb{N}}$ are chosen as in (20).

Firstly, Lemmas B.1 and B.2 provide deterministic upper bounds. Then the rest of the Appendix aims at obtaining probabilistic bounds of $(f_t)_{t\in\mathbb{N}}$, based on a decomposition of $f_t - f_\rho$ into two parts in (B-2): g_t is purely deterministic and is upper bounded in $\mathscr{L}^2_{\rho_\mathscr{X}}$-norm in Lemma B.3, and h_t is studied in detail in Lemmas B.4 and following. Lemma B.7 yields logarithmic estimates with large probability.

Lemma B.1: For any $\lambda > 0$,

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

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

Proof: (A) By definition,

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

The term we minimize on the right-hand side takes the value $\|f_\rho\|_\rho^2$ at $f = 0$, so that

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

which yields the result.

(B) Using (6),

$$\begin{aligned} \|f_\lambda\|_\rho &= \|(L_K + \lambda I)^{-1}L_K f_\rho\|_\rho \\ &\leq \|(L_K + \lambda I)^{-1}L_K\|\|f_\rho\|_\rho \leq \|f_\rho\|_\rho \leq M_\rho. \end{aligned}$$

$\blacksquare$

Lemma B.2: Assume $t_0^\theta \geq a(\kappa^2 + b)$. Then, for all $t \in \mathbb{N}$,

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

Proof: Recall that $f_t = (I - \gamma_t A_t) f_{t-1} + \gamma_t y_t K_{x_t}$. Now assume $t_0^\theta \geq a(\kappa^2 + b)$: using Lemma III.8 (see also (21)),

$$\begin{aligned}\|f_t\|_K &\leq \|1 - \gamma_t A_t\| \|f_{t-1}\|_K + \gamma_t \|y_t K_{x_t}\|_K \\ &\leq (1 - \gamma_t \lambda_t) \|f_{t-1}\|_K + \gamma_t \kappa M_\rho.\end{aligned}$$

By induction on t, we deduce

$$\begin{aligned}\|f_t\|_K &\leq \kappa M_\rho \sum_{j=1}^{t} \gamma_j \prod_{i=j+1}^{t} (1 - \gamma_i \lambda_i) \\ &\leq \max_{1 \leq j \leq t} \left(\frac{1}{\lambda_j}\right) \sum_{j=1}^{t} \gamma_j \lambda_j \prod_{i=j+1}^{t} (1 - \gamma_i \lambda_i) \leq \frac{1}{\lambda_t},\end{aligned}$$

since

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

■

In the rest of Appendix, we prove probabilistic bounds of $(f_t)_{t \in \mathbb{N}_0}$. First, using $L_t = \langle\ , K_{x_t}\rangle K_{x_t}$ (Eq. (19)), observe that the definition of the online learning sequence (7) can be rewritten as,

$$\begin{aligned}f_t - f_\rho = [I - \gamma_t (L_t + \lambda_t I)](f_{t-1} - f_\rho) \ldots \\ \ldots + \gamma_t (y_t K_{x_t} - L_t f_\rho) - \gamma_t \lambda_t f_\rho.\end{aligned}$$

Let us now define the following $(\mathcal{F}_t)_{t \in \mathbb{N}_0}$-adapted processes $(g_t)_{t \in \mathbb{N}_0}$ and $(h_t)_{t \in \mathbb{N}_0}$ recursively by

$$g_0 := -f_\rho, \quad h_0 := 0,$$

and

$$\begin{aligned}g_t &: = [I - \gamma_t (L_K + \lambda_t I)] g_{t-1} - \gamma_t \lambda_t f_\rho, \\ h_t &: = [I - \gamma_t (L_t + \lambda_t I)] h_{t-1} \ldots \\ &\quad \ldots + \gamma_t (y_t K_{x_t} - L_t f_\rho) + \gamma_t (L_K - L_t) g_{t-1}.\end{aligned}$$

We can easily prove by induction that

$$f_t - f_\rho = g_t + h_t, \tag{B-2}$$

using $f_0 = 0$.

Lemma B.3: Assume $t_0^\theta \geq a(\kappa^2 + b)$. Then, for all $t \in \mathbb{N}_0$,

(A) $\|g_t\|_\rho \leq\ M_\rho$;

(B) $\|g_t + f_\rho\|_K \leq 3 M_\rho / \sqrt{\lambda_t}$.

