A Dynamic Theory of Learning

by

Yuan Yao

B.S.E. (Harbin Institute of Technology) 1996
M.S.E. (Harbin Institute of Technology) 1998
M.Phil. (City University of Hong Kong) 2002

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Mathematics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Stephen Smale, Chair
Professor Steven N. Evans
Professor Peter Bartlett

Fall 2006

The dissertation of Yuan Yao is approved:

Chair Date

Date

Date

University of California, Berkeley

Fall 2006

A Dynamic Theory of Learning

Abstract

A Dynamic Theory of Learning

by

Yuan Yao

Doctor of Philosophy in Mathematics

University of California, Berkeley

Professor Stephen Smale, Chair

In this thesis, a dynamic theory of learning, also called *online learning* in computer science, is presented as stochastic approximations of the regression function from reproducing kernel Hilbert spaces (RKHS). We show by probabilistic upper bounds that these algorithms may achieve the same convergence rates as “batch learning”, and thus asymptotically reach the optimal rates in some senses.

Professor Stephen Smale
Dissertation Committee Chair

To my mum,

Yuhua Chang,

and,

my supervisor,

Steve Smale

Contents

Acknowledgments

I am greatly indebted to Professor Steve Smale, without whom this thesis never comes into reality. In my Odyssey to pursue Mathematics as an Engineering student, from Hong Kong to Berkeley and Chicago, he shaped my life as an applied mathematician by his insights, encouragement, collaboration and friendship.

I would like to extend special thanks to Professor Partha Niyogi, who has offered me a great deal of support during this work. Our interaction has had a significant influence on my research. Thanks to Professor Peter Bartlett, for his serving in my qualifying exam and thesis committee, and for various discussions on research. Professor Steve Evans also served as a thesis reader, who introduced me to an interesting topic of constructing reproducing kernel Hilbert spaces (RKHS) from Markov chains which becomes one of my future direction.

Much of this thesis work was done at Toyota Technological Institute at Chicago, at the campus of University of Chicago. I want to thank TTI-C and U of Chicago for their facilities and hospitality. In particular, I thank Adam Kalai for helpful discussions on online learning, David McAllester for his endeavor to reduce the lost in translations. Thanks to Carol, Don and Katherine, for administrative and technical supports. To Minh for deepening my understanding on RKHS and to Andrea for his collaboration and friendship.

Thanks to Jiangang Yao, Yong Wang, and all my Chinese friends at Berkeley and Chicago, for their friendship and giving me a home away home.

This work was partially supported by the National Science Foundation under the grant 0325113.

Chapter 1

Introduction

1.1 Online Learning Algorithms

Supervised learning, or *learning from examples*, is to find a function in a hypothesis space $\mathscr{H}$, which associates an input $x \in \mathscr{X}$ to an output $y \in \mathscr{Y}$, by drawing examples $(x_t, y_t)_{t\in\mathbb{N}}$ from $\mathscr{Z} := \mathscr{X} \times \mathscr{Y}$. By *online learning*, we mean a sequential decision process $(f_t)_{t\in\mathbb{N}}$ in the hypothesis space, where each f_t is decided by the current observation $z_t = (x_t, y_t)$ and f_{t-1} which only depends on previous examples, i.e. $f_t = T_t(f_{t-1}, z_t)$. As a contrast, *batch learning* refers to a decision utilizing the whole set of examples available at time t, i.e. a mapping $\mathscr{Z}^t \ni (z_i)_1^t \mapsto f \in \mathscr{H}$ (e.g. see [Vapnik 1998; Cucker and Smale 2002b]). Examples of such online learning algorithms include *Perceptrons* [Rosenblatt 1958] and *Adaline* (or *Widrow-Hoff algorithm*) [Widrow and Hoff 1960].

Why shall we study online learning algorithms?

- The sampling process could be dependent, e.g. Markov sampling where the examples are drawn from a Markov chain or competitive sampling where the environment plays a game against the learner;

- The computational cost of online learning, linear as $O(t)$ if without considering the evaluation cost in $f(x)$, is typically lower than "batch learning".

Although online learning could deal with dependent sampling, however, for the ease of comparison with existing "batch learning" results, in this thesis we take the standard assumption in statistical learning that the sample sequence $(z_t)_{t\in\mathbb{N}}$ is of independent and identically distributed (i.i.d.) according to a probability measure ρ on $\mathscr{Z}$. Our object is to minimize over some hypothesis space $\mathscr{H}$ the following functional

$$V(f) = \int_{\mathscr{Z}} l(f,z)d\rho \tag{1.1}$$

where $l : \mathscr{H} \times \mathscr{Z} \to \mathbb{R}$ is a *loss function*, which measures the cost of decision $f \in \mathscr{H}$ against the observation $z \in \mathscr{Z}$. The loss function is often assumed to be convex in $f \in \mathscr{H}$ such that convex optimization technique can be used to solve the problem. Typical examples of loss functions include

- *Hinge loss*: $l(f,z) = (1 - yf(x))_+$ where $(x)_+ := \max(x,0)$ and $y \in \mathscr{Y} = \{\pm 1\}$;
- *Square loss*: $l(f,z) = (f(x) - y)^2$ where $y \in \mathscr{Y} = \mathbb{R}$.

The hinge loss is used in Support Vector Machines for classifications [e.g. Cristianini and Shawe-Taylor 2000] and the square loss leads to Least Mean Square method, such as Adaline and its variations [Widrow and Lehr 1990].

The online learning algorithms considered in this thesis are constructed from stochastic gradient descent algorithms[1]

$$f_t = f_{t-1} - \gamma_t \nabla V(f_{t-1}) + \gamma_t \epsilon_t \qquad \text{for some } f_0 \in \mathscr{H}, \tag{1.2}$$

where the step size $\gamma_t > 0$ and $\epsilon_t \in \mathscr{H}$ is a random perturbation of zero mean. For example, the Robbins-Monro procedure takes $\epsilon_t = \nabla V(f_{t-1}) - \nabla_f l(f_{t-1}, z_t)$ [Robbins and Monro 1951; Kiefer and Wolfowitz 1952]. For a wider background on stochastic algorithms, see [Duflo 1996].

Among a variety of choices on the loss l and the hypothesis space $\mathscr{H}$, it leads to a simple structure but deeper understanding by selecting the square loss $l(f,z) = (f(x) - y)^2$ and the hypothesis space $\mathscr{H} = \mathscr{H}_K$, the reproducing kernel Hilbert space (RKHS) associated with a Mercer kernel K (see Appendix C).

With the square loss, it is well-known that the minimizer of (1.1) in $\mathscr{L}^2_{\rho_\mathscr{X}}$, the Hilbert space of square integrable functions with respect to $\rho_\mathscr{X}$ (i.e. the marginal probability measure on $\mathscr{X}$), is the *regression function*,

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

i.e. the conditional expectation of y given x.

With a RKHS, one could have a Hilbert space which is large enough as dense in $\mathscr{L}^2_{\rho_\mathscr{X}}$, and also small enough with well-controlled sampling properties (reproducing property). Now instead of minimizing (1.1), consider the following *Tikhonov* regularized least square problem

$$\min_{f \in \mathscr{H}_K} V_\lambda(f) = \int_{\mathscr{X} \times \mathscr{Y}} (y - f(x))^2 d\rho + \lambda \|f\|_K^2 \tag{1.3}$$

[1] When the loss function is non-differentiable, one may use subgradient method in stead.

where $\| \ \|_K$ denotes the norm in $\mathscr{H}_K$. Such a choice avoids the estimation of covering numbers of $\mathscr{H}$, which is difficult in most cases [Cucker and Smale 2002b; Zhou 2003]; it provides a simple estimate of optimal upper bounds asymptotically meeting lower bounds [Caponnetto and De Vito 2006; Smale and Zhou 2006a]; it bridges over the linear inverse problem toward other regularization schemes [Engl, Hanke, and Neubauer 1996; De Vito, Rosasco, Caponnetto, Giovannini, and Odone 2004]; and more interestingly, it takes an especially simple form in online learning algorithms [Smale and Yao 2006].

In fact the gradient of $l(f,z)$ with respect to f, $\nabla_f l(f,z) : \mathscr{H}_K \to \mathscr{H}_K$, is simply

$$\nabla_f l(f,z) = 2(f(x) - y)K_x + 2\lambda f.$$

With this observation, the stochastic gradient descent algorithm (1.2) with $V_\lambda(f)$ becomes

$$f_t = f_{t-1} - \gamma_t[(f_{t-1}(x_t) - y_t)K_{x_t} + \lambda f_{t-1}], \qquad \text{for some } f_0 \in \mathscr{H}_K, \text{ e.g. } f_0 = 0 \tag{1.4}$$

where

(A) for each t, (x_t, y_t) is independent and identically distributed (i.i.d.) according to ρ;

(B) the step size $\gamma_t > 0$ and $\sum_t \gamma_t = \infty$;

(C) the regularization parameter $\lambda > 0$.

The equilibrium for this algorithm is the unique minimizer f^*_λ for (1.3), satisfying the following linear equation,

$$(L_K + \lambda I)f = L_K f_\rho, \tag{1.5}$$

where the integral operator $L_K : \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{L}^2_{\rho_\mathscr{X}}$ is defined by $L_K(f) = \int_\mathscr{X} K(x,t)f(t)d\rho_\mathscr{X}$ (see [Cucker and Smale 2002b] or Appendix C for detail). Since $L_K + \lambda I$ ($\lambda > 0$) is invertible, we may write $f^*_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho$, which is called in this thesis as the *Tikhonov regularization equilibrium*.

λ_t	γ_t	Convergence	Reference
$\lambda_t = \lambda > 0$	$\gamma_t \to 0$	$f_t \to f_\lambda$	[Smale and Yao 2006; Yao 2006]
$\lambda_t \to 0$	$\gamma_t \to 0$	$f_t \to f_\rho$	[Tarrès and Yao 2006]
$\lambda_t = 0$	$\gamma_t \to 0$	$f_t \to f_\rho$	[Ying and Pontil 2006]

Table 1.1: Convergence and Regularization Schemes

Part I contributes to the study of the convergence of algorithm (2.2) and its variations to the Tikhonov regularization equilibrium f^*_λ. In particular, it shows that strengthened by an averaging process, we may have online learning algorithms competitive with “batch learning” algorithms with the same convergence rate.

In learning theory a fundamental goal is to approximate the regression function f_ρ, instead of f^*_λ. For this purpose, algorithm (1.4) can be extended to,

$$f_t = f_{t-1} - \gamma_t[(f_{t-1}(x_t) - y_t)K_{x_t} + \lambda_t f_{t-1}], \qquad \text{for some } f_0 \in \mathscr{H}_K, \text{ e.g. } f_0 = 0 \tag{1.6}$$

where the regularization parameter $\lambda_t > 0$ and $\lambda_t \downarrow 0$.

Part II contributes to the study of the convergence of this algorithm to f_ρ by following the Tikhonov regularization path $(f^*_{\lambda_t})_{t\in\mathbb{N}}$. In particular the asymptotic convergence rates in the probabilistic upper bounds are competitive with “batch learning” under the same prior assumptions on f_ρ, and actually are optimal in some senses.

As a summary on the influence of the choice of λ_t and γ_t, in Table 1.1 we list the convergence results and the recent literature studying these choices.

1.2 Notation and Assumptions

Let $\mathbb{N}$ be the set of natural numbers and $\mathbb{Z}_+ = 0 \cup \mathbb{N}$ be the set of nonnegative integers. Let $\mathscr{X} \subset \mathbb{R}^n$ be closed, $\mathscr{Y} = \mathbb{R}$, and $\mathscr{Z} = \mathscr{X} \times \mathscr{Y}$. Let ρ be a probability measure on $\mathscr{Z}$, $\rho_{\mathscr{X}}$ and $\rho_{\mathscr{Y}|x}$ be the marginal and the conditional probability measure induced by ρ, respectively. Given a random sequence $(z_t)_{t\in\mathbb{N}}$ drawn according to ρ, let $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathbb{N}} \subseteq X \times \mathbb{R}$ be the filtration $\mathcal{F}_t = \sigma\{(z_k) \;:\; 1 \leq k \leq t\}$. By $\mathbb{E}_0 = \mathbb{E}$ and $\mathbb{E}_t = \mathbb{E}[\ |\mathcal{F}_t]$ we denote the expectation and conditional expectation w.r.t. $\mathcal{F}_t$.

Let the function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ be a Mercer kernel and universal such that the reproducing kernel Hilbert space associated with K, $\mathscr{H}_K$, is dense in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$, the Hilbert space of square integrable functions w.r.t. the measure $\rho_{\mathscr{X}}$. Denote by $\langle\ ,\ \rangle_\rho$ the inner product in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ and by $\langle\ ,\ \rangle_K$ the inner product in $\mathscr{H}_K$. There is an isometry $L_K^{1/2} : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{H}_K$ such that $\langle f, g\rangle_\rho = \langle L_K^{1/2} f, L_K^{1/2} g\rangle_K$ for all $f, g \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$, which will be used throughout the thesis. For more details on RKHS and related, see Appendix C.

Throughout this thesis, 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

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

When $n < m$, the product and summation, $\prod_{i=m}^n x_i$ and $\sum_{i=m}^n x_i$, are understood to be 1 and 0, respectively.

Part I

Stochastic Approximation of Regularization Equilibrium

Chapter 2

Main Results

Consider the following *Tikhonov* regularization problem,

$$\min_{f \in \mathscr{H}_K} \int_{\mathscr{X} \times \mathscr{Y}} (f(x) - y)^2 d\rho + \lambda \|f\|_K^2, \qquad \lambda > 0. \tag{2.1}$$

In this setting there exists a unique minimizer f_λ^*, satisfying the following linear equation [see e.g. Cucker and Smale 2002b],

$$(L_K + \lambda) f = L_K f_\rho. \tag{2.2}$$

For $\lambda > 0$, $L_K + \lambda : \mathscr{H}_K \to \mathscr{H}_K$ is an isomorphism, whence $f_\lambda^* = (L_K + \lambda)^{-1} L_K f_\rho$, which will be called as *Tikhonov regularization equilibrium* in this thesis.