Proof: We prove (A) by induction: $\|g_0\|_\rho = \|f_\rho\|_\rho \leq M_\rho$ and, for all $t \in \mathbb{N}$, if we assume $\|g_{t-1}\|_\rho \leq M_\rho$ then, using (22),

$$\begin{aligned}\|g_t\|_\rho &\leq \|I - \gamma_t (L_K + \lambda_t I)]\| \|g_{t-1}\|_\rho + \gamma_t \lambda_t \|f_\rho\|_\rho \\ &\leq (1 - \gamma_t \lambda_t) \|g_{t-1}\|_\rho + \gamma_t \lambda_t M_\rho \leq M_\rho.\end{aligned}$$

To prove (B), observe that, for all $t \in \mathbb{N}$,

$$\begin{aligned}g_t + f_\rho &= [I - \gamma_t (L_K + \lambda_t I)] g_{t-1} + (1 - \gamma_t \lambda_t) f_\rho \\ &= [I - \gamma_t (L_K + \lambda_t I)](g_{t-1} + f_\rho) + \gamma_t L_K f_\rho \\ &= [I - \gamma_t (L_K + \lambda_t I)](g_{t-1} + f_\rho) + \gamma_t (L_K + \lambda_t I) f_{\lambda_t},\end{aligned}$$

so that

$$g_t + f_\rho - f_{\lambda_t} = [I - \gamma_t (L_K + \lambda_t I)](g_{t-1} + f_\rho - f_{\lambda_t}).$$

Let, for all $t \in \mathbb{N}$,

$$w_t := g_t + f_\rho - f_{\lambda_t}.$$

Then it is easy to show by induction that

$$w_t = \bar{\Pi}_1^t w_0 + \sum_{k=1}^{t} \bar{\Pi}_k^t (f_{\lambda_k} - f_{\lambda_{k-1}})$$

which implies, using Theorem IV.1 (D) with $r = 0$, and Lemma B.1 (A) ($w_0 = -f_{\lambda_0}$) that

$$\|w_t\| \leq 2 M_\rho / \sqrt{\lambda_t}.$$

This enables us to conclude, using again Lemma B.1 (A). ■

For all $t \in \mathbb{N}_0$, $M_t \in \mathbb{R}_+ \cup \{\infty\}$, let

$$\begin{aligned}\overline{L}_t &: = \mathbf{1}_{\{|h_{t-1}(x_t)| \leq M_t\}} L_t, \quad & \tilde{L}_t &:= \mathbf{1}_{\{|h_{t-1}(x_t)| > M_t\}} L_t, \\ \overline{L}_K &: = \mathbb{E}_{t-1}[\overline{L}_t], & \tilde{L}_K &:= \mathbb{E}_{t-1}[\tilde{L}_t].\end{aligned}$$

Note that $L_t = \overline{L}_t + \tilde{L}_t$ and $L_K = \overline{L}_K + \tilde{L}_K$.

For all $t \in \mathbb{N}$, let

$$\begin{aligned}\overline{h}_t &:= [I - \gamma_t (\overline{L}_t + \lambda_t I)] h_{t-1} + \gamma_t (y_t K_{x_t} - L_t f_\rho) \\ &\quad + \gamma_t (L_K - L_t) g_{t-1} \\ &= h_t + \gamma_t \tilde{L}_t h_{t-1} \qquad \text{(B-3)} \\ k_t &:= \overline{h}_t - (1 - \gamma_t \lambda_t) h_{t-1} \\ &= \gamma_t [-\overline{L}_t h_{t-1} + (y_t K_{x_t} - L_t f_\rho) + (L_K - L_t) g_{t-1}] \qquad \text{(B-4)} \\ &= \gamma_t [-\overline{L}_t h_{t-1} + y_t K_{x_t} + L_K g_{t-1} - L_t (f_\rho + g_{t-1})]. \qquad \text{(B-5)}\end{aligned}$$

In Lemma B.4 we upper bound $\|\overline{h}_t\|_K^2$ in conditional expectation; note that the result still holds when $M_t = \infty$. We threshold h_t into $\overline{h}_t$ in order to limit its conditional variance, which will be necessary in order to obtain logarithmic estimates with large probability in Lemma B.7, using on the other hand Lemma B.6 showing that, if M_t is large enough, $\|h_t\|_K \leq \|\overline{h}_t\|_K$.

Lemma B.4: Assume $t_0^\theta \geq 2a(b + \kappa^2)$. For all $t \in \mathbb{N}$, $M_t \in \mathbb{R}_+ \cup \{\infty\}$, we have

$$\mathbb{E}_{t-1}[\|\overline{h}_t\|_K^2] \leq (1 - \gamma_t \lambda_t)^2 \|h_{t-1}\|_K^2 + 3 \kappa^2 M_\rho^2 \gamma_t^2.$$

In particular, assume moreover that $\theta \geq 1/2$, $\epsilon := ab - (\theta - 1/2) > 0$ and $t_0 \geq \max(2ab, 2\epsilon, \epsilon + (2\theta - 1)/\epsilon)$, and let $A := a \kappa M_\rho \sqrt{3/\epsilon}$. Then $\|h_{t-1}\|_K \geq A \bar{t}^{1/2 - \theta}$ implies

$$\bar{t}^{\theta - 1/2} \mathbb{E}_{t-1}[\|\overline{h}_t\|_K] \leq (\bar{t} - 1)^{\theta - 1/2} \|h_{t-1}\|_K.$$

Proof: For all $t \in \mathbb{N}$, let

$$\zeta_t := (\overline{L}_K - \overline{L}_t) h_{t-1} + (L_K - L_t) g_{t-1} + (y_t K_{x_t} - L_t f_\rho),$$

so that

$$\overline{h}_t = [I - \gamma_t (\overline{L}_K + \lambda_t I)] h_{t-1} + \gamma_t \zeta_t.$$

Using $\mathbb{E}_{t-1}[\zeta_t] = 0$, we deduce that

$$\begin{aligned}&\mathbb{E}_{t-1}[\|\overline{h}_t\|_K^2] \qquad \text{(B-6)} \\ &\quad = \|[I - \gamma_t (\overline{L}_K + \lambda_t I)] h_{t-1}\|_K^2 + \gamma_t^2 E_{t-1}[\|\zeta_t\|_K^2].\end{aligned}$$

Let us now upper bound the two summands in the right-hand side of equality (B-7). First,

$$\begin{aligned}&\|[I-\gamma_t(\overline{L}_K+\lambda_t I)]h_{t-1}\|_K^2\\ &\quad=(1-\gamma_t\lambda_t)^2\|h_{t-1}\|_K^2-2\gamma_t(1-\gamma_t\lambda_t)\ldots\\ &\qquad\ldots\cdot\mathbb{E}_{t-1}[|h_{t-1}(x_t)|^2\mathbf{1}_{\{|h_{t-1}(x_t)|\le M_t\}}]\\ &\qquad\ldots+\gamma_t^2\|\mathbb{E}_{t-1}[\overline{L}_t h_{t-1}]\|_K^2,\end{aligned}$$

and

$$\begin{aligned}&\|\mathbb{E}_{t-1}[\overline{L}_t h_{t-1}]\|_K^2\\ &\quad\le\left(\mathbb{E}_{t-1}[|h_{t-1}(x_t)|\mathbf{1}_{\{|h_{t-1}(x_t)|\le M_t\}}\|K_{x_t}\|_K]\right)^2\\ &\quad\le\kappa^2\mathbb{E}_{t-1}[|h_{t-1}(x_t)|^2\mathbf{1}_{\{|h_{t-1}(x_t)|\le M_t\}}],\end{aligned}$$

using conditional Jensen's inequality.

Second, using that $\mathbb{E}[y_t K_{x_t}-L_t f_\rho \mid \sigma(\mathcal{F}_{t-1},x_t)]=0$,

$$\begin{aligned}&\mathbb{E}_{t-1}[\|\zeta_t\|_K^2]\\ &=\mathbb{E}_{t-1}[\|(\overline{L}_K-\overline{L}_t)h_{t-1}+(L_K-L_t)g_{t-1}\|_K^2\ldots\\ &\quad\ldots+\|y_t K_{x_t}-L_t f_\rho\|_K^2]\\ &\le\mathbb{E}_{t-1}[\|\overline{L}_t h_{t-1}+L_t g_{t-1}\|_K^2+\|y_t K_{x_t}\|_K^2]\\ &\le\kappa^2\left(2\mathbb{E}_{t-1}[|h_{t-1}(x_t)|^2\mathbf{1}_{\{|h_{t-1}(x_t)|\le M_t\}}]+2\|g_{t-1}\|_\rho^2+M_\rho^2\right)\\ &\le\kappa^2\left(2\mathbb{E}_{t-1}[|h_{t-1}(x_t)|^2\mathbf{1}_{\{|h_{t-1}(x_t)|\le M_t\}}]+3M_\rho^2\right),\end{aligned}$$

where we use Lemma B.3 (A) in the last inequality.

In summary, we obtain that

$$\begin{aligned}&\mathbb{E}_{t-1}[\|\overline{h}_t\|_K^2]\\ &\quad\le(1-\gamma_t\lambda_t)^2\|h_{t-1}\|_K^2-\gamma_t(2-2\gamma_t\lambda_t-3\gamma_t\kappa^2)\ldots\\ &\qquad\ldots\cdot\mathbb{E}_{t-1}[|h_{t-1}(x_t)|^2\mathbf{1}_{\{|h_{t-1}(x_t)|\le M_t\}}]\ldots\\ &\qquad\ldots+3\kappa^2M_\rho^2\gamma_t^2.\end{aligned}$$

Now, the assumption $t_0^\theta\ge 2a(b+\kappa^2)$ implies $2-2\gamma_t\lambda_t-3\gamma_t\kappa^2\ge 0$ for all $t\in\mathbb{N}$, which completes the proof of the first statement.