Given an independent and identically distributed random sequence $(x_t, y_t)_{t \in \mathbb{N}}$ drawn from ρ, consider the following $\mathcal{F}_t$-adapted sequence

$$f_t = f_{t-1} - \gamma_t((f_{t-1}(x_t) - y_t) K_{x_t} + \lambda f_{t-1}), \qquad \text{for some } f_0 \in \mathscr{H}_K, \text{ e.g. } f_0 = 0 \tag{2.3}$$

where the step size $\gamma_t > 0$ satisfies $\sum_{t \in \mathbb{N}} \gamma_t = \infty$. Here we consider the step size with a power law decay, $\gamma_t = O(t^{-\theta})$ for some $\theta \in [0, 1)$. In this part, we present some probabilistic upper bounds for the convergence

$$\|f_t - f^*_\lambda\|_K \to 0.$$

The algorithm (2.3) can be regarded as either the stochastic approximation of the gradient descent method for (2.1) [Kiefer and Wolfowitz 1952], or the stochastic approximation of the linear equation (2.2) [Robbins and Monro 1951]. Traditional analysis on stochastic approximations has been focusing on convergence and asymptotic rates. A convergence result often used in applications, known as the Robbins-Siegmund Theorem [Robbins and Siegmund 1971], imposes a condition on the step size that $\sum_t \gamma_t = \infty$ and $\sum_t \gamma_t^2 < \infty$, and leads to the almost sure convergence (with probability one). For the step size chosen here, $\gamma_t = O(t^{-\theta})$, this requires $\theta \in (1/2, 1)$. In this setting, the asymptotic rate has been shown as $O(\gamma_t^{1/2}) = O(t^{-\theta/2})$. Note that the condition $\sum_t \gamma_t = \infty$, is used to "forget" the error caused by initial choices. However the square summable condition, $\sum_t \gamma_t^2 < \infty$, is not necessary for the almost sure convergence. For example in [Duflo 1997] (or see the remarks in [Benaïm 1999]), to ensure the almost sure convergence it is enough that for all $c > 0$,

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

This even justifies the use of $\gamma_t = 1/\log^{1+\epsilon} t$ for some $\epsilon > 0$, which is however not pursued in this thesis. For more background on stochastic approximations, see for example [Duflo 1996; Kushner and Yin 2003], and references therein.

In [Smale and Yao 2006], we present a probabilistic upper bound based on the Markov inequality, that the following holds with probability at least $1 - \delta$ ($\delta \in (0, 1)$)

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

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

In [Yao 2006], we present three new probabilistic upper bounds by using exponential probabilistic inequalities for martingales in Hilbert spaces [Pinelis 1994], all of which lead to almost sure convergence and extend the rate of step size to $\theta \in [0,1)$, at some possible sacrifice of rates on λ.

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

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

This upper bound implies almost sure convergence for all $\theta \in (0,1)$, by changing $1/\delta$ to $\log 1/\delta$. Note that when $\theta = 0$, algorithm (2.3) is often called the *Adaline* or *Widrow-Hoff algorithm* ([Widrow and Hoff 1960], or see Chapter 5 in [Cristianini and Shawe-Taylor 2000]), which is not guaranteed to converge in this setting.

The second upper bound (as Theorem B) is given for the *averaging process* proposed in [Polyak 1990], [Ruppert 1988],

$$\bar{f}_t = \frac{1}{t}\sum_{j=1}^{t} f_j = \bar{f}_{t-1} - \frac{1}{t}(\bar{f}_{t-1} - f_t), \ \bar{f}_1 = f_1, \tag{2.4}$$

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

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

In contrast to "batch learning" case with a rate $O(\lambda^{-1}t^{-1/2})$ [Smale and Zhou 2006a], this upper bound achieves the same fixed rate in t for all $\theta \in [0,1)$, while losing the rate in λ. Note that the recursive representation in the second identity in (2.4) shows that the averaging process is a two-stage online algorithm, without keeping the decision history $(f_k)_{1\leq k\leq t}$.

It is possible to improve the rate in λ using variance-based probabilistic inequalities. In fact using the Pinelis-Bernstein inequality, we obtain the third bound (as Theorem B$^+$) for the averaging process,

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

which holds with probability at least $1-\delta$ ($\delta \in (0,1)$). This bound meets the same rate in batch learning for both regularization parameter λ and sample size t.

In the remaining we will present these three theorems with discussions, whose proofs will be given in the next chapter.

Before the formal statement of the theorems, we define a constant only depending on $\theta \in [0,1)$,

$$D_\theta = 1 + 2^{\frac{\theta}{1-\theta}}\left(1+\Gamma\left(\frac{1}{1-\theta}\right)\right) \geq 1. \tag{2.5}$$

where $\Gamma : \mathbb{R} \to \mathbb{R}_+$ is the gamma function (see Appendix B).

2.1 An Exponential Probabilistic Upper Bound: Theorem A

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

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

where

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

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

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

Here $\alpha = \lambda/(\lambda+\kappa^2)$ *and* $C_{\rho,\theta,K} = 16\sqrt{D_\theta}\kappa^{2-\theta}M_\rho(p+1)^{1/2(1-\theta)}$.

The proof of Theorem A will be given in the next chapter as a corollary of Theorem 3.2.

Remark 2.1. The second inequality is equivalent to

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

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

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

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

Remark 2.2. Note that when $\theta = 0$, the *Widrow-Hoff* algorithm [Widrow and Hoff 1960] can't ensure its convergence by this upper bound. However, it can be combined with the averaging process to achieve a convergence rate of $O(t^{-1/2})$, which will be discussed in the next subsection.

2.2 Averaging Process: Theorem B and B^+

It is natural to consider the average of the ensemble $\{f_1, \ldots, f_t\}$ up to time t, which might improve the convergence rate since by intuition averaging may reduce variance. In stochastic approximation, this acceleration by averaging was firstly observed independently by [Ruppert 1988] and [Polyak 1990] (or see [Polyak and Juditsky 1992]) based on asymptotic analysis; recently this phenomenon has also been noticed in learning theory society (see, e.g., [Cesa-Bianchi, Conconi, and Gentile 2004]). A recent result [Konda and Tsitsiklis 2004] studies this averaging process in a more general framework of two-time-scale linear stochastic approximations with asymptotic analysis. Below we show a probabilistic upper bound with a fixed rate $O(t^{-1/2})$ for all $\theta \in [0,1)$.

It should be noted that to implement the averaging process, we don't need to keep all the historical hypothesis f_k $(1 \leq k \leq t)$. In fact the update formula in (2.4) runs $\bar{f}_t$ in an online way parallel to f_t.

Theorem B. *Under the same condition of Theorem A, the following holds for all* $t \in \mathbb{N}$*,*

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

where

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

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

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

Here $C_1 = (p+1)\kappa^2 D_\theta \|f^*_\lambda\|_K$ *and* $C_2 = 2^{7/2+\theta}(p+1)\kappa^3 D_\theta M_\rho$*.*

The proof of Theorem B will be given in the next chapter as a corollary of Theorem 3.4.

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

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

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

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

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

Theorem B$^+$**.** *Under the same condition of Theorem A, the following holds for all* $t \in \mathbb{N}$*,*

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

where

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

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

$$\mathscr{E}_{samp}(t) \leq \frac{2^{2+\theta}(p+1)\kappa^3 D_\theta M_\rho}{3}\left(\frac{1}{\lambda^2 t}\right)\log\frac{2}{\delta} + 2^{1+\theta} D_\theta \sigma_\lambda \left(\frac{1}{\lambda}\right)\sqrt{\frac{1}{t}}\log\frac{2}{\delta}.$$

Remark 2.4. Typical choices of λ such as $\lambda = t^{-2r/(2r+1)}$ makes $\lambda\sqrt{t} \to \infty$, whence the rate $\mathscr{E}_{samp}(t) \sim O(\lambda^{-1}t^{-1/2}\log 1/\delta)$, the same rate as in batch learning.

Remark 2.5. Proposition 3.13-3 gives an estimate on σ_λ,

$$\sigma_\lambda \leq \sqrt{5(p+1)}\kappa M_\rho.$$

Remark 2.6. As in Remark 3.6, using the Markov inequality we may obtain

$$\mathscr{E}_{samp}(t) \leq 2^{\theta} D_{\theta} \sigma_{\lambda} \left(\frac{1}{\lambda}\right) \sqrt{\frac{1}{t}} \leq 2^{\theta} \sqrt{5(p+1)} \kappa D_{\theta} M_{\rho} \left(\frac{1}{\lambda}\right) \sqrt{\frac{1}{t}}.$$

In particular, if $\sigma_{\lambda} = 0$, the initial error gives

$$\|f_t - f_{\lambda}^{*}\|_K \leq (p+1)\kappa^2 D_{\theta} \|f_{\lambda}^{*}\|_K \left(\frac{1}{\lambda t}\right),$$

a upper bound for deterministic gradient descent algorithm.

2.3 Comparison with “Batch Learning” Results

Given a sample $\mathbf{z} = \{(x_i, y_i) : i = 1, \ldots, t\}$, “batch learning” means solving the following *regularized least square* problem (see, e.g., [Evgeniou, Pontil, and Poggio 1999], [Cucker and Smale 2002b])

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

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

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

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

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

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

A probabilistic upper bound for $\|f_{\lambda,\mathbf{z}} - f_{\lambda}^{*}\|_K$ is given in [Cucker and Smale 2002a], and this has been substantially improved by [De Vito, Caponnetto, and Rosasco 2004] using also some ideas from [Bousquet and Elisseeff 2002]. Moreover, [Zhang 2003] gives error bounds expressed in a different form. A recent result (Theorem 1 in [Smale and Zhou 2006a]) shows that,

Theorem 2.7. *With probability at least* $1-\delta$ *(*$\delta \in (0,1)$*) there holds*

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

Remark 2.8. A recent result [Caponnetto and De Vito 2006] shows that the rate $O(\lambda^{-1}t^{-1/2})$ is near-optimal in the sense that it leads to a rate asymptotically meeting the minimax lower bound in a weaker measure of convergence. Theorem B tells us that the averaging process achieves $O(\lambda^{-2}t^{-1/2})$, which is worse than batch learning in λ. Theorem B$^+$ improves this to $O(\lambda^{-1}t^{-1/2})$, the same rate as in batch learning.

Chapter 3

Linear Stochastic Approximation in Hilbert Spaces

In this chapter we study a more general problem, stochastic approximation of linear equations in Hilbert spaces. Some general upper bounds are given and they lead to the main theorems (A, B and B$^+$) in a special case.

Let W be a Hilbert space, $A(z) : W \to W$ a random positive operator and $b(z) \in W$ a random vector, both depending on $z \in \mathscr{Z}$. Denote their expectations by $\bar{A} = \mathbb{E}_z[A(z)]$ and $\bar{b} = \mathbb{E}_z[b(z)]$. Consider the following linear equation

$$\bar{A}w = \bar{b}, \tag{3.1}$$

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

In the sequel, we assume that almost surely,

Finiteness Condition. (A) $\underline{\alpha} I \leq A(z) \leq \overline{\alpha} I$ $(0 < \underline{\alpha} \leq \overline{\alpha} < \infty)$ and let $\alpha = \underline{\alpha}/\overline{\alpha} \in (0, 1]$;

(B) $\|b(z)\| \leq \beta < \infty$;

(C) $\mathbb{E}\|A(z)w^* - b(z)\|^2 = \sigma^2 < \infty$.

Remark 3.1. Condition A says that the random operator $A(z)$ has eigenvalues distributed within $[\underline{\alpha}, \overline{\alpha}]$. In particular, this puts a lower bound $\underline{\alpha} > 0$ on the eigenvalues, whence it is important to consider the condition number for the operator family $\{A(z) : z \in \mathscr{Z}\}$, $1/\alpha = \overline{\alpha}/\underline{\alpha}$, which controls the complexity of the algorithm. We shall see this point soon in the upper bounds.

Given an independent and identically distributed (i.i.d.) random sequence $(z_t)_{t\in\mathbb{N}}$, define a sequence $\{w_t\}_{t\in\mathbb{Z}_+}$ as successive stochastic approximations [Robbins and Monro 1951] of w^*,

$$w_t = w_{t-1} - \gamma_t(A_t w_{t-1} - b_t), \qquad \text{for some } w_0 \in W \tag{3.2}$$

where $A_t = A(z_t)$, $b_t = B(z_t) \in W$ and $\gamma_t = 1/\overline{\alpha}t^\theta$ for some $\theta \in [0, 1)$. It is clear that w_t is $\mathscr{W}$-valued random variable depending on $(z_k)_1^t$.

We also consider the averaging sequence

$$\bar{w}_t = \frac{1}{t}\sum_{j=1}^{t} w_j = \bar{w}_{t-1} - \frac{1}{t}(\bar{w}_{t-1} - w_t). \tag{3.3}$$

Note that the recursive representation in the second identity makes this averaging process an online algorithm, without memorizing previous decisions $(w_k)_{k=1}^t$.

Our purpose in this chapter is to give upper bounds for the *remainder* sequences,

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

and

$$\bar{r}_t = \bar{w}_t - w^*$$

for $t \in \mathbb{Z}_+$, which measure the distance between w_t (and $\bar{w}_t$) and the equilibrium w^*.

3.1 General Upper Bounds for Robbins-Monro Procedure and Averaging Process

The main results are shown in the following theorems.

Theorem 3.2. *Let* $\gamma_t = t^{-\theta}/\overline{\alpha}$ $(\theta \in [0,1))$ *and* $w_0 = 0$. *Then for all* $t \in \mathbb{N}$, *the following holds*

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

where

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

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

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

Remark 3.3. The sample error $\mathscr{E}_{samp}(t)$ decays at the rate $O(\alpha^{-(1+1/[2(1-\theta)])}t^{-\theta/2})$, in terms of the condition number α^{-1} and sample size t. Both of them will be improved in the next theorems.

The next result improves the sample error to the rate $O(\alpha^{-2}t^{-1/2})$, for the averaging process.

Theorem 3.4. *Let* $\gamma_t = t^{-\theta}/\overline{\alpha}$ $(\theta \in [0,1))$ *and* $w_1 = 0$. *Then for all* $t \in \mathbb{N}$ *the following holds*

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

where

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

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

$$\mathscr{E}_{samp}(t) \le \frac{2^{7/2+\theta} D_\theta \beta}{\overline{\alpha}} \left(\frac{1}{\alpha}\right)^2 \sqrt{\frac{1}{t}} \log^{1/2} \frac{2}{\delta}.$$

A further improvement on the rate of condition number $1/\alpha$ is given in the following theorem, which achieves the rate $O(\alpha^{-1}t^{-1/2})$.

Theorem 3.5. *Let* $\gamma_t = t^{-\theta}/\overline{\alpha}$ *(*$\theta \in [0,1)$*) and* $w_1 = 0$*. Define* $\alpha = \underline{\alpha}/\overline{\alpha} \in (0,1]$*. Then the following holds for all* $t \in \mathbb{N}$*,*

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

Here

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

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

$$\mathscr{E}_{samp}(t) \le \frac{2^{1+\theta} D_\theta}{\overline{\alpha}} \left(\frac{2\beta}{3\alpha\sqrt{t}} + \sigma\right) \left(\frac{1}{\alpha}\right) \sqrt{\frac{1}{t}} \log \frac{2}{\delta}.$$

Remark 3.6. Using the Markov inequality we can obtain

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

3.1.1 Proofs of Theorem A, B and B$^+$

Proof of Theorem A. We first show that the algorithm given by (2.3) can be derived from Equation (3.2); then Theorem A follows from Theorem 3.2.