Let us now prove the second statement:

$$\begin{aligned}\Delta_t:&=\mathbb{E}_{t-1}\left[\left(1-\frac{1}{\overline{t}}\right)^{1-2\theta}\|\overline{h}_t\|^2-\|h_{t-1}\|^2\right]\\ &\le\left(1-\frac{1}{\overline{t}}\right)^{1-2\theta}\left(1-\frac{ab}{\overline{t}}\right)^2\|h_{t-1}\|^2\ldots\\ &\quad\ldots-\|h_{t-1}\|^2+3\kappa^2M_\rho^2a^2\overline{t}^{-2\theta}.\end{aligned}\qquad\text{(B-7)}$$

Now, since $t_0\ge\max(2ab,2\epsilon,\epsilon+(2\theta-1)/\epsilon)$ and $\theta\ge 1/2$, we have

$$\begin{aligned}&\log\left[\left(1-\frac{1}{\overline{t}}\right)^{1-2\theta}\left(1-\frac{ab}{\overline{t}}\right)^2\left(1-\frac{\epsilon}{\overline{t}}\right)^{-1}\right]\\ &\quad\le-\frac{\epsilon}{\overline{t}}+\frac{2\theta-1+\epsilon^2}{\overline{t}^2}\le 0.\end{aligned}$$

using $\log(1-x)\le -x$ for all $x\in[0,1]$ and $\log(1-x)\ge -x-x^2$ for all $x\in[0,1/2]$.

Therefore (B-7) implies

$$\Delta_t\le-\frac{\epsilon}{\overline{t}}\|h_{t-1}\|^2+3\kappa^2M_\rho^2a^2\overline{t}^{-2\theta}\le 0.$$

The conclusion follows by conditional Jensen's inequality. ■

Lemma B.5: Assume $t_0^\theta\ge a(\kappa^2+b)$ and $t_0^{1-\theta}\ge b(2+M_tM_\rho^{-1})$; then

$$\|k_t\|_K\le 2\kappa M_\rho ab^{-1}\overline{t}^{1-2\theta}\text{ and }\mathbb{E}_{t-1}[\|k_t\|_K^2]\le 9\gamma_t^2M_\rho^2\kappa^2.$$

Proof: By definition (B-5), using $\|K_{x_t}\|_K\le\kappa$, $\|L_K f_\rho\|_K\le\kappa\|L_K^{1/2}f_\rho\|_K=\kappa\|f_\rho\|_\rho\le\kappa M_\rho$ and Lemma B.3 (A)-(B), we deduce

$$\begin{aligned}\|k_t\|_K&\le\gamma_t\left[\kappa(M_t+2M_\rho)+\|L_t(f_\rho+g_{t-1})\|_K\right]\\ &\le\kappa\gamma_t\left(M_t+2M_\rho+\frac{M_\rho}{\lambda_t}\right)\le\frac{2\kappa\gamma_tM_\rho}{\lambda_t}\\ &=\frac{2\kappa M_\rho ab^{-1}}{\overline{t}^{2\theta-1}},\end{aligned}$$

where we use $t_0^{1-\theta}\ge b(2+M_tM_\rho^{-1})$ in the last inequality.

Now, using (B-3), we obtain

$$\begin{aligned}&\mathbb{E}_{t-1}[\|k_t\|_K^2]\\ &\quad\le 3\gamma_t^2\Big[\mathbb{E}_{t-1}[\|\overline{L}_t h_{t-1}\|_K^2]\ldots\\ &\qquad\ldots+\mathbb{E}_{t-1}[\|y_tK_{x_t}\|_K^2]+\mathbb{E}_{t-1}[\|L_tg_{t-1}\|_K^2]\Big]\\ &\quad\le 3\gamma_t^2[2M_\rho^2\kappa^2+\|g_{t-1}\|_\rho\kappa^2]\le 9\gamma_t^2M_\rho^2\kappa^2.\end{aligned}$$

■

Lemma B.6: For all $t\in\mathbb{N}$, assume $M_t\ge 4\kappa M_\rho ab^{-1}\overline{t}^{1-2\theta}$, $t_0^\theta\ge 2a(\kappa^2+b)$ and $t_0^{1-\theta}\ge b(2+M_tM_\rho^{-1})$; then

$$\|h_t\|_K\le\|\overline{h}_t\|_K.$$

Proof: Assume $h_{t-1}(x_t)\ge M_t$ for instance; the other case is similar. By definition,

$$h_t=\overline{h}_t-\gamma_th_{t-1}(x_t)K_{x_t}$$

so that

$$\begin{aligned}\|h_t\|_K^2&=\|\overline{h}_t\|_K^2-2\gamma_th_{t-1}(x_t)\overline{h}_t(x_t)\ldots\\ &\quad\ldots+\gamma_t^2(h_{t-1}(x_t))^2K(x_t,x_t)\le\|\overline{h}_t\|_K^2\end{aligned}$$

if $\overline{h}_t(x_t)\ge\kappa^2\gamma_th_{t-1}(x_t)/2$.

But, using Lemma B.5,

$$\begin{aligned}\overline{h}_t(x_t)&=(1-\gamma_t\lambda_t)h_{t-1}(x_t)+k_t(x_t)\\ &\ge(1-\gamma_t\lambda_t)h_{t-1}(x_t)-2\kappa M_\rho ab^{-1}\overline{t}^{1-2\theta}\\ &\ge\kappa^2\gamma_th_{t-1}(x_t)/2,\end{aligned}$$

if $2\kappa M_\rho ab^{-1}\overline{t}^{1-2\theta}\le h_{t-1}(x_t)/2\le h_{t-1}(x_t)(1-\gamma_t\lambda_t-\kappa^2\gamma_t/2)$, since the assumption $t_0^\theta\ge 2a(\kappa^2+b)$ implies $1-\gamma_t\lambda_t-\kappa^2\gamma_t/2\ge 1/2$. ■

The following logarithmic upper bound holds under the assumptions $ab-(\theta-1/2)>0$, $\theta\in[1/2,1]$ and t_0 sufficiently large, but we assume $b=a^{-1}$ in its statement, for notational reasons.

Corollary B.7: Assume $\theta\in[1/2,1]$, $b=a^{-1}$, $t_0^\theta\ge 2+8\kappa^2a$ and $t_0^{1-\theta}\ge 4b$. Then, with probability at least $1-\delta$,

$$\begin{aligned}&\sup_{0\le k\le t}\|h_k\|_K(k+t_0+1)^{\theta-1/2}\\ &\quad\le\kappa M_\rho a\left[12at_0^{1/2-\theta}+15\sqrt{\log\overline{t}}\right]\log\frac{2}{\delta}.\end{aligned}$$

Proof: Let us first check that the assumptions of Lemmas B.4, B.5 and B.6 are satisfied, and apply these lemmas: $t_0^\theta \geq 3 + 8\kappa^2 a \geq 2a(\kappa^2 + b)$. Now $\epsilon = ab - (\theta - 1/2) \in [1/2, 1]$, and the hypothesis $t_0 \geq \max(2ab, 2\epsilon, \epsilon + (2\theta - 1)/\epsilon)$ is satisfied as long $t_0 \geq 3$, which is assumed here. We choose $M_t = 4\kappa M_\rho ab^{-1}\overline{t}^{1-2\theta}$; now $t_0^\theta \geq 8\kappa a$ and $t_0^{1-\theta} \geq 4b$ imply $t_0^{1-\theta} \geq b(2 + 4\kappa ab^{-1}t_0^{1-2\theta}) \geq b(2 + M_t M_\rho^{-1})$.

For all $i \in \mathbb{N}$, if $\|h_{i-1}\|_K \geq A(i + t_0)^{1/2-\theta}$, $A := a\kappa M_\rho \sqrt{3/\epsilon}$, then

$$\begin{aligned}\|h_i\|_K &\leq \|\bar{h}_i\|_K, && \text{(Lemma B.6)}\\ &\leq \|\bar{h}_i\|_K + \mathbb{E}_{i-1}(\|\bar{h}_i\|_K) - \mathbb{E}_{i-1}(\|\bar{h}_i\|_K)\\ &\leq \left(1 - \frac{1}{i + t_0}\right)^{\theta-1/2} \|h_{i-1}\|_K + \epsilon_i,\\ & && \text{(Lemma B.4)}\end{aligned}$$

where

$$\epsilon_i := \|\overline{h}_i\|_K - \mathbb{E}_{i-1}(\|\overline{h}_i\|_K)$$

satisfies

$$\begin{aligned}\|\epsilon_i\|_K &\leq 4\kappa M_\rho ab^{-1}(i + t_0)^{1-2\theta}, \text{ and}\\ \mathbb{E}_{i-1}[\|\epsilon_i\|^2] &\leq 9\gamma_i^2 M_\rho^2 \kappa^2.\end{aligned}$$

Let, for all $i \in \mathbb{N}$,

$$\eta_i := \sum_{k=1}^{i} \epsilon_k (k + t_0)^{\theta-1/2} \mathbf{1}_{\{\|h_{k-1}\|_K \geq A(k+t_0)^{1/2-\theta}\}}.$$

Fix $t \in \mathbb{N}$. For all $0 \leq i < t$, $\|\eta_{i+1} - \eta_i\| \leq 4\kappa M_\rho a^2 t_0^{1/2-\theta}$, and

$$\begin{aligned}\sum_{k=1}^{t} \mathbb{E}_{k-1}\|\eta_k\|^2 &\leq 9\kappa^2 M_\rho^2 a^2 \sum_{k=1}^{t} (k + t_0)^{-1}\\ &\leq 9\kappa^2 M_\rho^2 a^2 \log(1 + t/t_0).\end{aligned}$$

Let

$$\Delta := \left\{ \begin{aligned}&\sup_{1\leq i\leq t} \|\eta_i\| \ldots\\ &\ldots \leq 2\kappa M_\rho a \left[\frac{4at_0^{1/2-\theta}}{3} + 3\sqrt{\log\left(1 + \frac{t}{t_0}\right)}\right] \log\frac{2}{\delta}\end{aligned} \right\}.$$

By Proposition A.3, $P(\Delta) \geq 1 - \delta$.