Let $S_x : \mathscr{H}_K \to \mathbb{R}$ be the sampling operator (evaluation functional here) such that $S_x(f) = f(x)$. Let $S_x^* : \mathbb{R} \to \mathscr{H}_K$ be the adjoint of S_x defined by $\langle y, S_x(f)\rangle_{\mathbb{R}} = \langle S_x^*(y), f\rangle_{\mathscr{H}_K}$, whence by reproducing property $f(x) = \langle f, K_x\rangle$, we have $S_x^*(y) = yK_x$ for $y \in \mathbb{R}$. Now take $W = \mathscr{H}_K$, define $A(z) : \mathscr{H}_K \to \mathscr{H}_K$ by $f \mapsto S_x^* S_x(f) + \lambda$ and $b(z) = S_x^*(y)$. Then $\bar{A} = L_K + \lambda I$ and $\bar{b} = L_K f_\rho$. By this substitution, Equation (3.2) becomes (2.3).

Notice that $\overline{\alpha} = \lambda + \kappa^2$, $\underline{\alpha} = \lambda$, and $\beta = \kappa M_\rho$. Theorem A thus follows from Theorem 3.2. □

Proof of Theorem B. In a similar way to the proof of Theorem A, Theorem B follows from Theorem 3.4. □

Proof of Theorem B$^+$. Setting $\overline{\alpha} = \lambda + \kappa^2$, $\underline{\alpha} = \lambda$, $\alpha = \lambda/(\lambda + \kappa^2)$, $\beta = \kappa M_\rho$, and $\sigma = \sigma_\lambda$, the result follows from Theorem 3.5. □

3.2 Martingale Decomposition of Remainders

In this section, we decompose the remainder r_t and its average $\bar{r}_t$ into the sum of two parts: one is deterministic reflecting the error caused by initial choice, and the other is a martingale reflecting the fluctuation caused by random sampling. Upper bounds for them will be given in the next section. Such a decomposition is somehow close to the treatment in Robbins-Siegmund Theorem [Robbins and Siegmund 1971], where $\|r_t\|^2$ is transformed into a supermartingale. But our problem benefits from the linear structure and get a direct decomposition on r_t. We note that such a martingale decomposition can be extended to $\|r_t\|^2$ in nonlinear stochastic approximations and dependent sampling processes, which however are not pursued here.

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

$$\Pi_k^t = \begin{cases} (I-\gamma_t A_t)\cdot(I-\gamma_{t-1}A_{t-1})\dots(I-\gamma_k A_k)\,, & k \leq t; \\ I, & k > t. \end{cases} \tag{3.4}$$

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

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

Proposition 3.7. *For all $t \in \mathbb{N}$,*

$$r_t = \bar{\Pi}_1^t r_0 - \sum_{k=1}^{t} \gamma_k \bar{\Pi}_{k+1}^t \chi_k, \tag{3.5}$$

where $\chi_k = (A_k - \bar{A})w_{k-1} - (b_k - \bar{b})$ ($k \in \mathbb{N}_t$).

Proof. By Equation (3.2)

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

where the last step is due to that using $\bar{b} = \bar{A}w^*$,

$$\chi_t \;:=\; (A_t - \bar{A})w_{t-1} - (b_t - \bar{b}) = A_t w_{t-1} - \bar{A}(w_{t-1} - w^*) - b_t = A_t w_{t-1} - \bar{A} r_{t-1} - b_t$$

Then Equation (3.5) follows from induction on t. □

Note that $\bar{\Pi}_{k+1}^t$ is deterministic, w_{k-1} depends on $z_1, \dots, z_{k-1}$, $A_k - \bar{A}$ and $b_k - \bar{b}$ are both random variables of zero means depending only on z_k. Thus χ_k and $\gamma_k \bar{\Pi}_{k+1}^t \chi_k$ are random variables depending on $(z_i)_1^k$ whose conditional expectation $\mathbb{E}[\gamma_k \bar{\Pi}_{k+1}^t \chi_k | z_1, \dots, z_{k-1}] =$

0. Recall that given a sequence of random variables $(\xi_k)_{k\in\mathbb{N}}$ such that ξ_k depends on random variables $\{z_i : 1 \leq i \leq k\}$, (ξ_k) is called a *martingale difference sequence* if $\mathbb{E}_{z_k|z_1,\ldots,z_{k-1}}[\xi_k] = 0$. The sum of a martingale difference sequence is called a *martingale*. Thus we have the following martingale difference sequence,

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

With this Equation (3.5) can be written as,

$$r_t = \bar{\Pi}_1^t r_0 - \sum_{k=1}^{t} \xi_k. \tag{3.6}$$

Now consider the averaging process. Define

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

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

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

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

Proposition 3.8. *For all $t \in \mathbb{N}$,*

$$\bar{r}_t = \frac{1}{t}\left(\sum_{j=1}^{t} \bar{\Pi}_1^j\right) r_0 - \frac{1}{t}\sum_{k=1}^{t} \gamma_k \left(\sum_{j=k}^{t} \bar{\Pi}_{k+1}^j\right) \chi_k, \tag{3.7}$$

Proof. By Equation (3.5),

$$\bar{r}_t \;=\; \frac{1}{t}\sum_{j=1}^{t} r_j = \frac{1}{t}\left(\sum_{j=1}^{t} \bar{\Pi}_1^j\right) r_0 - \frac{1}{t}\sum_{j=1}^{t}\sum_{k=1}^{j} \gamma_k \bar{\Pi}_{k+1}^j \chi_k$$

where

$$\frac{1}{t}\sum_{j=1}^{t}\sum_{k=1}^{j} \gamma_k \bar{\Pi}_{k+1}^j \chi_k = \frac{1}{t}\sum_{k=1}^{t} \gamma_k \left(\sum_{j=k}^{t} \bar{\Pi}_{k+1}^j\right) \chi_k$$

which ends the proof. □

Let

$$\zeta_k = \begin{cases} \dfrac{\gamma_k}{t}(\displaystyle\sum_{j=k}^{t} \bar{\Pi}_{k+1}^{j})\chi_k, & 1 \le k \le t; \\ \\ 0, & k > t. \end{cases}$$

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

$$\bar{r}_t = \frac{1}{t}\left(\sum_{j=1}^{t} \bar{\Pi}_1^{j}\right) r_0 - \sum_{k=1}^{t} \zeta_k. \tag{3.8}$$

Now define an *initial error* by $\mathscr{E}_{init}(t) = \|\bar{\Pi}_1^{t-1} r_1\|$ (or, $\mathscr{E}_{init}(t) = \|\frac{1}{t}\left(\sum_{j=0}^{t-1} \bar{\Pi}_1^{j}\right) r_1\|$ in averaging process), which is deterministic and reflects the propagated effect of r_1; and a *sample error* by $\mathscr{E}_{samp}(t) = \|\sum_{k=1}^{t-1} \xi_k\|$ (or, $\mathscr{E}_{samp}(t) = \|\sum_{k=1}^{t-1} \zeta_k\|$ in averaging process), which is random and reflects the stochastic error caused by samples. The initial error can be bounded deterministically. For the sample error, we can obtain probabilistic upper bounds by using the exponential inequalities for martingale difference sequences in Hilbert spaces [Pinelis 1994].

For simplicity, we choose the Hoeffding inequality for martingale difference sequences in Hilbert spaces [Pinelis 1992]. Before presenting the proofs, we need some preliminary results. The following proposition collects some useful estimates.

Proposition 3.9. *Let $\alpha' = \alpha/(1-\theta)$. The following holds for all $t \in \mathbb{N}$,*

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

2. *For all $k \in \mathbb{N}$, the operator norm*

$$\left\|\sum_{j=k}^{t} \Pi_{k+1}^{j}\right\| \le 2^{\theta} D_\theta \frac{k^\theta}{\alpha},$$

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

3. $\|w^*\| \le \beta/\underline{\alpha}$;

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

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

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

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

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

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

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

2. Recall that the operator norm of a positive operator is bounded from above by its maximum eigenvalue. Then using Lemma B.3-2,

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

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

3. $\|w^*\| \le \|\bar{A}^{-1}\|\|\bar{b}\| \le \beta/\underline{\alpha}$.

4. $\|Y_t\| = \|A_t w^* - b_t\| \le \overline{\alpha}\beta/\underline{\alpha} + \beta \le 2\beta/\alpha$, since $\alpha = \underline{\alpha}/\overline{\alpha} \le 1$.

5. By Equation (3.2)

$$w_t \quad = \quad w_{t-1} - \gamma_t(A_t w_{t-1} - b_t) = (I - \gamma_t A_t)w_{t-1} + \gamma_t b_t = \Pi_1^t w_0 + \sum_{k=1}^{t}\gamma_k \Pi_{k+1}^t b_k,$$

whence

$$\|w_t\| \quad \le \quad \|\Pi_1^t\|\|w_0\| + \beta\sum_{k=1}^{t}\gamma_k\|\Pi_{k+1}^t\| \le e^{\alpha'(1-(t+1)^{1-\theta})}\|w_0\| + \frac{3\beta}{\underline{\alpha}},$$

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

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

7. Since $\|\chi_t\| = \|(A_t - \bar{A})w_t - (b_t - \bar{b})\| \leq 2\overline{\alpha}\|w_t\| + 2\beta$, apply part 5 and notice that $6\beta/\alpha + 2\beta \leq 8\beta/\alpha$, which gives the result. □

Now we are ready to give the formal proofs of Theorem 3.2 and 3.4.

3.2.1 Proof of Theorem 3.2

Proof of Theorem 3.2. By Equation (3.6) we have

$$\|r_t\| \quad \leq \quad \|\bar{\Pi}_1^t r_0\| + \|\sum_{k=1}^{t} \xi_k\| = \mathscr{E}_{init}(t) + \mathscr{E}_{samp}(t).$$

The upper bound on $\mathscr{E}_{init}(t)$ follows from Proposition 3.9-1. For the upper bound on $\mathscr{E}_{samp}(t)$, by Proposition 3.9-7 with $w_1 = 0$, $\|\chi_k\| \leq 8\beta/\alpha$, whence ξ_k is bounded by

$$\|\xi_k\| \quad \leq \quad \gamma_k \|\Pi_{k+1}^t\| \|\chi_k\| \leq \frac{8\beta}{\underline{\alpha}} \left[\frac{1}{k^\theta} \prod_{i=k+1}^{t} \left(1 - \frac{\alpha}{i^\theta}\right) \right] = c_k.$$

Applying Pinelis-Hoeffding inequality (Lemma A.1), we obtain

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

Let the right hand side equal δ, then

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

where

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

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

3.2.2 Proof of Theorem 3.4

Proof of Theorem 3.4. By Equation (3.8) we have

$$\|\bar{r}_t\| \quad \leq \quad \frac{1}{t}\left\| \left(\sum_{j=1}^{t} \bar{\Pi}_1^j \right) r_1 \right\| + \| \sum_{k=1}^{t} \zeta_k \| = \mathscr{E}_{init}(t) + \mathscr{E}_{samp}(t).$$

The initial error bound follows from Proposition 3.9-2 with $k = 0$. As to the sample error bound, by Proposition 3.9-2 and Proposition 3.9-7 with $w_1 = 0$, we obtain

$$\|\zeta_k\| \quad \leq \quad \frac{\gamma_k}{t}\left\| \sum_{j=k}^{t} \bar{\Pi}_{k+1}^j \right\| \|\chi_k\| \leq \frac{2^{\theta+3}\beta D_\theta \overline{\alpha}}{t\underline{\alpha}^2} = c_\zeta. \tag{3.9}$$

Applying Pinelis-Hoeffding inequality (Lemma A.1),

$$\mathbf{Prob}\left\{ \left\| \sum_{k=1}^{t} \zeta_k \right\| \geq \epsilon \right\} \leq 2\exp\left\{ -\frac{\epsilon^2}{2\sum_{k=1}^{t} c_\zeta^2} \right\},$$

and setting the right hand to be δ, we obtain

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

Plugging in (3.9) gives the sample error bound. □

3.3 Reversed Martingale Decomposition of Remainders

In this section we give a proof of Theorem B$^+$. Note that in the martingale decompostion in Section IV, $\chi_t = (A_t - \bar{A})r_t + \bar{b} - b_t$ whose variance grows in proportion to $\|r_t\|^2$, whence there is no improvement replacing the Hoeffding Inequality by the Markov inequality. However, we may avoid this by turning to the remainder decomposition used in [Smale and Yao 2006] where we directly deal with the variance, $\sigma^2 = \mathbb{E}\|Y_t\|^2$. Yet this approach leads to a reversed martingale decomposition for remainders, as we shall see soon.

The following lemma is taken from [Smale and Yao 2006], whose proof is included here for completeness.

Lemma 3.10. *For all $t \in \mathbb{N}$,*

$$r_t = \Pi_1^t r_0 - \sum_{k=1}^{t} \gamma_k \Pi_{k+1}^t Y_k.$$

Proof. Note that

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

using $Y_t = A_t w^* - b_t$. The result then follows from induction on $t \in \mathbb{N}$. □

It leads to the following decomposition for the averaged remainder.

Lemma 3.11. *For all $t \in \mathbb{N}$,*

$$\bar{r}_t = \frac{1}{t}\left(\sum_{j=1}^{t} \Pi_1^j\right) r_0 - \sum_{k=1}^{t} \frac{\gamma_k}{t}\left(\sum_{j=k}^{t} \Pi_{k+1}^j\right) Y_k.$$

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

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

Recall that a sequence of random variables (x_k) is called a *reversed martingale difference sequence* if (x_{-k}) is a martingale difference sequence. Then (ζ_k) is a reversed martingale difference sequence; since it depends on $\{z_k, \ldots, z_{t-1}\}$ and $\mathbb{E}_{z_k|z_{k+1},\ldots,z_{t-1}}[\zeta_k] = 0$, which implies that (ζ_{-k}) is a martingale difference sequence. We should note that although with this decomposition we obtain tighter bounds, but its reversed martingale structure can not be applied to dependent sampling process like Markov sampling.

3.3.1 Proof of Theorem 3.5

Proof of Theorem 3.5. The initial error bound follows from Proposition 3.9-2 with $k = 0$.