Now assume Δ holds. Let, for all $k \in \mathbb{N}$,

$$x_k := \|h_k\|_K (k + t_0 + 1)^{\theta-1/2}.$$

For all $k \leq t$, let

$$m := \max\{j \leq k : \|h_j\|_K < A(j + t_0 + 1)^{1/2-\theta}\}.$$

If $m < k$, then

$$\begin{aligned}x_{m+1} &\leq [A(m + t_0 + 1)^{1/2-\theta} + 2\kappa M_\rho a^2 \ldots\\ &\quad \ldots \cdot (m + t_0 + 1)^{1-2\theta}](m + t_0 + 2)^{\theta-1/2}\\ &\leq \frac{\sqrt{5}}{2}[A + 2\kappa M_\rho a^2 t_0^{1/2-\theta}]\\ &\leq \frac{\sqrt{5}}{2}[\sqrt{6}a\kappa M_\rho + 2\kappa M_\rho a^2 t_0^{1/2-\theta}];\end{aligned}$$

the second inequality comes from $[(m + t_0 + 2)/(m + t_0 + 1)]^{\theta-1/2} \leq \sqrt{5/4}$, since $t_0 \geq 3$.

On the other hand it is easy to prove by induction that, for all $k \leq t$,

$$x_k \leq x_{m+1} + \eta_k - \eta_{m+1}$$

and, therefore,

$$\begin{aligned}x_k &\leq \kappa M_\rho a \left[\left(\sqrt{5} + \frac{16}{3}\right) a t_0^{1/2-\theta} + \frac{\sqrt{30}}{2} \ldots\right.\\ &\quad \left.\ldots + 12\sqrt{\log\left(1 + \frac{t}{t_0}\right)}\right] \log\frac{2}{\delta}\\ &\leq 12\kappa M_\rho a \left[a t_0^{1/2-\theta} + \frac{1}{4} + \sqrt{\log\left(1 + \frac{k}{t_0}\right)}\right] \log\frac{2}{\delta}\\ &\leq \kappa M_\rho a \left[12 a t_0^{1/2-\theta} + 15\sqrt{\log t}\right] \log\frac{2}{\delta},\end{aligned}$$

using in the last inequality that, for all $t \geq 1$ and $t_0 \geq 2$, $\frac{1}{4} + \sqrt{\log(1 + t/t_0)} \leq \frac{1}{4} + \sqrt{\log(t + t_0)} \leq \frac{5}{4}\sqrt{\log(t + t_0)}$. ■

APPENDIX C
PROOF OF RESULTS OF SECTION III-B

Proof of Lemma III.8: Assume $t \geq j_0$. The spectral Theorem for compact operators implies that there is an orthonormal basis of $\mathscr{W}$ consisting of eigenvectors of A_t, so that, if $(\alpha_{t,k})_{k\in\mathbb{N}}$ are the eigenvalues of A_t, then

$$\begin{aligned}\|A_t^{-1}\|^{-1} &= \min_{k\geq 1} \alpha_{t,k} \geq \underline{\alpha}_t,\\ \|I - \gamma_t A_t\| &= \max_{k\geq 1}(1 - \gamma_t \alpha_{t,k}) \geq 0,\end{aligned}$$

where we use that, for all $k \in \mathbb{N}$, $\gamma_t \alpha_{t,k} \leq \gamma_t \overline{\alpha}_t \leq 1$.

But $\min_{k\geq 1} \alpha_{t,k} \geq \underline{\alpha}_t$ implies $\max_{k\geq 1}(1 - \gamma_t \alpha_{t,k}) \leq 1 - \gamma_t \underline{\alpha}_t$, thus (A).

The last claim follows from the inequality

$$\begin{aligned}\prod_{i=j}^{t}\left(1 - \frac{c}{i + t_0}\right) &\leq \exp\left(-\sum_{i=j}^{t} \frac{c}{i + t_0}\right)\\ &\leq \exp\left(-c \log\left(\frac{\overline{t} + 1}{j + t_0}\right)\right)\\ &= \left(\frac{j + t_0}{\overline{t} + 1}\right)^c.\end{aligned}$$

■

Proof of Theorem III.5: First, $\gamma_t \underline{\alpha}_t \to 0$ implies that there exists $j_0 \in \mathbb{N}$ such that $\gamma_t \underline{\alpha}_t \leq 1$ for all $t \geq j_0$. Hence Lemma III.8 (B) applies, so that

$$\|\Pi_j^t\| \leq \prod_{i=j}^{t}(1 - \gamma_i \underline{\alpha}_i). \tag{C-1}$$