To apply the Pinelis-Bernstein inequality A.4, notice that

$$\begin{aligned}\sum_{k=1}^{t} \mathbb{E}[\|\zeta_k\|^2 | z_{k+1}, \dots, z_t] &= \sum_{k=1}^{t} \frac{\gamma_k^2}{t^2} \|\sum_{j=k}^{t} \Pi_{k=k+1}^{j}\|^2 \mathbb{E}[\|Y_k\|^2] \\ &\leq \frac{2^{2\theta} D_\theta^2 \sigma^2}{\overline{\alpha}^2 \alpha^2} t^{-1}.\end{aligned}$$

where the last step is due to Proposition 3.9-2 and Finiteness Condition-C. Moreover

$$\|\zeta_k\| \quad \leq \quad \frac{\gamma_k}{t} \|\sum_{j=k}^{t} \Pi_{k=k+1}^{j}\| \|Y_k\| \leq \frac{2^{\theta+1} D_\theta \beta}{\overline{\alpha} \alpha^2} t^{-1}$$

using Proposition 3.9-2,4. The result then follows from Proposition A.4 with $M = \frac{2^{\theta+1} D_\theta \beta}{\overline{\alpha} \alpha^2} t^{-1}$, and $\sigma_t^2 = \frac{2^{2\theta} D_\theta^2 \sigma^2}{\overline{\alpha}^2 \alpha^2} t^{-1}$. □

Remark 3.12. Using the Markov inequality, we may obtain the following upper bound for the sample error. Note that

$$\mathbb{E} \|\sum_{k=1}^{t} \zeta_k\|^2 \quad \leq \quad \sum_{k=1}^{t} \frac{\gamma_k^2}{t^2} \mathbb{E}[\|\sum_{j=k}^{t} \Pi_{k=k+1}^{j}\|^2 \|Y_k\|^2] \leq \frac{2^{2\theta} D_\theta^2 \sigma^2}{\overline{\alpha}^2 \alpha^2} t^{-1}.$$

where the last is due to Proposition 3.9-2 and Finiteness Condition-C. The sample error bound then follows from the Markov inequality in Lemma A.5 by taking $X = \| \sum_{k=1}^{t} \zeta_k\|^2$.

The following proposition gives an estimate of σ_λ.

Proposition 3.13. *1.* $\|f_\lambda^*\|_K \leq M_\rho / \sqrt{\lambda}$*;*

2. $\|f_\lambda^*\|_\rho \leq 2M_\rho$*;*

3. $\sigma_\lambda \leq M_\rho \sqrt{5(\lambda + \kappa^2)}$*.*

Proof. 1. Note that

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

Taking $f = 0$, we have

$$\|f_\lambda^* - f_\rho\|_\rho^2 + \lambda \|f_\lambda^*\|_K^2 \leq \|f_\rho\|_\rho^2 \leq M_\rho^2, \tag{3.10}$$

which leads to the result.

2. From (7.10), we obtain $\|f_\lambda^* - f_\rho\|_\rho \leq M_\rho$. The result then follows from

$$\|f_\lambda^*\|_\rho \quad \leq \quad \|f_\lambda^* - f_\rho\|_\rho + \|f_\rho\|_\rho \leq 2M_\rho.$$

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

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

where the first term

$$\mathbb{E}[(f_\lambda^*(x) - y)^2 K(x,x)] \quad \leq \quad \kappa^2(\|f_\lambda^* - f_\rho\|_\rho^2 + \mathbb{E}[f_\rho(x) - y)^2] \leq 5\kappa^2 M_\rho^2,$$

using

$$\mathbb{E}[(f(x) - y)^2] \quad = \quad \mathbb{E}[(f(x) - f_\rho(x) + f_\rho(x) - y)^2] = \|f - f_\rho\|_\rho^2 + \mathbb{E}[(f_\rho(x) - y)^2],$$

the second term

$$\langle f_\lambda^*, L_K(f_\lambda^* - f_\rho)\rangle_K \quad \leq \quad \|L_K^{1/2} f_\lambda^*\|_K \cdot \|L_K^{1/2}(f_\lambda^* - f_\rho)\|_K = \|f_\lambda^*\|_\rho \cdot \|f_\lambda^* - f_\rho\|_\rho \leq 2M_\rho^2,$$

using the isometry $L_K^{1/2}$ such that $\|f\|_\rho = \|L_K^{1/2} f\|_K$, and the third term $\lambda^2 \|f_\lambda^*\|_K^2 \leq \lambda M_\rho^2$.

□

Chapter 4

Open Problems

In this part, we showed by probabilistic upper bounds that a two-stage online learning algorithm, the stochastic approximation of the gradient descent method followed by an averaging process, can achieve the almost sure convergence with the same fast rate as "batch learning", i.e. $O(\lambda^{-1}t^{-1/2}\log 1/\delta)$. We introduce two structural decompositions for the remainders, where the fast rate above is only achieved via the reversed martingale decomposition, which is however not applicable in dependent sampling settings.

Some open problems are still left, including:

(A) Can one achieve the rate $O(\lambda^{-1}t^{-1/2}\log 1/\delta)$, using the martingale decomposition approach which is extendable to dependent sampling settings? This issue will be important in the study of dependent sampling generalizations.

(B) The upper bounds in this part hold for all $\theta \in [0,1)$ but not include $\theta = 1$, i.e. $\gamma_t = O(t^{-1})$. We conjecture that for one-stage stochastic gradient descent algorithm with step size $\gamma_t = O(t^{-1})$, one can also achieve the convergence rate $O(\lambda^{-1}t^{-1/2})$.

Part II

Stochastic Approximation of Regularization Path

Chapter 5

Main Results

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

$$f_t = f_{t-1} - \gamma_t[(f_{t-1}(x_t) - y_t)K_{x_t} + \lambda_t f_{t-1}], \qquad \text{for some } f_0 \in \mathscr{H}_K, \text{ e.g. } f_0 = 0 \qquad (5.1)$$

where

(A) for each t, (x_t, y_t) is independent and identically distributed (i.i.d.) according to ρ;

(B) the step size $\gamma_t > 0$ such that $\sum \gamma_t = \infty$;

(C) the regularization parameter $\lambda_t > 0$ such that $\lambda_t \to 0$.

We are going to give probabilistic upper bounds for the distance $\|f_t - f_\rho\|$ in $\mathscr{L}^2_{\rho_\mathscr{X}}$ and $\mathscr{H}_K$. We start from the triangle inequality

$$\|f_t - f_\rho\| \le \|f_t - f^*_{\lambda_t}\| + \|f^*_{\lambda_t} - f_\rho\|,$$

where the second term $\|f^*_{\lambda_t} - f_\rho\|$ is called the *approximation error*, denoted by $\mathscr{E}^{(\rho)}_{approx}(t)$. The first term can be further decomposed into three parts,

$$\|f_t - f^*_{\lambda_t}\| \le \mathscr{E}_{init}(t) + \mathscr{E}_{samp}(t) + \mathscr{E}_{drift}(t).$$

Such a decomposition is given either by the martingale decomposition in Theorem 6.1, or by the reversed martingale decomposition in Theorem 6.5. In this way we may bound

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

To be precise, we use notations $\mathscr{E}_*^{(\rho)}$ or $\mathscr{E}_*^{(K)}$ to specify the underlying space of $\mathscr{L}^2_{\rho_\mathscr{X}}$ or $\mathscr{H}_K$. In the remaining of this chapter, we are going to provide upper bounds for each of the four errors under the assumption that $f_\rho \in L_K^r(\mathscr{L}^2_{\rho_\mathscr{X}})$, which, roughly speaking, are that if $r \in [1/2, 1]$,

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

and if $r \in (1/2, 3/2]$,

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

under the choice that

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

Finally setting $\theta = 2r/(2r+1)$ leads to upper bounds $\|f_t - f_\rho\|_\rho \leq O(t^{-r/(2r+1)}$ $(r \in [1/2, 1])$ (Theorem C) and $\|f_t - f_\rho\|_K \leq O(t^{-(r-1/2)/(2r+1)})$ $(r \in (1/2, 3/2])$ (Theorem D). These rates are the same as the best known results in "batch learning" [Smale and Zhou 2006a] and actually are optimal in some senses (see Remark 5.4).

5.1 Upper Bounds for $\mathscr{L}^2_{\rho_{\mathscr{X}}}$-convergence: Theorem C

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

$$\|f_t - f_\rho\|_\rho \leq C_1 t^{-1} + \left(C_2 \sqrt{\frac{1}{\delta}} + \left(C_3 + C_4 \sqrt{\frac{1}{\delta}} \right) \|L_K^{-r} f_\rho\|_\rho \right) t^{-r/(2r+1)},$$

where

$$C_1 = t_0 M_\rho, \quad C_2 = \sqrt{6}\kappa M_\rho (1 + \kappa(\kappa+1)), \quad C_3 = \frac{5r+1}{r(r+1)}, \quad C_4 = 2\sqrt{2}\kappa.$$

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

Its proof will be given in the next chapter via the martingale decomposition in Theorem 6.1.

Remark 5.1. A special case is $r = 1/2$, which is equivalent to say $f_\rho \in \mathscr{H}_K$. In this case $\gamma_t = \lambda_t = (t+t_0)^{-1/2}$, whence it does not satisfy the Path Following Condition (B) in Theorem A. But Theorem C suggests a weaker notion that f_t follows the regularization path, *i.e.* $\lim_{t\to\infty} \mathbb{E}[\|f_t - f^*_{\lambda_t}\|_\rho] = 0$, which in fact converges at a rate of $O(t^{-1/4})$ uniformly for all $f_\rho \in \mathscr{H}_K$.

Remark 5.2. It is still open whether the upper bound above can be improved by replacing $1/\delta$ with $\log 1/\delta$. For details, see more discussions in Remark 7.8, on the problem of using Bernstein's type inequalities here.

For the ease of comparisons with existing results, consider the generalization error [e.g. see Cucker and Smale 2002b],

$$\mathscr{E}(f) := \int_{\mathscr{X}\times\mathscr{Y}} (f(x)-y)^2 d\rho = \|f - f_\rho\|_\rho^2 + \mathscr{E}(f_\rho)$$

which is often used to evaluate the performance of learning algorithms in literature. We have the following corollary of Theorem C.

Corollary 5.3. *Under the same condition of Theorem C, there holds with probability at least* $1-\delta$ *(*$\delta \in (0,1)$*), for all* $t \in \mathbb{N}$*,*

$$\mathscr{E}(f_t) - \mathscr{E}(f_\rho) \le 2C_1 t^{-2} + 2(C_2\sqrt{\frac{1}{\delta}} + C_3\|L_K^{-r} f_\rho\|_\rho + C_4\sqrt{\frac{1}{\delta}}\|L_K^{-r} f_\rho\|_\rho)^2 t^{-2r/(2r+1)},$$

where $C_1, \ldots, C_4$ *are the same constants in Theorem C.*

Remark 5.4. 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. To be precise, let $\mathscr{P}(b,r)$ ($b > 1$ and $r \in (1/2,1]$) be the set of probability measure ρ on $\mathscr{X}\times\mathscr{Y}$, such that: (A) almost surely $|y| \le M_\rho$; (B) $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_\mathscr{X}}$; (C) the eigenvalues $(\mu_n)_{n\in\mathbb{N}}$ of $L_K : \mathscr{L}^2_{\rho_\mathscr{X}} \to \mathscr{L}^2_{\rho_\mathscr{X}}$, arranged in a nonincreasing order, are subject to the decay $\mu_n = O(n^{-b})$. Then the following minimax lower rate was given as Theorem 2 in [Caponnetto and De Vito 2006],

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

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

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 [Györfi, Kohler, Krzyżak, and Walk 2002] suggests a kind of individual lower rates for learning problems. The following individual lower rate was obtained as Theorem 3 in [Caponnetto and De Vito 2006]: 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, its eigenvalues are summable. Therefore by taking $b = B = 1$, one may obtain an eigenvalue-independent lower rate $O(t^{-2r/(2r+1)})$ for all possible L_K. In this way, the upper bound in Corollary 5.3 reaches both the minimax and the individual lower rates.

For $r > 1$, the convergence rates in the upper bounds will be no faster than the case of $r = 1$, which is often refered as the *saturation* issue of Tikhonov regularization in inverse problems [Engl, Hanke, and Neubauer 1996]. It is hoped that using other regularization schemes such as iterative regularization one may overcome this saturation issue. For $r \leq 1/2$, however, it is unclear if we can achieve the optimal rates.

5.2 Upper Bound for $\mathscr{H}_K$-convergence: Theorem D

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

$$\|f_t - f_\rho\|_K \le \frac{D_1}{t} + (D_2 \log\frac{2}{\delta} + D_3 \|L_K^{-r} f_\rho\|_\rho) \left(\frac{1}{t}\right)^{\frac{2r-1}{4r+2}},$$

where

$$D_1 = t_0^{3/2} M_\rho, \quad D_2 = (5\kappa+1) M_\rho, \quad D_3 = \frac{20r-2}{(2r-1)(2r+3)}.$$

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

Its proof will be given in the next chapter via the reversed martingale decomposition in Theorem 6.5.

Remark 5.5. The asymptotic rate $O(t^{-(2r-1)/(4r+2)})$ is the same as the batch learning algorithms [Theorem 2, in Smale and Zhou 2006a].

Remark 5.6. 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 $O(t^{-(2r-1)/(4r+2)})$, reflects the error caused by random fluctuations by the i.i.d. sampling. The third term at a rate $O(\|L_K^{-r} f_\rho\|_\rho t^{-(2r-1)/(4r+2)})$, collects contributions from both drifts along the regularization path $f^*_{\lambda_t} - f^*_{\lambda_{t-1}}$ and the approximation error $f^*_{\lambda_t} - f_\rho$, since they share the same rate upto different constants.

Chapter 6

Stochastic Algorithms for Linear Ill-posed Problems

Let $\mathscr{W}$ be a Hilbert space, $\bar{A} : \mathscr{W} \to \mathscr{W}$ be a positive operator and $\bar{b} \in \mathscr{W}$. Assume that $\bar{A}$ has an *unbounded* inverse. The following linear equation

$$\bar{A}w = \bar{b}. \tag{6.1}$$

is thus *ill-posed* since its solution, $w^* = \bar{A}^{-1}\bar{b}$, is *discontinuous.*

As in the standard setting of stochastic approximation [Robbins and Monro 1951], assume that $\bar{A} = \mathbb{E}[A(z)]$ and $\bar{b} = \mathbb{E}[b(z)]$, i.e. the expectations of random operator $A : \mathscr{Z} \to SL(\mathscr{W})$ and random vector $b : \mathscr{Z} \to \mathscr{W}$. However since $\bar{A}$ has infinite condition number, the analysis in Part I will fail.

6.1 Regularization Paths of Linear Ill-posed Problem

To solve this ill-posed problem with unbounded $\bar{A}^{-1}$, one may construct a sequence $\bar{A}_t \to \bar{A}$ and $\bar{b}_t \to \bar{b}$, where each $\bar{A}_t$ has bounded inverse. Then one has a sequence $w_t^* = \bar{A}_t^{-1}\bar{b}_t$ which is expected to converge to the solution of (6.1), and each w_t^* is continuous with respect to $\bar{A}_t$ and $\bar{b}_t$. Such a sequence (w_t^*), will be called a *regularization path* of the solution of Equation (6.1).

The following examples include typical regularization schemes [Engl, Hanke, and Neubauer 1996].