Let us use the reversed martingale decomposition of $r_\cdot$, from times j_0 to t:

$$\|r_t\| \leq \mathscr{E}_{init}(t) + \mathscr{E}_{samp}(t) + \mathscr{E}_{drift}(t),$$

where

$$\begin{aligned}\mathscr{E}_{init}(t) &:= \|\Pi^t_{j_0+1} r_{j_0}\|,\\ \mathscr{E}_{samp}(t) &:= \|\sum_{j=j_0+1}^{t} \gamma_j \Pi^t_{j+1}(A_j \bar{w}_j - b_j)\|,\\ \mathscr{E}_{drift}(t) &:= \|\sum_{j=j_0+1}^{t} \Pi^t_j \Delta_j\|.\end{aligned}$$

Now, by (C-1),

$$\mathbb{E}(\mathscr{E}_{init}(t)^2) \leq \exp\left(-2\sum_{i=j_0+1}^{t} \gamma_i \underline{\alpha}_i\right)\mathbb{E}(\|r_{j_0}\|^2) \to_{t\to\infty} 0$$

since $\sum_t \gamma_t \underline{\alpha}_t = \infty$, and

$$\mathscr{E}_{drift}(t) \leq \sum_{j=j_0+1}^{t} \|\Delta_j\| \prod_{i=j}^{t}(1-\gamma_i\underline{\alpha}_i) \to_{t\to\infty} 0$$

by assumption (C). Now consider the sample error. Using the independence of $(z_t)_{t\in\mathbb{N}}$ (see Remark III.2),

$$\begin{aligned}\mathbb{E}(\mathscr{E}_{samp}(t)^2) &= \mathbb{E}\left\|\sum_{j=j_0+1}^{t} \gamma_j \Pi^t_{j+1}(A_j\bar{w}_j - b_j)\right\|^2\\ &= \sum_{j=j_0+1}^{t} \gamma_j^2 \mathbb{E}\|\Pi^t_{j+1}(A_j\bar{w}_j - b_j)\|^2\\ &\leq C\sum_{j=j_0+1}^{t} \gamma_j^2 \prod_{i=j+1}^{t}(1-\gamma_i\underline{\alpha}_i)^2,\end{aligned}$$

where $C := \sup_{t\in\mathbb{N}} \mathbb{E}\|A_t\bar{w}_t - b_t\|^2 < \infty$ by assumption. This completes the proof, using (B). ■

Proof of Lemma III.6: Let $\epsilon > 0$. The assumptions $\limsup_{t\to\infty} a_t/b_t = 0$ and $b_t \to_{t\to\infty} 0$ imply that there exists $N \in \mathbb{N}$ such that $a_t \leq \epsilon b_t/2$ and $b_t \leq 1$ for all $t > N$. On the other hand, $\sum_{t\in\mathbb{N}} b_t = \infty$ implies that there exists $N_1 \in \mathbb{N}$ such that, for all $n \geq N_1$,

$$\sum_{k=1}^{N} a_k \prod_{i=k+1}^{n}(1-b_i) < \frac{\epsilon}{2}.$$

Now

$$\sum_{k=N+1}^{n} a_k \prod_{i=k+1}^{n}(1-b_i) \leq \frac{\epsilon}{2}\sum_{k=N+1}^{n} b_k \prod_{i=k+1}^{n}(1-b_i),$$

and we can write the right-hand side of this last inequality as a telescopic sum, i.e.

$$\begin{aligned}\sum_{k=N+1}^{n} b_k \prod_{i=k+1}^{n}(1-b_i) &= \sum_{k=N+1}^{n}[1-(1-b_k)]\prod_{i=k+1}^{n}(1-b_i)\\ &= 1-\prod_{i=N+1}^{n}(1-b_i) \leq 1,\end{aligned}$$

which enables us to conclude. ■

References

[1] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright, "Information-theoretic lower bounds on the oracle complexity of convex optimization," *IEEE Trans. Inform. Theory*, vol. 58, no. 5, pp. 3235–3249, May 2012.

[2] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, no. 3, pp. 337–404, 1950.

[3] F. R. Bach and E. Moulines, "Non-asymptotic analysis of stochastic approximation algorithms for machine learning," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2011, pp. 451–459.

[4] M. Benaim, "Dynamics of stochastic approximation algorithms," in *Séminaire de Probabilités, XXXIII* (Lecture notes in Mathematics). Berlin, Germany: Springer-Verlag, pp. 1–68.

[5] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Boston, MA, USA: Kluwer, 2004.

[6] D. P. Bertsekas, "Incremental proximal methods for large scale convex optimization," *Math. Program.*, vol. 129, no. 2, pp. 163–195, Oct. 2011.

[7] A. Caponnetto and E. De Vito, "Optimal rates for regularized least squares algorithm," *Found. Comput. Math.*, vol. 7, no. 3, pp. 331–368, Jul. 2007.

[8] C. Carmeli, E. De Vito, and A. Toigo, "Vector valued reproducing kernel hilbert spaces integrable, functions and mercer theorem," *Anal. Appl.*, vol. 4, pp. 377–408, 2006.