Example 6.1.1 (Tikhonov Regularization). Let $\bar{A}_t = \bar{A} + \lambda_t$ with $\lambda_t \to 0$ and $\lambda_t > 0$, $\bar{b}_t = \bar{b}$. This implements the *Tikhonov regularization*,

$$w_t^* = (\bar{A} + \lambda_t)^{-1}\bar{b}.$$

Example 6.1.2 (Landwebter Iteration Regularization). Define

$$w_t^* = \|\bar{A}\|^{-1} \sum_{i=0}^{t-1} (1 - \|\bar{A}\|^{-1}\bar{A})^i \bar{b}.$$

One can regard that $\bar{A}_t$ is defined implicitly via

$$\bar{A}_t^{-1} := \|\bar{A}\|^{-1} \sum_{i=0}^{t-1} (1 - \|\bar{A}\|^{-1}\bar{A})^i.$$

This scheme is often called as the *Landwebter iteration* in classical inverse problems.

Example 6.1.3 (General Regularizations). Let $g_t : \mathbb{R} \to \mathbb{R}$ be a function such that as $t \to \infty$, $g_t(\sigma) \to \sigma^{-1}$ for $\sigma \in (0, \|\bar{A}\|)$. Define

$$w_t^* := g_t(\bar{A})\bar{b}$$

which realizes many regularization schemes. In particular, for Tikhonov regularization $g_t(\sigma) = (\sigma + \lambda_t)^{-1}$ ($\lambda_t \downarrow 0$) and for Landwebter iteration $g_t(\sigma) = \gamma \sum_{i=0}^{t-1} (1 - \gamma\sigma)^i$ where $\gamma \leq \|\bar{A}\|^{-1}$ is a constant. Another famous example is the truncated spectrum regularization

$$g_t(\sigma) = \begin{cases} \sigma, & \text{if } \sigma \geq \lambda_t \\ 0, & \text{otherwise} \end{cases}$$

for some $\lambda_t \downarrow 0$.

6.2 Stochastic Approximation of Tikhonov Regularization Paths

Define

$$w_t = w_{t-1} - \gamma_t((A_t + \lambda_t)w_{t-1} - b_t), \qquad w_0 = 0, \tag{6.2}$$

where

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

(B) step size $\gamma_t > 0$ and $\sum_{t \in \mathbb{N}} \gamma_t = \infty$;

(C) Tikhonov regularization parameter $\lambda_t > 0$ and $\lambda_t \downarrow 0$.

Let $\bar{A}_t = \bar{A} + \lambda_t$. Consider the *Tikhonov regularization path*

$$w_t^* := \bar{A}_t^{-1}\bar{b} = (\bar{A} + \lambda_t)^{-1}\bar{b}.$$

In the remaining we present two structural decompositions of the *remainder*,

$$r_t := w_t - w_t^*. \tag{6.3}$$

Both ways decomposes r_t into three parts: one depending on r_0; one depending on the following defined *drifts* along the regularization path (w_t^*),

$$\Delta_t := w_t^* - w_{t-1}^*; \tag{6.4}$$

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

Example 6.2.1. Let $\mathscr{W} = \mathscr{H}_K$, $\bar{A} = L_K$, $\bar{b} = L_K f_\rho$, $\bar{A}_t = L_K + \lambda_t$, $A_t = \langle\ , K_{x_t}\rangle_K K_{x_t} + \lambda_t$, $\bar{b}_t = L_K f_\rho$, and $b_t = y_t K_{x_t}$. Then in learning we are going to approximate the solution of the following linear ill-posed equation,

$$L_K f = L_K f_\rho, \qquad f \in \mathscr{H}_K.$$

The Tikhonov regularization path is

$$w_t^* = f_{\lambda_t}^* := (L_K + \lambda_t)^{-1} L_K f_\rho,$$

whose stochastic approximation is

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

6.2.1 Martingale Decomposition

Theorem 6.1 (Martingale Decomposition). *Let* $\chi_t = (A_t - \bar{A}_t)w_{t-1} - (b_t - \bar{b})$, $\Delta_t = w_t^* - w_{t-1}^*$, *and*

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

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

$$r_t = \bar{\Pi}_{t_0+1}^t r_{t_0} - \sum_{j=t_0+1}^t \gamma_j \bar{\Pi}_{j+1}^t \chi_j - \sum_{j=t_0+1}^t \bar{\Pi}_j^t \Delta_j \tag{6.5}$$

Remark 6.2. This decomposition was proposed in [Yao 2006]. Note that in this decomposition only the second term is random. The operator $\bar{\Pi}_{j+1}^t$ is deterministic and χ_j is a zero mean random variable depending on $z_1, \ldots, z_j$. Therefore the conditional expectation $\mathbb{E}[\gamma_j \bar{\Pi}_{j+1}^t \chi_j | z_1, \ldots, z_{j-1}] = 0$, whence for each t, $\gamma_j \bar{\Pi}_{j+1}^t \chi_j$ is a *martingale difference sequence* for all $t \in \mathbb{N}$, whose sum is a *martingale sequence* of zero mean. Note that this martingale property holds even for dependent sampling $z_t(z_1, \ldots, z_{t-1})$.

Setting $\mathscr{W} = \mathscr{H}_K$, $\bar{A} = L_K$, $\bar{b} = L_K f_\rho$, $b_t = y_t K_{x_t}$, and $t_0 = 0$, we obtain the following corollary.

Corollary 6.3. *Let $L_t = \langle \ , K_{x_t} \rangle_K K_{x_t}$, $\chi_t = (L_t - L_K) f_{t-1} - (y_t K_{x_t} - L_K f_\rho)$, $\Delta_t = f_{\lambda_t}^* - f_{\lambda_{t-1}}^*$, and*

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

Then for all $t \in \mathbb{N}$ and $f_0 = 0$,

$$f_t - f_{\lambda_t}^* = -\bar{\Pi}_1^t f_{\lambda_0}^* - \sum_{j=t_0+1}^t \gamma_j \bar{\Pi}_{j+1}^t \chi_j - \sum_{j=t_0+1}^t \bar{\Pi}_j^t \Delta_j \tag{6.6}$$

Remark 6.4. An importance feature of this decomposition used in this paper, lies in that the operator $\bar{\Pi}_j^t$ is deterministic and when taking $\bar{A}_i = L_K + \lambda_i$, it has a spectral decomposition by the eigenfunctions of $L_K : \mathscr{L}_{\rho_\mathscr{X}}^2 \to \mathscr{L}_{\rho_\mathscr{X}}^2$. This feature plays a key role in the proof of Theorem C. But a disadvantage is that the term χ_t depends on w_{t-1}, which increases the difficulty to bound χ_t. In fact, the open problem how to improve Theorem C by replacing $1/\delta$ to $\log 1/\delta$, depends on how to get a tighter bound on $\|\chi_t\|$, see Remark 7.8 for details.

Proof of Theorem 6.1. By definition,

$$\begin{aligned} r_t &= w_t - w_t^* \\ &= w_{t-1} - \gamma_t(A_t w_{t-1} - b_t) - w_t^* \\ &= (I - \gamma_t \bar{A}_t)(w_{t-1} - w_{t-1}^*) - \gamma_t[(A_t - \bar{A}_t)w_{t-1} - (b_t - \bar{A}_t w_t^*)] - (I - \gamma_t \bar{A}_t)[w_t^* - w_{t-1}^*] \\ &= (I - \gamma_t \bar{A}_t) r_{t-1} - \gamma_t \chi_t - (I - \gamma_t \bar{A}_t)\Delta_t, \end{aligned}$$

where the last step is due to $\bar{A}_t w_t^* = \bar{b}$. The result then follows from induction on $t \in \mathbb{N}$. □

6.2.2 Reversed Martingale Decomposition

Theorem 6.5 (Reversed Martingale Decomposition). *Define a random operator on* $\mathscr{W}$*,*

$$\Pi_j^t(z_j, \ldots, z_t) = \begin{cases} (I - \gamma_t A_t - \gamma_t \lambda_t) \ldots (I - \gamma_j A_j - \gamma_j \lambda_j), & j \le t; \\ I, & j > t. \end{cases}$$

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

$$r_t = \Pi_{t_0+1}^t r_{t_0} - \sum_{j=t_0+1}^{t} \gamma_j \Pi_{j+1}^t (A_j w_j^* - b_j) - \sum_{j=t_0+1}^{t} \Pi_j^t \Delta_j \tag{6.7}$$

Remark 6.6. Note that Π_{j+1}^t is a random operator depending on $z_{j+1}, \ldots z_t$, and $A_j w_j^* - b_j$ is a zero mean random variable depending on z_j. By independence of $(z_t)_{t\in\mathbb{N}}$, the conditional expectation $\mathbb{E}[\gamma_j \Pi_{j+1}^t (A_j w_j^* - b_j) | z_{j+1}, \ldots, z_t] = 0$, whence for each t, $\gamma_j \Pi_{j+1}^t (A_j w_j^* - b_j)$ is a *reversed martingale difference sequence* whose sum is a *reversed martingale sequence* with zero mean. For more background on reversed martingale, see for example [Neveu 1975]. This decomposition will be used to derive Theorem D.

Setting $\mathscr{W} = \mathscr{H}_K$, $\bar{A} = L_K$, $\bar{b} = L_K f_\rho$, $b_t = y_t K_{x_t}$, and $t_0 = 0$, we obtain the following corollary.

Corollary 6.7. *Let* $L_t = \langle\ , K_{x_t}\rangle_K K_{x_t}$ *and* $Y_t = L_t f^*_{\lambda_t} - y_t K_{x_t} = (f^*_{\lambda_t}(x_t) - y_t)K_{x_t}$. *Define a random operator on* $\mathscr{W}$*,*

$$\Pi^t_j(z_j, \ldots, z_t) = \begin{cases} (I - \gamma_t L_t - \gamma_t \lambda_t) \ldots (I - \gamma_j L_j - \gamma_j \lambda_j)\,, & j \le t; \\ I, & j > t. \end{cases}$$

Then for all $t \in \mathbb{N}$ *and* $f_0 = 0$*,*

$$f_t - f^*_{\lambda_t} = -\Pi^t_{t_0+1} f^*_{\lambda_0} - \sum_{j=t_0+1}^{t} \gamma_j \Pi^t_{j+1} Y_j - \sum_{j=t_0+1}^{t} \Pi^t_j \Delta_j \qquad (6.8)$$

Proof of Theorem 6.5. By definition,

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

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

6.3 Stochastic Iterative Regularization

Note that in the algorithm (6.2), setting the regularization parameter $\lambda_t \equiv 0$ realizes iterative regularization with

$$g_t(\sigma) = \sum_{i=0}^{t-1} \gamma_i (1 - \gamma_i \sigma)^i.$$

To see this, define a new iteration $(\bar{w}_t)_{t\in\mathbb{Z}_+}$,

$$\bar{w}_t = \bar{w}_{t-1} - \gamma_t(\bar{A}\bar{w}_{t-1} - \bar{b}), \qquad \bar{w}_0 = 0, \tag{6.9}$$

which can be regarded as taking conditional expectation $\mathbb{E}_{t-1}$ on both sides of (6.2), By induction (6.9) gives the *iterative regularization path*

$$\bar{w}_t = \sum_{i=0}^{t} \gamma_i (1 - \gamma_i \bar{A})^i \bar{b}$$

The counterpart of (6.9) in $\mathscr{H}_K$ is

$$f_t^* = f_{t-1}^* - \gamma_t(L_K f_{t-1}^* - L_K f_\rho), \qquad \text{where } \bar{f}_0 = 0 \tag{6.10}$$

whose stochastic approximation is

$$f_t = f_{t-1} - \gamma_t(f_{t-1}(x_t) - y_t)K_{x_t}, \qquad \text{where } f_0 = 0 \tag{6.11}$$

where the step size $\gamma_t > 0$ such that $\sum \gamma_t = \infty$.

The regularization effect of $(f_t^*)_{t\in\mathbb{Z}_+}$ can be seen from the following bound.

Theorem 6.8. *Suppose $f_\rho \in L_K^r(\mathscr{L}^2_{\rho_{\mathscr{X}}})$ for some $r > 0$ and $f_0^* = 0$. Then for all $t \in \mathbb{N}$,*

$$\|f_t^* - f_\rho\|_\rho \le \|f_\rho\|_\rho \left(\frac{r}{e}\right)^r \left(\sum_{i=0}^{t} \gamma_i\right)^{-r};$$

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

$$\|f_t^* - f_\rho\|_K \le \|f_\rho\|_K \left(\frac{r - 1/2}{e}\right)^{r-1/2} \left(\sum_{i=0}^{t} \gamma_i\right)^{-(r-1/2)}.$$

Remark 6.9. Hence for $\sum_t \gamma_t = \infty$, we have $\|f_t^* - f_\rho\| \to 0$ in both $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ and $\mathscr{H}_K$. This theorem can be regarded as the *approximation error* for the iterative regularization (6.10).

Proof of Theorem 6.8. Let $f_\rho = L_K^r g$ with $\|g\|_\rho \leq R$. By induction with $f_0 = 0$, Equation (6.10) leads to

$$f_t^* - f_\rho = g_t(L_K) L_K f_\rho - f_\rho = -r_t(L_K) f_\rho,$$

whence

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

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

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

But the function

$$g(x) = -\sum_i \gamma_i x + r \log x, \qquad x > 0,$$

is maximized at $x^* = r/(\sum_i \gamma_i)$ with $g(x^*) = -r + r\log r - r \log \sum_i \gamma_i$. Taking $\gamma_t = (t+1)^{-\theta}/\kappa^2$, we obtain

$$\|L_K^r r_t(L_K)\| \quad \leq \quad (r/e)^r (\sum_{i=0}^{t-1} \gamma_i)^{-r}.$$

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

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

Replacing r by $r - 1/2$ above leads to the second bound. □

In [Ying and Pontil 2006] the convergence of $(f_t)_{t\in\mathbb{N}}$ in Equation (6.11) was studied without considering the iterative regularization scheme above. A thorough study on iterative regularization in this setting is beyond the scope of this thesis. Here we only present a martingale decomposition theorem which is the foundation for further developments.

6.3.1 Martingale Decomposition

Theorem 6.10 (Martingale Decomposition). *Let* $\xi_t = (\bar{A} - A_t)w_{t-1} + (b_t - \bar{b})$*, and*

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

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

$$w_t - w_t^* = \bar{\Pi}_{t_0+1}^t (w_{t_0} - w_{t_0}^*) + \sum_{j=t_0+1}^{t} \gamma_j \bar{\Pi}_{j+1}^t \xi_j \tag{6.12}$$

Similarly setting $\mathscr{W} = \mathscr{H}_K$, $\bar{A} = L_K$, $\bar{b} = L_K f_\rho$, $b_t = y_t K_{x_t}$, and $t_0 = 0$, we obtain the following corollary.

Corollary 6.11. *Let*

$$\bar{\Pi}_j^t = \begin{cases} \prod_{i=j}^{t} \left(I - \gamma_i L_K - \gamma_i \lambda_i\right), & j \leq t; \\ I, & j > t. \end{cases}$$

For all $t \in \mathbb{N}$*, the following holds*

$$f_t - f_t^* = -\sum_{j=1}^{t} \gamma_j \bar{\Pi}_{j+1}^t \xi_t$$

where $\xi_t = (L_t - L_K)f_t - (y_t K_{x_t} - L_K f_\rho)$*.*

Chapter 7

Stochastic Tikhonov Regularization Paths in Learning

In this chapter we are going to give probabilistic upper bounds for

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

in $\mathscr{L}^2_{\rho_\mathscr{X}}$ or $\mathscr{H}_K$, where f_t is defined as Equation (5.1), i.e.

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

Throughout this chapter, we assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_\mathscr{X}}$ for some $r > 0$.

To be precise, for the convergence in $\mathscr{L}^2_{\rho_\mathscr{X}}$, we use the martingale decomposition in Theorem 6.1,

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

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

$$\bar{\Pi}_j^t = \begin{cases} \displaystyle\prod_{i=j}^{t} \left(I - \gamma_i(L_K + \lambda_i I)\right), & j \le t; \\ I, & j > t. \end{cases} \tag{7.1}$$

The reason of using such a decomposition, is that due to the isometry $L_K^{1/2} : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{H}_K$ such that $\|r_t\|_\rho = \|L_K^{1/2} r_t\|_K$, one can benefit from the spectral decomposition of $L_K^{1/2}\bar{\Pi}_j^t$ to get a tighter estimate. However such a nice feature is lost in the reversed martingale decomposition in that the operator $L_K^{1/2}\Pi_j^t$ below can't be diagonalizable. But this decomposition has shortcomings as well: due to χ_t depends on f_{t-1}, which increases the difficulty to estimate $\|\chi_t\|_\rho$. In fact just for this reason, we can not directly apply the Pinelis-Bernstein inequality to improve Theorem C by replacing $1/\delta$ with $\log 1/\delta$.

We make the following definitions for convenience.

[Definitions of Errors]

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

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

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

(D) *Approximation Error*: $\mathscr{E}^{(\rho)}_{approx}(t) := \|f_{\lambda_t} - f_\rho\|_\rho$, which measures the distance between the regression function and the regularization path at time t.

For convergence in $\mathscr{H}_K$, we use the reversed martingale decomposition in Theorem

6.5,

$$r_t = -\Pi_1^t r_0 - \sum_{j=1}^{t} \gamma_j \Pi_{j+1}^t (A_j \bar{w}_j - b_j) - \sum_{j=1}^{t} \Pi_j^t \Delta_j \tag{7.2}$$

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

$$\Pi_j^t(x_j, \dots, x_t) = \begin{cases} (I - \gamma_t(L_t + \lambda_t I)) \cdot (I - \gamma_{t-1}(L_{t-1} + \lambda_{t-1} I)) \dots (I - \gamma_j(L_j + \lambda_j I)), & j \le t; \\ I, & j > t, \end{cases}$$

For convenience, we make the following definitions.

[Definitions of Errors]

(A) *Initial Error*: $\mathscr{E}_{init}^{(K)}(t) := \|\Pi_1^t r_0\|_K$, which reflects the propagation error by the initial choice f_0;

(B) *Sample Error*: $\mathscr{E}_{samp}^{(K)}(t) := \|\sum_{j=1}^t \gamma_j \Pi_{j+1}^t (A_j \bar{w}_j - b_j)\|_K$, where $\xi_j = \gamma_j \Pi_{j+1}^t (A_j \bar{w}_j - b_j)$ is a reversed martingale difference sequence, reflecting the random fluctuation caused by sampling;

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

(D) *Approximation Error*: $\mathscr{E}_{approx}^{(K)}(t) := \|f_{\lambda_t} - f_\rho\|_K$, which measures the distance between the regression function and the regularization path at time t.

7.1 Drifts along Regularization Paths and Approximation Error

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 7.1.

Theorem 7.1. *Let $\lambda > \mu \geq 0$. If $\mu = 0$ we define $f_\mu = f_\rho$. Assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_\mathscr{X}}$ for some $r > 0$.*

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

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

(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

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

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

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

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

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

Remark 7.2. From (A) and (B) (or (D) and (E)) we can see that $\|f_\lambda - f_\mu\|_\rho \leq O(|\lambda^{\min(r,1)} - \mu^{\min(r,1)}|)$ (or $\|f_\lambda - f_\mu\|_K \leq O(|\lambda^{\min(r-1/2,1)} - \mu^{\min(r-1/2,1)}|)$). In this way the upper bounds 'saturate' in the rates when f_ρ has large enough regularity indexed by $r > 1$ (or $r > 3/2$).

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

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

which yields

$$f_\lambda - f_\mu = (\mu - \lambda)(L_K + \lambda I)^{-1} (L_K + \mu I)^{-1} L_K f_\rho. \tag{7.3}$$

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

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

where

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

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

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

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

(B) For any $s \leq r$, $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_\mathscr{X}}$ implies $L_K^{-s} f_\rho \in \mathscr{L}^2_{\rho_\mathscr{X}}$. If $s \geq 1$, by Equation (7.3),

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

(C) In particular if $r \geq 1/2$, this implies $f_\rho \in \mathscr{H}_K$, whence by (7.3)

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

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

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

where

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

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

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

7.1.1 Approximation Error

Theorem 7.3 (Approximation Error). *For* $r \in (1/2, 3/2]$,

(A) $\|f_{\lambda_t} - f_\rho\|_\rho \le C_1(t+t_0)^{-r(1-\theta)}$, *where* $C_1 = r^{-1}\|L_K^{-r} f_\rho\|_\rho$.

(B) $\|f_{\lambda_t} - f_\rho\|_K \le C_2(t+t_0)^{-(r-1/2)(1-\theta)}$, *where* $C_2 = (r-1/2)^{-1}\|L_K^{-r} f_\rho\|_\rho$.

Proof. The first inequality follows from Theorem 7.1(A) with $\lambda = \lambda_t$ and $\mu = 0$. The second follows from Theorem 7.1(D) with $\lambda = \lambda_t$ and $\mu = 0$. □

7.1.2 Drift Error

Theorem 7.4 (Drift Error). *Let* $r > 0$ *and* $t_0^\theta \ge \kappa^2 + 1$. *Assume* $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$.

(A) For $r \in (0,1]$, $\mathscr{E}^{(\rho)}_{drift}(t) \le C_3(t+t_0)^{-r(1-\theta)}$, *where* $C_3 = \dfrac{4(1-\theta)}{1-r(1-\theta)}\|L_K^{-r} f_\rho\|_\rho$.

(B) For $r \in (1/2, 3/2]$, $\mathscr{E}^{(K)}_{drift}(t) \le C_4(t+t_0)^{-(r-1/2)(1-\theta)}$, *where* $C_4 = \dfrac{4(1-\theta)}{1-(r-1/2)(1-\theta)}\|L_K^{-r} f_\rho\|_\rho$.

Proof. (A) By Theorem 7.1(A), it follows that

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

where we use

$$\begin{aligned} |\lambda_t^r - \lambda_{t-1}^r| &= |(t+t_0)^{-r(1-\theta)} - (t+t_0-1)^{-r(1-\theta)}| \\ &\le r(1-\theta)(t+t_0-1)^{-r(1-\theta)-1} \\ &\le 4r(1-\theta)(t+t_0)^{-r(1-\theta)-1}, \quad \frac{a+1}{b+1} \ge \frac{a}{b} \text{ if } b > a > 0, \end{aligned} \tag{7.4}$$

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

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

By Lemma 7.11(B),

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

whence

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

since

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

(B) Similar to Part (A) by replacing r with $r - 1/2$ using Theorem 7.1(D). □

7.2 Initial Error

Theorem 7.5 (Initial Error). *Let $t_0^\theta \geq \kappa^2 + 1$. Then for all $t \in \mathbb{N}$,*

(A) $\mathscr{E}_{init}^{(\rho)}(t) \leq C_5 (t+t_0)^{-1}$ where $C_5 = t_0 M_\rho$.

(B) $\mathscr{E}_{init}^{(K)}(t) \leq C_6 (t+t_0)^{-1}$, where $C_6 = t_0^{3/2} M_\rho$.

Proof. (A) By Lemma 7.11(D) with $j = 1$ we obtain

$$\mathscr{E}_{init}^{(\rho)}(t) \leq \|\bar{\Pi}_1^t\| \|r_0\|_\rho \leq \frac{t_0}{t+t_0} \|r_0\|_\rho.$$

For $f_0 = 0$, using Lemma 7.13(B), $\|r_0\|_\rho = \|f_{\lambda_0}\|_\rho \leq M_\rho$.

(B) By Lemma 7.11(B) with $j = 1$,

$$\mathscr{E}^{(K)}_{init}(t) \leq \|\Pi_1^t\| \|r_0\|_K \leq \frac{t_0}{t + t_0} \|r_0\|_K.$$

For $f_0 = 0$, using Lemma 7.13(A), $\|r_0\|_K = \|f_{\lambda_0}\|_K \leq t_0^{(1-\theta)/2} M_\rho \leq t_0^{1/2} M_\rho$ as $\theta \in (0, 1]$. □

7.3 Sample Error

7.3.1 Sample Error in $\mathscr{L}^2_{\rho_\mathscr{X}}$

Theorem 7.6 (Sample Error in $\mathscr{L}^2_{\rho_\mathscr{X}}$). *Assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_\mathscr{X}}$ for some $r \in [1/2, 1]$ and $t_0^\theta \geq \kappa^2 + 1$. Then with probability at least $1 - \delta$ ($\delta \in (0, 1)$), there holds for all $t \in \mathbb{N}$,*

$$\mathscr{E}^{(\rho)}_{samp}(t) \leq C_7 t^{-\theta/2}$$

where

$$C_7 = \sqrt{\frac{2}{\delta}} \kappa (\sqrt{3} M_\rho + 2\|L_K^{-r} f_\rho\|_\rho + \sqrt{3}\kappa(\kappa + 1) M_\rho)$$

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

Lemma 7.7. *Assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_\mathscr{X}}$ for some $r \in [1/2, 1]$ and $t_0^\theta \geq \kappa^2 + 1$. Then for all $t \in \mathbb{N}$, there holds*

$$\mathbb{E}\|\chi_t\|_K^2 \leq C_8.$$

where

$$C_8 = 2\kappa^2(3M_\rho^2 + 4\|L_K^{-r} f_\rho\|_\rho^2 + 3(\kappa^2 + 1)\kappa^2 M_\rho^2)$$

Proof. By definition

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

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

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

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

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

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

where using Lemma 7.13, $\|f_\lambda\|_\rho \leq M_\rho$, and Corollary 7.16 for the bound on $\mathbb{E}\|f_{t-1} - f_{\lambda_{t-1}}\|_\rho^2$,

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

Putting two terms together gives the result. □

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

Proof of Theorem 7.6. Denote $X_j = \gamma_j \bar{\Pi}_{j+1}^t \chi_j$, which is a martingale difference sequence. It suffices to show

$$\mathbb{E}[\|\sum_{j=1}^t X_j\|_\rho^2] \leq C_8 (t + t_0)^{-\theta}. \tag{7.5}$$

Then it follows from the Markov inequality

$$\mathbf{Prob}\left\{(z_i)_1^t \in \mathscr{X}^t : \|\sum_{j=1}^t X_j\|_\rho \geq \epsilon\right\} \leq \frac{\mathbb{E}[\|\sum_{j=1}^t X_j\|_\rho^2]}{\epsilon^2} \leq \frac{C_8}{\epsilon^2} t^{-\theta}.$$

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

$$\sqrt{\frac{C_8}{\delta}} = \sqrt{\frac{2}{\delta}}\kappa(3M_\rho^2 + 4\|L_K^{-r} f_\rho\|_\rho^2 + 3(\kappa^2+1)\kappa^2 M_\rho^2)^{1/2} \leq \sqrt{\frac{2}{\delta}}\kappa(\sqrt{3}M_\rho + 2\|L_K^{-r} f_\rho\|_\rho + \sqrt{3}\kappa(\kappa+1)M_\rho)$$

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

It remains to prove (7.5). By isometry $L_K^{1/2} : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{H}_K$,

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

Using Lemma 7.7, we have $\mathbb{E}\|\chi_j\|_K^2 \le C_8$. To estimate $\sum_{j=1}^t \gamma_j^2 \|\bar{\Pi}_{j+1}^t L_K \bar{\Pi}_{j+1}^t\|$, we use the spectral decomposition of $L_K : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{L}^2_{\rho_{\mathscr{X}}}$. Let $(\mu_\alpha, \phi_\alpha)_{\alpha\in\mathbb{N}}$ be an orthonormal eigen-system of L_K. For simplicity, denote $a_i = \gamma_i\lambda_i + \gamma_i\mu_\alpha$, then

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

where for large enough t_0,

$$\sup_j \gamma_j \prod_{i=j+1}^t (1-a_i) \le \sup_j \gamma_j \prod_{i=j+1}^t (1-\gamma_i\lambda_i) \le \sup_j \frac{1}{(j+t_0)^\theta} \cdot \frac{j+t_0}{t+t_0+1} \le (t+t_0)^{-\theta}, \quad (7.6)$$

and

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

which gives (7.5). □

Remark 7.8. It is still an open problem, if we can improve this bound to replace $1/\delta$ with $\log 1/\delta$, by using the Pinelis-Bernstein inequality for the martingale difference sequence. The difficulty seems that, the Pinelis-Bernstein inequality needs a uniform bound on $\|\gamma_j \bar{\Pi}_{j+1}^t \chi_j\|_\rho$, which so far is only $O(t^{-(1-2\theta)})$, by Lemma 7.12 such that $\|\chi_t\|_K \leq O(\|f_{t-1}\|_K) \leq O(1/\lambda_t)$ and $\|\gamma_j L_K^{1/2} \bar{\Pi}_{j+1}^t\| \leq O(t^{-\theta})$ as in the proof above. Then using the Pinelis-Bernstein inequality in Proposition A.4, we have

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

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

7.3.2 Sample Error in $\mathscr{H}_K$

Theorem 7.9 (Sample Error in $\mathscr{H}_K$). *Let $a = 1$, $t_0^\theta = (\kappa+1)^2)$ and $\gamma_0 = t_0^{-\theta}$. The following holds with probability at least $1-\delta$ ($\delta \in (0,1)$) in the space Z^t,*

$$\mathscr{E}_{samp}^{(K)} \leq C_9 (t+t_0)^{1/2-\theta}$$

where $C_9 = (5\kappa+1)M_\rho \log \frac{2}{\delta}$.