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

[10] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA, USA: SIAM, 1992.

[11] M. Duflo, *Algorithmes Stochastiques*. Berlin, Germany: Springer-Verlag, 1996.

[12] B. Efron, I. Johnstone, T. Hastie, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–451, 2004.

[13] H. W. Engl, M. Hanke, and A. Neubauer, "Regularization of inverse problems," in *Mathematics and its Applications*. Boston, MA, USA: Kluwer, 1996.

[14] T. Evgeniou, M. Pontil, and T. Poggio, "Regularized networks and support vector machines," *Adv. Comput. Math.*, vol. 13, no. 1, pp. 1–50, Apr. 2000.

[15] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York, NY, USA: Springer-Verlag, 2002.

[16] R. P. Halmos and V. S. Sunder, *Bounded Integral Operators in L^2 Spaces* (Results in Mathematics and Related Areas), vol. 96. Berlin, Germany: Springer-Verlag, 1978.

[17] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Ann. Math. Statist.*, vol. 23, no. 3, pp. 462–466. 1952.

[18] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2165–2176, Aug. 2004.

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

[20] M. Loéve, "Fonctions alèatoires du second ordre," in *Processus Stochastiques et Mouvement Brownien*, P. Lèvy, Ed. Paris, France: Gauthier-Villars, 1948.

[21] J. Neveu, *Discrete-Parameter Martingales*. Amsterdam, The Netherlands: North Holland, 1975.

[22] E. Parzen, "An approach to time series analysis," *Ann. Math. Statist.*, vol. 32, no. 4, pp. 951–989, 1961.

[23] I. Pinelis, "Optimum bounds for the distributions of martingales in Banach spaces," *Ann. Probab.*, vol. 22, no. 4, pp. 1679–1706, 1994.

[24] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.

[25] S. Rosset and J. Zhu, "Piecewise linear regularized solution paths," *Ann. Statist.*, vol. 35, no. 3, pp. 1012–1030, 2007.

[26] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[27] S. Smale and Y. Yao, "Online learning algorithms," *Found. Comput. Math.*, vol. 6, no. 2, pp. 145–170, 2006.

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

[29] S. Smale and D.-X. Zhou, "Learning theory estimates via integral operators and their approximations," *Constructive Approx.*, vol. 26, no. 2, pp. 153–172, 2007.

[30] S. Smale and D.-X. Zhou, “Online learning with Markov sampling,” *Anal. Appl.*, vol. 7, no. 1, pp. 87–113, Jan. 2009.

[31] P. Tarrès and Y. Yao, “Learning as stochastic approximation of regularisation paths,” in *Proc. Learn. Theory Approx. Workshop, Math. Forschungsinstitut Oberwolfach*, 2008.

[32] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Roy. Statist. Soc, Ser. B.*, vol. 58, no. 1, pp. 267–288, 1996.

[33] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA, USA: SIAM, 1990.

[34] D. Williams, *Probability with Martingales*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[35] Y. Yao, “A dynamic theory of learning,” Ph.D dissertation, Dept. Math., Univ. Calfornia, Berkeley, CA, USA, 2006.

[36] Y. Yao, “On complexity issue of online learning algorithms,” *IEEE Trans. Inform. Theory*, vol. 56, no. 12, pp. 6470–6481, Dec. 2010.

[37] Y. Ying and M. Pontil, “Online gradient descent learning algorithms,” *Found. Comput. Math.*, vol. 5, no. 5, pp. 561–596, 2008.

[38] D.-X. Zhou, “Density problem and approximation error in learning theory,” in *Proc. Abstract Appl. Anal.*, vol. 2013, 2013, Article ID 715683, doi:10.1155/2013/715683.

Pierre Tarrès received his MSc in mathematics from the Ecole Normale Supérieure, Paris, in 1998. In october 2001, he received his PhD from the Ecole Normale Supérieure, Cachan, with a thesis on stochastic approximation algorithms and reinforced random walks. He was appointed Fellow of the Centre National de la Recherche Scientifique in 2002, and is an Associate Professor in the Mathematical Institute, University of Oxford, since 2005. His scientific interests encompass self-interaction and learning in random structures, with a particular emphasis on the study of the long-term behavior of self-interacting random walks.

Yuan Yao received the B.S.E and M.S.E in control engineering both from Harbin Institute of Technology, China, in 1996 and 1998, respectively, M.Phil in mathematics from City University of Hong Kong in 2002, and Ph.D. in mathematics from the University of California, Berkeley, in 2006. Since then he has been with Stanford University and in 2009, he joined the School of Mathematical Sciences, Peking University, Beijing, China, as a professor of statistics in the Hundred Talents Program. His current research interests include topological and geometric methods for high dimensional data analysis and statistical machine learning, with applications in computational biology, computer vision, and information retrieval.