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

Lemma 7.10. *Let $A_t w_t^* - b_t = (f_{\lambda_t}(x_t) - y_t)K_{x_t} + \lambda_t f_{\lambda_t}$.*

(A) $\|A_t w_t^* - b_t\|_K \leq (\kappa+1)^2 M_\rho / \sqrt{\lambda_t}$;

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

Proof. Recall that

$$A_t w_t^* - b_t = (f_{\lambda_t}(x_t) - y_t)K_{x_t} + \lambda_t f_{\lambda_t}.$$

Then

(A) Using $\|f_\lambda\|_K \leq M_\rho/\sqrt{\lambda}$ in Lemma 7.13(A),

$$\|A_t w_t^* - b_t\| \leq \|f_{\lambda_t}(x_t)K_{x_t}\|_K + |y_t|\|K_{x_t}\|_K + \lambda_t\|f_{\lambda_t}\|_K \leq M_\rho\kappa^2/\sqrt{\lambda_t} + M_\rho\kappa + M_\rho\sqrt{\lambda_t}$$

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

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

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

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

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

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 7.13(B). □

Now we are ready to give the proof of the sample error bounds, Theorem 7.9.

Proof of Theorem 7.9. Denote $X_j = \gamma_j \Pi_{j+1}^t (A_j w_j^* - b_j) = \gamma_j \Pi_{j+1}^t \xi_j$, which is a reversed martingale difference sequence. It suffices to prove that

$$\|X_j\| \leq 2\gamma_t \lambda_t^{-1/2}(\kappa+1)^2 M_\rho \tag{7.7}$$

$$\mathbb{E}_{j-1}\|X_j\|^2 \leq 4\gamma_t^2\kappa^2 M_\rho^2 \tag{7.8}$$

Then by Proposition A.4, a varied form of Pinelis-Bernstein inequality, we have

$$\begin{aligned}
\mathscr{E}^{(\rho)}_{samp}(t) &= \|\sum_{j=1}^{t} X_j\| \le 2\left(\frac{1}{3}\gamma_t\lambda_t^{-1/2}(\kappa+1)^2 M_\rho + 2\sqrt{t}\gamma_t\kappa M_\rho\right)\log\frac{2}{\delta} \\
&\le M_\rho\left(\frac{2}{3}(\kappa+1)^2(t+t_0)^{-\theta/2} + 4\kappa\right)(t+t_0)^{1/2-\theta}\log\frac{2}{\delta} \\
&\le M_\rho\,(5\kappa+1)\,(t+t_0)^{1/2-\theta}\log\frac{2}{\delta}, \quad \text{since } t_0^\theta \ge (\kappa+1)^2.
\end{aligned}$$

To see (7.7) and (7.8), note that for $t_0^\theta \ge (\kappa+1)^2 \ge \kappa^2+1$, $\|\Pi_j^t\| \le \prod_{i=j}^t (1-\gamma_i\lambda_i)$, whence

$$\begin{aligned}
\|\gamma_j \Pi_{j+1}^t\| &\le \frac{1}{\lambda_j}\gamma_j\lambda_j \prod_{i=j+1}^{t}(1-\gamma_i\lambda_i) = \frac{1}{\lambda_j}\left[\frac{1}{j+t_0}\prod_{i=j+1}^{t}(1-\frac{1}{i+t_0})\right], \\
&= \frac{1}{\lambda_j}\left[\frac{1}{j+t_0}\cdot\frac{j+t_0}{t+t_0}\right] = \frac{1}{\lambda_j(t+t_0)}.
\end{aligned}$$

Then it follows from Lemma 7.10,

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

and

$$\mathbb{E}_{j-1}\|X_j\|^2 \le \frac{1}{\lambda_j^2(t+t_0)^2}\mathbb{E}\|A_j w_j^* - b_j\|_K^2 \le \frac{4\kappa^2 M_\rho^2}{\lambda_t^2(t+t_0)^2} = 4\gamma_t^2\kappa^2 M_\rho^2,$$

as desired. □

7.4 Total Error

7.4.1 Proof of Theorem C

Proof of Theorem C. By triangle inequality,

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

Combining Theorem 7.3(A), 7.4(A), 7.5(A), and 7.6, and setting $\theta = 2r/(2r+1)$, we obtain

$$\|f_t - f_\rho\|_\rho \leq (C_1 + C_3 + C_7)(t + t_0)^{-r/(2r+1)} + C_5(t + t_0)^{-1}$$

where

$$C_1 + C_3 = \left(\frac{1}{r} + \frac{4}{r+1}\right) \|L_K^{-r} f_\rho\|_\rho = \frac{5r+1}{r(r+1)} \|L_K^{-r} f_\rho\|_\rho$$

whence

$$C_1 + C_3 + C_7 \quad = \quad \sqrt{\frac{6}{\delta}} \kappa M_\rho (1 + \kappa \sqrt{\kappa^2 + 1}) + \left(2\sqrt{2}\kappa \sqrt{\frac{1}{\delta}} + \frac{5r+1}{r(r+1)}\right) \|L_K^{-r} f_\rho\|_\rho,$$

which ends the proof. □

7.4.2 Proof of Theorem D

Proof of Theorem D. By

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

with Theorem 7.3(B), 7.4(B), 7.5(B), and 7.9, we obtain

$$\|f_t - f_\rho\|_K \leq (C_2 + C_4)(t + t_0)^{-(r-1/2)(1-\theta)} + C_6(t + t_0)^{-1} + C_9(t + t_0)^{1/2-\theta}.$$

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

$$\|f_t - f_\rho\|_K \leq (C_2 + C_4 + C_9) t^{-(2r-1)/(4r+2)} + C_6 t^{-1}. \tag{7.9}$$

Noticing that

$$C_2 + C_4 = \left(\frac{2}{2r-1} + \frac{8}{2r+3}\right) \|L_K^{-r} f_\rho\|_\rho = \frac{20r-2}{(2r-1)(2r+3)} \|L_K^{-r} f_\rho\|_\rho.$$

We end the proof by plugging these constants into (7.9). □

7.5 Basic Estimates

Lemma 7.11. *If $t_0^\theta \geq \kappa^2 + 1$, then the following holds for all $t \in \mathbb{N}$,*

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

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

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

(D) $\|\bar{\Pi}_j^t\| \leq \frac{j+t_0-1}{t+t_0}$.

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

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

Therefore

$$\|I - \gamma_t A_t\| = \|I - \gamma_t L_t - \gamma_t \lambda_t\| \leq 1 - \gamma_t \lambda_t = 1 - \frac{1}{t+t_0},$$

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

(B) To show this,

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

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

7.5.1 Estimates of Path Radius

Lemma 7.12. *If $f_0 = 0$, then for all $t \in \mathbb{N}$,*

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

Proof. Since

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

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

$$\|f_t\|_K \leq \|1 - \gamma_t\lambda_t - \gamma_t L_K^{x_t}\| \|f_{t-1}\|_K + \gamma_t \|y_t K_{x_t}\|_K \leq (1-\gamma_t\lambda_t)\|f_{t-1}\|_K + \gamma_t \kappa M_\rho.$$

By induction on t, we have

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

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

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

since

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

This gives the bound. □

Lemma 7.13. *For any $\lambda > 0$,*

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

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

Proof. (A) Note that

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

Taking $f = 0$, we have

$$\|f_\lambda - f_\rho\|_\rho^2 + \lambda\|f_\lambda\|_K^2 \leq \|f_\rho\|_\rho^2 \leq M_\rho^2, \tag{7.10}$$

which leads to the result.

(B) By definition we obtain

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

□

7.5.2 Estimates of Path Gap

In this section we derive some direct estimates for the remainder variance $\mathbb{E}\|f_t - f_{\lambda_t}\|^2$, in $\mathscr{H}_K$ norm or $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ norm. The result in Lemma 7.15 will be used to derive a constant upper bound for $\mathbb{E}\|f_t - f_{\lambda_t}\|^2_\rho$ when $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ with $r \in [1/2, 1]$, i.e. Corollary 7.16.

Lemma 7.14. *Let $f \in \mathscr{H}_K$, and $\gamma, \lambda \in \mathbb{R}_+$. For all $z = (x, y) \in X \times \mathbb{R}$, let*

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

Then

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

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

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

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

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

Therefore

$$\mathbb{E}\|f^z - f_\lambda\|^2_\star = \|[I - \gamma(L_K + \lambda I)](f - f_\lambda)\|^2_\star + \gamma^2 \zeta(f) \tag{7.11}$$

where

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

since

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

Let us now study the two terms in Equation (7.11). First,

$$\|[I - \gamma(L_K + \lambda I)](f - f_\lambda)\|_\star^2$$

$$= (1-\gamma\lambda)^2\|f - f_\lambda\|_\star^2 - 2\gamma(1-\gamma\lambda)\langle L_K(f - f_\lambda), f - f_\lambda\rangle_\star + \gamma^2\|L_K(f - f_\lambda)\|_\star^2$$

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

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

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

where if $\star = K$,

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

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

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

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

Lemma 7.15. *Let*

$$\pi_k^t = \begin{cases} \prod_{i=k}^{t}(1 - \gamma_i\lambda_i), & k \leq t; \\ I, & k > t. \end{cases} \tag{7.12}$$

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

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

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

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

Proof. Using Lemma 7.14, for $\gamma_t(\lambda_t + 2\kappa^2) \leq 1$,

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

Note that

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

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

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

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

This gives the iteration formula,

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

which, by induction, leads to

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

which ends the proof. □

Corollary 7.16. *Assume that $L_K^{-r} f_\rho \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ with $r \in [1/2, 1]$ and $t_0^\theta \geq \kappa^2 + 1$. Then*

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

Proof. For $t_0^\theta \geq \kappa^2 + 1$, we have $\gamma_t \kappa^2 + \gamma_t \lambda_t \leq 1$, whence by Lemma 7.15 with $f_0 = 0$,

$$\mathbb{E}\|f_t - f_{\lambda_t}\|_\rho^2 \leq \pi_1^t \|f_{\lambda_0}\|_\rho^2 + \sum_{j=1}^t \gamma_j \lambda_j \pi_{j+1}^t \frac{\|f_{\lambda_j} - f_{\lambda_{j-1}}\|_\rho^2}{\gamma_j \lambda_j} + C_\rho \sum_{j=1}^t \frac{\gamma_j}{\lambda_j} \gamma_j \lambda_j \pi_{j+1}^t \tag{7.13}$$

The first term is not larger than M_ρ^2, using $\pi_1^t \leq 1$ and $\|f_\lambda\|_\rho \leq M_\rho$ in Lemma 7.13(B).

Now consider the second term. By Theorem 7.1(A) with $r \in [1/2, 1]$,

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

where we use

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

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

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

Using the telescope sum

$$\sum_{j=1}^t \gamma_j \lambda_j \pi_{j+1}^t = 1 - \pi_1^t \leq 1, \tag{7.14}$$

we have a bound for the second term

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

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

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

Together with the telescoping sum (7.14), the third term is not larger than C_ρ. This completes the proof. $\square$

Chapter 8

Open Problems and Future Directions

(A) It is not clear if Theorem C can be improved by replacing $1/\delta$ with $\log 1/\delta$, using exponential probabilistic inequalities instead of the Markov inequality. As in Remark 7.8, the difficulty may lie in a sharper estimate on the growth of f_t, or improve the Bernstein-type inequalities for nonuniformly bounded random variables.

(B) Iterative regularizations and their stochastic approximations are closely related with Boosting [Yao, Rosasco, and Caponnetto 2006] and await further explorations.

(C) The theory developed in this thesis takes the standard i.i.d. sampling assumption as the main stream of statistical learning theory. It is a large un-explored field for online learning with dependent sampling, such as Markov sampling [e.g. Aldous and Vazirani 1990; Smale and Zhou 2006b], mixing process [e.g. Meir 2000], and competitive settings like games [e.g. Vovk 2001; Vovk 2005]. The reference above is rather incomplete and just for a rought taste. An important application of online learning with Markov sampling might be stochastic algorithms to solve dynamic programming, also called reinforcement learning or Neuro-Dynamic Programming [e.g. Bertsekas and Tsitsiklis 1996; Van Roy 1998]. It should be noted that a RKHS can be induced from a Markov process via its Green's function or Dirichlet form [Diaconis and Evans 2002], where the latter might be useful to semi-supervised learning.

Bibliography

ALDOUS, D. and U. VAZIRANI (1990). A markovian extension of valiant's learning model. In *Proceedings of the* 31^{st} *Symposium on Foundations of Computer Science*, pp. 392–396.

BENAÏM, M. (1999). Dynamics of stochastic approximations. In *Le Seminaire de Probabilites, Lectures Notes in Mathematics, Vol 1709*, pp. 1–68. Springer-Verlag.

BERLINET, A. and C. THOMAS-AGNAN (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers.

BERTSEKAS, D. P. and J. N. TSITSIKLIS (1996). *Neuro-Dynamic Programming*. Belmont, Massachusetts: Athena Scientific.

BOUSQUET, O. and A. ELISSEEFF (2002). Stability and generalization. *Journal of Machine Learning Research* (2), 499–526.

CAPONNETTO, A. and E. DE VITO (2006). Optimal rates for regularized least squares algorithm. *Foundations of Computational Mathematics*. accepted.

CESA-BIANCHI, N., A. CONCONI, and C. GENTILE (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory 50*(9), 2050–2057.

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

Cucker, F. and S. Smale (2002a). Best choices for regularization parameters in learning theory. *Foundations Comput. Math. 2*(4), 413–428.

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

De Vito, E., A. Caponnetto, and L. Rosasco (2004). Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*. preprint.

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

Diaconis, P. and S. Evans (2002). A different construction of gaussian fields from markov chains: Dirichlet covariances. *Ann. Inst. Henri Poincare B, 38*, 867–838.

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

Duflo, M. (1997). Cibles atteignables avec une probabilité positive d'aprés m. benaim. *Unpublished manuscript*.

Engl, H. W., M. Hanke, and A. Neubauer (1996). *Regularization of Inverse Problems.* Kluwer Academic Publishers.

Evgeniou, T., M. Pontil, and T. Poggio (1999). Regularization networks and support vector machines. *Advances of Computational Mathematics 13*(1), 1–50.

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

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

Kiefer, J. and J. Wolfowitz (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics 23*, 462–466.

Konda, V. R. and J. N. Tsitsiklis (2004). Convergence rate of linear two-time-scale stochastic approximation. *The Annals of Applied Probability 14*(2), 796–819.

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

Meir, R. (2000). Nonparametric time series prediction through adaptive model selection. *Machine Learning 39*(1), 5 – 34.

Minh, H. Q. (2006). *Reproducing Kernel Hilbert Spaces in Learning Theory.* Ph. D. thesis, Brown University.

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

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

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

POLYAK, B. T. (1990). New method of stochastic approximation type. *Automation and Remote Control 51*, 937–946.

POLYAK, B. T. and A. B. JUDITSKY (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal of Control and Optimization 30*(4), 835–855.

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

ROBBINS, H. and D. SIEGMUND (1971). A convergence theorem for nonnegative almost supermartingales and some applications. In J. S. Rustagi (Ed.), *Optimizing Methods in Statistics*, pp. 233–257. Academic Press, New York.

ROSENBLATT, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review 65*(6), 386–408.

RUPPERT, D. (1988). Efficient estimators from a slowly convergent robbins-monro procedure. Technical report, Technical Report 781, School of Operations Research and Industrial Engineering, Cornell University.

SMALE, S. and Y. YAO (2006). Online learning algorithms. *Foundation of Computational Mathematics 6*(2), 145–170.

SMALE, S. and D.-X. ZHOU (2006a). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*. accepted.

SMALE, S. and D.-X. ZHOU (2006b). Online learning with markov sampling. in preparation.

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research 2*, 67–93.

Tarrès, P. and Y. Yao (2006). Online learning as stochastic approximations of regularization paths. preprint.

Van Roy, B. (1998). *Learning and Value Function Approximation in Complex Decision Processes*. Ph. D. thesis, Massachusetts Institute of Technology.

Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wiley & Sons, Inc.

Vovk, V. (2001). Competitive on-line statistics. *International Statistical Review 69*, 213–248.

Vovk, V. (2005, November). On-line regression competitive with reproducing kernel hilbert spaces. *eprint arXiv:cs/0511058*.

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

Widrow, B. and M. Hoff (1960). Adaptive switching circuits. *IRE WESCON Convention Record* (4), 96–104.

Widrow, B. and M. A. Lehr (1990). 30 years of adaptive neural networks: Perceptron, madaline, and backpropagation. *Proceedings of the IEEE 78*(9), 1415–1442.

Yao, Y. (2006). On complexity issue of online learning algorithms. *IEEE Transactions on Information Theory*. Accepted.

Yao, Y., L. Rosasco, and A. Caponnetto (2006). On early stopping in gradient descent learning. *Constructive Approximation*. Accepted.

Ying, Y.-M. and M. Pontil (2006). Online gradient descent learning algorithms. preprint.

Zhang, T. (2003). Leave-one-out bounds for kernel methods. *Neural Computation 15*, 1397–1437.

Zhou, D.-X. (2003). Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory 49*(7), 1743–1752.

Appendix A

Probabilistic Inequalities for Hilbert-valued Martingales

The following inequality is due to Iosif Pinelis [Pinelis 1992] (see also Theorem 3.5 in [Pinelis 1994]).

Lemma A.1 (Pinelis-Hoeffding). *Let $(\xi_i)_{i\in\mathbb{N}} \in \mathscr{H}$ be a martingale difference sequence in a Hilbert space $\mathscr{H}$ such that for all i almost surely $\|\xi_i\| \le c_i < \infty$. Then for all $t \in \mathbb{N}$,*

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

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

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

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

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.3 (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*

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

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

Proposition A.4. *Let ξ_i be a martingale difference sequence in a Hilbert space. Suppose that almost surely $\|\xi_i\| \leq M$ and $\sum_{i=1}^t \mathbb{E}_{i-1}\|\xi_i\|^2 \leq \sigma_t^2$. Then the following holds with probability at least $1-\delta$ ($\delta \in (0,1)$),*

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

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

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

Note that $\epsilon > 0$, then

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

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

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

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

Lemma A.5 (Markov). *Let X be a nonnegative random variable. Then for any real number $\epsilon > 0$, we have*

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

Appendix B

Some Estimates based on Gamma Function

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

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

The *gamma function* is defined by $\Gamma(a) = \Gamma(a,0)$.

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

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

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

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

and if $\tau < 0$,

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

Here

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

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

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

$$\begin{aligned} e^{-at^{1-\theta}} \int_1^t x^{-(\theta+\tau)} e^{ax^{1-\theta}} dx &= \frac{1}{1-\theta}\int_0^{t^{1-\theta}-1} (t^{1-\theta}-y)^{-\tau/(1-\theta)} e^{-ay} dy \\ &= \frac{t^{-\tau}}{1-\theta}\int_0^{t^{1-\theta}-1} \left(1 - \frac{y}{t^{1-\theta}}\right)^{-\tau/(1-\theta)} e^{-ay} dy. \qquad \text{(B.1)} \end{aligned}$$

1. (*For $A_{\theta,a}$*) For $0 \le y \le t^{1-\theta} - 1$, $1 - \dfrac{y}{t^{1-\theta}} \le 1$. Thus if $\tau \ge 0$, Equation (B.1) has

$$\begin{aligned} r.h.s. &\ge \frac{t^{-\tau}}{1-\theta}\int_0^{t^{1-\theta}-1} e^{-ay} dy = \frac{t^{-\tau}}{a(1-\theta)}(1 - e^{-a(t^{1-\theta}-1)}) \\ &\ge \frac{t^{-\tau}}{a(1-\theta)}(1 - e^{-a(2^{1-\theta}-1)}), \qquad \text{for } t \ge 2. \end{aligned}$$

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

$$r.h.s. \le \frac{t^{-\tau}}{1-\theta}\int_0^{t^{1-\theta}-1} e^{-ay} dy = \frac{t^{-\tau}}{a(1-\theta)}(1 - e^{-a(t^{1-\theta}-1)}) \le \frac{t^{-\tau}}{a(1-\theta)}.$$

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

$$\frac{1}{1-s} = 1 + \frac{s}{1-s} \le 1 + \frac{y/t^{1-\theta}}{1 - (t^{1-\theta}-1)/t^{1-\theta}} = 1 + y. \qquad \text{(B.2)}$$

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

$$\begin{aligned} r.h.s. &\le \frac{t^{-\tau}}{1-\theta}\int_0^{t^{1-\theta}-1} (1+y)^{\tau/(1-\theta)} e^{-ay} dy \\ &\le \frac{2^{\tau/(1-\theta)} t^{-\tau}}{1-\theta}\int_0^1 e^{-ay} dy + \frac{2^{\tau/(1-\theta)} t^{-\tau}}{1-\theta}\int_1^{t^{1-\theta}-1} y^{\tau/(1-\theta)} e^{-ay} dy \\ &\le \frac{2^{\tau/(1-\theta)}}{1-\theta} t^{-\tau}\left\{\int_0^\infty e^{-ay} dy + \int_0^\infty y^{(1+\tau-\theta)/(1-\theta)-1} e^{-ay} dy\right\} \\ &\le \frac{2^{\tau/(1-\theta)}}{a(1-\theta)}\left(1 + a^{-\tau/(1-\theta)}\Gamma\left(\frac{1+\tau-\theta}{1-\theta}\right)\right) t^{-\tau}. \end{aligned}$$

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

$$\begin{aligned} r.h.s. &\geq \frac{t^{-\tau}}{1-\theta}\int_0^{t^{1-\theta}-1}(1+y)^{\tau/(1-\theta)}e^{-ay}dy \\ &\geq \frac{2^{\tau/(1-\theta)}t^{-\tau}}{1-\theta}\int_0^1 e^{-ay}dy + \frac{2^{\tau/(1-\theta)}t^{-\tau}}{1-\theta}\int_0^{t^{1-\theta}-1} y^{\tau/(1-\theta)}e^{-ay}dy \\ &\geq \frac{2^{\tau/(1-\theta)}t^{-\tau}}{1-\theta}\int_0^1 e^{-ay}dy = \frac{2^{\tau/(1-\theta)}(1-e^{-a})}{a(1-\theta)}t^{-\tau}. \end{aligned}$$

This completes the proof. □

Lemma B.2. *Let $\theta \in (0,1)$. Then for all $t \geq 1$,*

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

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

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

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

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

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

where the last step is due to $t > \zeta > x$, whence

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

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

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

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

To show (B.3), by setting $s = x + \tau$,

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

This completes the proof. □

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

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

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

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

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

$$\psi_\theta^1(t,\alpha) := \sum_{k=1}^{t}\frac{1}{k^\theta}\prod_{i=k+1}^{t}\left(1 - \frac{\alpha}{i^\theta}\right) \leq \frac{3}{\alpha};$$

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

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

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

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

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

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

Thus

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

which equals

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

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

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

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

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

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

whence

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

3. Notice that

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

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

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

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

where

$$\begin{aligned} \sum_{k=1}^{t-1} \frac{1}{k^\theta} e^{\frac{\alpha}{1-\theta}(k+1)^{1-\theta}} &\le 2^\theta \sum_{k=1}^{t-1} \left(\frac{1}{k+1}\right)^\theta e^{\frac{\alpha}{1-\theta}(k+1)^{1-\theta}} \\ &\le 2 \int_1^{t+1} e^{\frac{\alpha}{1-\theta}x^{1-\theta}} x^{-\theta} dx \\ &\le \frac{2}{\alpha} e^{\frac{\alpha}{1-\theta}(t+1)^{1-\theta}}, \end{aligned}$$

which gives $2/\alpha$ by multiplying $e^{-\frac{\alpha}{1-\theta}(t+1)^{1-\theta}}$.

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

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

where if $\alpha = 1$,

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

and if $0 < \alpha < 1$,

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

4. Notice that

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

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

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

Then

$$\begin{aligned}
\sum_{k=1}^{t-1}\frac{1}{k^{2\theta}}\prod_{i=k+1}^{t}\left(1-\frac{\alpha}{i^{\theta}}\right)^2 &\leq 2^{2\theta}\sum_{k=1}^{t-1}\frac{1}{(k+1)^{2\theta}}e^{2\alpha'[(k+1)^{1-\theta}-(t+1)^{1-\theta}]}\\
&\leq 2^{2\theta}e^{-2\alpha'(t+1)^{1-\theta}}\int_1^{t+1}x^{-2\theta}e^{2\alpha' x^{1-\theta}}dx\\
&\leq \frac{2^{\theta/(1-\theta)+2\theta-1}}{\alpha}\left\{1+(2\alpha')^{-\theta/(1-\theta)}\Gamma\left(\frac{1}{1-\theta}\right)\right\}(t+1)^{-\theta},
\end{aligned}$$

where the last step is due to Lemma B.1 with $\tau=\theta$. Now by

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

we obtain

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

Combining the two terms together, we obtain

$$\begin{aligned}
\psi_\theta^2(t,\alpha) &\leq t^{-2\theta}+2^{\theta/(1-\theta)+2\theta-1}\left\{1+\Gamma\left(\frac{1}{1-\theta}\right)\right\}\left(\frac{1}{\alpha}\right)^{1/(1-\theta)}t^{-\theta}\\
&\leq 2^{\theta}\left\{\alpha^{1/(1-\theta)}(2t)^{-\theta}+2^{\theta/(1-\theta)}\left[1+\Gamma\left(\frac{1}{1-\theta}\right)\right]\right\}\left(\frac{1}{\alpha}\right)^{1/(1-\theta)}t^{-\theta}\\
&\leq 2D_\theta\left(\frac{1}{\alpha}\right)^{1/(1-\theta)}t^{-\theta},
\end{aligned}$$

for $t\in\mathbb{N}$, as desired. □

Appendix C

Reproducing Kernel Hilbert Spaces (RKHS)

In this appendix we provide some basic background on RKHS necessary for this thesis. For a thorough treatment on this topic, see [Berlinet and Thomas-Agnan 2004] for general interests in statistics, [Wahba 1990] for spline models, and [Minh 2006] for learning theory, as well as the references therein.

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

$\| \|_K$. The most important property of RKHS is the *reproducing property*: for all $f \in \mathscr{H}_K$ and $x \in X$, $f(x) = \langle f, K_x \rangle_K$.

Denote $\mathbf{x} = (x_k)_1^t$, $\mathbf{y} = (y_k)_1^t$, and $\mathbf{z} = (z_k)_1^t$. With the reproducing property, we may define the *sampling operator* $S_{\mathbf{x}} : \mathscr{H}_K \to l_2(\mathbf{x})$ by $S_{\mathbf{x}}(f) = (f(x_k))_{k=1}^t$. In particular, S_x is the evaluation functional such that $S_x f = f(x)$. Denote by $S_{\mathbf{x}}^* : l_2(\mathbf{x}) \to \mathscr{H}_K$ its adjoint, such that $S_{\mathbf{x}}^*(\mathbf{y}) = \frac{1}{t}\sum_{i=1}^{t} y_i K_{x_i}$. Clearly the operator norm $\|S_{\mathbf{x}}\| = \|S_{\mathbf{x}}^*\| \leq \kappa$ and $\|S_{\mathbf{x}}^* S_{\mathbf{x}}\| \leq \kappa^2$.

Let $\mathscr{C}(\mathscr{X})$ be the Banach space of continuous functions on $\mathscr{X}$. Define a linear map $L_K : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{C}(\mathscr{X})$ by $L_K(f)(x) = \int_X K(x,t) f(t) d\rho_X$. Together with the inclusion $J : \mathscr{C}(\mathscr{X}) \to \mathscr{L}^2_{\rho_{\mathscr{X}}}$, $J \circ L_K : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{L}^2_{\rho_{\mathscr{X}}}$ is a compact operator on $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ [e.g. Halmos and Sunder 1978], which by abusing the notation is also denoted as L_K.

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

$$\begin{aligned} L_K^r : \quad & \mathscr{L}^2_{\rho_{\mathscr{X}}} && \to \mathscr{L}^2_{\rho_{\mathscr{X}}} \\ & \sum_\alpha a_\alpha \phi_\alpha && \mapsto \sum_\alpha a_\alpha \mu_\alpha^r \phi_\alpha \end{aligned} \tag{C.1}$$

In particular, $L_K^{1/2} : \mathscr{L}^2_{\rho_{\mathscr{X}}} \to \mathscr{H}_K$ is an isometrical isomorphism between the quotient space $\mathscr{L}^2_{\rho_{\mathscr{X}}} / \ker(L_K)$ and $\mathscr{H}_K$. For simplicity we assume that $\ker(L_K) = \{0\}$, which happens when K is a universal kernel [Steinwart 2001] such that $\mathscr{H}_K$ is dense in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$. With $L_K^{1/2}$, $\langle \phi_\alpha, \phi_{\alpha'} \rangle_K = \langle L_K^{-1/2} \phi_\alpha, L_K^{-1/2} \phi_{\alpha'} \rangle_\rho = \mu_\alpha^{-1} \langle \phi_\alpha, \phi_{\alpha'} \rangle_\rho$, whence (ϕ_α) is a bi-orthogonal system in $\mathscr{H}_K$ and $\mathscr{L}^2_{\rho_{\mathscr{X}}}$.

The restriction of L_K on $\mathscr{H}_K$ induces an operator $L_K|_{\mathscr{H}_K} : \mathscr{H}_K \to \mathscr{H}_K$, which is the *covariance operator* of $\rho_{\mathscr{X}}$ in $\mathscr{H}_K$, since $L_K|_{\mathscr{H}_K} = \mathbb{E}[S_x^* S_x]$ by the reproducing property. When the domain is clear from the context, $L_K|_{\mathscr{H}_K}$ is also denoted by L_K. Note that all the three operators have norm bound $\|L_K\| \leq \kappa^2$.

Index